

This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (<https://dspace.lboro.ac.uk/>) under the following Creative Commons Licence conditions.



For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

A Treatise on Web 2.0 with a Case Study from the Financial Markets

A Doctoral Thesis

By

Martin D. Sykora

Submitted in partial fulfillment of the requirements
for the award of
Doctor of Philosophy of Loughborough University

April 2012

© Martin Sykora 2012

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgments or in references, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (Date)

Abstract

There has been much hype in vocational and academic circles surrounding the emergence of web 2.0 or social media; however, relatively little work was dedicated to substantiating the actual concept of web 2.0. Many have dismissed it as not deserving of this new title, since the term web 2.0 assumes a certain interpretation of web history, including enough progress in certain direction to trigger a succession [i.e. web 1.0 \rightarrow web 2.0]. Others provided arguments in support of this development, and there has been a considerable amount of enthusiasm in the literature. Much research has been busy evaluating current use of web 2.0, and analysis of the user generated content, but an objective and thorough assessment of what web 2.0 really stands for has been to a large extent overlooked. More recently the idea of collective intelligence facilitated via web 2.0, and its potential applications have raised interest with researchers, yet a more unified approach and work in the area of collective intelligence is needed.

This thesis identifies and critically evaluates a wider context for the web 2.0 environment, and what caused it to emerge; providing a rich literature review on the topic, a review of existing taxonomies, a quantitative and qualitative evaluation of the concept itself, an investigation of the collective intelligence potential that emerges from application usage. Finally, a framework for harnessing collective intelligence in a more systematic manner is proposed.

In addition to the presented results, novel methodologies are also introduced throughout this work. In order to provide interesting insight but also to illustrate analysis, a case study of the recent financial crisis is considered. Some interesting results relating to the crisis are revealed within user generated content data, and relevant issues are discussed where appropriate.

Key words: Web 2.0, Social Media, Collective Intelligence, Mass Collaboration, Information Retrieval, Web-mining, Text-mining, Sentiment-analysis, Financial Markets, Financial Crisis, Collaborative Systems

Acknowledgements

This thesis would not have been possible without the input, and support from numerous individuals. In this section I would like to express my sincere gratitude to all those who have supported me throughout this journey. First of all I am indebted to my supervisor Dr. Helmut Bez. His advice and our lengthy discussions concerning my work, and his comments on my drafts were insightful and invaluable. He was always very approachable, and ready to listen to any of my problems, complaints, worries, or ideas. Other academics in the department, including Prof. Paul Chung, Dr. Iain Phillips, Dr. Eran Edirisinghe and many others have helped me along the way.

I am also grateful to my colleagues, Dr. Richard Forsyth, Roman Kingsland, Dr. Xiaoming Wang, Marek Panek, Peter Holotik, and Markus Schmid. Richard's advice regarding natural language processing and text-mining in general has been very useful. Roman and Xiaoming shared some of their experience in data-mining with me. Several issues with web 2.0 systems were tackled thanks to discussions with Marek and Peter. Markus has been a wonderful colleague and friend in the later stages of my PhD, discussing together, many perils of academic life, and more general research problems.

I would like to express my gratitude to my mother, brother, sister and the rest of my family. I wish I could also thank my Dad for all his support, without his encouragement I would most likely never have discovered computer science in the first place, and this thesis is dedicated to him. I hope he can see these very lines, and understand my heartfelt appreciation for everything he ever did for me. Thanks to this PhD, I have also met my wonderful wife, Suzanne Elayan. She inspires me every day, and she is the joy in my life. Many of my good friends have provided much support along the way, and I must hence express my sincere thanks to them all.

List of Abbreviations

AJAX	Asynchronous JavaScript and XML
AMH	Adaptive Market Hypothesis
API	Application Programming Interface
BOW	Bag of Words
CI	Collective Intelligence
CSS	Cascading Style Sheets
DF	Document Frequency
DHTML	Dynamic HyperText Markup Language
DOM	Document Object Model
DSS	Decision Support System
EMH	Efficient Market Hypothesis
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
KNN	K Nearest Neighbour
ML	Machine Learning
NE	Named Entities
NLTK	Natural Language Toolkit
NN	Neural Network
POS	Part of Speech
RDF	Resource Description Framework
REST	Representational State Transfer
RSS	Really Simple Syndication
SOAP	Simple Object Access Protocol
SVM	Support Vector Machine
TF	Term Frequency
UGC	User Generated Content
URL	Uniform Resource Locator
WM	Wayback Machine
WWW	World Wide Web
XHTML	eXtensible HyperText Markup Language
XML	eXtensible Markup Language

Table of Contents

1 Introduction	1
1.1 Background and Motivation	1
1.2 Research Design	2
1.2.1 Research Aim	2
1.2.2 Research Objectives	2
1.3 Contribution Highlights	5
1.4 Thesis Structure	6
1.4.1 Chapter Outlines	7
2 Background	10
2.1 Web 2.0 Defined	10
2.1.1 Information Accuracy and Trust (Wikipedia)	11
2.1.2 Ushahidi, NASA and Ebird	12
2.2 Web 2.0 in Historical, Social and Economic Context	14
2.2.1 Historical Perspective	14
2.2.2 Social Perspective	18
2.2.3 Economic Perspective	20
2.2.4 Critical Perspective	21
2.3 Web 2.0 Taxonomies	27
2.3.1 Existing Taxonomies	27
2.3.2 The Proposed Web 2.0 Taxonomy	32
2.4 Current Applications of Web 2.0	37
2.4.1 Clinical Practice	37
2.4.2 Corporate Use	41
2.4.3 Politics, Public Service and Education	42
2.4.4 Journalism and Geography	43
2.5 Summary	45
3 Defining Web 2.0	46
3.1 Background	47
3.2 Proposed Methodology	49
3.2.1 Validating Step-2	50
3.2.2 Technical Implementation of Step-2	51
3.2.3 Technical Implementation of Step-3	52
3.3 Results	52
3.3.1 Step-1: Total Tag Use	52
3.3.2 Step-2: Bookmarking over Time	53
3.3.3 Step-3: Associated Comments and Tags	55
3.4 Discussion	56
3.4.1 Limitations	57
3.5 Summary	57
4 Survey based Web 2.0 Investigation	58
4.1 Motivation for the Survey Study	58

4.2 Survey design	61
4.2.1 General Considerations	61
4.2.2 Question Order, Format and Wording	62
4.2.3 Sample Design.....	66
4.2.4 Survey Distribution Method.....	67
4.2.5 Self-selection Bias	67
4.3 Evaluation of the Survey-response Data for Analysis.....	68
4.3.1 Avoiding Misrepresentation of the Population's Sample	68
4.3.2 Geographical Survey-response Distribution	68
4.3.3 Ensuring Overall Reliability.....	69
4.3.4 Missing Responses (Unanswered Questions).....	69
4.3.5 Demographic Distributions of Responses	70
4.4 Data Pre-processing.....	72
4.4.1 Representing Responses Numerically	72
4.4.2 Representing Factors Numerically	73
4.4.3 Factor Analysis – Confirmation of Factor Choice.....	73
4.4.4 Frequency Distributions and Normality Tests	75
4.5 Survey Results and Analysis	76
4.5.1 Correlation Analysis	76
4.5.2 Web 2.0 Related Questions.....	77
4.5.3 Business Activity	86
4.5.4 Trust.....	88
4.5.5 Time.....	92
4.5.6 Motivation	97
4.5.7 Wikipedia.....	104
4.6 Limitations of the Study	106
4.7 Summary	107
5 Historical Evolution towards Web 2.0	109
5.1 Background	109
5.2 Evolution of Web 2.0 and Web-design Considerations	110
5.3 Methodological Details and Limitations	113
5.4 Results	117
5.4.1 Dataset Overview	117
5.4.2 Analysis of Web-design Changes	120
5.5 Discussion	130
5.5.1 Criticism: Breadth of the Study.....	131
5.5.2 Reflections on the Methodology	133
5.6 Summary	136
6 Collective Intelligence Data Sources	137
6.1 Overview and Previous Literature.....	139
6.2 Youtube.....	141
6.2.1 Youtube Background Information	141
6.2.2 Methodology	143
6.2.3 Youtube, the Financial Crisis and Authors of Content	146
6.2.4 Information Efficiency on Youtube	151

6.2.5 Summary and Limitations	154
6.3 Delicious.....	156
6.3.1 Methodology	156
6.3.2 A brief Summary of the Financial Crisis	157
6.3.3 Analysis and Results.....	159
6.3.4 Summary and Limitations	165
6.4 Amazon.....	167
6.4.1 Methodology	169
6.4.2 Results: Competing Product Reviews (part i)	176
6.4.3 Results: Trading Books Reviews (part ii).....	187
6.4.4 Summary and Limitations	200
6.5 Wikipedia.....	201
6.5.1 The Wikipedia Culture	201
6.5.2 Useful Wikipedia Features	203
6.6 Community Finance and Prediction Websites.....	204
6.6.1 Trend Prediction Websites	205
6.7 Summary	207
7 Newsmental; A Custom Source of Collective Intelligence	208
7.1 Motivation	209
7.2 Design.....	212
7.2.1 Architecture	213
7.2.2 News Presentation	218
7.2.3 Rating News and Related Issues	219
7.2.4 Miscellaneous Features	222
7.3 Launch and Maintenance.....	224
7.3.1 Issues and Maintenance.....	224
7.3.2 Traffic	225
7.3.3 Feedback from Users.....	226
7.4 Results	227
7.4.1 Agreement among Readers.....	230
7.4.2 Pre-processing	232
7.4.3 Collective News Analysis.....	232
7.5 Discussion	237
7.5.1 Limitations.....	238
7.5.2 Future Work.....	240
7.6 Summary	241
8 Collective Intelligence Overview and Characterisation.....	243
8.1 Motivation	244
8.2 Background	245
8.2.1 Collective Intelligence.....	245
8.2.2 Value of the Underlying Data	248
8.3 The Framework	249
8.3.1 Extracting Data (Technical Perspective)	254
8.3.2 Example Applications.....	257
8.4 Discussion and Limitations	262

8.5 Summary	263
9 Conclusions and Future Work	265
9.1 Conclusions	265
9.2 Future Work.....	269
References	270
Appendix A: Defining Web 2.0.....	291
Appendix B: Survey Analysis and Results.....	296
Appendix C: Historical Website Study	326
Appendix D: Collective Intelligence Sources	329
Appendix E: Custom Source of Collective Intelligence	334
Appendix F: Publications.....	338

1 Introduction

1.1 Background and Motivation

This thesis represents a contribution to the body of work on web 2.0, and the collective intelligence that emerges from participatory web based applications. Over the last years there has been much excitement and hype in vocational and academic circles surrounding the emergence of web 2.0. Chi (2008) wrote; *“It’s clear web 2.0 isn’t just a fad, but a fundamental transformation of the web into a true collaborative and social platform.”*, and Barbry (2007) labelled web 2.0 as intensely groundbreaking for the web. However, relatively little work was dedicated to systematically substantiating what web 2.0 really stands for, although the concept has become widely used and references to it have been on the increase in academic literature (e.g. Van De Belt, 2010). This thesis identifies a number of characteristics of web 2.0 that are then systematically investigated with the help of several carefully designed studies (see section 1.2 for methodology, and section 1.3 for originality). There is still some confusion about what the term web 2.0 refers to, and in fact more recently Jackson and Lilleker (2009) have gone as far as suggesting that there is ground for what they would call “web 1.5”. This is clearly a dangerous area to be headed into and is another reason why the presented work in this thesis is considered to be of value to the research community. It is important that before new terminology is brought into the field that the old is understood well enough, otherwise unnecessary complexity is introduced. With the prevalence of web 2.0, the collective intelligence that emerges from participatory web applications has also received some attention in the research community. This work looks at how web 2.0 systems can be leveraged in aggregate to help provide collective intelligence solutions to various disparate domains of application. There is a need for a systematic step-by-step framework, since there is a lack of standardisation and much research without consideration for higher-level abstractions. A novel systematic framework to guide such work is hence presented, and a study of the financial markets, specifically the recent financial crisis is addressed in this context.

In order to highlight the significance of web 2.0 further, it ought to be mentioned that the European Union has commissioned several scientific studies to investigate the importance of web 2.0, and its technological and economic implications (Osimo 2008). More recently a European Union funded project, to last ten years, the FutureICT¹ project, with a budget of 1 billion EUR was introduced, and will be aimed at various aspects of social sciences relying among other tools, heavily on web 2.0 based technologies and the emergent collective

¹ <http://www.futurict.eu/the-project>

intelligence (Bishop and Baudains 2010).

Further background to web 2.0 is provided in Chapter 2, and background on collective intelligence in the context of web 2.0 is given in chapter 6, and chapter 8. Motivation for work is also highlighted throughout the thesis.

1.2 Research Design

1.2.1 Research Aim

The aim of this research is to define and evaluate web 2.0 and its wider context, such as factors that lead to its popularity, and to investigate its potential, in particular within collective intelligence applications, which includes a characterisation of collective intelligence that describes how web 2.0 systems can be leveraged in aggregate to that end.

This is to be achieved through a mixture of quantitative methodology with elements of qualitative research, with the help of a case-study from within financial markets, where deemed appropriate, although it is expected that web 2.0 based collective intelligence will have potential applications in a variety of other disparate fields of study. A survey study will be conducted in one of the chapters, and in another chapter a rarely used but valuable research tool (Wayback Machine – web based archives) will be employed in a historical study of web 2.0 evolution. Throughout the work, several elements and standard techniques from text-mining and web based information retrieval / data-mining will also be employed, and necessary details will be provided as and where relevant throughout the chapters. In order to achieve our research aim, a set of objectives which are implemented to ensure that the research aim is met were formulated, and are discussed in the next section (1.2.2).

1.2.2 Research Objectives

The aim of this research was reached through the research objectives, listed below, which were investigated by applying the research methodologies briefly mentioned in the previous section.

Objective 1 – To critically analyse literature on web 2.0 and social media

The main purpose of this objective is to provide a synthesis of web 2.0 literature. In particular from a historical, social, economic and critical perspective, in order to investigate why the concept of web 2.0 emerged, what its social and economic implications are, and also what drawbacks and criticisms it has spawned. The synthesis of literature will further focus on

substantiating the actual concept “web 2.0” and “social media”, with a presentation of existing taxonomies, and provide an overview of current applications of web 2.0 systems in a variety of vocational fields, ranging from clinical practice to education and journalism.

Objective 1.1 – To propose a new web 2.0 taxonomy

This sub-objective takes existing taxonomies that were identified in objective 1 into consideration and proposes a new two-step taxonomy, which is as generic but allows addressing specific web 2.0 concerns directly, that weren’t explicit in previous taxonomies. The rationale behind this taxonomy is that website elements which typically facilitate the sharing of content on web 2.0 should be used from the ground-up to categorise web 2.0 applications, rather than a direct category assignment.

Objective 2 – To quantitatively investigate and substantiate web 2.0 related terms

Propose, validate and use a new methodology for investigating emergence of neologisms in the English language, in order to quantify the prevalence of new web 2.0 related “2.0” buzz-words, such as “law 2.0” or “education 2.0”. Similar to the initial enthusiasm surrounding the internet – the so called “e-words”, e.g. e-commerce or e-health... – web 2.0 terms began to emerge in recent years and within this objective our goal is to employ a quantitative methodology to assess these terms.

Objective 3 – To investigate factors considered to be essential for web 2.0 and to add support for elements of the taxonomy (developed in objective 1.1), using a survey study

In objective 1, based on the synthesis of web 2.0 literature, the likely reasons / factors for the uptake of web 2.0 were revealed. A survey on a relatively large sample of respondents is undertaken in order to establish how important these reasons are as of today, how users perceive them and how they relate to each other. The survey will further be used to assess the prevalence of the actual terms “web 2.0” and “social media”, the popularity of web 2.0 applications by their usage, assess the motivations that drive people to use them and activities that are commonly performed on web 2.0 systems. The latter is relevant to the taxonomy proposed in objective 1.1, and helps in justifying it.

Objective 4 – To analyse and understand the historical evolution that has lead for web 2.0 to emerge from the so called “web 1.0”

There are several specific features that are commonly associated with web 2.0 systems. In this objective we plan to use web based archives available for research, in order to perform a

historical study that investigates how various features were taken up on the web. Web 2.0 implies a designated version and a discrete evolution, although presumably web 2.0 applications likely emerged gradually with a new type of practice. This research objective will try to answer the above, and acts as a complementary study to the survey, which instead provided a current snapshot, rather than a historical perspective on web 2.0.

Objective 5 – To define collective intelligence and investigate possible connections between collective intelligence from web 2.0 systems and the financial markets

The simple use of web 2.0 creates another level, or type of information that has far reaching implications and presents an area of research in itself; this is the so called collective intelligence, generated by web 2.0 application use. An analysis of literature, which will be extended in research objective 7, will be used to help define collective intelligence. In order to drive our investigation, the case of financial markets will be used to assess collective intelligence in a number of different web 2.0 applications. Some interesting results relating to the recent credit crisis are also revealed and discussed.

Objective 5.1 – To establish whether efficient information transfer within web 2.0 systems exists, using the concept of EMH (Efficient Market Hypothesis)

This sub-objective is concerned with establishing whether and to what degree efficient information transfer occurs on web 2.0 applications. In particular the degree of efficiency in propagation of financial news within a higher entry-barrier social media is to be established. Web 2.0 based user generated content is to be correlated with financial stockmarket price data, effectively using stockmarkets as a proxy for efficient information transmission. The Efficient Market Hypothesis from financial markets (Fama 1965) is used as the central idea to drive this investigation.

Objective 6 – To illustrate and design a new web 2.0 system and to assess whether it can act as a custom source of collective intelligence

This objective addresses the idea of custom collective intelligence sources, based on custom web 2.0 systems design. The design of a custom web 2.0 system is developed and delivered. It is argued that it can be feasible and indeed highly desirable to introduce a new web 2.0 platform, where the aim is to satisfy a collective intelligence acquisition need for sharing specific type of knowledge, at a desired granularity. A financial news analysis and opinion sharing web 2.0 site is developed within this objective and subsequently the collective intelligence amassed during the system's use is explored and evaluated. The outcomes of this

objective and the previous research objective 5 pave the way for the final research objective 7.

Objective 7 – To propose a mathematical characterisation of collective intelligence that can act as a set of guidelines or framework for collective intelligence applications

A literature review of collective intelligence (in addition to literature from research objective 5) will be undertaken, and in addition to outcomes and observations from objectives 1, 5 and 6, a framework for harnessing collective intelligence in a more systematic manner is proposed. The aim is to provide a formal characterisation of how to leverage collective intelligence in the web 2.0 environment. Also, in order to stimulate further work in the area, a variety of possible applications of the framework are introduced.

1.3 Contribution Highlights

In terms of web 2.0, many related concepts have been investigated in separate studies (Cormode and Krishnamurthy 2008), whereas in this thesis an emphasis was placed on discussing web 2.0 and related collective intelligence in its entirety. First of all, a valid contribution to the existing body of literature was made (in chapter 2) by providing a detailed discussion of web 2.0 itself and wider historical, social, economic, and other issues. An alternative two-step Taxonomy for categorising web 2.0 applications was introduced, after several existing taxonomies from prior literature were discussed. A novel methodology for investigating generic neologism emergence in the English language was proposed (see chapter 3), and subsequently the use of “2.0” terms, related to web 2.0, that have cropped up in recent years were studied. This study was the first to look at such a wider set of terms.

In order to better understand the current use of web 2.0, but also as this relates to various features of importance that were identified in previous work, but not yet investigated in this capacity, a large web based survey in chapter 4 was conducted. The study also provided much sought after survey design and reliability evaluation details, often left out in prior literature in this area. The survey study presents a series of novel insights, such as the characteristic differences between the perceptions of the terms “web 2.0” and “social media”, but it also looks at the atomic collaborative activities that were identified within the two-step Taxonomy for categorising web 2.0 applications, and which are later used in chapter 8.

In addition to the online survey, a historical study that looked at the evolution of web 2.0 on a selection of carefully chosen webpages, was conducted in chapter 5. This study used a relatively underused tool in a way not used before. In terms of novelty, such a study was non-existent in previous research, and it is significant in that it quantifies and substantiates certain

features that are regarded as characteristic of web 2.0 applications over a prolonged historical time-period, reaching as far back as the late 90s on some of the websites studied.

Another contribution of the thesis was an in depth study of Youtube in relation to the financial crisis. The data under investigation was novel, and the study reported a first investigation of its kind to the use of collaborative media sharing website for stock market analysis. A convincing case for efficiency of information transfer was also made, and represents a valid contribution to previous work in this area. Studies of the financial crisis were also conducted with the link sharing application Delicious and with User Generated Content (UGC) from Amazon books and leading manufacturer's products pages. Especially the study of Amazon provides interesting insights that confirm existing work from prior literature, and new conclusions about Amazon based UGC in relation to financial markets are drawn.

Chapter 7 demonstrated an exemplary case study of a web 2.0 application design, with the aim of ultimately providing UGC that other, existing web 2.0 applications do not provide, or at least not at the desired granularity. The contribution of this chapter was in terms of web 2.0 system engineering for collective intelligence capture, as we have not come across a similar walkthrough in prior work. Chapter 8 reviews collective intelligence work in the literature, and a strong argument is made in favour of a guide on harnessing collective intelligence from web 2.0. The chapter presents a novel and abstract framework that is hoped will aid future researchers in the area. Overall the thesis integrates an in-depth study of web 2.0 with work on harnessing UGC within collective intelligence systems.

The next section presents the structure of the thesis document and introduces an outline for each chapter.

1.4 Thesis Structure

The thesis can be roughly broken down into two main parts. Most of the literature review and background information to web 2.0 and related concepts are introduced in chapter 2. The chapter provides a valuable critical evaluation of the main reasons behind development of web 2.0, and several current applications of web 2.0 are presented. Given this background the first part of the thesis is concerned with substantiating and objectively quantifying web 2.0 as a concept, and its current use is investigated in the context of the thesis in chapters 3 to 5. The second part of the thesis focuses on collective intelligence and examples from the financial crisis are investigated in chapters 6 to 8.

The next section (section 1.4.1) provides a break-down overview of each significant chapter within the thesis. It must also be noted, that on the first page of every chapter there is a Tag-

cloud included. This is a stylistic element, since it is a tool often associated with web 2.0 applications, and actually provides a brief graphical summary of the chapter's contents. Each Tag-cloud contains 100 most frequent keywords (after the removal of common English words, e.g. and, the, by...), size of words represents the increasing word counts, and is generated from the entire chapter text.

1.4.1 Chapter Outlines

1.4.1.1 Chapter 2: Background

Section 2.1 provides a definition of web 2.0, based on prior literature, and by introducing several example web 2.0 applications. Section 2.2 discusses web 2.0 in a wider but highly important context of historical, social, and economic developments, heavily drawing on prior literature. A large subsection is also dedicated to various criticisms, and motivations for social media use are discussed in detail. Previous taxonomies described in literature are presented in section 2.3, and sub-section 2.3.2 presents a new, two-step taxonomy for categorising web 2.0 applications. Finally, section 2.4 illustrates numerous examples of web 2.0 use in different vocational areas.

1.4.1.2 Chapter 3: Defining Web 2.0

Section 3.1 introduces the problem of neologism term emergence, the recent uptake in “2.0”, web 2.0 based terms, and background prior-literature, which adds support for a new methodology to study neologism term emergence. The proposed methodology that consists of three steps is presented in section 3.2. Then in section 3.3, actual results of the “2.0” neologism study, based on the suggested methodology, are discussed. Finally section 3.4 points out limitations of the proposed methodology.

1.4.1.3 Chapter 4: Survey based Web 2.0 Investigation

The chapter begins with sections 4.1, 4.2, and 4.3, presenting motivation for the survey study, survey design, and an initial respondent evaluation, respectively. Section 4.4 discusses the pre-processing of the survey data, and sub-section 4.4.3 presents support for thematic factors for survey evaluation, using Principal Component Analysis. The most elaborate section of the chapter is section 4.5, which presents the entire survey results, analysis, and discussion. Each factor, as identified in section 4.4.3 is discussed in turn, separately. Finally section 4.6 concerns itself with limitations of the study.

1.4.1.4 Chapter 5: Historical Evolution towards Web 2.0

Section 5.1 introduces the dataset used within the study in this chapter, and related background literature. The next section establishes several characteristic features, which are often associated with web 2.0. Chapter 2 is used for this purpose. In section 5.3, the actual methodology of the historical study is discussed, and justification is given for the choice of websites, i.e. Youtube, Amazon, Flickr, Twitter, Craigslist, Digg, and Yahoo. First, the overall results of the historical study are discussed in section 5.4, followed by an individual sub-section dedicated to each website. Finally section 5.5 summarises some of the results and discusses issues with the study. One of the highlighted issues is addressed in sub-section 5.5.1, in which a similar analysis on over 50 top websites is conducted, since the initial study only considered a narrow selection of 7 websites. Sub-section 5.5.2 discusses the use of the Wayback Machine based methodology, and some of the issues encountered.

1.4.1.5 Chapter 6: Collective Intelligence Data Sources

Chapter 6 briefly introduces collective intelligence, although more details on CI are provided in chapter 8. The rest of the chapter provides a detailed description of existing sources of CI; in particular Youtube, Delicious, Amazon, Wikipedia, trend prediction, and financial community web 2.0 sites. Section 6.1 contains a literature review of previous studies of the UGC generated from web 2.0 systems. The well known Youtube media sharing website is analysed in much detail within section 6.2, with sub-section 6.2.3 presenting some relationships to the financial crisis, efficient information transfer on Youtube is investigated within sub-section 6.2.4, and limitations are discussed in sub-section 6.2.5. The potential of Delicious for achieving similar results as with Youtube are investigated in section 6.3, and various links to the financial markets and the crisis are made throughout this section, with a discussion of limitations in sub-section 6.3.4. Section 6.4 presents a relatively unique study of Amazon based UGC. The section is split into two main parts, part i (sub-section 6.4.2), which looks at UGC associated with leading products of several manufacturers, and part ii (sub-section 6.4.3), which looks at reviews of finance related books. Sub-sections 6.4.1 and 6.4.4, respectively cover in detail the methodology employed, and the various limitations of the study. Finally, section 6.5 presents Wikipedia, and section 6.6 community websites and trend prediction websites.

1.4.1.6 Chapter 7: Custom Source of Collective Intelligence

Section 7.1 evaluates the various motivations for and against a custom source of CI. The design of an example custom source of CI, namely the Newsmental web 2.0 system is presented in section 7.2. This is an extensive section and references good design practices and various issues are carefully described and considered. The launch of the web 2.0 system and problems encountered are illustrated in section 7.3. Section 7.4 describes the UGC generated on Newsmental, and the CI potential is evaluated on several examples relating to the financial crisis, in particular on an example of an important IMF vote, and the case of the Greek Debt (both in sub-section 7.4.3). Finally section 7.5 discusses limitations and possible future work at length.

1.4.1.7 Chapter 8: Collective Intelligence Framework

Chapter 8 introduces a framework for harnessing CI from web 2.0 systems. Motivations for the work are introduced in section 8.1, followed by background and prior-literature on CI in section 8.2, and also the value of the underlying data from web 2.0 systems use is highlighted. The actual framework is introduced formally in section 8.3. The sub-section 8.3.1 discusses several technical particularities in the data retrieval stages, and sub-section 8.3.2 introduces five example framework applications, to help illustrate the potential of the framework and stimulate research in the area. A discussion and limitations of the framework are presented in section 8.4.

2 Background



2.1 Web 2.0 Defined

Who, these days, has not heard of Wikipedia, Facebook, Twitter¹ and terms such as Open source, Blogs or Wikis. These applications have taken the internet community by storm (Surowiecky 2005, Pascu 2008, Li and Bernoff 2008, Tapscott and Williams 2008, Howe 2009, Leadbeater 2009, Shirky 2009, Radwanick 2010, Shirky 2010)², but what do they really stand for, and how, and in what ways might they be of value in a wider field of practical problems? These applications are often referred to by the somewhat vague but yet eloquent term – “web 2.0”. The term refers to a perceived second generation of world wide web that facilitates communication, information sharing, interoperability, and most importantly collaboration on the web. It stands for the paradigm that a considerable portion of the web is read and write (Vosen and Hagemann 2007), where instead of a single agent (i.e. administrator) updating a website, everybody can now interactively update, upload and collaborate on content of web-pages (Tapscott and Williams 2008), whether these are file, personal detail or news sharing

1 <http://www.twitter.com>, <http://www.wikipedia.com>, <http://www.facebook.com> all fall into the top 11 most frequently visited websites on the entire world wide web; as measured by the Alexa Internet Inc. service – see <http://www.alexa.com/topsites> It should be noted that there are currently approx. as many as 2 billion internet users worldwide – <http://www.internetworldstats.com/stats.htm>

2 Some of these sources are best-selling *popular science* books and highlight not just the academic interest but also the interest of the wider public into new web 2.0 media.

applications.

The term web 2.0 was introduced by the president of O'Reilly Media (Tim O'Reilly) in his influential blog post in 2005 (O'Reilly 2005), following a spontaneous conference debate at which O'Reilly and other industry and academic domain experts discussed the changes in development and usage of world wide web taking ubiquitous hold at the time. Their conclusion was that the world wide web is becoming a collaborative platform, and web 2.0 could be understood to represent this platform in its own right. In other words, collaboration needed little to no human intervention as it became mostly automated thanks to wide adoption of database backed programming techniques, asynchronous client-server communication (AJAX), or HTML / CSS standardization among other things. The original definition was given by example³ which makes it intrinsically subjective, and despite all his [O'Reilly's] efforts it became frequently misunderstood. Still O'Reilly managed to convey main elements of the “new” conventions and principles expected in web-sites that would be characteristic of a symbolic second generation of the web. Essentially his web 2.0 is based on a set of seven principles which the reader can also study further in O'Reilly (2005), and Anderson (2007). The primary principle that resonates throughout the definition is that previously most systems were characterised by a small number of pre-approved participants and were not fully interactive, while web 2.0 by contrast is characterised by online communities, open and very easy sharing, interactivity and collaboration (O'Reilly 2005, Vosen and Hagemann 2007, Shirky 2009). While technical aspects such as shorter release cycles, using innovative approaches of perpetual beta testing, and lightweight programming models, based on RSS, Rest APIs and other loosely coupled systems are important side developments.

2.1.1 Information Accuracy and Trust (Wikipedia)

It ought to be emphasised that web 2.0 stands for the collective nature of sharing, among many participants of an online system. An example of this, often given in most literature to exemplify web 2.0, is Wikipedia⁴. It is a radical experiment in trust, since even anonymous users are allowed to and in fact are encouraged to edit and re-edit this web-based encyclopedia in the communal hope of producing an immense and “complete” body of encyclopedic knowledge. Critics, such as Keen (2007) point out the seemingly intrinsic problem that is such a vast text would clearly have to be riddled with inaccuracies. Quite surprisingly however Wikipedia was found to be an accurate resource and is now arguably even becoming a standard encyclopedic

3 e.g. mp3.com → Napster / P2P, Britannica Online → Wikipedia, Personal Websites → Blogging, Ofoto → Flickr

4 Wikipedia is an online, publicly maintained encyclopedia. It covers millions of topic definitions. A concise and relatively complete analysis of Wikipedia is provided in the book by O'Sullivan (O'Sullivan 2009).

reference text. A comparison with encyclopedia Britannica (Giles 2005) suggests a similar level of information accuracy in both encyclopedias. Interestingly, 70%-80% of inaccurate edits on Wikipedia get corrected almost instantly (Adler et al. 2008a, Adler et al. 2008b). This can be attributed to the dynamic nature and self-managing environment of collaborative participation.

2.1.2 Ushahidi, NASA and Ebird

Since Wikipedia is a common example provided in literature, let us consider three other case studies, namely; Ushahidi, NASA Clickworkers and Ebird. Ushahidi is a more recent example of mass collaboration facilitated via web 2.0 principles. Ushahidi was developed to help citizens track outbreaks of ethnic violence in Kenya after the disputed 2007 presidential elections (Adewumi 2008). Since the government banned mainstream media from reporting on the violence, Ory Okolloh, a Kenyan political activist decided to blog about it on her political blog and asked her readers to email and send comments of any violence breaking out in their area. However this method of communication became so popular that her blog became a critical source of first-person reporting, to the extent Okolloh became unable to cope with categorising and processing all the incidents on her own. She therefore asked for volunteer help. Subsequently two programmers, Erik Hersman and David Kobia decided to provide assistance by writing (a first version of open source web-application) Ushahidi which would automatically aggregate and categorise violence reports. Reports could also be sent from mobile phones and via text messages and this is when the application really took off. The reports would then be presented by Ushahidi in a number of formats – including a geographical mapping of the attacks in near-real time. Categorised and filtered violence reports were freely available to anyone. Access to this information finally brought some clarity and sense of order into the prevailing chaotic post-election events. NGOs were able to use the data to target humanitarian response (Shirky 2010), and the Kenyan government also adopted a softer approach, since governments all around the world tend to act less violently towards their citizens when they are being observed. A Harvard study (Meier 2008) concluded that Ushahidi was more efficient at reporting localised acts of violence as opposed to other active media in the region. Since then Ushahidi was used to collectively amass knowledge in a number of politically motivated violence hot-spots, and natural disasters, including recent violence outbreaks in the Dem. Republic of Congo, earthquakes in Haiti, and Chile, or Washington D.C.'s winter storms during the early 2010 (Ushahidi 2010).

The NASA Clickworkers project was a pilot study by several NASA employees to assess whether public volunteers, each working for a few minutes here and there, could perform some

repetitive and routine scientific analysis tasks⁵. The work consisted in marking craters (by marking four points on a crater rim to draw a circle) in the imagery data from the Mars Viking Orbiter. A second task was also set, in which users had to categorise the age of craters (Barlow 2000 as cited in Kanefsky et al. 2001). The two sub-goals of the study were, (1) to find whether people are interested in volunteering their free time for routine scientific tasks, and (2) whether the public has the training and motivation to produce accurate results in a scientifically important task. The results were reported upon in Kanefsky et al. (2001). In conclusion, the quality of markings showed that the computed consensus of a large number of Clickworkers was virtually indistinguishable from the inputs of a geologist with years of experience in identifying Mars craters⁶. The important element in this application was the sheer number of participants – over 85,000 users visited the site within the first six months of the sites operation. Over 1.9 million entries ensured high redundancy and averaged out any errors made by individuals, and effectively the consensus opinion of what would make out a crater on the imagery data would be collected. Indeed, NASA Clickworkers is an early example of web 2.0 application, at a time when web 2.0 still wasn't a defined concept. It took even NASA over 6 years to act upon this initial success. NASA is now launching the collaborative citizen scientist initiative – the interested reader can look up further details at <http://www.nasa.gov/open/plan/peo.html>.

Ebird⁷ is a web 2.0 based project launched back in 2002 by Cornell University, in order to gather basic data on bird abundance and distribution at various geographical locations throughout time. The individual amateur bird-watchers / users are motivated to share their observation data by having access to their past observations and the collective observations generated by others in the region. Initially met with small success the application grew rapidly and became much more popular with bird watchers in the recent years. This may be attributed to the improvements in the sites interactivity and responsiveness, as well as internet accessibility. Quality of the data is ensured by client side validation checks which ensure that data is of sufficient accuracy in order to be used in scientific research. Bird observations are now accepted for both hemispheres and users can share observation reports on their bird sightings from a number of web enabled mobile devices, and 'in the wild' bird watching kiosks. Quite recently Ebird was used to track the Brown Pelican population after the gulf Oil spill in

5 Problems chosen had properties of being time-consuming to solve, difficult to automate and scientifically important.

6 A systematic comparison of thousands of individual Clickworker inputs to the known, already catalogued craters showed the Clickworkers coming within a few pixels of the accepted catalogue positions (essentially within the precision of the catalogue itself). Accuracy could further be improved by cross-checking redundant inputs from different clickworkers. Faint craters classed as having little to no detectable “ejecta blanket” were detected with an impressive 95% accuracy on a sample – see (Kanefsky et al. 2001).

7 <http://www.ebird.org> also see <http://ebird.org/ebird/eBirdReports?cmd=Start> for an exploratory data analysis.

the gulf of Mexico.

As can be seen the collaborative potential that can be facilitated by employing web 2.0 principles and techniques is quite conspicuous from the four case studies presented above. In order to understand web 2.0 fundamentals and how these recent developments in technologies and principle can be used, it is essential to set the concept of web 2.0 into its historical, social and economic context – hence we present an elaborate discussion in this respect thorough the next section. We build upon academic work from fields ranging from economics and media studies to philosophy, in order to explain effects that web 2.0 has had, and to better understand what its boundaries and implications are.

2.2 Web 2.0 in Historical, Social and Economic Context

In order for the wider academic community to grasp the significance of the collaborative web, we felt it important to illustrate it in its interdisciplinary context. A reflection over the historical developments of the second generation of world wide web is presented, and reasons for its uptake are postulated. An analysis of web 2.0 in terms of its influence onto Media Studies and Economics is also provided.

2.2.1 Historical Perspective

Tim Berners-Lee, the innovator of the world wide web⁸, in an IBM developer works interview (Laningham 2006) when asked what opinion he had of web 2.0⁹, argued that there was little need for the term at all, since [as he understood] it would imply that collaboration on a wide and interactive scale was not the original goal set out with the world wide web, which in fact was his intention from the beginning anyway¹⁰. At the same time in Berners-Lee and Fischetti (1999), he acknowledges that all his efforts in this direction went astray and also explains the reasons for this; “[...]. *Part of the reason, I guessed was that collaboration required much more*

8 Berners-Lee submitted the first proposal for the WWW during March 1989 while working as a fellow at CERN.

9 Question asked by Laningham: “You know, with Web 2.0, a common explanation out there is Web 1.0 was about connecting computers and making information available; and Web 2.0 is about connecting people and facilitating new kinds of collaboration. Is that how you see Web 2.0?”

10 Berner-Lee's answer: “Totally not! Web 1.0 was all about connecting people. It was an interactive space, and I think web 2.0 is of course a piece of jargon, nobody even knows what it means. If web 2.0 for you is Blogs and Wikis, then that is people to people. But that was what the web was supposed to be all along. And in fact, you know, this web 2.0, means using the standards which have been produced by all these people working on web 1.0 [...]. So web 2.0 for some people it means moving some of the thinking client side so making it more immediate, but the idea of the Web as interaction between people is really what the web is. That was what it was designed to be as a collaborative space where people can interact. Now, I really like the idea of people building things in hypertext, [...] and I think that Blogs and Wikis are two things which are fun, I think they've taken off partly because they do a lot of the management of the navigation for you and allow you to add content yourself. [...]

of a social change in how people worked [...] As a medium it grew very global and became more a publication medium but less of a collaboration medium.”. It took many years for the web to evolve into what it is today, and even though human nature is a social and collaborative one by default (Shirky 2010), the collaborative usage of world wide web did not really take hold until the early 2000s¹¹. This is not to say that earlier efforts didn't exist.

Here is an excerpt (Berners-Lee and Fischetti 1999) describing Berners-Lee's efforts for a read and write web in the very early 90s; *“Although browsers were starting to spread, no one working on them tried to include writing and editing functions. There seemed to be a perception that creating a browser had a strong potential for payback, since it would make information from around the world available to anyone who used it. Putting as much effort into the collaborative side of the web didn't seem to promise that millionfold multiplier.”* (on p. 61) , he further concludes; *“Without a hypertext editor, people would not have the tools to really use the Web as an intimate collaborative medium.”*¹² It is noteworthy that as early as 1995 there were efforts by W3C¹³ to understand the web in terms of its collaborative potential – a workshop on WWW and Collaboration (W3C 1995) was held¹⁴. Unfortunately many early efforts by W3C received little attention from the wider academic community. This was, in part, due to the nature of W3Cs' work – a commitment to develop WWW technology standards – as adoption of these standards was exceptionally slow, especially during the 90s (Zeldman 2007). But also, in part, due to the initial experiments with WWW collaborative tools not gaining enough traction, for example Gramlich (1995) listed major technical barriers and quite simply the lack of user participation as failures to his project. It is clear that ideas behind collaborative use of WWW, i.e. what web 2.0 stands for, were not recent but have been around for a while. Tim Berners-Lee's vision of the world wide web was for a tool which created, gathered and allowed to share knowledge through human interaction and collaboration. It seems that Berners-Lee himself did not quite understand how this would exactly become possible, therefore it might be useful to look at the web as Charles Leadbeater (Leadbeater 2009) suggests – “Web 2.0 is simply a stage of development in which the web is progressing towards this goal”.

11 Early examples of first collaborative use successes were Ebird or NASA Clickworkers and many others followed in the early 2000s, for further early case-studies we recommend Tapscott and Williams 2008.

12 We would like to note; Berners-Lee describes a relatively simple hypertext editor and not page edit functionalities directly plugged to storage systems of web-applications which would not become more common until late 90s.

13 World Wide Web Consortium (<http://www.w3.org/>) is the main international standards organisation for the world wide web, founded and headed by Tim Berners-Lee.

14 For further very early work on collaboration we refer the interested reader to the relevant working group archive, accessible at <http://www.w3.org/Collaboration/>

In many respects the burst of the dot-com bubble¹⁵ had long lasting effects on the internet, ranging from economical to trust issues. Even to this date the NASDAQ¹⁶ index has not yet recovered to its average values from before the dot-com bust. The crisis had at least three profound effects on the world wide web; 1- Trust in online businesses was lost, 2- Investments were pulled out from technology companies hence innovation slowed down, 3- The companies that survived the bubble were considered to be doing something better than the rest. As O'Reilly notes in his influential blog post (O'Reilly 2005) – *“the need for web 2.0 was spawned by an apparent realisation that the dot-com collapse marked some kind of turning point for the web”*. Indeed, to some extent just jargon, but nevertheless web 2.0 is often used to convey this new situation in which the world wide web will not “fail us again”, like it had in March 2000. Slight technological advances, but mainly improvements in trust, and emphasis on economically sustainable business models were behind a worldwide internet recovery. In the bullet points below, a number of causes for the uptake in collaborative web usage are suggested – in terms of what it is that has changed since the events of the dot com crash:

- Trust in the web grew with better legal regulations and numerous security improvements of online cash transactions. Also the emergence and growing popularity of online services such as PayPal, Amazon or Ebay and national store chains (Tesco, John-Lewis, etc...) had the effect of large number of online users becoming more confident in online transactions and in some way making these more commonly accepted within wider society (Hoffman et al. 1999, Chen and Barnes 2007). Finally trust also emerged due to standardisation of the underlying web mark-up technology (HTML, CSS, etc...), which made the web experience more consistent among browsers (Zeldman 2007). Questions related to trust are investigated empirically in chapters 4 and 5.
- Viability of online business models became a tangible reality (Li and Bernoff 2008). Online advertising became a business model capable of supporting practically any websites with enough visitor numbers. Assuming a common click-through rate of around 2%, most higher to average traffic websites can usually generate considerable profits purely from online advertising¹⁷. Other web based income generating revenue models are discussed in Lindmark (2009). Further to this, many niche and smaller businesses which built an online presence were now able to implement their proved and tested business models with online cash payment facilities, largely thanks to the increased trust into the web. Innovative new business models also developed, such as the group buying model (Wang 2009), where a number of consumers act together in an

15 A highly speculative period covering approximately 1995-2000 during which Internet related companies became significantly overvalued. The turning point (bubble burst) is taken to be NASDAQ's price peak from March 10th 2000.

16 A good measure for the performance of technology companies is the NASDAQ composite stockmarket-index, due to its bias towards technology stocks.

17 It has been recently, convincingly shown that display based online advertising is actually very effective. The interested reader can consult this comScore Inc. research paper (Fulgoni et al. 2010). Also, as of 2007, 67% of the top 137 web 2.0 websites were using advertisement as their business revenue model (Slot and Frissen 2007).

online community to negotiate better deals and discounts for products and/or services¹⁸. This business model became hugely popular with a large number of localised group buying websites appearing all over the world. Finally let us not forget the overall internet economics – that is web-design, online advertising, infrastructure, database and data modelling and more recently cloud computing, and cloud data storage have made the internet industry very substantial in terms of economic scale and turnover (Lindmark 2009, pp. 28-29). Also, see (Lindmark 2009, pp. 47-56) for a discussion on web 2.0 employment and venture capital investments, both of which are substantial. Popularity of business models on web 2.0, by the public, is analysed in chapter 4.

- Standardisation of web mark-up made it also possible to employ more advanced and powerful client-side scripting techniques. One of these is the well known AJAX (Asynchronous JavaScript and XML) technology which made it possible for web-pages to only send partial-page server requests to a server¹⁹. In other words, the entire web-page in a browser did not need updating, as a small part of the page could be updated with new data from the server asynchronously based on, for example, some user activity in the rest of the page. This programming technique had a major effect on interactivity within web applications, and perceived latency for web-pages was significantly reduced. This also meant that input / output complexity – multiple levels of input became possible (atomic elements described in section 2.3.2.2, can be nested). Emergence of the above, and of the elements of the two bullet points below are analysed empirically in chapter 5.
- Development became easier – productivity tools and easy to use server side scripting languages as well as client side scripting libraries appeared²⁰. These are often open-source and therefore initial web development start up costs can be small. Further to this, the default settings in these frameworks allow to use many elements of web 2.0, “out of the box”.
- Architecture / Integration advances made the web a true application platform, since these advances make it possible to loosely integrate a number of entirely separate web pages together. At the core of this are Web Services technology and RSS (Really Simple Syndication) which made information exchange between websites possible, and hence these applications gained a global (above a single web-application) scope (see Schroth and Janner 2007, Wilde 2010). Thanks to these integration technologies but also to the standardisation efforts, world wide web content became compatible over a heterogeneous set of mobile platforms. In more technical terms, browsers act as abstractions for platform, this is extremely significant since it stipulates a shift to completely online applications – in fact this is already happening with laptops, often

18 There has been relatively little academic research into online group buying applications; however, what stands out is the major difference between Asian, European and North American group buying. Examples of successful group buying applications include: www.Groupon.com, www.MyCityDeal.com, www.BuyWithMe.com, www.twangoo.com

19 Frames have also been used to this effect; however, their implementation tends to be cumbersome and inconsistent. Flash (the ubiquitous plug-in from Macromedia, now Adobe), and Microsoft's Silverlight are also alternatives to AJAX. More recently HTML 5 has introduced several features which facilitate AJAX like behaviours. Google Gears technology has also provided some of this functionality; however, with the adoption of HTML 5 it will be obsolete.

20 Ruby (Ruby On Rails), Python (CherryPython), PHP (Drupal, Wordpress), JavaScript (jQuery, Scriptalicious). In brackets are examples of open source frameworks and libraries, widely used in web 2.0 application development.

referred to as net-books being produced partly to satisfy this new demand.

Due to the latency, interactivity, trust, standardisation and architectural developments as highlighted in the points above, collaborative and social use of the world wide web has emerged to be worthwhile. This is important since in aggregate it makes social media possible (social media and web 2.0 are considered synonymous throughout this thesis²¹). It is not to say that collaborative-capable elements materialised all at once, instead they transpired progressively. In order to understand the historical development of web 2.0, a quantitative and systematic investigation of the evolution of websites throughout time is undertaken in chapter 5.

2.2.2 Social Perspective

Recent years have seen a proliferation of web 2.0 applications, an increased usage by the masses, and not just by the younger generations²². In fact due to the large participation of web 2.0 users, that is, users who themselves often contribute by sharing links, commenting, rating or performing other social actions (since these are the sort of participatory activities web 2.0 applications facilitate), these web 2.0 applications also became better known as “social media”²³. Numerous prominent and influential journalists are now blogging and tweeting on the web. Well known celebrities, politicians and entire political parties joined social networks such as facebook or micro-blogging application Twitter. News agencies, traditional TV and radio stations are also building up a heavy presence on various web 2.0 applications, such as Reuters, ABC, FT, CNN news (Sykora and Panek 2009). It is quite clear now, that social media has taken a share from traditional mass media as the wider public is increasingly becoming engaged, and mass media demographic usage patterns are now starting to change (Mares and Woodard 2006, Bond 2008). A definition of social media is provided in Kaplan and Haenlein (2010) – “*social media is a group of Internet-based applications (web 2.0 applications), [or in other words] applications that build on the ideological and technological foundations of web 2.0, which allow the creation and exchange of user generated content.*”

Numerous academics have studied and analysed how web 2.0 based social media fits into the

21 Social media tends to be used by media scientists more frequently than the term web 2.0, which tends to be more common with other vocational fields. Although Kaplan and Haenlein (2010) tried to define a distinction between the terms, they did not succeed to make a strong point. When the term media is constrained to web based media, then as far as the author is aware there is nothing to suggest that social media and web 2.0 should not be treated as synonymous terms.

22 According to Forrester Research, 75% of Internet surfers used web 2.0 applications in 2008 (instead of 56% in 2007) by joining social network websites, reading Blogs, or contributing reviews to shopping sites. The research (accessible here: <http://bit.ly/9jIkB6>) also found that 33-45 year-olds increasingly participate in web 2.0 usage. Age related and many other world wide web statistics are available in Pascu (2008).

23 “Social media are media for social interaction, using highly accessible and scalable publishing techniques. Social media use (often web-based) technologies to transform and broadcast media monologues into social media dialogues.” - Wikipedia (http://en.wikipedia.org/wiki/Social_media)

existing picture of mass media. Traditional media allowed for one-to-one conversations (e.g. telephones, mail) or broadcasts to groups (e.g. television, newspapers), but not both. Now the two modes of personal media (letters and phone calls made by ordinary citizens) and public media (visual or print communications made by a small group of professionals) have been fused together through social media (Shirky 2009). It has been said that this empowers creativity, democratises media production, and celebrates the individual (Zimmer 2008). Subsequently, social media is having rather interesting effects in a number of areas. Researchers within sports science are finding that the traditional parasocial²⁴ relationship that used to exist between fans and sports athletes is changing to a more direct, equal and informal one (Kassing and Sanderson 2010). Twitter and other social media provide enhanced access in the athlete–fan relationship. This has been considered “good” and “bad” in some cases. For example: on August 26, 2009, Lance Armstrong invited his Twitter followers to join him for a ride that evening around a local park via a seemingly innocuous tweet (i.e., “*Good morning Dublin. Who wants to ride this afternoon? I do. 5:30pm at the roundabout of Fountain Road and Chesterfield Avenue. See you there*”). Amazingly, over 1,000 people showed up to join him for the ride (Kassing and Sanderson 2010). Ways in which athletes have experimented with Twitter have also raised questions of privacy, appropriate disclosure and governance of Twitter by sports organisations. For example NBA player Michael Beasley tweeted about checking into a Houston rehabilitation facility after he tweeted suicidal thoughts. Other players have tweeted during games, which has now been disallowed by NFL (Pegoraro 2010). The question of how organisations respond to employees’ use of social media remains, and will be interesting to observe as a clash of privacy boundaries will likely occur. Communication over social media, and the relationship between customers and manufacturers / service providers, is also changing expectations of consumers. There is evidence that companies or political parties are seemingly under pressure to listen to individual concerns more than they have in the past (Li and Bernoff 2008, Jackson and Lilleker 2009). Other specific examples of social media use with consideration for its implications to society are discussed in section 2.4.

Social media is further characterised by its global nature – nearly 2 billion users all around the world are online. It is social in that the media is participatory in terms of social engagement. It is ubiquitous – computers, mobile devices, broadband internet and wireless networks have become cheaper and more accessible than ever before and finally social media is cheap – publishing barriers in terms of costs have practically disappeared, e.g. virtually anyone can start a Blog (with no technical knowledge) which is instantly accessible to a mass worldwide

24 A term used by social scientists to describe a one-sided, “parasocial” interpersonal relationship in which one party knows a great deal about the other, but the other does not.

audience. David Silver (Silver 2008) concludes his critical paper on impact of web 2.0, by aptly pointing out that, “*we are witnessing the birth of a new writeable generation, a generation of young people who think of media as something they read and something they write – often simultaneously.*”

2.2.3 Economic Perspective

Let us now consider the value and the impact of these developments onto the field of economics itself. As some academics like to put it, during the 19th century onwards, distribution of information, knowledge and culture became industrialised (Benkler 2006, Shirky 2010). The steam powered printing press and other expensive machinery and methodologies were required to run, print and distribute the necessary volumes of newspapers. Later with television there was a need for highly qualified workforce and expensive studios. This created a professional class of producers and a large group of (mostly passive) consumers. With social media we have now gained the ability to balance consumption with sharing and our own content production, hence the internet effectively thins the line of separation between “amateurism” and “professionalism”²⁵. Publishing costs, online, have virtually disappeared. Costs associated with collaborating in groups or coordinating groups have also collapsed, examples of this are Wikipedia, Ushahidi, Ebird, or the open source Apache or Linux movements²⁶.

The virtual disappearance of group coordination costs is the basis behind “social production”²⁷, a model of economic production first suggested by Harvard professor Yochai Benkler (Benkler 2002), and later made popular in his book *The Wealth of Networks: How Social Production Transforms Markets and Freedom* (Benkler 2006). In 1937, the economist Coase asked – if markets are efficient why and under what circumstances do people organise themselves into managed groups or firms, given that production could be carried out without any organisation; why would an entrepreneur hire help instead of contracting out for some particular task on the free market. It turns out that the transaction costs on the market may become a barrier (Coase 1937), so that, where the cost of achieving a certain outcome through organisational means is lower than the cost of achieving that same result through implementation of the price system, organisations will emerge to attain that result. Benkler postulated that under certain circumstances, non proprietary, or commons-based peer production may be less costly in some

25 For further discussion on the idea of amateurs vs. professionals, see Shirky (2010), pp. 56-62, 152-155 or Keen (2007).

26 Linux and Apache are open-source projects. Open-source coordination has been facilitated for a long time through non web based protocols. It was therefore possible for technically minded individuals to reap the benefits of collaboration via the Internet long before world wide web has developed the numerous characteristics of web 2.0

27 The terms social production and peer production will be used interchangeably within this thesis.

dimension than either markets or managed hierarchies (firms). One could say that when the cost of organising an activity on a peered basis is lower than the cost of using the market, and lower than the cost of hierarchical organisation, then peer production will emerge (Benkler 2002), see table 2.1.

Table 2.1 – Organisational forms as a function of firm-based management vs. market vs. peering (source: adapted from Benkler 2006)

Market exchange of x cheaper than organising / peering x	Pure market
Organising x cheaper than market exchange or peering of x	Market with firms
Peering cheaper than both market exchange and organisation	Proprietary “open source” / “peer production” efforts

The idea of peer production as an alternative or complementary economic mechanism for achieving economic goals is an attractive one, but more importantly it highlights the impact that proliferation of web 2.0 has had. Peer production is not of main interest within this Thesis, but where relevant to web 2.0, it will be mentioned throughout parts of the thesis (for example in the next section, 2.2.4).

2.2.4 Critical Perspective

Despite much enthusiasm there are many problems and critical opinion surrounding web 2.0. The issue of children's vulnerability has been raised in the past (Marwick 2008) and privacy in general is a major concern, with the increased flow of personal information across social media (Gross and Acquisti 2005, Zimmer 2008). However, the set of issues is much wider, and ranges from criticism concerned with information saturation, to “slave labour” in peer production. In fact an entire issue of the internet journal FirstMonday (Zimmer 2008) was dedicated to critical views of social media, and its unintended consequences. This section presents a critical overview of some of these issues.

2.2.4.1 Information Overload

Information overload (i.e. threat of paralysis due to information overload) has received considerable attention. Kassing and Sanderson (2010) studied the communication over social media (Twitter) between athletes and fans. They concluded that it is unclear how fans will keep up with their many sports idols – basically, at what point will Twitter saturate users' tolerance for receiving communication. Kassing and Sanderson conclude that there is likely to be a burnout factor for fans, i.e. too much twitter accounts being followed will prove problematic.

In technical literature the related field of email filtering and spam detection has been well

researched, e.g. (Androutsopoulos et al. 2000). Other practical research with intention of combating information overload included work on quality assessment of web-page content (Blumenstock 2008), or detecting disputed topics (Ennals et al. 2010), where the authors developed browser plug-ins that would scan web-page content against topics in a disputes database and highlight these in the browser, hence shortening the amount of time for readers to background-check certain information, and help cope with information overload. Simpler tools such as Tag-clouds or RSS feeds also help to cope with information overload, since only the important topics and news-items can be followed by choice. In line with peer production, Benkler describes Slashdot's²⁸ elaborate peer review system (Benkler 2002), and points out that the same dynamic that is used in peer production of content can be implemented to produce relevance and accreditation, and effectively peer review content. This seems to be a reasonable solution for larger online communities; however, integrating effective information filtering techniques and tools within social media and web 2.0 platforms is a major challenge. Several online services facilitate the detection and recommendation of individuals on social media worth following (e.g. Klout²⁹); however, in the sports example, above, the problem is not in whom to follow (fans know their favourites) but what information (i.e. tweets) should be filtered as mundane or uninteresting. Unfortunately what some may find mundane, others may care about; hence, there is a need for more intelligent data filtering.

A related set of issues are concerned with data-items that are not intended for sharing and publication – this concerns the wider issue of privacy (Solove 2008). Since the uptake in social media should not imply that sharing is necessarily a good thing, at all cost. Despite being difficult to implement in practice, a salient reduction in publishing (public vs. private information) might help combat information overload to some degree³⁰.

2.2.4.2 Motivation and “Slave Labour”

It was suggested by several academics (Petersen 2008, Scholz 2008, Shirky 2010) that the treatment of individuals participating in peer production, by corporations, in some instances is

28 Slashdot is a technology related news website that features user submitted stories that appear as discussion threads for comments. The comments and news item selection is peer reviewed by the Slashdot members themselves in a rather intricate peer review system, closely described in Benkler (2002).

29 A social media analytics service that computes influence scores based on social media activity of users and their popularity <http://www.klout.com> (also see <http://en.wikipedia.org/wiki/Klout>). Similar services, such as <http://www.peerindex.com/>, <https://www.twentyfeet.com/>, <http://www.postrank.com/> and <http://wefollow.com/> also exist.

30 See an example, relating to social networks at <http://youopenbook.org>. The quantity of data that would be saved from being published is difficult to estimate (for numerous reasons), but it might arguably be quite sizeable in aggregate, if numerous high traffic social media are considered. An interesting recent proposition made by Prof. Viktor Mayer-Schonberger concerns the idea of assigning expiration dates to data-files to reduce information overload and in order to combat certain social trends, highlighted in more detail within Mayer-Schonberger (2009).

equivalent to “slave labour”. Petersen states that web 2.0 represents, “*an architecture of exploitation that capitalism can benefit from.*” The criticism put forward is serious. Corporations have been known to claim ownership over content produced by users, which is a very explicit form of exploitation. Alternatively corporations lock user data within an interface, and allow user ownership of data, but effectively this has the same effect as direct data ownership³¹. Petersen (2008) provides several specific examples of such exploitation, but finds (based on interviews) that users whom he may see as being exploited, do not see themselves as such. Hence, it is important to discuss the motivations behind user generated content contributions within web 2.0.

In 1972, Edward L. Deci investigated motivations in voluntary engagement by letting a group of students play a puzzle game and experimenting with various rewards to motivate game play, in order to observe voluntary game engagement (Deci 1972). His findings on intrinsic motivations led to the emergence of a field known within psychology as “self-determination theory”. Intrinsically motivated activities are such where an individual expects no external reward, the activity is a reward in itself – examples of such activities would be what a person does as a hobby or in their free time. It can hence be extrapolated that motives to participate in collaborative projects such as Wikipedia or NASA Clickworkers are mostly intrinsically motivated. Further studies (Deci et al. 1999, Cameron et al. 2001, Tomasello et al. 2008) have shown that in environments with high degree of freedom to choose an activity, payment can crowd out other kinds of motivations – that is when an extrinsic reward is tied to an activity people like, and then the reward is taken away, then the intrinsic motivations for that activity generally disappear. According to this line of thought, sometime monetary reward can actually be counter-productive to a voluntary activity. Benkler (2006) observes two useful dimensions for measuring social collaboration efforts: 1-modularity and 2-granularity. By modularity, Benkler means a property of a project that describes the extent to which it can be broken down into smaller components, or modules, that can be independently produced before they are assembled into a whole. By granularity, Benkler means the size of the modules, in terms of the time and effort that an individual must invest in producing them. It has been suggested that given low-participation costs in social media, contributions from many individuals who are intrinsically motivated becomes practical and highly significant when the granularity and modularity are small enough (Benkler 2006, Tapscott and Williams 2008, Brabham 2008, Shirky 2010). There have been also numerous instances of extrinsically motivated collaborative web 2.0 activities, some of these instances are referred to as “crowdsourcing”. This is where an extrinsic (usually cash) reward is used to motivate and outsource a problem solving task to a

31 Issues of copyright from a legal perspective are also discussed in Barbry (2007).

web based “crowd”. Brabham studies crowdsourcing as a field in its own right (Brabham 2008)³². As valuable as some of his observations are, crowdsourcing is mostly a buzz-word. Open-source bounties and reverse open-source bounties³³ in, for example, software feature requests are fairly common, and crowdsourcing can in many ways be seen as instances of existing open-source and web 2.0 processes or simple calls for participation in the classical sense of public tenders. Nevertheless sometime, extrinsic motivations have been used in web 2.0 systems to motivate participation (e.g. in the form of weekly random prize draws, or participation based points competitions).

Of course it is highly interesting to understand what actual intrinsic motivations push people to contribute UGC on web 2.0 systems (User Generated Content, or UGC is the acronym for any content that is contributed by users of a web 2.0 system; be it a Blog-post, social profile, pictures, comments, etc.). Indeed, recently, numerous researchers have tried to identify and understand different kinds of intrinsic motivations (Forte and Bruckam 2005, Kuznetsov 2006, Wagner and Prasarnphanich 2007, Bishr 2009). A relatively early effort was by Forte and Bruckam (2005) who interviewed 22 Wikipedians about their motives. The interview subjects were top, voluntary contributors – many of whom spent 30 hours a week working on Wikipedia. Kuznetsov (2006), took a more systematic approach, and successfully identified several kinds of motivations³⁴. These are; Altruism (*charity or generosity without any expectation of improving ones individual welfare*), Reciprocity (*reciprocity, or reciprocal altruism is the process by which a person who commits an altruistic act receives a benefit in return, perhaps at a later time*), Community (*acting for the benefit of a group of people, who regularly interact with each other and share a common set of values and needs*), Reputation (*acting for a personal benefit, to gain respect, trust and appreciation by one’s peers*), Autonomy (*the desire for freedom of independent decision making / knowledge, being in control*). Wagner and Prasarnphanich (2007) identify similar motives to Kuznetsov, and further discuss altruism and individualistic / reputation based behaviours on cooperative platforms, specifically Wikis; using a survey of 35 Wikipedians to develop their ideas. However, things get more complicated, as there are for example many altruistic motives. Bishr discusses different forms of altruism and interestingly suggests that users who contribute to Wikipedia may be acting out of self-interest / authority, rather than with a strong sense of common good, i.e. they act to compete in building

32 See Brabham (2008) for a review of several interesting extrinsically motivated web 2.0 examples. The case-studies presented in Brabham (2008) include; iStockPhoto, Threadless, InnoCentive and several advertising competition campaigns, where cash incentives were used to motivate users / producers.

33 Cash rewards are sometime offered for completing certain open-source activities, http://en.wikipedia.org/wiki/Open_source_bounty, http://en.wikipedia.org/wiki/Reverse_bounty, <https://www.bountysource.com> see for more information.

34 Kuznetsov (2006) devised these motives mainly based on a survey of over 100 undergraduate and postgraduate, university students, see her paper for more details.

an altruistic reputation. Intrinsic motivations were also investigated in a commercial / business setting (Paroutis and Saleh 2009). Arguably the intrinsic motivations may not always be strong, and prevalent. Some have criticised that most users of web 2.0 consume rather than contribute content, also known as the problem of "free-riders" (Berlanga et al. 2011; van Dijk Keynote 2008). Wikipedia is often given as an example, since it was found that only around 2-3% of the community actually contributes content (Benkler 2006, O'Sullivan 2009). Although, little other convincing empirical evidence has shown free-riding to be prevalent throughout web 2.0 applications. In fact Antin and Cheshire (2010), present compelling arguments against the hypothesis of free-riding on Wikipedia. They explain how Wikipedia readers, are more often not free-riders and in fact fulfil an important service to the community (their argument is supported by results from a 20-minute long survey of 165 respondents). Claims of free-riding are dismissed by the authors as largely unsubstantiated and vague.

Finally, as motivations within a web 2.0 environment are still not fully understood, Bishr (2009) for example, calls for studies that would investigate percentages of people with certain motives and what applications such people tend to use. In chapter 4 of this thesis, an empirical study using a survey of over 700 respondents is conducted, with a section of the survey dedicated to answering this and related questions, regarding web 2.0 motivations.

2.2.4.3 Experts vs. Amateurs

Many have criticised the democratisation of media production (e.g. Keen 2007, Anderson 2007, Lanier 2010). Publishing information to a wide audience of the public became easy and cheap. Arguably this increase in freedom to publish likely lowers the quality of the produced work. To illustrate this point further, Shirky suggests that before Gutenberg invented his press, the average book was a masterpiece, after Gutenberg, people got throwaway erotic novels, dull travelogues, and hagiographies of the landed gentry, of interest to no-one but a few historians. Today this may seem somewhat amusing but in 1569 Martin Luther observed: *"The multitude of books is a great evil. There is no measure of limit to this fever for writing, everyone must be an author; some out of vanity, to acquire celebrity and fame, others for the sake of mere gain"* (Shirky 2010). It seems that publishing freedom and quality are conflicting goals. Hence with social media the definition of quality becomes more variable within one community to the next. Nevertheless diversity expands, and it is hoped that the best work becomes better than what went before. Koltay (2011) also argues that this variability in quality does not necessarily matter, at least while users are aware of who has produced the piece of work in question. Hence identity of content producers or groups of content producers can be quite important.

2.2.4.4 Time spent on Social Media

A recent study by Nielsen Company found that the average monthly time spent on Facebook by a user is around seven hours and forty-five minutes (as of August 2011; Nielsen 2011). Two earlier studies confirm that the average time spent on Facebook has been on the increase; seven hours, one and a half years earlier (as of January 2010; Nielsen 2010) and only four hours and thirty-nine minutes two years earlier (as of June 2009; Nielsen 2009). Facebook is a particularly good example of an informal social networking site; however, what about more explicit peer production web 2.0 websites. In order to understand the significance of time spent on web 2.0 websites, let us consider the amount of time spent in front of the television per year. Assuming 20 hours / week³⁵ of television are watched by an average person, then that represents more than 1,000 hours per year and more than 50,000 hours in 50 years of a hypothetical lifetime. In comparison, 100,000,000 hours were spent on aggregating knowledge into Wikipedia to date (as of April 2008 – source: Shirky 2009³⁶). To put this into perspective, assuming 1,000 hours of television watching a year, for an average adult US citizen, this would amount to a total of about two hundred billion hours of television each year, in the US alone. Compared to hundred million hours spent on developing the entire Wikipedia content, the time people spent on social media is still negligible³⁷ in comparison to television. Since social media is becoming more of a leisure activity (such as television watching), it is possible the trend of spending more time on social media will keep increasing (Mares and Woodard 2006, Bond 2008, Radwanick 2010). At the moment it is unclear and difficult to estimate the rate of such trends; however, some companies also encourage their employees in the use of social media at work (Dwyer 2007, DiMicco et al. 2008). On the other hand, as was briefly pointed out in section 2.2.2, participation in social media is not always encouraged by employers and responsible organisations. It will therefore be interesting to see how much time in various demographics people will be willing and happy to engage with social media in future.

With the excitement over collaborative abilities facilitated through web 2.0, numerous buzzwords began appearing throughout blogs, forums, twitter feeds and other media, but unfortunately also within academia, executive circles and professional industry bodies

35 This is a lower estimate, the amount of time is actually higher on average, for more complete statistics see the Sourcebook for teaching science website; <http://www.csun.edu/science/health/docs/tv&health.html>

36 This estimate considered total amount of time people spent on every edit in an article, every argument about those edits and for every language that Wikipedia supported. See Shirky (2010), pp. 9-11 for an explanation of this simple comparison between Wikipedia contribution and TV consumption.

37 It must be pointed out that TVs are located in certain places only, limiting access to them (physically). While social media is available on laptops / mobile devices, and a lot of people include social media use in their work routine. The potential reach of social media is hence noteworthy.

(Surowiecky 2005, Brabham 2008, Jackson and Lilleker 2009). Terms such as perpetual beta, collabularies, neogeography, crowdsourcing, web 1.5, and other, contribute towards a certain state of confusion. This can be expected with an emerging field that is in the process of establishing itself, and as concepts within it are being defined. However, it is essential that these developments are integrated with previous academic work³⁸. As it was already expressed in chapter 1, it is hoped that this thesis contributes towards a better understanding of web 2.0 and the related concepts of web based collaboration. To that end, the next section (2.3) addresses the various categorisations of web 2.0 that have emerged over time.

2.3 Web 2.0 Taxonomies

Since there is a multitude of social media applications³⁹ there is a desire to categorise these in a systematic manner (O'Reilly 2006, Hearst 2009, Leadbeater 2009, Shirky 2010, Kaplan and Haenlein 2010). In this section the various taxonomies that have been proposed in literature, in order to categorise and organise web 2.0 applications in a meaningful way, will be reviewed. It is concluded that a more systematic taxonomy may be beneficial, in view of existing work, and hence, a new 2-step based taxonomy is presented in sub-section 2.3.3.

In order to aid discussion of the presented taxonomies, the following web 2.0 applications will be used as examples; Picassa, Flickr, Ebay, Youtube, Craigslist, Ushahidi, Digg, Slashdot, Reddit, Facebook, Foursquare, Twitter, Amazon and Writely⁴⁰. These are just some examples of web based systems which allow user contributions and social sharing compatible with web 2.0 characteristics described so far.

2.3.1 Existing Taxonomies

2.3.1.1 Degree of Connectedness, Collaboration, and Collective Intent

The meaning and definition of what type of web-applications are representative of web 2.0 has had the tendency to be misunderstood (Fallows 2006, O'Reilly 2006). This is because the

38 The appearance of jargon has severely limited other areas of research (e.g. Technical Analysis within the financial markets) hence it is crucial that any potential jargon is put within context of previous research (Schwager 1993).

39 There are a number of comprehensive online based lists of web 2.0 systems, see Lindmark (2009), pp. 73.

40 Picassa, Flickr are picture sharing and Youtube video publishing applications, Ebay is an online auction site, Craigslist is a network of online classified advertisements, Digg, Slashdot and Reddit are social news websites where the community votes news items up or down (effectively democratising the news publishing process), Facebook and Foursquare are social networks with the latter a location-based social network website, Twitter a well known micro-blogging website (only messages of 140 characters can be sent, it is also a social networking capable system), Writely / Google Docs is a free web-based word editor and allows collaborative document editing and document tagging, Amazon is a major online retailer with social sharing capabilities.

original definition was proposed by a set of subjective examples, over 7 major characteristics (O'Reilly 2005), and it therefore tends to occur that some characteristics are given undue weight⁴¹. For example AJAX (Asynchronous JavaScript and XML) technique⁴² is frequently considered to be essential for any web 2.0 application; however, this is not necessarily the case⁴³. Due to this criticism and since the definition of web 2.0 can seem somewhat unbounded, O'Reilly proposed a scale based hierarchy of “*Web 2.0-ness*” (O'Reilly 2006). *Web 2.0-ness* is a scale ranging from 0 (least web 2.0) to 3 (most web 2.0). Classifying an application on this scale involves asking the question of connection indispensability; that is, *whether connection to the web is indispensable for the given application*. The categories (or levels) are:

- **Level 3** – could *only exist on the web*, fully essential is the network and the connections it makes between applications or people, e.g. Ebay, Craigslist, Ebird, NASA Clickworkers, Ushahidi, Twitter, Facebook, Wikipedia,...
- **Level 2** – *Unique advantage of being online* is gained, but application could exist offline as well, e.g. Flickr, Picassa,...
- **Level 1** – Gains *additional features by being online*, but can mostly justify its existence offline, e.g. Writely, Google-Docs,...
- **Level 0** – Would *work offline* with all the data in a local cache, e.g. MapQuest, Yahoo!Local, Google Maps,...

This categorisation is not precise by any means, and as O'Reilly (2006 – personal online discussion) points out it wasn't his intention to make a fit-all categorisation of web 2.0-ness; “*Words are pointers, and yes, they have some baggage. I tend to think not of narrow and precise boundaries to a concept like this [the web 2.0 categories, suggested], but rather, a gravitational core. And a metaphor is just that: an aid to perception and thought, not a bounding box.*”

Leadbeater (2009) builds on O'Reilly's definition and suggests a scale (level 1 to 3), based on the level of collaboration rather than the connectedness. The level of collaboration for *level 1* emphasises that; collaboration is a useful, indirect and unintentional by-product of a singular activity, at *level 2*; collaboration is a deliberate purposeful activity, and is concerned with creating content for a fairly well-defined end-goal, at *level 3*; collaboration itself is the primary purpose of activity.

41 For a good discussion of the seven characteristics as proposed by O'Reilly (2005), we recommend Anderson (2007), pp. 14-26.

42 AJAX actually isn't a new technology in the classical sense, but rather a combination of client side scripting (usually JavaScript) on the Document Object Model of a web-page, XML / JSON or some other data serialisation type for data transmission, and XMLHttpRequest object are used together, to implement partial page-requests – i.e. AJAX.

43 AJAX contributed to the uptake of collaborative web usage, since web interaction became more responsive and complex web interfaces became feasible; however, this is not to say that AJAX is indispensable for web based collaboration. Other characteristics such as sharing, platform interoperability, lightweight programming models and loosely coupled infrastructures are also important.

It would seem that in different situations it may be useful to apply other dimensions for classifying various 2.0 applications. For example, where online community potential is an important element of the web 2.0 tools being classified, a classification based on the degree of collective intention, rather than collaboration itself would be more useful. The philosopher Gilbert (2006) explains that collective action is interpreted as a matter of people doing something (a task) together. It is assumed that this involves their having a collective intention to do that task together, where the parties are jointly committed, and this joint intent tends to be binding. A useful and thorough explanation of joint intent and joint commitment is given in Gilbert (2006). The degree of collective intent can be useful in assessing and auditing the usability and feasibility of a web 2.0 application. A scale with three levels could look as follows; *level 1* – no collective intent seems to be explicitly defined, *level 2* – there is some explicit collective intent, *level 3* – the collective intent is strong and very well defined, i.e. collective intent and elements of joint commitment can be arranged collectively by the users themselves.

The above dimensions of collaboration, connectedness, and collective intent might be helpful in describing the nature of some web 2.0 applications; however, ultimately they must only serve as an aid, and if used carelessly may in fact contribute towards more confusion than clarity.

2.3.1.2 Other Taxonomies

Marti Hearst in Hearst (2009) proposes eight types of social technologies⁴⁴, which facilitate: *1-Recruitment of Outside Expertise, 2-Crowdsourcing, 3-Data sharing, 4-Shared virtual world platforms, 5-Collaborative creation, 6-Social networking, 7-Idea market trading, or 8-Implicit contributions*. Most public use of social media is concentrated in three groups of activities; 3-Data sharing, 5-Collaborative creation and 6-Social networking (Pascu 2008). Similar classifications of web 2.0 applications were also proposed by Anderson (2007, pp. 7-13) and Lindmark (2009, pp. 15-18)⁴⁵. Nevertheless, only three categories appear consistently throughout these taxonomies, i.e. **Wikis (collaborative creation)**, **Blogs / social tagging / media sharing (data sharing)** and **Social networking** (Anderson 2007, Lindmark 2009, Hearst 2009). Anderson further suggested that since original web 2.0 applications are appearing with a high frequency, they deserve a miscellaneous (other or uncategorised) category; hence,

⁴⁴ Marti Hearst uses the term social technology to refer to social media in her work. We consider these terms equivalent, as our definition of social media is wide and this is preferred in favour of introducing new terms.

⁴⁵ Another categorisation is offered on <http://www.web20searchengine.com/web20/web-2.0-list.htm> where groups are based on the function of the web 2.0 applications – i.e. music, time management, document processing, etc. An extensive list of 1,000 major applications is classified. For a simple and basic tutorial on Blogs, Wikis, and other web 2.0 applications, see Anderson (2007). Also other extensive lists of web 2.0 applications available online, are provided by Lindmark (2009), pp. 73.

Anderson suggested; *1-Blogs, 2-Wikis, 3-Social tagging, 4-Multimedia sharing, 5-Audio blogging and Podcasting, 6-RSS and Syndication, 7-Others (Crowdsourcing, Social Networking, Mashups, Uncategorised)*. Lindmark proposed the following; *1-Blogs, 2-Wikis, 3-Social tagging, 4-Social networking, 5-Multimedia sharing, 6-Social gaming (virtual online worlds, and MMOGs) and 7-Other applications (RSS, Mashups, Podcasting, Micro-Blogging, standard apps. with web 2.0 features)*.

Kaplan and Haenlein (Kaplan and Haenlein 2010) propose six different groups of categories for social media which rely on research in the field of media studies and social processes – the two key elements of social media. The media-related component of the classification relies on social presence theory (Short et al. 1976) and the closely related idea of media richness (Daft and Lengel 1986). In brief⁴⁶ social presence is concerned with the acoustic, visual, and physical contact that can be achieved, dependent on ideas of intimacy, and immediacy of a communication medium. Media richness is the amount of information that the medium allows to be transmitted in a given time interval. With respect to the social component of the classification, the concept of self-presentation (Goffman 1959) and self-disclosure (Schau and Gilly 2003) are important. In brief, self-presentation assumes that in any type of social interaction people have the desire to control the impressions other people form of them, whereas self-disclosure is the conscious or unconscious revelation of personal information that is consistent with the image one would like to portray. Combining both (social and media) dimensions together leads to a classification of social media which is presented in table 2.2 (Kaplan and Haenlein 2010).

46 For a much more detailed explanation of the definition Kaplan and Haenlein (2010) is recommend.

Table 2.2 – Classifications of social media by self-presentation / self-disclosure and Social presence / Media richness

	Social presence / Media richness			
		<i>Low</i>	<i>Medium</i>	<i>High</i>
Self-presentation / Self-disclosure	<i>High</i>	Blogs	Social networking sites (e.g. Facebook)	Virtual social worlds ⁴⁷ (e.g. Second Life)
	<i>Low</i>	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. Youtube)	Virtual game worlds ⁴⁸ (e.g. World of Warcraft)

Kaplan and Haenlein's classification is based on the study of social media in media sciences, and is naturally insightful, and more systematic than the other taxonomies presented.

Somewhat related to Benkler's economic ideas of peer production (Benkler 2006), Shirky presents a delightful and equally utopian thesis of cognitive surplus (Shirky 2010). The concept of cognitive surplus is based on the idea that aggregated free time of many individuals who collaborate or simply participate in social media (and who display an amount of goodwill in their social interactions) – can be regarded as a commodity, which Shirky terms “cognitive surplus”. Shirky argues that before the internet era, managing our free time used to be a largely personal issue, more a matter of using it up than actually using it. Whereas the “wiring of humanity” lets us treat free time as a shared global resource, and lets us design new kinds of participation and sharing that take advantage of that resource, by fusing means, motive and opportunity to create cognitive surplus out of the raw material of accumulated free time (see Ushahidi for an example). In the light of this thesis Shirky presents four categories of social media use, based on the scope of sharing that occurs with the use of a particular instance of social media application: 1 – personal sharing, 2 – communal sharing, 3 – public sharing, and 4 – civic sharing. This spectrum explains the degree of value created for participants versus non-participants. A more detailed explanation of these groups is given in Shirky (2010, pp. 174-181).

2.3.1.3 Web 2.0 Users Role Classification

In his report for the European Union, David Osimo defines a simple model for the diverse user roles of web 2.0 applications (Osimo 2008). He identifies four different user roles of web 2.0 systems. The first set of users are the core users of web 2.0, those generating fully fledged

⁴⁷ Virtual social worlds are similar to virtual game worlds, however there are no rules restricting the range of possible interactions. Social worlds allow their inhabitants to chose their behaviour more freely and essentially live a virtual life similar to their real life.

⁴⁸ Virtual game worlds are platforms that replicate a game related three-dimensional environment in which a game character can appear and interact with other characters as they would in real life (constrained to rules of the game). A well known example of this is the “World of Warcraft”, and takes place within the Warcraft world of Azeroth, where game characters are elves, orcs, dwarves or humans, currently with more than 11million players worldwide.

content (e.g. Blog posts, Wikipedia articles, etc...). The second set of users are people who provide feedback, comments and reviews for existing content. The third set of users is composed of web users who access, read and watch the content produced by the first two sets of users. The fourth and last group of users represents those that don't deliberately use web 2.0 applications, but unknowingly provide input into web 2.0 driven applications (for example, some search engines convert search terms into tags which are displayed on web-pages for others to use, or when buyers purchase products on Amazon, this is exploited by Amazon to provide buying recommendations on the website to other users). This user-role classification is presented as set of concentric circles, where each stands for a different user activity, see figure 2.1 below.

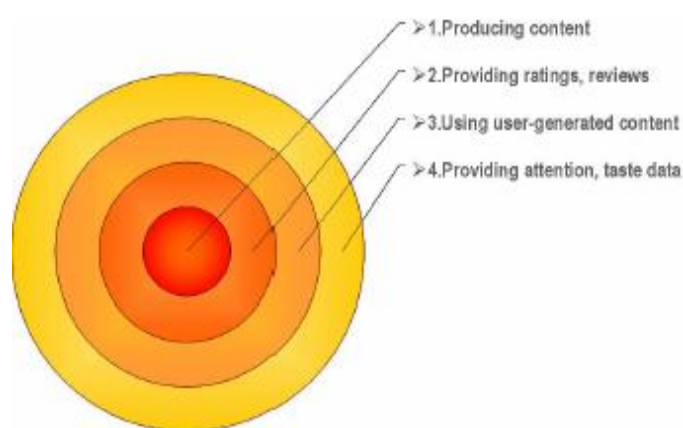


Figure 2.1 – The different user-roles in web 2.0 application use. See Osimo (2008) and Lindmark (2009) for details.

2.3.2 The Proposed Web 2.0 Taxonomy

In view of existing work discussed above, this section presents a more systematic taxonomy for classifying the plethora of web 2.0 applications in existence. The proposed web 2.0 application type taxonomy is based on a two-steps approach, which answers two simple and fundamental questions; *what* is being shared and *how* it is being shared. It is hoped that the proposed classification is more transparent than schemes by Anderson (2007), Hearst (2009), and Lindmark (2009). The taxonomy introduced by Kaplan and Haenlein (2010) is biased towards uses in media sciences, whereas the proposed taxonomy is directly relevant to web 2.0 applications in general, and the main task of sharing that it facilitates. Therefore it is more widely applicable, and in addition it caters for new web 2.0 applications more appropriately than the taxonomies presented by Anderson (2007) and Lindmark (2009).

2.3.2.1 What is being shared?

To begin with, the following (not necessarily exclusive⁴⁹) categorisation of web 2.0 applications, which is based around sharing (the core concept of web 2.0), is put forward.

1. **Link Sharing** – Main purpose of these applications is to collect and share links within an online community. Often tools, such as bookmarklets⁵⁰ or browser plug-ins are used in conjunction with these applications to facilitate more stream-lined bookmarking. Examples: Delicious.com, Connotea.com, Blogmarks.com, FeedMarker.com,...
2. **Multimedia Sharing** – Primarily concerned with image, video-clips, presentation-slides, audio or document uploading. Simple tagging, commenting, rating or more sophisticated social interactions are usually possible. Examples: Youtube.com, Revver.com, Flirck.com, Picassa.com, Slideshare.com, Writely.com, Jamendo.com,...
3. **Review Sharing** – Concerned with sharing recommendations, product reviews and ratings. Examples: Amazon.com, Epinions.com, Consumersearch.com, Ebay.com
4. **Information Sharing** – Primarily concerned with news and other information sharing, where usually anybody can share certain information (generally textual information) with the rest of a community. These shared items can often be ranked up or down and further annotated or edited by other users. Examples: Wikipedia.com, Ushahidi, eBird.com, NASA Clickworkers, Digg.com, Slashdot.com, ...
5. **Social Networking** – (Social / Profile Sharing) The main purpose of these applications is to essentially share personal profile information and to engage in a range of useful or “enjoyable” social interactions online. Examples: Facebook.com, Myspace.com, Orkut.com, ...
6. **Market Places & Auctions** – (Transactions [Bid / Offer] Sharing) Mainly concerned with sharing different modes of information and ultimately facilitating a trading or auctioning mechanism of products, services, commodities or other more abstract concepts. Examples: Craigslist.com, Ebay.com, Trendio.com, Amazon marketplace,...

The categories above would mostly fit into Hearst's 3-Data sharing, 5-Collaborative creation and 6-Social networking activities (discussed in section 2.3.1.2). This is reasonable, since most public use of social media is concentrated in those three groups of activities (Pascu 2008).

⁴⁹ An application may fall under several categories at once, since it facilitates several social interactions.

⁵⁰ A small script / application, stored as the URL of a bookmark. Generally JavaScript is used for this purpose. Bookmarklets provide a relatively platform / browser independent way of adding programmatic functionality to a simple bookmark, and are in common use. Bookmarklets can be used to manipulate web-pages, display custom ad-hoc interfaces, and communicate data across web among other capabilities.

2.3.2.2 *How is it being shared?*

There exist certain elements of re-occurring atomic activities within web 2.0 applications that facilitate participation. A systematic identification of these atomic activities is a useful initial step in assessing collaborative ability of a web 2.0 system (identification of atomic activities is important in our proposed framework within chapter 8, and atomic activities are further mentioned within chapters 4, 5, 6 and 7).

The proposed list of atomic activities which facilitate collaboration on web 2.0 is presented below.

- **comment** – commenting a blog or Flickr post in the classical sense, or a simple text description usually associated with a resource. However this feature can occur in a number of contexts and forms, for example, within Youtube; video responses can be submitted on a given video – essentially a (multimedia) comment in the context of the response (Sykora and Panek 2009).
- **tag** – a (key)word or term associated with or assigned to a piece of information / resource. Tags are generally chosen informally, based on personal choice; however, some systems may enforce tags from a set of predefined tags.
- **like / unlike** – this is usually a simple binary vote, in favour of an already submitted user contribution; however, the target of the binary vote can be practically any resource / content. A variation on this is a “report this” or “off topic” button, usually positioned adjacent to some user generated content. The former allows to vote against inappropriate content and the latter is commonly used to vote against relevance.
- **rate** – usually on a scale of 1 to 5; however, the scale can be arbitrary. This is a type of vote just like the previous item (above), yet due to its very wide use and for the purposes of this classification, it deserves its own status of an atomic activity.
- **link** (create a link) – links can be part of some content submissions, such as within submission of comments or posts (or link submissions on their own). A link indicates some semantic connection to another resource and hence is of interest.
- **post / micro-post** – even though related to comment, a post is semantically different in that it is a submission that represents the main content on which it may be possible to comment on. The post is (on a technical level) in the form of a well known textual character set (i.e. UTF-8, ASCII, EBCDIC) and its length can be but, might not be limited. A micro-post refers to smaller posts, on micro-blogging websites, usually limited in length of characters.
- **file upload** (video/picture/music files) – often has the same semantic role as a post, however does not have to (i.e. can occur in comments for example). File uploads are often limited in size and can generally be on any topic of interest (and in a number of video / file formats).
- **edit** – often one may be allowed to edit some shared resource, where that resource has either been submitted by someone else or was submitted by the editor. Edits are changes made to content in the scope of a submission (e.g. a specific post, file upload, etc...). In other words the event of editing a shared resource and the change itself; be it a file

upload, post or any editable resource (where the given system permits collaborative editing). Systems that allow shared editing generally save the pre-edit and post-edit versions of the document. For example, Wikipedia provides an elaborate versioning system for the entire edit histories of its resources.

- **community / group / category** – this might seem to be quite an abstract concept, but is very common within web 2.0 applications (groups on Facebook, Wikipedia, Amazon, Flickr, or Youtube categories). They usually act as a relatively free (mutually non-exclusive) way of grouping users and / or their contributions (posts, file uploads) under one name or concept. Groupings and categories may be decided on by users themselves (this is often the case), but depending on the system they may be pre-set, as is the case with Youtube categories, which are also mutually exclusive.
- **alert / message** – web 2.0 systems tend to have extensive notification infrastructures, which generate message flows conditioned on collaborative atomic activities, such as Facebook news feeds, or standard RSS (Really Simple Syndication) feeds which communicate alerts about recent edits, posts, rating, etc... Common activities include the act of setting up such feeds by users, or the act of subscribing to feeds.
- **open APIs** – one of the building blocks of web 2.0 is openness and connectivity of loosely coupled systems (O'Reilly 2005). This becomes often possible thanks to web 2.0 systems offering application programming interfaces (APIs) that in turn facilitate access to data and / or services. Facebook, Twitter or Amazon all have very well known APIs which allow many account processing tasks to be automated.

Numerous elements mentioned above were studied extensively and subjected to empirical and theoretical analysis. For example, edits, specifically the type of Wikipedia edits and their frequencies and characteristics were studied in much detail (Ehmann et al. 2008, Scheider 2010). Further, attention was dedicated to file uploads (Cha et al. 2007), binary voting of the type like / unlike (Danescu-Niculescu-Mizil 2009), links (Brin and Page 1998), blog posts (Chesley et al. 2006, Mishne and Rijke 2006, Ali-Hasan and Adamic 2007), ratings (Matsuo and Yamamoto 2009), RSS (Nanno and Okumura 2006, Gruhl et al. 2006), APIs (McCown and Nelson 2007) and a more general discussion of several of these elements is provided in Wunsch-Vincent and Vickery (2007), and Ochoa and Duval (2008).

The work in chapter 4 substantiates, and helps to better understand the popularity, and prevalence of the atomic activities described above, on a significant survey sample of over 700 respondents.

2.3.2.3 The Final Taxonomy

The final taxonomy is based on a two-step approach, in which first the atomic collaborative elements (see section 2.3.2.2) of a web 2.0 application are identified, and then the dominant categories of the type of shared content are chosen (see section 2.3.2.1) web 2.0 applications:

1. **First step (1)**; the atomic collaborative elements of the system are identified, using the atomic elements we have systematically described above (section 2.3.2.2).
 - Once these are decided upon, we have a good idea of what the nature of the data is that is being shared, which will help with step two – i.e. what the main type of information shared is.
2. **Second step (2)**; based on what the dominant data object(s) of the sharing process are, we can establish the most fitting category (or several categories, section 2.3.2.1) for a given system.
 - Once one or more of the sharing categories for an application have been picked, the web 2.0 application is effectively classified in a more meaningful way. See the discussion below.

The two-step web 2.0 application type classification answers two simple and fundamental questions, *what* is being shared (**Step 2**) and *how* it is being shared (**Step 1**). The categories above are defined around the core web 2.0 concept of sharing, and this categorisation can be extended to cope with other applications, such as Virtual World Games (or MMOGs, Massive Multiplayer Online Games), by categorising such, simply under Social and Information Sharing. Of course the atomic collaborative elements might be quite complex with these types of 3D environment applications, but the basic categorisation still applies. Hearst (2009) for example, suggested a special separate category for MMOG applications.

An application, such as Twitter (a Micro-Blogging system) is primarily concerned with sharing information and social profiles, hence we could categorise Twitter as a social and information sharing web 2.0 application. This would be much more meaningful classification than “Micro-Blogging”, which was put into an *Other Applications* category by Lindmark (2009), simply under microblogging, which does not explicitly or readily imply any characteristic of the system.

Social media also vary in complexity, in terms of possible interactions. Facebook for example allows users to keep in touch with friends and colleagues in a sophisticated manner (i.e. using a multitude of atomic collaboration elements) – via contact lists, notifications, news-feeds, groups, events, and one can share pictures, videos, links, send micro-posts, full length posts, customise profile walls, find friends based on their interests, compare scores in online games, tag friends in pictures and numerous other ways of keeping in touch. On the other hand many web 2.0 applications can be simplified into basic agent-to-agent mediation which only brings people together to help exchange basic goods and / or information. Examples such as; Craigslist, lolCats, PickupPal or CouchSurfing⁵¹, allow exchange of goods or services to occur

51 lolCats on www.icanhascheezburger.com is a website for sharing comical pictures of cats with captions and other animal images online, PickupPal is a ride sharing web-site to coordinate car-pooling and CouchSurfing allows to find and offer peer places to stay for free (usually in the participants own house/apartment).

online, or in person (in the physical world), as is the case with PickupPal and CouchSurfing. Generally, the more atomic collaborative elements a system has, the richer the sharing experience and user generated content on a 2.0 system will be.

2.4 Current Applications of Web 2.0

Although web 2.0 adoption is now widespread and has been for several years, the systems available are still often used in rather basic and / or common ways within various vocational areas. This section provides a review of several fields in which web 2.0 is used. The review is mostly based on prior literature and case studies. Its aim is to illustrate practical web 2.0 usage and the benefits, challenges, and opportunities it brings.

2.4.1 Clinical Practice

Within the domain of clinical practice, there has been some criticism that many contemporary healthcare professionals still use the Internet in an old fashioned way (McLean et al. 2007), and few will probably know about or have used health related Podcasts⁵², Blogs, Wikis⁵³ and other web 2.0 tools. McLean et al. present a useful summary of the current state of web 2.0 usage in healthcare communities and encourage greater user participation in developing and managing content within web 2.0 applications, by highlighting benefits of several pilot projects.

To list but a few efforts; there are now Wikipedia style applications for knowledge collaboration, such as FluWiki, which is dedicated to helping communities prepare for treatment and to avoid flu epidemics, a range of Wikis for healthcare professionals such as Wiki Surgery, Healtheva or Sermo (a knowledge exchange Wiki community, only healthcare practitioners can join). It was found in a study (Murray et al. 2005), that interactive health communication applications have positive effects for people with chronic illnesses. There are also examples of social networking applications, such as a completely user-generated database of reviews and networked patients and health professionals on PatientOpinion, CarePages, CureTogether, PatientsLikeMe, MySpace Cure Diabetes Group, Facebook Diabetes Support Group, and other Facebook, and MySpace based support groups now in existence. Clearly better informed patients and practitioners will be able to make more informed treatment decisions, and the benefits of participatory or even collaborative communication facilitated via

⁵² Podcasts are somewhat synonymous to multimedia based Blogs, and are usually syndicated via the RSS protocol – see <http://en.wikipedia.org/wiki/Podcast>

⁵³ A Wiki is an expandable collection of interlinked web-pages that allows any user to quickly and easily add, remove, or edit content. Wiki platforms were mostly pioneered by online encyclopedia project Wikipedia.

web 2.0 are significant, as illustrated by existing pilot studies (Wright et al. 2009). Wright et al. provide a useful starting point for literature review of clinical decision support systems in the pre – web 2.0 era, and it is suggested by the authors that web 2.0 implementations of clinical support systems might prove useful in future. This conclusion is based on an examination of three cases studies, specifically; Clinfowiki, Partners HealthCare eRooms and Epic Community Library. A recent study (Blumenthal et al. 2010), provides an overview of a collaborative effort⁵⁴ to train staff and integrate web 2.0 technologies into the work of two local public health departments in Michigan (US), and discusses this deployment process in its entirety, from careful planning to implementation options and staff training. The study identified several areas of possible improvement in efficiency and effectiveness within public health practice using qualitative and quantitative methodologies. Surveys, brainstorming sessions, needs assessment focus groups and interviews of sizeable staff samples were conducted. It was found that over two thirds of staff was already using web 2.0 tools unknowingly, however less than 10% were actively contributing content and 37% felt either uncomfortable or very uncomfortable learning new technologies. Implementation of these web technologies carry numerous risks which were also identified and addressed in this study.

2.4.1.1 CureTogether Walkthrough

In order to help illustrate how a community web 2.0 application other than the more obvious Wiki works in practice, this sub-section provides a brief walkthrough, of one such health issues community. CureTogether is an anonymous and discreet health problems web 2.0 community that anyone can join. It allows users to choose their suffered illness from a set of conditions (see figure 2.2). For each condition the user can specify the symptoms suffered, severity of those symptoms, likely or suspected causes (e.g. cold weather for asthma), the treatments one has undergone with an indication of how effective they were, and any side effects they have had. Providing these details is the initial stage that needs to be completed when joining the community. Once these details are provided, the user's conditions are matched against the community (see the top of figure 2.2, the conditions “Stomach Pain” and “Allergies” are compared with rest of the user-base who are suffering the same health problems).

54 Between the University of Michigan Health Sciences Libraries, the Prevention Research Centre of Michigan and the Genesee County Health Department and Monroe County Health Department.

Condition	Score (?)	Percentile (?)
Stomach Pain stats	55	50
Allergies stats infographic	54	81

Do you have any of these conditions?

Check Y for the conditions you have, then take the surveys to get your score and see how you compare.

Fatigue	<input type="checkbox"/> Y <input type="checkbox"/> N
Anxiety	<input type="checkbox"/> Y <input type="checkbox"/> N
Depression	<input type="checkbox"/> Y <input type="checkbox"/> N
Back pain	<input type="checkbox"/> Y <input type="checkbox"/> N
Insomnia	<input type="checkbox"/> Y <input type="checkbox"/> N
Lower Back Pain	<input type="checkbox"/> Y <input type="checkbox"/> N
Neck Pain	<input type="checkbox"/> Y <input type="checkbox"/> N

Figure 2.2 – The choice of suffered conditions. At the top of figure, the conditions “Stomach Pain” and “Allergies” are compared with the community, for which more detailed statistics can be brought up

Once a user becomes a member, community details concerning a particular condition can be brought up in the system. Figure 2.3 illustrates the symptoms that are suffered by asthmatics (1,049 registered users) and their breakdown by severity. Similarly in figure 2.4, treatments and whether and how much they have helped within the community of asthmatics, are presented. Similar summaries are available for treatment side effects, and likely causes. The so called “surveys” that a user can take to provide further background details of their condition, are dynamic, as new items regarding any condition can be added by individual users dynamically.

Asthma (1,049 members)

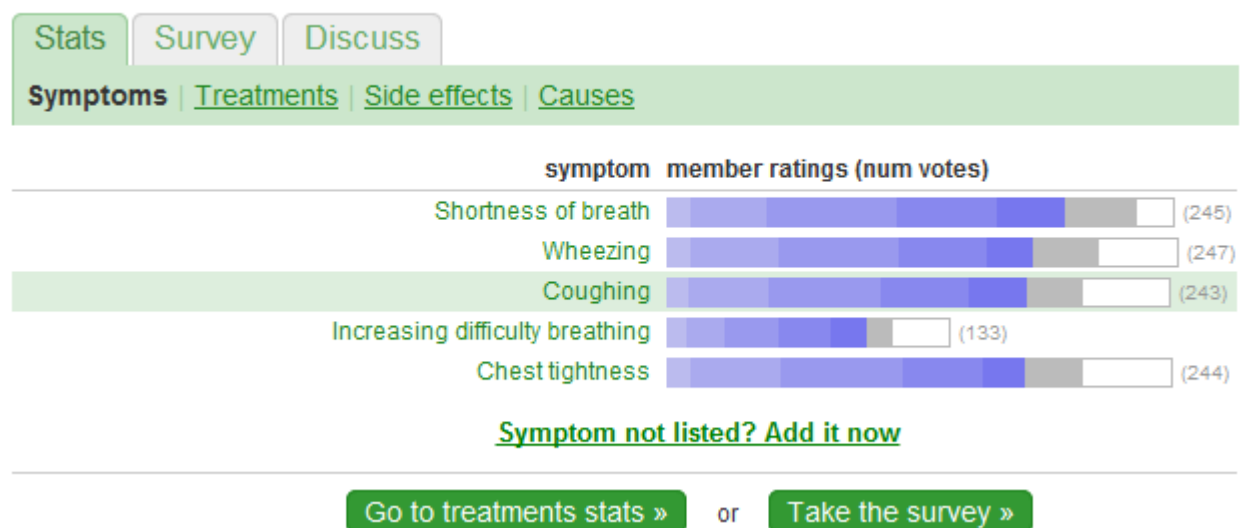


Figure 2.3 – Symptoms suffered by asthmatics and breakdown of its severity (for 1,049 community members)

In addition to viewing the community assessment of a condition, a discuss tab (visible in figures 2.3 / 2.4) allows a public discussion regarding any aspects of the condition to take place. The system also facilitates private discussions (message sending) between users.

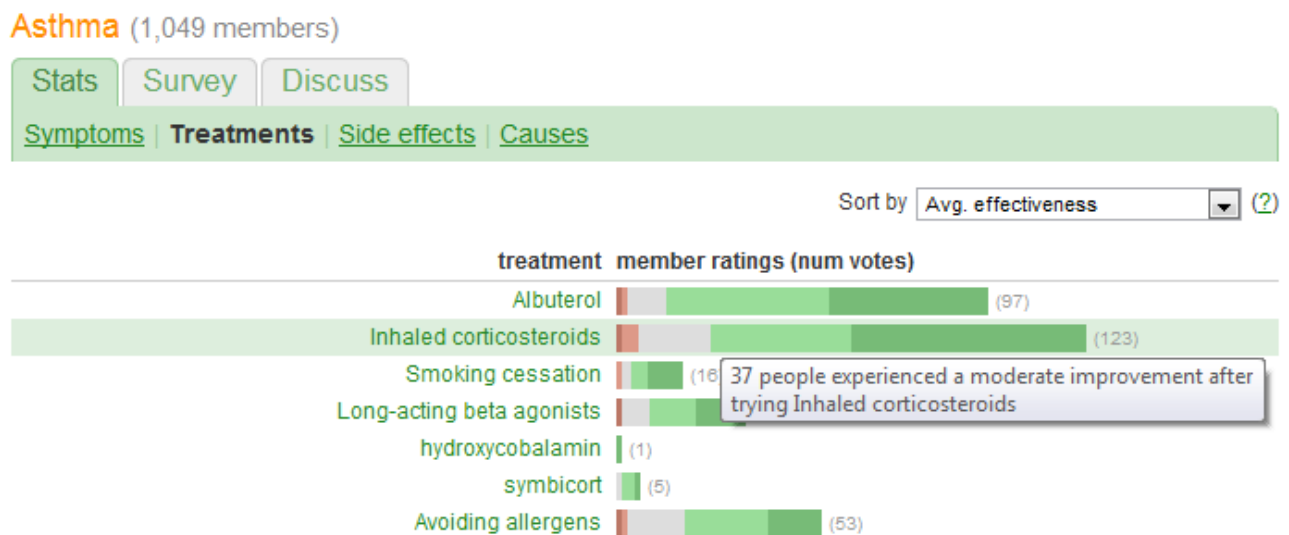


Figure 2.4 – Treatments for asthmatics and their effect

The advantages of sharing health problems, having a condition compared with a community of people suffering the very same health problems, being able to access a number of statistics and discuss specific treatments, side-effects, symptoms, and to contact other members privately in order to discuss targeted health issues, are arguably beneficial. The CureTogether community facilitates an effective way of finding patients suffering from a very wide range of common but also rare diseases⁵⁵. The potential of early diagnosis thanks to a system like CureTogether certainly does exist, although numbers of confirmed actual patient diagnoses where the initial diagnosis originated from CureTogether are not available to confirm the extent of this. Sufferers who are not suffering alone tend to cope better with their ailment and the benefits of a supportive community have been shown to be significant (Murray et al. 2005, Wright et al. 2009). Given the above considerations, it must be concluded that web 2.0 communities such as these can have some positive benefits. However, there is also scope for misdiagnosis and some risk of trivialising certain conditions, since the average user will not be a qualified health professional. Although it seems that the CureTogether community takes health problems seriously, it would be highly appropriate to only use CureTogether once a health professional has been visited.

Since most of CureTogether functionality was discussed in this sub-section, it will be useful at this point to briefly evaluate CureTogether within the taxonomy proposed in section 2.3.2. The

⁵⁵ A list of many of the health problems covered by CureTogether, <http://curetogether.com/findall.php?opt=showall>

atomic collaborative elements will be identified (sub-section 2.3.2.2) and finally CureTogether will be assigned to its relevant categories (sub-section 2.3.2.1). First of all there is a frequent use of the “like / unlike” element (i.e. binary vote; yes / no – when choosing suffered conditions, etc.), the “rate” element (i.e. on a 1-5 scale – when assessing severity of symptoms and benefits of treatments), and the possibility to comment (i.e. participate in discussion and sent text based replies). In addition a strong emphasis is placed on communities / groups, and also an alert / messaging system for direct messaging between users is supported. Hence, there are clearly at least five main atomic collaborative elements, which means that the collaborative footprint is relatively large, and predominantly allow the sharing of information, organised by groups, and a sense of community is further supported by direct messaging. Based on this assessment of atomic activities, it would seem appropriate to place CureTogether into the “information sharing” category. Keeping in mind that the second most sensible classification for CureTogether would be that of the “social networking” category, due to the sharing of personal information and potential for engagement in social interactions between users, although most of it being anonymous and hence the “information sharing” categorisation seems most fitting.

To conclude this walkthrough, web 2.0 applications such as PatientsLikeMe, MoodScope⁵⁶ or other, are similar to CureTogether. In that they offer an interactive way to collaborate and learn from a community that shares a set of similar health problems, and has to deal with related issues.

2.4.2 Corporate Use

In the corporate world Blogs are becoming increasingly important and are starting to be used quite extensively in customer relationship management (Dwyer 2007). Many companies also encourage their employees to Blog on their work, and collect any valuable feedback via readership commentary (Efimova and Grudin 2007). Companies are actively participating in social networking and a variety of social media (DiMicco et al. 2008, Li and Bernoff 2008). In fact there is now a range of commercial services, which analyse public relationship related chatter on social media (e.g. Sysomos, Brandwatch, Attensity, Vocus, and DowJones Insight Media Services⁵⁷). A study by Li and Bernoff (2008) identified several unsuccessful businesses use cases of social media. The authors provide a thorough analysis of the common detrimental

56 MoodScope (<http://www.moodscope.com/>) is website, which allows a user to help deal with depression. It allows users to submit their mood regularly and to effectively measure, track, and share it with the community by nominating trusted friends, to act as buddies, and help a user deal with mood swings.

57 Further companies leading this area are; Socialradar, Radian 6, Ecairn, Simplify360, Engagor, Lithium, ReputeMe and EmailVision.

features and practices that might lead to successful web 2.0 integration within existing core business processes (i.e. research and development, marketing, sales, customer support, operations). It is further argued throughout the paper, that; *“...with the increase in social participation among consumers and the growing sophistication of the underlying technologies, it's now possible to put social applications on an equal footing with other business projects. That is, they can deliver measurable progress towards significant, strategic business goals.”* The importance of web 2.0 implementation in the corporate world is further highlighted in the OECD commissioned report (Wunsch-Vincent and Vickery 2007)⁵⁸.

2.4.3 Politics, Public Service and Education

Political discourse has been a major topic within Blog posts and other social media debates for some time now (Adamic and Glance 2005, Farrell and Drezner 2008). Jackson and Lilleker (2009) document and evaluate party use of web 2.0 in the UK, and found a presence on social media for virtually all parties concerned. Recently a more direct use of web 2.0 within government has been proliferating (Osimo 2008, Huijboom et al. 2010, Kuzma 2010, Parycek 2010). For example, the EU Joint Research Centre (JRC-IPTS) investigated web 2.0 for purposes of e-government and a number of reports on this research are available (Osimo 2008, Pascu 2008, Lindmark 2009). The use of web 2.0 within government can be broken down into use within front office domains and back office domains. The government's front office domains where web 2.0 could and arguably should be utilised are service provision, political participation and transparency, law enforcement, and the back office domains include regulation, cross-agency collaboration and knowledge management (Osimo 2008). Governments now use Social Media, Blogs and even Micro-Blogging services (e.g. Twitter) to communicate with their citizens or to “openly” discuss policies. Web 2.0 based volunteered collaborative initiatives by governments which ask their citizens to help monitor elections and submit their election observations to various web 2.0 systems also became more common. US, Australian or European Union patent offices experimented with peer2patent projects, which have been highly successful in simplifying, speeding up, and keeping costs of patent processing tasks down⁵⁹. Intellipedia, a Wiki platform managed by the CIA, enables direct collaboration between the analysts of 14 US Intelligence agencies, and has been used successfully in a number of intelligence detection tasks⁶⁰. Code for America⁶¹ is a civic initiative (in the USA),

58 Other documents in the series are available from <http://www.oecd.org/sti/> or <http://www.oecd.org/digitalcontent>

59 See <http://www.peertopatent.org/>

60 The project hosts around 900,000 pages edited by 100,000 users with 5,000 page edits per day. See “Wikipedia for Spies: The CIA Discovers Web 2.0”, <http://bit.ly/c1JJ0h>, Last Accessed: 3rd July 2010

61 <http://www.codeforamerica.com>

whose goal is to facilitate code sharing and collaboration via web 2.0 platforms, to avoid unnecessary costs for local governments. Nearly every city performs much of the same functions for its citizens, yet there is significant duplication of spending by cities and municipalities as each build their own IT platforms. It is hoped that this initiative will help to lower costs for the municipalities in general (this project is still in early stages). Social production (Benkler 2002), in the form of coordinated calls for action, as well as more sophisticated collaborations are more common in government, than in other domains. Shirky (2010, pp. 161-213) provides much background and a useful debate on public and civic social media usage for peer managed government use.

The usage of web 2.0 within education, higher academic institutions, and libraries has also been on the increase (Alexander 2006, Boulos 2007, Anderson 2007, McLean et al. 2007). It has been suggested that students of all ages learn best when immersed within a culturally and socially rich environment in which learners and peers are committed to achieving the same goals and can regulate each others' performance. Therefore it would seem that the web 2.0 tools have potential to both liberate and tie learners together in dynamic learning communities (McLean et al. 2007). A number of use cases of web 2.0 in education have been studied (Boulos 2007, Anderson 2007) with some encouraging success stories. Within libraries the use of web 2.0 technology has become quite common, with a number of established good practices (Casey and Savastinuk 2007). It was noted that Wikis can be useful writing tools that aid composition practice (Alexander 2006), and their use has been quite prolific within teaching and learning environments. Anderson (2007) is an excellent source of critical case-studies of web 2.0 use in learning and teaching, scholarly research, and libraries or archiving.

2.4.4 Journalism and Geography

Since web 2.0 largely concerns itself with facilitating efficient data sharing, naturally it has already found many interesting applications within journalism. One such example is CNN iReport, a web 2.0 system allowing any registered user to submit and edit news stories within a community of "citizen reporters". Essentially, any news can be uploaded since the contributions are neither edited, nor fact-checked, or screened. Certain (urgent or timely) stories may get vetted and cleared by CNN, and these would be subsequently used in CNNs mainstream broadcast⁶². iReport is defined by a distinctive news-friendly community, which seems to find pleasure in reporting news. The community of contributors consists of around 20,000 enthusiasts who get ranked based on their site activity and value of contributions. Clearly the

62 See <http://ireport.cnn.com/> and <http://ireport.cnn.com/faq.jspa> for more information.

chance of an iReport item being selected by CNN for broadcast acts as strong motivation itself (31,800 out of 485,000 reports were vetted by CNN, to date). Other well known examples of web 2.0 journalism are the use of Blogs⁶³, peer-reviewed online news systems such as Digg, SlashDot or WikiNews. Certain instances of Youtube use have also played a role in web 2.0 journalism. Mobile phone videos of post election riots in Iran, where coverage of the riots was exclusive to mobile phone video-clips uploaded by demonstrators themselves to Youtube is a good example. Moreover press agencies, newspapers and TV stations like BBC, Financial Times, or Bloomberg duplicate their content from traditional distribution channels on Youtube (Sykora and Panek 2009), Twitter messages and Facebook discussions (Tapscott and Williams 2008). Such behaviour by mainstream media encourages social media engagement of users with the news. The enthusiasm about utilising communities by enabling average people to enrol as news correspondents has been at times excessive (Shirky 2010). And it ought to be noted that in many cases “citizen journalists” are not much more than “citizen news gatherers”⁶⁴. One must consider that professional journalists cover fires, floods, crime, the legislature and the Government (i.e. Downing Street, the White House) every day. A citizen journalist, an amateur, will most of the time, simply not have access, and have to be on the outside. Some have also argued (Reese et al. 2007, Thurman 2008) that traditional and web 2.0 based journalism is complementary to one another, and both forms of journalism will come closer together, rather than one pushing out the other.

Surprisingly, web 2.0 has found much application in Geography as well (Scharl and Tochtermann 2007). Geographic information systems (GIS) have been around for a long time, and thanks to GPS and satellite technology there are now vast amounts of data. However, a lot of geographic information is not visible within this data (Goodchild 2007). Therefore, a number of web 2.0 systems have appeared recently to facilitate collection of volunteered geographic information (VGI⁶⁵). Wikimapia, a web 2.0 service allowing users to contribute descriptions of places of interest, along with geographic coordinates, or submission of georeferenced photographs to media sharing websites such as Flickr, and other pilot studies of VGI collection, such as OpenStreetMap or Inrix, have grown much in popularity. The idea that a large population (potentially many millions) of users could act as “sensors” who understand the importance of local knowledge and who would be easily capable to annotate geographic data, has been responsible for much excitement in the field of geography (Turner 2006, Goodchild 2007, Scharl and Tochtermann 2007).

63 Now commonly used by professional journalists, or amateur blog writers who turned professional.

64 See <http://digitaljournalist.org/issue0912/lets-abolish-citizen-journalists.html>

65 See (Flanagin and Metzger 2008)

2.5 Summary

In this chapter *web 2.0* was defined and introduced with several examples. Prior literature has emphasised the significance of web 2.0 and social media (for the purpose of this thesis, both terms are considered synonymous). The related historical, social and economic background is rarely discussed in literature, and hence this chapter presents a valuable discussion of some wider issues. Next, various web 2.0 taxonomies that are in existence were reviewed, and a new categorisation of web 2.0 applications which is more transparent than schemes by Anderson (2007), Hearst (2009), and Lindmark (2009), not biased towards uses in media sciences (Kaplan and Haenlein 2010), and is more widely applicable (Anderson 2007 and Lindmark 2009), was proposed. As web 2.0 adoption is now becoming widespread, it was felt that work was needed to examine the existing efforts, and hence the final section of this chapter reviewed web 2.0 application uses throughout vocational fields. In order to illustrate the wide relevance, applications in clinical practice, corporate use, politics, public service, education, journalism and geography were reviewed. The review was mostly based on prior literature and case studies.

This chapter illustrates the impact, reach, but also issues associated with social media. Clearly some change and development in the web and its use has brought on significant changes. Nevertheless, in nearly all prior literature web 2.0 was defined subjectively or simply by use of examples. There is a real need to provide a more objective or at least a more quantitative way of defining the concept and meaning of web 2.0. The next chapter tackles this issue, and considers, to what extent web 2.0 is simply a buzz-word, and whether there is some quantifiable substance behind the concept.

3 Defining Web 2.0



The last chapter illustrated how web 2.0 has had a major effect on a whole range of information services in different domains such as; education, medicine, corporate marketing or government. For example, “library 2.0” has evolved to represent applications of web 2.0 based technologies used within library services (Boulos et al. 2006, Casey and Savastinuk 2007). Similarly, “government 2.0”, “education 2.0”, “law 2.0” and “medicine 2.0” have all emerged within a flurry of “2.0” buzz-words. The “2.0” refers to web 2.0 technologies as the “new version” of web usage in the respective disciplines. Buzz-words or neologisms often cause some amount of confusion. A neologism; from Greek (neos “new” + logos “speech”) is a newly coined word or expression, that may be in the process of entering common use, but has not yet been accepted into mainstream language. For instance; *how is library 2.0 really different from well established library processes, and is the term even justified?* This can be open to some debate; however, library 2.0 has seen wide and common use within popular and academic literature. To this end it is interesting to investigate as to what degree some “2.0” terms have taken hold in common use, as it may also indicate the prevalence of web 2.0 technology in those fields. It should also be noted that a similar wave of neologism creation occurred in the late 1990s. With the large initial enthusiasm surrounding internet, the so called “e-words” began appearing, i.e.; e-commerce, e-business, e-solutions or e-health, see (Eysenbach 2001). Hughes et al. (2008) argues, for

example, that e-health and medicine 2.0 are two separate and legitimate topics in their own right¹.

3.1 Background

Numerous aspects of neologisms within the English language have been studied from a linguistics point of view for many years (Stekauer and Lieber 2006). A suitable framework however for our investigation is the basic theory of sensemaking, as understood within organisational studies². This is the process by which people give meaning to experience. *“Collaborative sensemaking takes place over a certain time-window over which terms can evolve, get redefined, or completely new terms indeed emerge”* (Weick et al. 2005). This happens when new concepts come into existence that represent an innovation which changed the way some people thought and spoke about the concept.

An online collaborative bookmarking system called Delicious (delicious.com)³ seems suitable for a quantitative investigation of neologism emergence. This is for a number of reasons related to sensemaking theory, which will be discussed. It is further suggested that the method applied in this work is superior to another technique used in the past.

In Delicious a *term* or concept is essentially represented by the content of a bookmarked URI, with which one or more keywords or tags are associated. All bookmarks with their tags and comments on Delicious are publicly accessible, and hence the aggregate consensus on term representation via tag annotations is readily available. The collaborative sensemaking over a *time-window* can be observed since all bookmarks on Delicious also have a temporal dimension.

It could be said that in aggregate, bookmark tags on Delicious can be assumed to be representative of prevailing user interests. The set of tags used in bookmarks, as well as frequency of tag use within that set represents the collective description of that URL by many users. In fact a remarkable stability in the relative proportions of tags for a given URL was

1 Neither the stakeholders nor the principal tool used (the Internet) distinguishes Medicine 2.0 from eHealth. However, the principles of generation of content by users, the power of networks, open source, personalized health care, and the focus on collaboration across all stakeholders are not always highlighted by eHealth and suggest that these fields have different emphasis.

2 For an alternative definition of sensemaking theory see <http://en.wikipedia.org/wiki/Sensemaking>

3 Delicious is a web 2.0 application which allows a user to mark content (web-pages, pictures, or any other content with a valid URI) with descriptive terms, also called keywords, tags or social annotations. This is a common way in web 2.0 systems to organise content for navigation, filtering and search. Since all bookmarks are public, users are allowed to browse other user's bookmarks and view their tagging habits. Most users are motivated to use an application like Delicious for their own gain - the service has advantages over browser based bookmarks, e.g. being accessible and searchable from any location. There are no limitations on who can view bookmarks (all bookmarks are public) hence the second advantage is that of being able to see bookmarks of other people. Delicious provides aggregation views and recommendations based on this data. Delicious.com is used by individuals as well as in professional and business circles.

found to exist on delicious.com (Golder and Huberman 2006). Golder and Huberman found empirically on (a large) set of bookmarked URL links that usually after first 100 or so bookmarks, each tag's frequency is a nearly fixed proportion of the total frequency of all tags used. Figure 3.1 exemplifies this stable pattern, where after about 100 bookmarks the tag proportions become very stable. Golder and Huberman explain this stable pattern by resorting to the dynamics of a stochastic urn model, and discuss a couple further reasons for this relatively unexpected phenomenon in their paper (Golder and Huberman 2006, pp. 205-206).

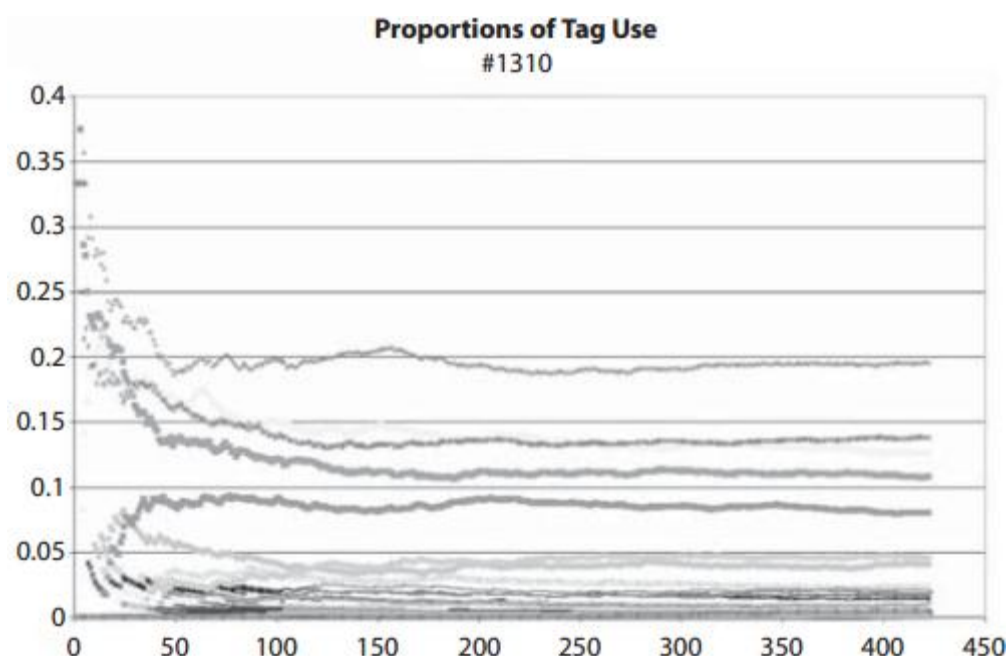


Figure 3.1 – Stabilisation of tags' relative proportions for an example URL. The y-axis denotes fractions and the x-axis time in units of bookmarks added [source: Golder and Huberman 2006, pp. 205]

Their results essentially mean that (mostly) independent URL bookmarkers tend to use the same proportion of tags to describe content on web-pages. Due to these tag proportions being highly consistent, one can infer that spontaneous mutual consensus emerges for a sizeable group of users and resources. Yet the links do not have to become highly popular to be useful, since already after a 100 bookmarks of a resource the pattern becomes stable. It must also be noted that user tags associated with Delicious bookmarks have been found to be highly representative of the URI resource. Bao et al. (2007) looked at using Delicious tag annotations in page indexing for search algorithms, improving on the well known Page-Rank algorithm (Brin and Page 1998)⁴. Their final algorithm performed extremely well on an experimental dataset extracted from Delicious. Their work highlights the fact that Delicious tag annotations are usually very good summaries of the web page content they represent. Nevertheless tags are still susceptible to a range of issues, such as polysemy, synonymy, homonymy and basic level

4 The successful Google search engine was predominantly based on the page-rank algorithm.

variation, see (Golder and Huberman 2006 – pp. 199-200) for a discussion of these issues. A final consideration of this exploratory study is the user base of Delicious and its demographic profile, which overall suits the web 2.0 focus of this investigation.

3.2 Proposed Methodology

As was already mentioned this chapter concerns itself with a quantitative investigation of the “2.0” neologism emergence, in order to better understand the nature and hype surrounding these terms. Although term frequency generated through social media use, has been applied widely in various areas as a tool to investigate social or economic questions (Thelwall 2009, also see chapter 6), here the use of social bookmarking is suggested as a new methodology, to investigate neologism emergence within the English language in particular. The proposed methodology is well founded on previous research, and results reported in prior literature – i.e. term proportion stability, accuracy of content representation by tags, and the compatibility of collective bookmarking with the theory of sensemaking (all discussed in section 3.1). A need for the new methodology is evident from work such as Hughes et al. (2008), or Van De Belt (2010) who would have clearly benefited from a more automated and quantitative methodology in their studies of term prevalence. Both studies have also investigated academic journal papers for the prevalence of terms such as Medicine 2.0; however, the examination of grey literature is where the proposed technique could be of benefit.

The proposed methodology consists of three steps. These will be discussed within this section, at each step technical considerations, expected output, and validation of the methodology step are discussed. Then in section 3.3, actual results of the “2.0” neologism study based on the methodology suggested here, are discussed. The three methodology steps are described below. Each step accomplishes a different task in evaluating a term. Step one is used to indicate the current prevalence (i.e. popularity) of a term / neologism in relation to similar (non-neologism) terms. Step two helps assess trends in popularity, and step three helps to discover associations between terms.

1. Total number of times a tag or set of tags was used (**current term prevalence**)
 - a. *Output:* Simple totals of tag use frequency over all the links and all bookmarkers, where the total tag occurrence is compared to each other.
 - b. *Validation:* The accuracy of content annotated by tags and tag use stability has been demonstrated in prior literature (section 3.1).
 - c. *Technical implementation:* Straightforward use of Delicious’ tag search interface which reports back the totals.
2. Bookmarking over time (**term popularity trends**)

- a. *Output*: Monthly, weekly, or daily frequency time-series of bookmarkers' bookmarks for all resources on a particular topic. This can be used to track popularity trend(s).
 - b. *Validation*: See sub-section 3.2.1
 - c. *Technical implementation*: See sub-section 3.2.2
- 3. Associated comments and tags (**discovery of related or synonymous terms**)
 - a. *Output*: List of top, most common key terms, occurring in comments, and other top tags, on a particular topic, grouped by year, or month. This is used to discover associations and related terms.
 - b. *Validation*: Since comments left by bookmarkers present further insight into the content, or provide reasons for bookmarking a resource, the use of key terms from comments to discover associations with a tag, seems reasonable. Also, text in comments is normalised, and pre-processed using accepted text-processing technique. Other tags used by bookmarkers represent related concepts, and using these to provide further insight seems valid, support for this was already given in Bao et al. (2007), see section 3.1.
 - c. *Technical implementation*: See sub-section 3.2.3

3.2.1 Validating Step-2

In order to validate this simple technique of “popularity” trend measurement, bookmarking frequencies related to an arbitrarily selected event of high socio-economic importance was examined. Specifically the controversial (\$60bn) Bernie Madoff fraud scheme that weighted heavily onto the already ongoing financial crisis of 2008 was chosen. It was expected that Delicious bookmarks would be highly congruent to real events. A timeline of the events, following discovery of the fraud on 9th December 2008 is presented chronologically in the lower part of figure 3.2, and “Madoff” tag-related bookmarking frequencies in the upper part. The chart in the figure illustrates that “Madoff” was not a focus of interest among Delicious users until the fraud became public knowledge during early December 2008. The initial spike of interest was followed by a through in January (yet nearly 200 bookmarkers bookmarked Madoff related resources during this month), and subsequently the months of February, March, June and July showed renewed interest as victim customer accounts became public (February), Madoff pleaded guilty (March), Madoff was sentenced to 150 years in prison (June), and as Madoff began serving his prison sentence (July). The items highlighted in blue within the figure are more specific to the Madoff fraud and will not be discussed in more detail here. Clearly the Delicious data correlates with interest-in or popularity of a topic, which adds some support to this methodology.

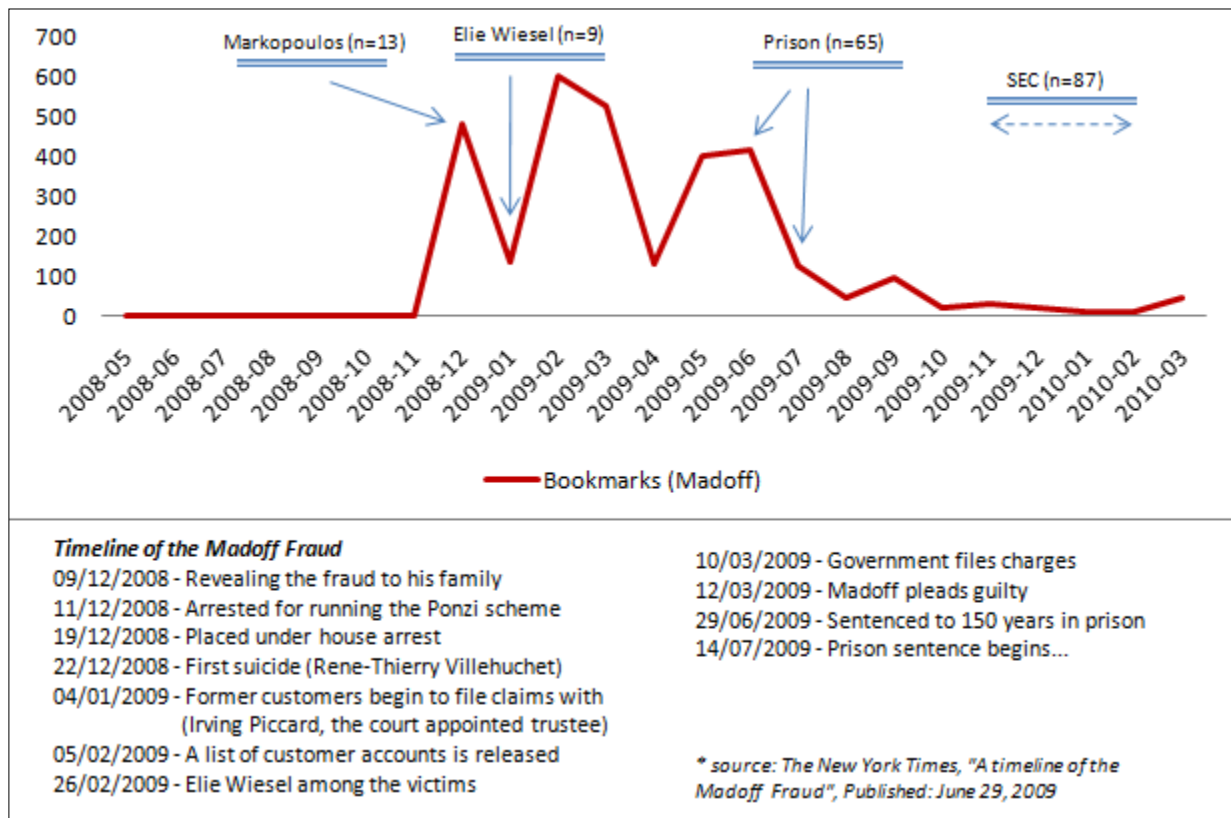


Figure 3.2 – Bernie Madoff fraud case (2008-05...2010-03, monthly data)

3.2.2 Technical Implementation of Step-2

A tag search on Delicious returns bookmarkers' bookmarks for all resources on a particular topic, i.e. the tag(s) searched. Unfortunately this result list is paginated and in an unusable form for further processing, necessary for this and the next step in the methodology. Therefore a simple http request wrapper was built in C# to extract the results into a local database. The method of screen scraping was employed to parse page mark-up structure of the link / bookmark pages and the extracted data was stored in a MySQL database (over 300MB). The unique URL of the bookmarked link, its associated page-title, all user-names of bookmarkers of the URL, with date-time stamps when bookmarked, the tags, and comments these users used were retrieved and saved for each URL. Data from Delicious was filtered so that only links that have been bookmarked by ≥ 3 individual users were downloaded in order to avoid insignificant bookmarks, and to ensure that "lower quality" resources were left out from the analysis. Unfortunately, Delicious imposes numerous limits on amount of returned links and bookmarkers hence the dataset was limited in quantity by these imposed restrictions⁵. Finally as the extracted data is stored in a local database, this database can be readily queried for all

⁵ Restrictions like this tend to be common with many web 2.0 systems. There are usually certain ways to mitigate such restrictions, see chapter 6 or also Thelwall (2009).

counts of bookmarkers' bookmarks on a particular topic aggregated by week (or any other unit), to generate the popularity time-series.

3.2.3 Technical Implementation of Step-3

Since data is in a usable form in a local database store (see previous sub-section), comments and tags used by bookmarkers for particular links can be readily processed. However, it was found that only 20% of all bookmarks have descriptions (i.e. comments) on Delicious. A large fraction of descriptions were also found to repeat for the same unique links – on average 22% of all bookmark comments per link are repeat comments. This is due to Delicious' bookmarking tool supplying a textual description by default. As will be illustrated in sub-section 3.3.3, the top tags and top keywords from comments are automatically identified and presented for further analysis, grouped by a long enough time-period (i.e. year or month), in order to facilitate detection of any changes or trends in keyword associations. Standard text processing techniques were applied to the textual data (*Python scripts and NLTK - Natural Language Toolkit libraries were used*), i.e. word tokenisation, stop-word removal (based on an English dictionary), Part of Speech detection and Lemmatisation, to discover top frequent comments and tags (Bird et al. 2009). Terms were lemmatised, which is a text-processing technique to group together different inflected forms of a word, based on Part-of-speech of the words. This aids in a much more accurate word counting process, since various inflected forms of the same word are registered under the same computation. Please note that repeat comments were excluded from the analysis (around 22% on average per link).

3.3 Results

3.3.1 Step-1: Total Tag Use

In order to quantify the prevalence of web 2.0 neologisms the number of bookmarks per tag in existence is discussed in this section. Table 3.1 presents an overview of these findings.

Table 3.1 – Delicious tag popularity (count of all URLs with the given tag, as of 16th July 2010)

Topic (<i>domain of interest</i>)	New Term (<i>the new 2.0 term</i>)	Old Terms (<i>two separate tags</i>)
Library	library2.0 (94,605)	library + web2.0 (66,367)
Medicine	medicine2.0 (1,374)	medicine + web2.0 (3,488)
Law	law2.0 (389)	law + web2.0 (3,749)
Education	education2.0 (5,255)	education + web2.0 (249,900)
Business	business2.0 (2,363)	business + web2.0 (209,551)

It can be appreciated from the first row in the table that since there were 94,605 bookmarks tagged with “library2.0”, and 66,367 bookmarks tagged with two individual tags “library” and “web2.0”, this indicates that the term “library2.0” has taken precedence over simply speaking about the topic in the context of web 2.0 (“library” and “web 2.0”). Clearly, “library 2.0” is a concept that (at least) Delicious users have collaboratively agreed on. This collaborative agreement on new terms; however, occurred less for other concepts. For example the tags “medicine” with “web2.0” were used 3,488 times, compared to 1,374 for “medicine2.0”. Hence it can be said, medicine in the context of web 2.0 is less frequently the focus of interest to Delicious users, nevertheless the usage of “medicine2.0” relative to the two tags (“medicine” and “web 2.0”) is still quite high, indicating that chances are this term is used more often by a smaller (more expert) group of people. The term “law2.0” occurred 389 times, whereas “law” and “web2.0” separately, as many as 3,749 times. Interestingly “education2.0” tag, occurred as many as 5,255 times compared to “education” and “web2.0” (249,900 times together), which hints the term education 2.0 isn't wide-spread, actually far-from it. However, a significant interest into education and how the new web 2.0 tools are related to it certainly exists within the Delicious user-base. Similarly “business2.0” was retrieved on 2,363 bookmarks, the two tags “business” and “web2.0” were retrieved from 209,591 bookmarks. This shows that the term “business2.0” does not carry much if any significance within the business community, however it turns out “enterprise2.0” tag is much more popular and appears as many as 87,088 times.

3.3.2 Step-2: Bookmarking over Time

In this sub-section library 2.0 and medicine 2.0 related bookmarks were selected, and aggregated into two time-series as per methodology (see sub-section 3.2.2). The results presented here essentially illustrate evolution of these concepts in terms of bookmarking frequencies over time, and show clear trends in popularity. From figures 3.3 and 3.4 we can appreciate the emergence of web 2.0 related medical and librarian trends. A relatively far reaching time-window is captured by the data, ranging from week one in 2006 to week 32 in 2010. The popularity in library 2.0 and medicine 2.0 tag use increased to some extent over the years; however, there is clearly a much stronger up trend in popularity of the library 2.0 concept, than there is for medicine 2.0. During 2009-2010 the library 2.0 concept has been far more pronounced ($\mu=1020$, $\sigma=389$; on weekly data), whereas the interest into medicine 2.0 faded for the same period ($\mu=311$, $\sigma=122$). In fact, for the time period 2009-2010, there was a significant negative correlation between library 2.0 and medicine 2.0, $r=-.322$ ($N=88$), p (2-tailed) $< .01$. Although from table 3.2, a significant relationship between library and medicine

frequencies does exist.

Overall these results indicate that the use of web 2.0 in the library and medical domains as well as the actual terms library 2.0 and medicine 2.0 have been on the increase. This increase seems to have been quite consistent during 2006 to 2008 (figures 3.2 and 3.3). Nevertheless the data strongly indicates that the term library 2.0 is still being used, while use of medicine 2.0 has faded in popularity. Although not conclusive, it may indicate that medicine 2.0 is still a neologism, while library 2.0 is becoming a well accepted term.

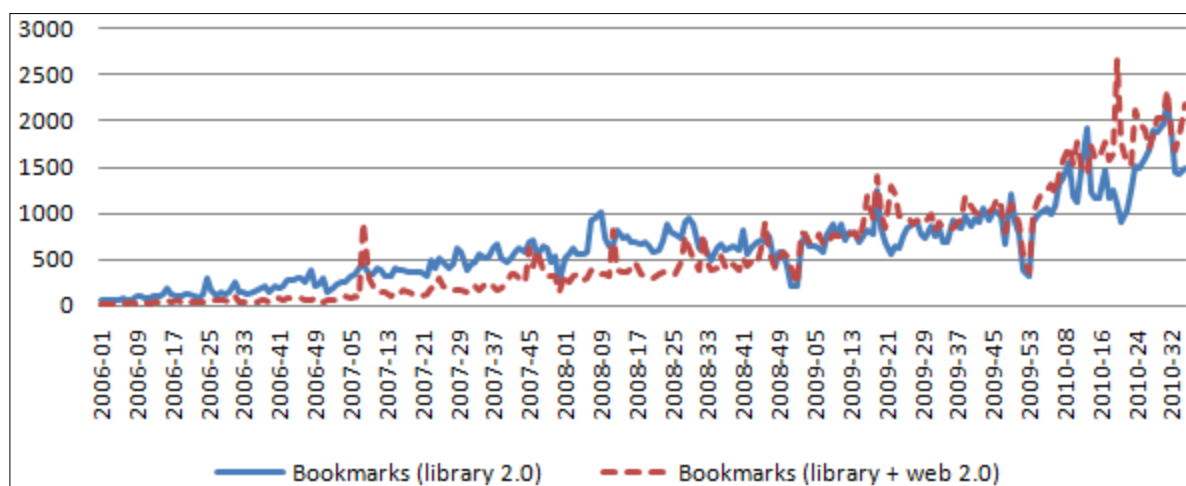


Figure 3.3 – Library tagged bookmarks (2006-01...2010-32, weekly data)

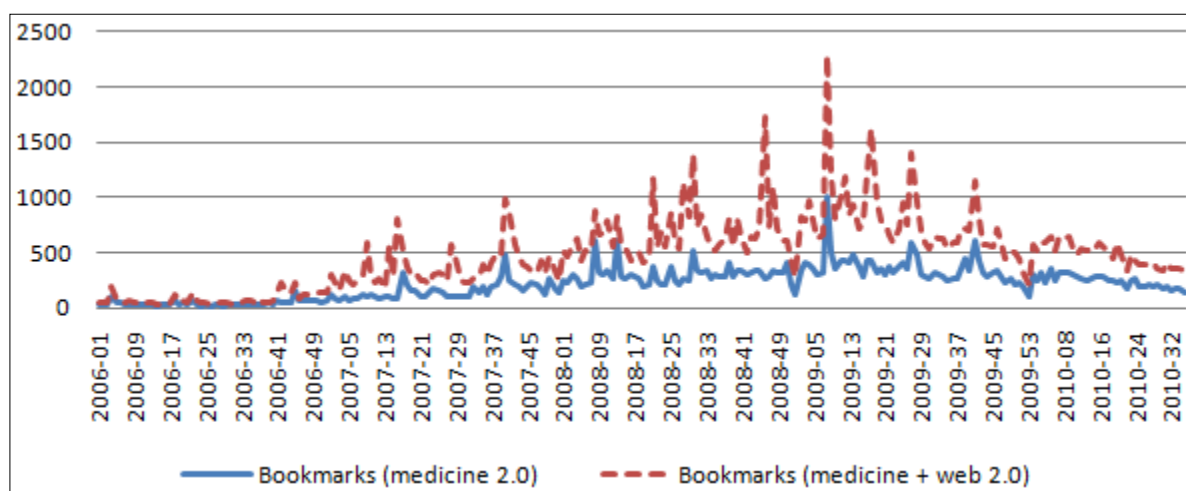


Figure 3.4 – Medicine tagged bookmarks (2006-01...2010-32, weekly data)

Table 3.2 – Bivariate Pearson correlation coefficients, on bookmarks (2006-01...2010-32, weekly data)

	Library 2.0	Library + Web 2.0	Medicine 2.0	Medicine + Web 2.0
Library 2.0	1	.899**	.510**	.466**
Library + Web 2.0		1	.444**	.384**
Medicine 2.0			1	.898**
Medicine + Web 2.0				1

****.** Correlation is significant at the 0.01 level (2-tailed).

3.3.3 Step-3: Associated Comments and Tags

An analysis of comments⁶ used by all bookmarkers over the period of 2005-2010 revealed that in the top 10 most occurring words many technology related terms were used (e.g. Javascript, Ajax, or Framework). This is probably due to the somewhat more technically minded Delicious user-base. Further it was found that web 2.0 related terms (e.g. Twitter, Flickr, Blogs, or Wikis) tend to occur consistently throughout most years (except for 2005), see tables A.1-A.6 and figures A.1 and A.2 in appendix A. More interestingly, Medicine 2.0 bookmarked resources from 2009 onwards included the term “Health 2.0” in top 10 terms of the comment text frequency tables⁷, highlighting the competing trend of “Health 2.0” vs. “Medicine 2.0”, which were also discussed in two Medical Journal papers (Hughes et al. 2008, and Van De Belt 2010). Both papers looked at the differences and concluded that based on an analysis of content, Health 2.0 and Medicine 2.0 were not substantially different and refer mostly to the same concepts. It was further suggested that Health 2.0, may be more widely used and accepted than Medicine 2.0 (Van De Belt 2010), which is also confirmed using our methodology, since 16,002 Delicious users bookmarked Health 2.0 related resources as opposed to only 1,374 Medicine 2.0 related resources.

Table 3.3 – Top 10 words (text normalised as per methodology) based on all available data for the year 2010

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Book	1.Health	1.Health	1.Health	1.Health
2.Book	2.Book	2.Book	2.Library	2.Healthcare	2.Medical	2.Medicine	2.Medical
3.Blog	3.Web	3.Tool	3.Web	3.Medicine	3.Http	3.Science	3.Information
4.Education	4.Use	4.Education	4.Free	4.Medical	4.Information	4.Medical	4.Site
5.Technology	5.Site	5.Reference	5.Site	5.Packrati.us	5.Twitter	5.Healthcare	5.Http
6.Tool	6.Information	6.Resource	6.Use	6.Reference	6.Free	6.Reference	6.Search
7.Resource	7.Online	7.Web	7.Read	7.Research	7.Patient	7.Research	7.Free
8.Packrati.us	8.Librarian	8.Technology	8.Online	8.Community	8.Social	8.Blog	8.Web
9.Research	9.Resource	9.Javascript	9.Tool	9.Health2.0	9.Online	9.Community	9.Patient
10.Librarian	10.Blog	10.Webdesign	10.Search	10.Twitter	10.Site	10.Video	10.Online

Finally it was found that top tags tend to represent concepts, whereas top comments also represent actions, i.e. share, read, search, or make. This is probably due to the nature of the actual descriptions for bookmarked links. Please see appendix A (tables A.1-A.6), for all top 10

⁶ An analysis of the related comments and all the tags used for each topic over the years 2005-2010, where each year was treated separately, was undertaken. The analysis involved the top word frequencies of comments and tags in each year. Terms were lemmatised (a text-processing technique to group together different inflected forms of a word based on Part-of-speech of the words) to facilitate analysis, after tokenising and removal of stop-words (see sub-section 3.2.3).

⁷ In a study by Hughes et al. (2008), a review of over 2405 papers indicated that “2.0” was associated with Health 2.0, Medicine 2.0, Physician 2.0, Nursing Education 2.0, Medical Librarian 2.0, and Physician Learning 2.0 which is some way a tribute to the extreme popularity of such neologisms. Only Medicine2.0 and Health 2.0 were however deemed to have enough importance to be studied further.

terms over the period 2005-2010, and Table 3.3 above for 2010. The keyword ranks of these terms tend to be relatively stable hence there are no significant or interesting patterns evident, except for the discovery of “Health 2.0” term.

3.4 Discussion

Results in this chapter indicate towards the terms Library 2.0, Enterprise 2.0 and Health 2.0 being more prevalent than their counterparts, such as Medicine 2.0, or Business 2.0. Arguably all these terms are neologisms; however, there is some indication that Library 2.0 is possibly evolving into a well accepted term / concept. The results do not indicate conclusively to what degree a term is a neologism, a number of issues have to be taken into account, including the limitations of the proposed methodology (sub-section 3.4.1). For example, any results stemming from the methodology assume that the Delicious user-base is somehow representative enough. Hence, an important consideration is the population of the Delicious user-base from which the sample (*extracted dataset*) was taken (Moore and McCabe 2001, pp. 230-277). It can be assumed with some degree of confidence that the collection of bookmarks, tags and comments aggregated on Delicious were predominantly generated by a middle aged group of users who are more technically minded than the average population, and who are more likely to have had higher education⁸. It is also assumed that a relatively wide breadth of topics would be covered by the bookmarkers, since daily Delicious usage encourages such browsing and data collection (Orchard 2006).

Nevertheless, results are indicative, and the methodology is more efficient than manual studies of gray literature by Hughes et al. (2008), or Van De Belt (2010). In addition to the ability of evaluating current overall popularity of terms, which are accurately represented by online content, the popularity of trends can be readily tracked, usually as far back as 5 years⁹. This is very advantageous, compared to for example, Google or other search engines. Finally, key terms in frequency tables that are obtained from an automated analysis of Delicious comments and related tags, considerably speeds up a classical content analysis and can be used to discover potentially interesting, related sub-topics, represented within associated comments and tags. Clearly this method is by far, more rudimentary than a classical content analysis, yet it has been used successfully to detect Health 2.0 as an important concept used in concurrence with Medicine 2.0, and this was shown by far on a larger dataset that would be feasible for analysis

8 <http://www.quantcast.com/delicious.com>, - some approximate demographic statistics from Quantcast and demographic and traffic statistics from Alexa service <http://www.alexa.com/siteinfo/icio.us#trafficstats>

9 Usually data from 2006 onward is available. Although the Delicious service was launched in September 2003, during the earlier years the system wasn't yet well known or heavily used. By 2006, user volume and user contribution volume, is sizeable, based on extracted data.

by a human expert.

3.4.1 Limitations

This study presents a number of limitations. The results from the methodology are only as good as the accuracy of the Delicious dataset. We have come across minor inconsistencies when paginating through and extracting the data from search returned result lists, but more importantly, the problematic issue was the limitation on the number of accessible result list items. Hence the second and third step of the methodology is somewhat constrained by Delicious' data access limitation, although retrieved bookmarked items are still in the thousands.

Another limitation is that only Delicious was used in this study, when it may have been sensible to extend the methodology to other bookmarking services. For instance it would be clearly of advantage to confirm results obtained from Delicious, with results obtained through one or more other web based bookmarking services. It must be pointed out that Blog based data is not equivalent to bookmarking data, in that in a blog post only the author assigns tags to content, whereas within bookmarking a single resource has tags assigned by many individuals. This is a major difference and hence, although Blog posts in aggregate can be used to measure general popularity, they do not readily lend themselves for a neologism analysis.

Finally, any conclusions drawn, based on the methodology, regarding the question of whether terms are neologisms should be drawn carefully. In fact, just as was the case in studies by Hughes et al. (2008), and Van De Belt (2010), the results should only be considered indicative.

3.5 Summary

It is hoped that not only some light has been shed on web 2.0 based term emergence but also that future investigations of neologism emergence may benefit from the herein proposed methodology. The methodology consists of three parts, each concerned with a slightly different task. First the current term prevalence, secondly trends of term's usage over the past, and finally detection of potentially interesting and related terms is investigated by each methodology step. Finally the results and limitations were discussed.

Given that in the last two chapters, web 2.0 was introduced, and related terms assessed quantitatively, in the next section a more in-depth understanding of web 2.0 systems will be sought with the help of an extensive questionnaire study and subsequent analysis.

4 Survey based Web 2.0 Investigation



4.1 Motivation for the Survey Study

Chapter 2, amongst other things, discussed notable elements of web 2.0 systems and their wider contextual significance and implications. Several important facets that played a catalyst role in web 2.0 adoption were also identified. The term web 2.0, itself was discussed in some detail, with ample reference to academic literature. This chapter presents a large survey with over 700 responses with the sole aim to substantiate and elaborate the concept of web 2.0 and social media, as it is publicly perceived. Some indication of how users relate to various issues, such as time spent, motivations, or trust awareness in the web 2.0 context, will also be given. This type of understanding is missing from current body of literature, and a survey in the web based user population seems called for, and appropriate.

Among previous work is a detailed survey of the Blogging community, from Technorati, better known as the “State of the Blogosphere”¹, which is a yearly in-depth survey of Blog related developments, and has been conducted each year from 2004 onwards. It investigates questions such as how many Bloggers make a living from blogging, in what topics they blog, what

1 <http://technorati.com/state-of-the-blogsphere/> - contains links to results of every survey since 2004. The survey in 2011 was conducted on a sample of 4,114 bloggers, and in 2010 it was 7,205, for example. On <http://technorati.com/social-media/article/state-of-the-blogsphere-2011-part1/> one may find a summary of some demographic results from the 2011 survey.

advertising strategies they use, etc. – this is probably the most complete regular survey of Blogging. Market research organisations such as Pew Internet Research Center², also conduct a range of valuable surveys, often on sample sizes of 1,000 or more respondents; however, these surveys tend to focus on marketing and advertising goals. As for academic research, Kennedy et al. (2007) conducted a large survey of over 2000 first-year students from three universities on web and web 2.0 technologies within an educational setting. However, surveys using smaller samples are much more common in academia. Berlanga et al. (2011) conducted a survey of social sharing application use (e.g. Facebook, LinkedIn), on a sample of 47 respondents. Daugherty et al. (2008) investigated the generation of UGC with a survey of 325 respondents. Researchers have often used surveys to investigate motivations for web 2.0 application use (Kuznetsov 2006, Wagner and Prasarnphanich 2007, Antin and Cheshire 2010). Kuznetsov (2006) surveyed over 100 undergraduate and postgraduate students, Wagner and Prasarnphanich (2007) conducted a survey of 35 active Wikipedia users, and Antin and Cheshire (2010) collected 20-minute long survey data from 165 respondents (their survey took place just before the respondents were awaiting reimbursements from other experiments). Despite these surveys, web 2.0 user motives are still not well researched, Bishr (2009) for example, called for the need of studies that investigate percentages of people with certain motives and what applications such people tend to use.

It is believed that the study in this chapter, specific to the needs of the thesis, will provide novel and worthy contribution to the body of academic knowledge. Indeed insights from this study assist the overall thesis in at least two respects. *First*; novel observations on the sample allow to postulate some interesting extrapolations about the overall population of web 2.0 users, which can be put into context of previously reported findings and help explain the phenomenon of web 2.0 and social media. *Second*; the survey provides empirical support for various characteristics of web 2.0 discussed throughout chapter 2, and the results also lend support for the collective intelligence aggregation framework in chapter 8, and helps to establish the robustness of the overall model and framework.

Particularly the following issues raised earlier in the thesis motivated the formulation of questions in constructing the survey.

- A significant part of chapter 2 was concerned with the concept of web 2.0, with experiments in chapter 3 attempting to quantify the prevalence of related terms. This survey hopes to provide quantitative indication as to the prevalence of web 2.0 as a concept among different demographic backgrounds of web users, and also how users relate to various issues, such as time, or trust in the web 2.0 context. This type of

2 Pew Internet Research Centre, web 2.0 related reports - <http://www.pewinternet.org/Topics/Topic-Category-3/Web-20.aspx?x=x,x&start=1>

understanding is missing from current body of literature, and a survey in the web based user population seems called for, and appropriate.

- Social media was identified within section 2.2.2³ as a popular *re-definition* from old style media, the survey aims to provide some indication as to the significance of the term and how it compares in use to the concept of web 2.0.
- A list of basic atomic activities which facilitate collaboration on web 2.0 systems was proposed in section 2.3.2.2, as part of a web 2.0 taxonomy. Different application types and specific web 2.0 website systems were also presented. The survey aims to measure the usage of some of these atomic activities, taking into consideration the activity types and characteristics of users who use them.
- In sections 2.2.1 and 2.2.3, it was mentioned that viability of online business models is becoming a tangible reality. The survey aimed to provide some indication of the prevalence of business models amongst the population of web 2.0 users, with some insights into the characteristic profiles of users.
- In section 2.2.1 trust was raised as an important factor in the re-emergence of the web revolution under the flagship term web 2.0, hence another aim of the survey was to provide some insight into the importance of trust as it relates to a web 2.0 environment.
- It has been suggested by various researchers (see section 2.2.4.4) that time spent on the web is instrumental in the potential, but also the current popularity of social media. The survey provides some indication of time spent on web 2.0 applications and how this affects other characteristics of respondents in the context of a web 2.0 environment.
- Motivation behind web 2.0, collaboration and UGC (User Generated Content) has been researched extensively (see section 2.2.4.2 for details), and a survey study which does not relate the motivations of individuals to the factors identified in this thesis would be incomplete. Motivations as they relate to individual elements of activities on web 2.0 and a wider range of applications haven't been looked at before (as far as the author is aware) and this survey may provide some much needed insight.

In summary, the primary objective of the survey was to shed more light onto web 2.0 usage habits and the main elements behind web 2.0 related features as they were identified and discussed in much detail within chapter 2 of this thesis. Due to limitations in terms of insights that a survey-response sample can provide (Passmore et al. 2002), naturally not all research questions could be answered; however, every effort was made to limit any negative consequences within the survey (limitations of the survey are highlighted in section 4.5). The actual elements of survey design are discussed next.

³ For example, based on the discussion in chapter 2.2, social-media was expected to be a far better known term amongst the wider general public than web 2.0. Without undertaking a survey-style research to investigate further, such a claim cannot be substantiated.

4.2 Survey design

The main aim of the survey design was to facilitate the collection of as large a sample, as accurate, and as representative of the overall population as possible⁴. Ultimately the goal is to draw conclusions from a sample that will to some extent be valid for the population of web 2.0 users. In this section, the actual phrasing of questions, the main design considerations for an online survey, and various issues connected to the sampling procedure are presented. A systematic and careful design, which closely follows the guidelines from recent literature, was followed, in order to ensure that collected survey answers are of genuine and accurate nature.

4.2.1 General Considerations

By increasing a respondents' willingness to answer a survey, reliability and validity of survey-responses tends to increase, and less error equals better data (Hill 2009). One way to increase respondents' willingness to answer a survey is to keep the questionnaire short. Not only does this increase the potential of collecting significantly more unique responses, but it also increases the reliability and validity of individual responses. Since there was a need for a sizeable sample, the survey was designed to be brief, with the aim of the final survey design not to take more than 1-2 minutes to complete, for an average respondent. Hence the survey contained only ten questions with an optional question asking for demographic information; the entire survey roughly fits onto an A4 sheet. All questions were closed type questions, since open questions are often too vague or general to meet question objectives and they take more effort for the respondent to answer, whereas closed questions are easier to analyse and to compare across survey responses (Martin 2006). In order to provide a valid answer it was only required to click on checkboxes and radio-buttons throughout the survey. The online version of the survey form is still available on <http://www.newsmental.com/survey.aspx>, for inspection, and a print copy is also included in appendix B, figure B.3.

In addition to the already mentioned advantages of a brief survey design, the short design also made snowball sampling more feasible (see section 4.2.3, on sample design), and since completion of the survey was not time consuming, responses from a wider set of demographic backgrounds were expected⁵. An important issue which is closely tied together with survey

4 Large – in terms of the number of responses from unique individuals; Accurate – in terms of responses to be representative of the question's actual intent; Representative – in terms of reaching various strata of the population of all web users so that at least some indicative conclusions can be drawn about the entire population from the sample. These are common requirements for surveys; Moore and McCabe (2001).

5 Busy individuals, such as businessman, doctors, professors and other full time professionals are more likely to sacrifice their time when they are presented with the promise of a very short survey. Many individuals with

design is how one motivates individuals to take out the time to complete online surveys. Motivation was offered through three modes: **1**-Already mentioned low, 1-2 minute survey completion time, **2**-Guarantee of anonymity, the promise of anonymous surveys generally increases response rates⁶, **3**-A non-monetary incentive was to appeal to potential respondents by stressing the fact that only 1-2 minutes of their time would contribute to research, and that the study outcomes will be openly published at completion – out of 726 respondents as many as 250 provided their email addresses for the purpose of receiving the study outcomes. These three points were used in advertising, to motivate participation in the survey.

4.2.2 Question Order, Format and Wording

Martin (2006) pointed out that small changes in question wording, or order can substantially affect responses. Respondents do not necessarily respond to the literal meaning of a question, but rather to what they infer to be its intended meaning. A so called conversational perspective to survey design has recently been advocated (Martin 2006), in which considerations of the influences that one question may have on interpretations of subsequent ones, are carefully considered. A quote from Martin (2006) elaborates; *“The argument is that when people are asked to form a judgment they must retrieve some cognitive representation of the target stimulus, and also must determine a standard of comparison to evaluate it. Some of what they call to mind is influenced by preceding questions and answers, and this temporarily accessible information may lead to context effects.”*

The survey was designed to ask questions that would help answer main aspects underlying the web 2.0 phenomenon. First the stage for the topic of the questionnaire would be set, by asking a relatively generic, but topical question (*e.g. Have you heard of the term web 2.0?*). This would be followed-up by a question asking for the applications that respondents have used and the basic actions that they have performed previously on the web. Once these questions are dealt with, further questions can concern themselves with specific sub-topics of interest that now assume contextual awareness of the respondents – towards example applications and activities on web 2.0 that respondents indicated, they have used. Respondents are able to relate to

busy lifestyles however will avoid surveys of 10-15 minutes or more in length. This may introduce demographic bias into the sample. Our study stands out from among other published studies since it does not inconvenience potential respondents based on time to fill out (time costs are minimal). Despite losing some detail in response information, it is believed that much is gained in reducing the bias.

⁶ The survey did offer an option to provide email address of the respondent at the completion of the survey (see http://www.newsmental.com/survey_thank_you.aspx or figure B.3 in appendix B) which would be used to inform the respondent of the final outcomes of the study. This email however was not linked in any way with the questionnaire answers and could not be directly related to any particular set of responses, hence anonymity was maintained.

possible answers by the context introduced in earlier questions. In order to provide structure to the overall questionnaire, visual cues were used to group similar questions together⁷. The survey also grouped questions into logical groups that were made accessible by a right-hand side fixed navigation menu (visible at all times), which would allow to scroll / jump within the page. The questionnaire question groups were, **1-Background** questions, **2-Trust**, **3-Time**, **4-Motivation**, **5-Closing / Final** questions. Martin (2006) further recommends avoiding and minimising embedded clauses or complicated sentences, in order to elude cognitive overload due to complexity or ambiguity which may result in partial or variable interpretations and misinterpretations of questions⁸. Given all presented considerations, an initial survey draft design of the questionnaire was sent out to a number of PhD colleagues and a small number of test subjects for comments and a small pilot run; this is considered good design practice (Passmore et al. 2002). The feedback received was useful in clarifying question structure and wording.

Each question from the survey is now presented in turn, with design decisions explained as and where necessary. The order of the questions and factors they relate to is as follows:

1. **Web 2.0 competence: Q1, Q2 and Q3**
2. **Business models: Q4**
3. **Trust: Q5 and Q6**
4. **Time: Q7 and Q8**
5. **Motivations: Q9**
6. **Wikipedia: Q10** (as a notable *web 2.0* body of encyclopaedic reference)
7. **Demographic optional questions: Q10p** (*age, education level, expertise domain*)

1. Q1 (Web 2.0 awareness), asks to “Check each statement that you agree with:” where four multiple response checkboxes are provided. These four responses can be regarded as a pair of two Guttman scale questions⁹. The responses are: **a-**“I have heard of web 2.0”, **b-**“I have a rough idea / understanding about web 2.0”, and relating to social media: **a-**“I have heard of social-media”, **b-**“I have a rough idea / understanding about social-media”. **Q2 (Web 2.0 apps)**, simply asks the respondent to “Check the applications that you use or have used in the past”. The multiple response checkbox choices are; Twitter (Micro-blogging), Youtube, Facebook / Myspace (Social networks), Delicious.com (link sharing), Flickr / Picassa (Picture sharing), Wikipedia (shared resource encyclopaedia), Digg / Reddit (news sharing), Craigslist, Ebay, Amazon (commercial websites based on web 2.0 elements). **Q3 (Web 2.0 activities)**, requests

⁷ Visual cues made use of colours, borders and spatial separation – see <http://www.newsmental.com/survey.aspx>

⁸ Martin (2006) contains a number of useful examples of actual question phrasing issues.

⁹ On a Guttman scale, items are arranged in an order so that an individual who agrees with a particular item also automatically agrees with items of lower rank-order.

respondents to “Check all the activities that you have done on the web at some point”. Possible responses represent 10 different multiple response checkboxes. The answers indicate type of activities such as submitting Blog posts, commenting on content, editing shared resources...

2. Q4 (Business model), simply asks respondents to tick all checkboxes that apply. The multiple response checkboxes are: whether respondents have followed online adverts on purpose; bought something online (trip/hotel-booking, book, movie, music, clothes, membership...); used PayPal, webmoney or other web-based payment system; and whether they used a group buying website (such as Groupon, BuyWithMe, Twango...).

3. Q5 (Trust): Two questions (Q5 and Q6) are intended to find out about people’s attitudes towards trust on the web. The first question (Q5) suggests the statement “I trust most websites that I use on a regular basis.” and asks of the respondent to rate their agreement with the statement, i.e. strongly agree, agree, disagree, or strongly disagree. This is known as a Likert style response-item¹⁰. All Likert style questions in this survey use five points in their scale. This is based on (Martin 2006) as she suggests that the recommended number of categories in a scale should be seven, plus or minus two, however Passmore et al. (2002) further warns of the so called “floor” or “ceiling” effects, where subjects tend to choose responses that cluster at either the top or bottom of the scale¹¹, hence it was decided to keep the number of categories in a likert scale relatively small and to use a five point scale. **Q6 (Trust)** asks respondents to tick all the trust-level related statements that they agree with (multiple responses are allowed). a- “I feel comfortable sharing my personal details (email, pictures, opinions ...) on web-pages”, b- “I feel comfortable sharing my personal details on web-pages that use all the appropriate security precautions and measures”, c- “I feel comfortable purchasing products online”, d- “I feel comfortable purchasing from online stores that I know”, and e- “Looking at a web-page I can usually judge whether it is a trustworthy page or not”. Responses a, b, and c, d, are both separate Guttman-scale responses, where agreeing with b and d automatically means the user also agrees with a and c. Response e, is an additional piece of information that helps to provide some indication of a respondents ability to judge trustworthiness of sites.

4. Q7 (Time), is a Likert-scale question, as is **Q8 (Time)**, and they both ask the same question, with a small yet significant difference. Often when asking time related questions, surveys may ask respondents to report time in terms of absolute frequencies (e.g. “Up to ½ hour, ½ to 1 hour ...”), however it was found experimentally that providing a frame of reference to the respondent via an absolute frequency scale can lead to biased frequency reports purely based on

10 When responding to a Likert questionnaire item, respondents specify their level of agreement or disagreement on a symmetric agree-disagree (*often 5, 7 or 9 point*) scale (Likert 1932).

11 As a result of this clustering, the instrument may not capture a significant amount of the true variability in opinion among respondents (Passmore et al. 2002).

the survey designers choice of the scale (Schwarz and Sudman 1996)¹². Hence Q7 and Q8 both ask for time indication on a relative scale. Q7 asks respondents to compare the time spent online to their peers¹³, whereas Q8 provides the statement “I spent too much time on Twitter, Facebook, Youtube, Wikipedia...” and asks for a likert scale agreement. Both likert scale responses are expected to be highly correlated with each other, and this is used to validate the surveys reliability later within the chapter.

5. Q9 (Motivation), asked respondents “Would you contribute to any of the websites mentioned because (tick all that apply best to you):”, with the following possible (multiple answers allowed) responses, “I want to contribute content for the greater common good” (Altruism), “I want to contribute content for greater good but I expect similar action in return” (Reciprocal Altruism), “I want to contribute to my community and to help raise awareness within it from my actions” (Community / Sense of belonging), “I want to build my online profile (i.e. web reputation)” (Self presentation), “I want to show my experience and autonomy / knowledge in a certain topic” (Autonomy / Knowledge). Each response is a possible motivational factor, adapted from literature, specifically Kuznetsov (2006). Also, see section 2.2.4.2, which discusses previous work in motivation on web 2.0, at length. Given the somewhat ambiguous nature of the question’s request, the use of “don’t know” as an explicit response option was considered. However, according to literature on this topic, including such options in surveys results in loss of data and it was also suggested that offering the option does not improve data quality or reliability (Martin 2006).

6. Q10 (Wikipedia): I decided to provide an additional question to the survey, a simple Likert scale question – i.e. “I consider Wikipedia.com to be a useful body of encyclopaedic reference:”. It was of particular interest to us to find whether respondents consider Wikipedia to be a serious text of encyclopaedic reference. This has been a heated topic of ongoing debate among academia and media. This result is expected to contribute to one side of the argument.

7. Q10op (Demographic), the last question in the survey asks users to provide the age-group they are in, qualification level achieved and expertise area (i.e. field of humanities, business, engineering, or computer science), via a set of three drop down menus. Passmore et al. (2002), recommends placing questions of more sensitive nature towards the end of surveys, since by that time, the subject is feeling comfortable and familiar with the survey format and is more

12 This is, as Martin (2006) explains, because there is a strong normative expectation, where respondents are influenced by what is perceived to be the normative or average response.

13 “Compared to your friends, in your free time how much time per day do you spent on sites such as Facebook, Digg, Youtube, LinkedIn, Amazon, Ebay, Craigslist, Twitter, Myspace, Reddit, Delicious, MySpace, Flickr or Picassa:”, and the likert-styled responses are: 5-“Probably much more than most of my friends”, 4-“Probably more than some of my friends”, 3-“Same as majority of my friends”, 2-“Less than most of my friends”, 1-“No time at all, or nearly no time”.

likely to respond.

As far as validation of questions is concerned, only Likert-scale questions were required by default, i.e. questions 5, 7, 8 and 10. Missing responses from multiple-answer (checkbox-style) questions are allowed. This is intentional since otherwise the respondent would be forced to select answers in questions that they might ultimately not have an answer for; e.g. the respondent never having used a web 2.0 app would not have to pick one in question 2 of the survey.

4.2.3 Sample Design

In several previous studies, sample design, which is a crucial survey design issue, received disproportionately little attention (e.g. Daugherty et al. 2008). A clear understanding of sample design is necessary in order to appreciate the sample's relationship to the population that the sample is intended to represent. It was envisaged for the survey to investigate as wide a sample as possible in terms of age-group, skill levels (education level) and technical expertise. Since the web 2.0 phenomenon is of relevance to anyone with an internet connection, sampling was limited to online-only responses. Unfortunately a readily available approximate sampling frame from which to draw samples of the population to construct an SRS, stratified SRS sample¹⁴, or similar, was not available. Any such sampling frame for all practical purposes would likely still include a number of biases, e.g. ISPs representing certain types of customers. Since SRS sampling was ruled out, a variation of snowball sampling (also sometimes referred to as respondent-driven sampling) was employed (Goodman 1961), in order to collect as many answers as possible from a coherent group of people. This is a non-probability sampling method where existing respondents are asked to recruit other respondents from their own acquaintances; Salganik and Heckathorn (2004) describe how this method can be bias-free. More recently, Fowler and Christakis (2010) further elaborated how information gets propagated in social network from person to person in an online setting. They identified the mechanism of mimicry to be a main underlying concept behind such information propagation, in which people (network-connections) significantly mimic the sharing of information they consider 'cool' and / or otherwise useful. It is this effect, we observed, as at some point people were passing the survey around with literally groups of people filling it out at various locations (around campus / departments, companies, and families / friends, or using facebook).

¹⁴ A useful but brief overview of basic sampling designs is available on pages 256-262 of Moore and McCabe, 2001.

4.2.4 Survey Distribution Method

Initially the survey was shared via email to all university departments (staff and students alike) at Loughborough University, various contacts and friends. Respondents were actively encouraged to share the survey further. Since it was felt that older, young and especially less technically minded people might be under-represented in the survey response set, potential respondents were encouraged to spread the survey voluntarily and without any discrimination based on age, skill, etc. to their colleagues, friends and family – suggesting they present the survey as a short, anonymous, simple, 1-minute long survey, to be filled out by anyone, in order to “help research on the evolution of World Wide Web”. This turned out to work well, since the range of demographic spread in age, technical ability and education-levels was quite wide (see section figure B1 in appendix B, for details of the survey’s demographic variation). This was the main intention in the snowball sample design (Salganik and Heckathorn 2004)¹⁵; however, due to the survey being primarily advertised within the academic community, there is clearly a bias towards this stratum of population, which is further discussed in the limitations section 4.6.

4.2.5 Self-selection Bias

A self-selected sample (also known as voluntary response sample bias, see Moore and McCabe 2001) is a real problem, for example in the survey by Daugherty et al. (2008) this likely produced some inaccurate results¹⁶. This survey study tried to minimise this effect by making it simple to fill out (*e.g. only check-box and radio button clicks are necessary to answer any questions; sections of the questionnaire can be reviewed and accessed easily using a fixed menu on the right side of the survey form*) and quick to complete (*e.g. 1-2 minutes only – brief survey; just a single page with 10 questions; demographic questions were optional as not to discourage respondents from having to divulge ‘personal’ details; questions were phrased carefully to minimise ambiguity*). This would help not to discriminate against less computer literate respondents and avoid discrimination of highly qualified individuals who potentially have less time to spare for answering surveys.

¹⁵ Snowball sampling is considered very effective in having a wide reach, and when sampling hidden populations (i.e. sampling from populations not represented in the sampling frame), see Salganik and Heckathorn 2004.

¹⁶ Despite producing valuable insights, given the low response rate and the overall length of their survey (23% of respondents did not complete the survey because of its length), it is likely there was a considerable self-selection bias in their sample. Therefore it is believed that in addition to presenting more recent results, the survey presented in this thesis carries less chance of self-selection bias and hence response values should be more realistic and accurate.

4.3 Evaluation of the Survey-response Data for Analysis

In total 726 survey responses were analysed, from the time interval 11th April 2011 to 22nd May 2011. Altogether 736 responses were collected, although 10 responses were discarded as they were identified to be duplicates. Several issues associated with the responses are discussed in this section.

4.3.1 Avoiding Misrepresentation of the Population's Sample

Since only one answer per person is considered valid, a number of measures were taken to avoid miss-representation of the population's sample. Throughout the advertisement of the study it was explicitly stressed that only one answer per person is allowed and providing more answers would be invalid. The same was pointed out after the survey's completion – on the *thank you* page that appeared after survey submission. Responses with same answers, submitted from the same unique IP, and within a very short interval were excluded as duplicate and / or fraudulent submissions. In the final set of 726 survey responses, there are 30 instances where the IP was shared with at least two or more other survey responses. Dynamic IP allocation is common with many ISP providers, making shared IPs rather common. A closer investigation of those responses showed that, firstly – they were unique answer responses, and secondly – they were coming through organisational ISPs, hence after close inspection these 30 responses were considered valid and 726 responses were evaluated throughout this chapter.

4.3.2 Geographical Survey-response Distribution

Since the IP of each response was recorded we can to some extent assign an approximate geographical area to each response. It was found that 146 unique responses (20% of all responses) originated from an IP range reserved for Loughborough University¹⁷. The remaining responses were mostly from Britain and other parts of Europe, including Austria, Switzerland, Germany, France, Slovakia, and from several countries outside of Europe. This represents a wide sample from across numerous geographical areas. Clearly the response sample is limited to areas with internet connectivity, yet this is not issue since the thematic focus of the survey requires online connectivity. Since the internet is free of political borders we did not want to limit responses to individual countries, and to consider web 2.0 as a World-wide phenomenon, and capture the current habits of its users along varied age-groups, skill-levels, technical expertise and geographical regions.

¹⁷ The IP range for Loughborough University is (158.125.0.0–158.125.255.255)

4.3.3 Ensuring Overall Reliability

Typically a robust survey evaluation should establish reliability of the survey, i.e. are the survey's answers to the questions measuring the item of interest consistently? Cronbach's alpha (Cronbach 1951) is commonly used for this task, since the survey is very short (only 10 questions) and the interpretation is intuitive, a bivariate correlation on questions 7 and 8 to help ascertain reliability was used. Questions 7 and 8 represent time spent on social applications – it is expected the pair strongly correlates over the entire 726 response survey set. Spearman's rho correlation indeed revealed a strong correlation between the relative-assessment and self-assessment of time spent on websites, i.e. $r = 0.635$ (726), p (two-tailed) $< .01$. This provides indication that respondents acted consistently, and reliability seems to be satisfied.

Throughout an exploratory analysis of survey responses, it was possible to further ascertain reliability of the survey (see section 4.3.4). Outlier responses were checked for integrity, and it was found that outliers could generally be explained. In particular a positive answer to Q4-4 (previous participation in group buying) would imply a very confident web user, and it would be expected that such a user would clearly be familiar with web 2.0 / social media or at least have used numerous such applications before. Indeed out of the 104 respondents who used group buying, there was statistically significant tendency in users being more experienced.

4.3.4 Missing Responses (Unanswered Questions)

Since only Likert-scale questions were required (as per default validation checking), and missing responses from multiple-answer (checkbox-style) questions were allowed by design, it was possible that some questions were not answered at all. Table 4.1 provides an overview of the number of unanswered questions in the survey.

Table 4.1 – Unanswered multiple-answer (checkbox-style) questions, out of all 726 responses

Question	Q1	Q2	Q3	Q4	Q6	Q9
Unanswered	89	0	28	7	21	213

An unanswered question can mean that the user simply did not associate with any of the statements, or hasn't used or done any of the activities. Of course there is also the risk that a user accidentally ignored the question. The bullet point list below attempts to provide some explanations for the missing answers.

- (Q1) – 89 respondents have simply never heard of web 2.0 before.
- (Q2) – All respondents answered question two, i.e. every single respondent has used at least one web 2.0 application, the average being five applications.

- (Q3) – 28 people did not answer question three, however from these 28 people all have used a web 2.0 application (Q2), but more than half of them (13) have never heard of social media or web 2.0 (Q1).
- (Q4) – Seven individuals quite simply did not relate to question four, presumably because they simply haven't necessarily bought anything online. This assumption is also supported by the below average web 2.0, trust and time scores for these seven users.
- (Q6) – 21 did not answer question six on trust awareness; however, for this respondent group the median and mode score for Q5 (likert-scale question on trust – we can use it to check the outcome of Q6), was 3, whereas for the entire sample it was 4. This indicates that the 21 missing respondents for Q6, did not necessarily have an opinion on trust-awareness, and that is likely why they ignored answering Q6.
- (Q9) – As many as 213 people did not answer question nine on motivation, this is partly explained by the more ambiguous nature of the question. It might be relatively hard for somebody to identify what actually drives and motivates them. Also some people were unable to answer Q9 since it does not apply to them – i.e. they do not contribute content in a significant way (see section 4.5.6).
- (demographic Qs) – Only one single response with all the optional demographic values at their default drop-down menu selections¹⁸ was submitted in the 726 responses. Even this response seems to be legitimate response given the plausible answers in other questions within this response (i.e. responses were other, than default responses). This also highlights that all respondents who filled out the survey also supplied their demographic information.

4.3.5 Demographic Distributions of Responses

A variety of age, qualification and skill-set groups are present in our final sample. The frequency tables for age-group, education-level and area of expertise in tables 4.2, 4.3 and 4.4, present the 726 survey responses as they are distributed along these demographic variables.

Table 4.2 – Age group – Frequency table for the 726 survey-responses

	Frequency	Percent	Cumulative Percent
≤ 19 years	89	12.3	12.3
20 - 30 years	396	54.5	66.8
31 - 40 years	123	16.9	83.7
41 - 50 years	51	7.0	90.8
51 - 60 years	55	7.6	98.3
≥ 61 years	12	1.7	100.0
Total	726	100.0	

About 83.70% of all respondents were 40 years and younger. However, as many as 118 responses from 41 year olds and older individuals were received. Unfortunately survey-answers from 61 year olds and older individuals are only limited to 12 responses. It thus is sensible to

¹⁸ These are [19 years or less–age, Still in school–qualification, Computer and Information Sciences–expertise]

group 51-60 and ≥ 61 year olds into a common group to deal with the small sample size. This means that some resolution for the age group gets lost, and since the oldest age group represents most likely retired people, who may exhibit quite different behaviour compared to younger age groups, this group was kept in most of the analysis. The small sample in this age group will only serve as weak indication of any possible pattern, and by no means will any substantive conclusions be made about the oldest group.

Table 4.3 – Qualification level – Frequency table for the 726 survey-responses

	Frequency	Percent	Cumulative Percent
Still in school	12	1.7	1.7
Finished school	37	5.1	6.7
Undergraduate	306	42.1	48.9
Postgraduate	231	31.8	80.7
PhD / Dr	140	19.3	100.0
Total	726	100.0	

Most of our respondents are well educated (*51.10% are studying towards or already received a postgraduate degree*), and respondents who consider themselves to have expertise in computing are the biggest expertise group in our sample (*32.50%*).

Table 4.4 – Expertise area – Frequency table for the 726 survey-responses

	Frequency	Percent	Cumulative Percent
Theoretical	62	8.5	8.5
Arts / humanities	168	23.1	31.7
Business / economics	106	14.6	46.3
Engineering	154	21.2	67.5
Computers	236	32.5	100.0
Total	726	100.0	

Arts and humanities is the second largest response group (23.10%). In fact Figure 4.1 illustrates that the area of expertise variable has a bi-modal distribution. The first mode, clusters around a technical crowd, and the second mode clusters around the more arts related respondents.

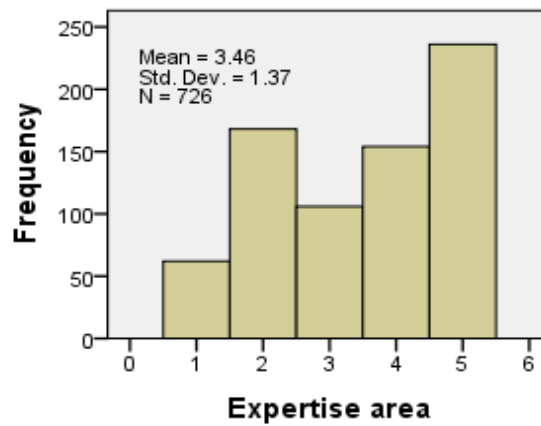


Figure 4.1 – Distribution histogram of respondents' expertise areas (higher x-axis no. can be interpreted as more IT technical skills)

The appendix B, figure B.1, contains the whole set of distribution histograms for all numerical response variables.

4.4 Data Pre-processing

4.4.1 Representing Responses Numerically

A simple score for each question was generated by summing up the count of all the ticked responses in an answer provided per question. For example for question Q2, a multiple response choice of web 2.0 applications used by the respondent had to be selected from a choice of 10 possible applications, hence a score between 0-10 would result for question Q2. This was applied to all multiple response questions (Q1 [0-4], Q2 [0-10], Q3 [0-10], Q4 [0-4], Q6 [0-5], Q9 [0-5]), with the exception of Q1 and Q6, where in Q1 the choice of the last two answers was exclusively weighted being twice as important¹⁹, and in Q6 the choice of answers 2 and 4 was weighted twice as important as their counterpart choices 1 and 3²⁰; this is because of the Guttman-scale in these two questions²¹. The Likert-scale questions (Q5, Q7, Q8, Q10) were all represented with a score in range of 1-5, where 5 stands for strong agreement and 1 for strong disagreement with a question's statement.

Finally the background question for the age-group was represented on a score ranging from 1 (youngest age-group) to 6 (oldest age group). Similarly qualification level and expertise area were ranked from 1 (lowest qualification / least technically-IT skilled expertise, respectively) to 5 (highest qualification / most technically-IT skilled expertise, respectively). Effectively,

¹⁹ Instead of having just heard about web 2.0 or social media, the respondents here indicated that they have actually a rough idea what the term means, hence a higher score would be given.

²⁰ Instead of using any websites to share personal data or conduct purchases, only secure websites are preferred.

²¹ A set of items that can be ranked in some order so that, for a rational respondent, the response pattern can be captured by a single index on that ordered scale. See http://en.wikipedia.org/wiki/Guttman_scale

converting all three background questions into ordinal values.

4.4.2 Representing Factors Numerically

The summed scores for questions Q1, Q2 and Q3 were aggregated into a new latent variable “web 2.0 competence” with range 0-24. Trust [range 1-10] was represented by the aggregate score of Q5 and Q6, and Time by summing Q7 and Q8 which resulted in a range of 2-10 (since both are compulsory likert scales). Since in both cases the lowest score was 1 and 2, the scores were transformed so that trust and time would be put into ranges of 0-9 and 0-8, respectively. This allows for an easier numerical interpretation, e.g. a score of 2 for time would confusingly mean that in both Q7 and Q8 the respondent indicated that they spent practically no time online.

4.4.3 Factor Analysis – Confirmation of Factor Choice

In order to add support to the choice of latent variables (Web 2.0 competence, Trust and Time; above) a factor analysis using PCA (Principal Component Analysis) was performed. Factor analysis is commonly used for feature reduction and for inferring latent variables by extracting so called principal components that best explain the variance in the data with each component having as high a variance as possible, constrained by being orthogonal (uncorrelated) to the preceding principal components (see Han and Kamber 2006 for more details). The final resulting component matrix after rotation (Varimax – orthogonal rotation method was used) of all the principal components where their Eigenvalues were higher than 1.0²² are shown in table 1.5 below (note: other output, relating to this Principal Component Analysis is detailed in appendix B).

²² It was suggested in Kaiser (1960) that a cutoff for Eigenvalues of 1.0 is generally appropriate, based on the idea that the Eigenvalues represent the amount of variation explained by a factor and that an eigenvalue of 1.0 represents a substantial amount of variation. Alternatively the point of inflection on a scree plot is often used.

Table 4.5 - Rotated Component Matrix^a

	Component			
	1	2	3	4
Q5 Trust (likert)	-.039	.053	.750	.245
Q7 Time (relative)	.250	.849	.085	.047
Q8 Time (likert)	.048	.906	.137	-.009
Q10 Wikipedia useful (likert)	.095	.018	.093	.915
Web2.0 competence	.957	.188	.124	.015
Q4 score (business)	.411	-.014	.405	-.171
Q6 score	.225	.178	.827	-.071
Trust	.158	.160	.956	.048
Time	.155	.974	.125	.018
Varied motives	.522	.230	.121	.265
Q1 score	.708	-.115	-.010	.173
Q2 score	.785	.176	.169	-.089
Q3 score	.824	.296	.122	-.016

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

The table displays four columns for each of the principal components, and a row for the loadings of each variable onto a principal component. The factor loadings that best associate each variable with the appropriate factor are highlighted. Looking at the second and third columns associated with factors 2 and 3, we can see high loadings for all the time and trust related questions respectively. Each represents a quite distinctive theme, i.e. trust and time – which was expected. In the first column associated with factor 1, we can see high loadings for web 2.0 competence, Q1 score and Q2 score, followed by Q9 score (variety of motives for using web 2.0), and possibly Q4 score (business model related). Q4, however also loads highly on factor 3 (related to trust), from which we infer that this business related variable does not quite fit into any single one factor, but is related to trust (factor 3) and web 2.0 (factor 1) awareness. Quite naturally it makes sense for business related activity to be related to trust of the individual and skill / awareness of the web 2.0 platforms used to accomplish the business related activity. The variety of motives (Q9), does not load quite as highly on web 2.0 awareness factor 1, as some other variables do, however this could be expected since motivations for collaborative participation are not necessarily influenced by web 2.0 awareness or web 2.0 applications usage (at least not at the resolution of the survey provided feedback). Finally the fourth column (factor 4), indicates a high loading for Q10 (Wikipedia usefulness), since this is thematically somewhat unrelated to any of the other factors, this distinctive high loading on factor 4 makes sense.

Overall it is found that the thematic factors extracted through PCA in this experiment, confirm the initial analysis design decision on the factors selection for this survey (in section 4.4.2).

4.4.4 Frequency Distributions and Normality Tests

In order to decide whether non-parametric or parametric tests are appropriate in analysis, consideration was given to variable distributions. Unfortunately most variables in the sample have a small range due to the nature of the survey – e.g. any Likert scale question ranges only between 1 to 5. This can be somewhat circumvented by aggregating variables into factors (as described in sections 4.4.2 and 4.4.3). This way we end up with latent variables, such as *web 2.0 competence*, with a larger range, by aggregating Q1, Q2 and Q3 scores.

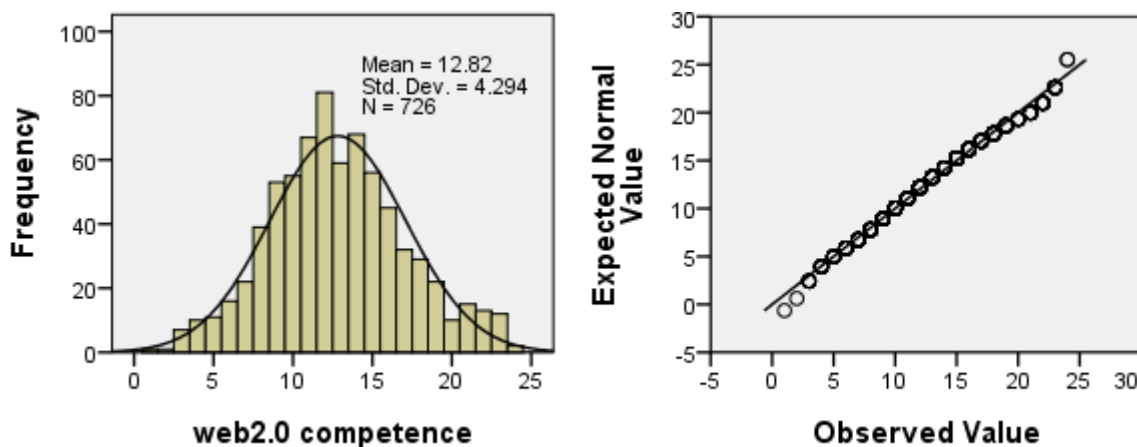


Figure 4.2 – distribution histogram (*left*) and normal q-q plot (*right*) of *web 2.0 competence*

The figure 4.2 shows the *web 2.0 competence* variable to be more or less normally distributed, also with its q-q plot confirming a normal distribution (see Moore and McCabe 2001 for further statistical details). However, none of the other factors or individual variables has a range of more than 10 values, which provide little resolution for a substantive decision about the normality of a distribution, even though the underlying variable might be normally distributed. Some variables also show multimodal or strongly skewed distributions (see histograms in appendix B, figure B.1). Hence where deemed appropriate Man Whitney U-Test, Kruskal-Wallis Test (non-parametric equivalent of one-way ANOVA), Chi- X^2 test, Spearman's rho, and Kendall's tau b, will be used. A good statistics text covering details of these procedures is Moore and McCabe (2001), or Conover (1998).

4.5 Survey Results and Analysis

4.5.1 Correlation Analysis

In table B4 of appendix B, a correlation matrix between all the factor scores and demographic variables is presented; this sub-section describes the statistically significant correlations, in order to provide an overview of interesting relationships between variables, and to confirm expected relationships. Kendall's tau b measure was used to measure correlations since it makes allowance for tied data, which will occur quite frequently since the variable scores only have a few ordinal values²³. For example a relationship was expected between time spent online and the average user age, which was found to be $-.337$, at p (two-tailed) $< .0005$ significance. In fact all correlations reported in this sub-section are significant at that level, although some of the discussed correlations, despite being statistically significant are relatively weak. Hence one cannot say the correlations found represent strong relationships between the variables; however, the correlation matrix does portray an indicative picture of a number of relationships, which are highlighted in the bullet points below and summarised in the subsequent paragraph.

- **Demographic variables:** Age and qualification level are positively correlated (.286); however, there is no significant relationship with either towards expertise area (IT technical skill). Age is negatively related to web 2.0 competence ($-.129$), trust ($-.240$), time ($-.337$), and the variety of motives ($-.140$). Eventhough age and qualification level are related, qualification level does not correlate with a decrease in web 2.0 awareness (.029). Only trust and time spent online are negatively correlated ($-.155$ and $-.142$, respectively). Finally, as expected, there is a systematic positive correlation between expertise area and web 2.0 competence (.207), which points to the likelihood of more competent web 2.0 users the more they are experienced with IT related skills.
- **Web 2.0 competence:** It was found that web 2.0 competence correlates significantly with business score (.266), trust (.229), time (.236) and motives (.365). Interestingly motivations are more heterogeneous with higher web 2.0 competence, and correlating individual scores of the web 2.0 competence factor reveals that web 2.0 activities (Q3 score) has the highest positive correlation towards motives variety. The degree to which people find Wikipedia useful doesn't correlate with web 2.0 competence.
- **Business (Q4):** The significant correlations exist with trust (.237) and as already mentioned web 2.0 competence (.266)
- **Trust and time:** Trust correlates significantly with time (.225), variety of motives (.213), and time spent also correlates with a higher number of web 2.0 motives (.229).
- **Motives and Wikipedia usefulness:** There is one significant correlation between Wikipedia's usefulness and the increasing heterogeneity of motives (.135).

²³ With other correlation measure variants the presence of a lot of ties will pull the coefficient towards zero, thus implying that a relationship is weaker than it really is.

Heterogeneity of motives is correlated with trust (.213) and time (.229) but most importantly with web 2.0 competence (.365), as already mentioned.

In summary then, it was found that the older people are generally more likely to spend less time online, put less trust into web applications, and show less web 2.0 competence. Older people also might have somewhat more single minded motivations for using the web 2.0 applications. Qualification level does not correlate negatively with web 2.0 competence, although more qualified individuals spent somewhat less time on web 2.0 and have less trust in general. Individuals from an IT skilled / technical background might show higher web 2.0 awareness. Web 2.0 competence was found to be one of the most important variables in terms of being positively correlated with business score, trust, time spent online and motives. People, who are heavier web 2.0 users, tend to have a higher trust score. They also use the web more significantly for business activities and spend more time on it. Heterogeneity of motives increases with the amount of web 2.0 applications used and the range of web 2.0 activities the individual undertakes, which makes a lot of intuitive sense. Wikipedia correlates significantly with heterogeneity of motives, this is probably due to the varied reasons that people have for contributing to Wikipedia; however, trust, time spent online, age, qualification, IT expertise and even competence doesn't have any noticeable relationship towards people's attitudes to Wikipedia.

4.5.2 Web 2.0 Related Questions

For the purposes of this thesis, and in support of chapter 2, some of the most interesting insights from the entire survey came from the analysis of the first three survey questions; discussed in this section. First the public awareness of web 2.0 and social media is quantified, then individual web 2.0 applications, and finally the types of activities undertaken by individuals on web 2.0 are presented and discussed.

4.5.2.1 Public Awareness of the Terms "Social Media" and "Web 2.0"

The survey provided an excellent opportunity to extend support for the usage of the mentioned terms. It was found that exactly 50% of the sample is aware of both, web 2.0 and social media. The more educated the respondent group was the more likely they were to know about both, web 2.0 and social media²⁴. 77% respondents of all computer-IT area of expertise were aware

²⁴ Aware of both terms (percentages are of each educational group), 33.3% (still in school), 35.1% (left school), 45.4% (undergraduate students), 53.7% (postgraduate students), 59.3% (PhD / Dr.)

of both terms, in comparison with only 28% of individuals within arts / humanities area. As for the other expertise areas, 41%, 44% and 40% of respondents within business / economics, engineering and theoretical areas of expertise, respectively, knew about both terms. Knowledge about both topics is more prevalent within individuals with IT and computing related knowledge. From the remaining respondents 76% (274 out of 363) were aware of at least web 2.0 or social media. However for these 274 respondents, social media is by far the better known concept since only 42 individuals have heard of web 2.0, as opposed to as many as 232 who knew the term social media only.

All in all, 87.7% (637 respondents) heard of social media or web 2.0 or both. Half of our sample (363) has heard of both (social media and web 2.0), 12% (89) from the entire sample have never heard of either term, and the remaining 37% (274) of the sample heard either only about web 2.0 (15%, 42) or social media (85%, 232). It was hence found that social media is many times more popular a term than web 2.0, in fact 5 times more popular.

4.5.2.2 Popularity of Web 2.0 Applications

Table 4.6 presents all 10 web 2.0 applications that respondents currently use or have used at some point in the past, ranked from top to bottom by their popularity in the overall sample²⁵.

Table 4.6 – Overall ranking of web 2.0 applications based on all survey responses

Web 2.0 Application	Respondents	overall %
<i>Youtube</i>	698	96.1%
<i>Wikipedia</i>	693	95.5%
<i>Amazon</i>	667	91.9%
<i>Facebook or Myspace</i>	645	88.8%
<i>Ebay</i>	583	80.3%
<i>Twitter</i>	309	42.6%
<i>Flickr or Picassa</i>	303	41.7%
<i>Digg or Reddit</i>	88	12.1%
<i>Craigslist</i>	70	9.6%
<i>Delicious</i>	50	6.9%

The rankings can be roughly broken down to; **1 – heavily used applications:** Youtube, Wikipedia, Amazon, Facebook / Myspace and Ebay; **2 – relatively popular applications:** Twitter and Flickr / Picassa; and **3 – tailing or niche applications:** Digg / Reddit, Craigslist and Delicious.

²⁵ Note: multiple choices per respondent were allowed for this question, and there wasn't a single respondent that did not select at least one application.

More interestingly; however, application popularity can now be broken down by different categories, such as by age-groups. Since age-group can be interpreted as an ordinal variable one can appreciate a number of popularity trends as the respondent group gets older²⁶. Figure 4.3 illustrates the variations in application popularity across different age groups. Facebook / Myspace for example is extremely popular with younger respondents; however, its popularity drops of rather steeply with older age-groups²⁷, who are probably less interested into the kinds of social interactions these web applications offer. This result corroborates with Berlanga et al. (2011), who found that 72% respondents in their survey used social networks, except that the average age of their respondents was 42 years. Youtube on the other hand picks up some popularity with the oldest age-group (≥ 61), where this could be due to older generation preferring more visual stimuli for example, instead of studying Wikipedia (which does drop off)²⁸. Twitter and picture sharing applications (i.e. Flickr / Picassa) are used by about 40% of respondents over most age-groups and their popularity ranks do not vary too much. Amazon is the most popular application with 31 year olds and older respondents, probably as older respondents enjoy reading more books, have more financial means to complete purchases, or simply that this is to some extent a symptom of the sample being slightly biased towards academics. Overall there is some tendency for popular web 2.0 applications to become less popular with older age-groups²⁹. As the older generations did not “grow up” on these kinds of applications, unless they are professionals that have an explicit requirement to use some of these applications, there is often very little reason for these people to change their existing behaviours. However, once the current, young generation gets older this pattern is going to change. Popularity rankings of web applications tend not to vary too much³⁰ over different age groups, i.e. the rankings over all the age groups tend to be quite stable. This is probably best illustrated by scatter plot of rankings in figure 4.3, which shows clear linear relationship with little variation for different age-groups.

²⁶ Note: the survey is a snapshot in time (April 2011), respondents of varying ages contributed their answers, and hence this survey is not tracking the habits of individuals over time.

²⁷ There is some popularity gain from the 51-60 year olds for Facebook / Myspace.

²⁸ Actually the oldest age group (≥ 61) only contains 12 individual respondents. This is too small a number to allow us to indicate significant evidence to draw any meaningful conclusions for this age group, yet it can act as some weak indication for this group.

²⁹ Some indication towards this can be observed in figure 4.3, but this is also discussed briefly and a correlation matrix of the application rankings between age-groups is provided in table B5, appendix B.

³⁰ See figure B.2 (in appendix B) for popularity rankings across age-groups, popularity magnitude changes are left out. This figure is based entirely on the same data as figure 4.3; however differences between closely trending applications (i.e. Youtube, Amazon, Ebay, Wikipedia) become more evident, in this simplified chart.

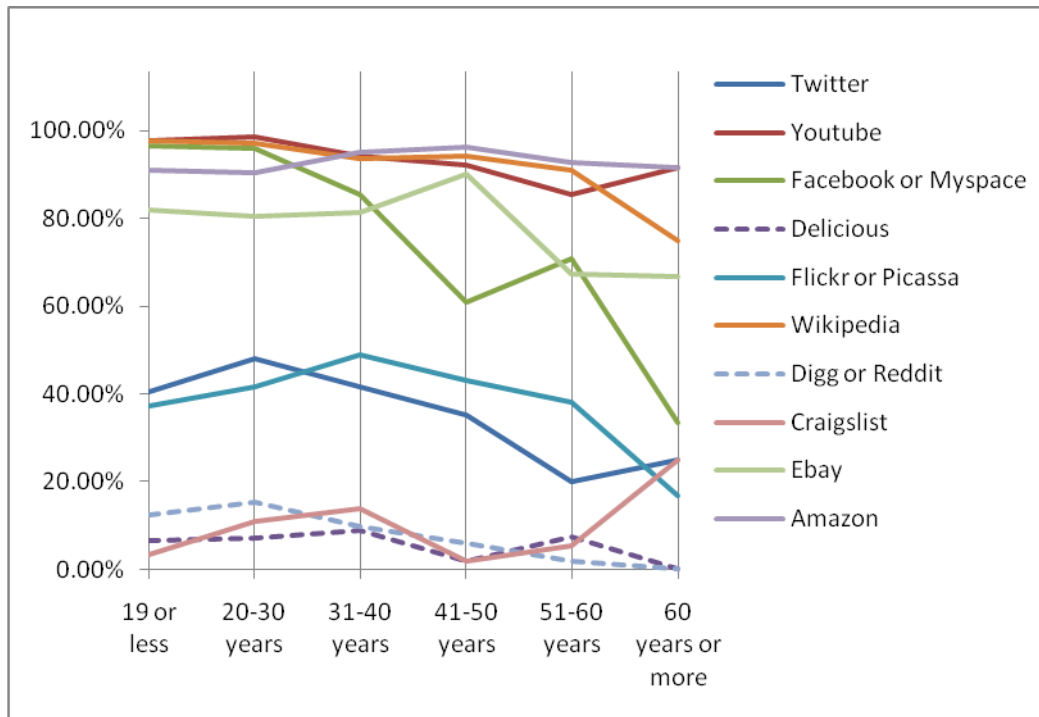


Figure 4.3 – Popularity of web 2.0 applications over increasing age-group (% of respondents confirming use)

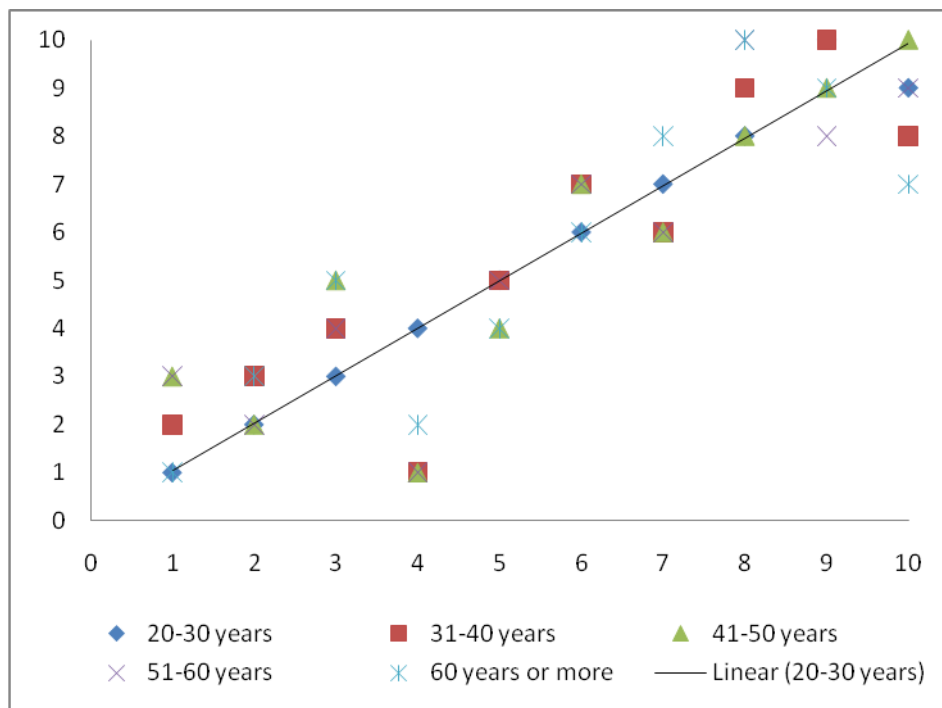


Figure 4.4 – Scatter plot between popularity (of web 2.0 applications) rankings of youngest (19+) age group and all other age groups (1 [highest] to 10 [lowest])

4.5.2.3 Popularity of Web 2.0 Applications by Education and Expertise

Web 2.0 application popularity can be broken down by any meaningful variable in the survey. In particular popularity patterns within different education levels and areas of expertise can be investigated in more detail, in order to provide further insight into general patterns of use preferences. Figure 4.5, presents popular rankings as they change over respondent groups of

increasingly more highly qualified individuals.

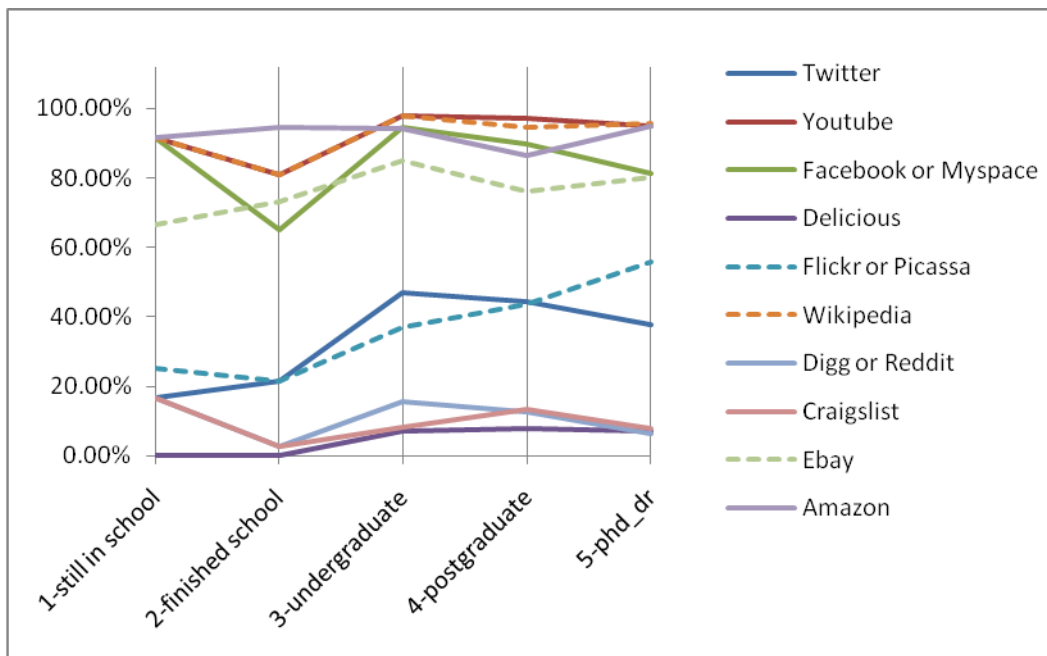


Figure 4.5 – Popularity of web 2.0 applications over increasing education levels (% of respondents)³¹

Rankings can be broken down, roughly into, heavily used applications, relatively popular applications and tailing or niche applications. These rankings do not change substantially – see earlier section, 4.2.5.2. However, a pronounced pattern relating to picture sharing applications (Flickr / Picassa), where popularity nearly doubles as respondent groups become more qualified, does clearly exist. It is not apparent why this pattern emerges. One possible explanation could relate to a desire for self-exposure with an increasing academic achievement level, but one may only speculate. Similarly, Twitter becomes substantially more popular with undergraduate, postgraduate students and PhD/Dr³² level educated individuals than with less qualified ones. Ebay is the most popular with undergraduates, which probably points to the resourcefulness of students required due to limited financial means and the “snap a cheap deal” mentality, so often prevalent amongst undergraduates. Interestingly a clear pattern of Delicious and Digg / Reddit usage is noticeable in that the undergraduate and more highly qualified groups use these web 2.0 applications considerably more than the two less educated groups. Facebook and Myspace are actually quite popular with PhD/Dr., and in fact only the respondents who finished their basic school education don’t find social networking applications as interesting, it seems.

³¹ Sample sizes for the different groups are available within table B.7 in appendix B.

³² Those who work towards a PhD or who finished and received their title, i.e. most highly educated respondents.

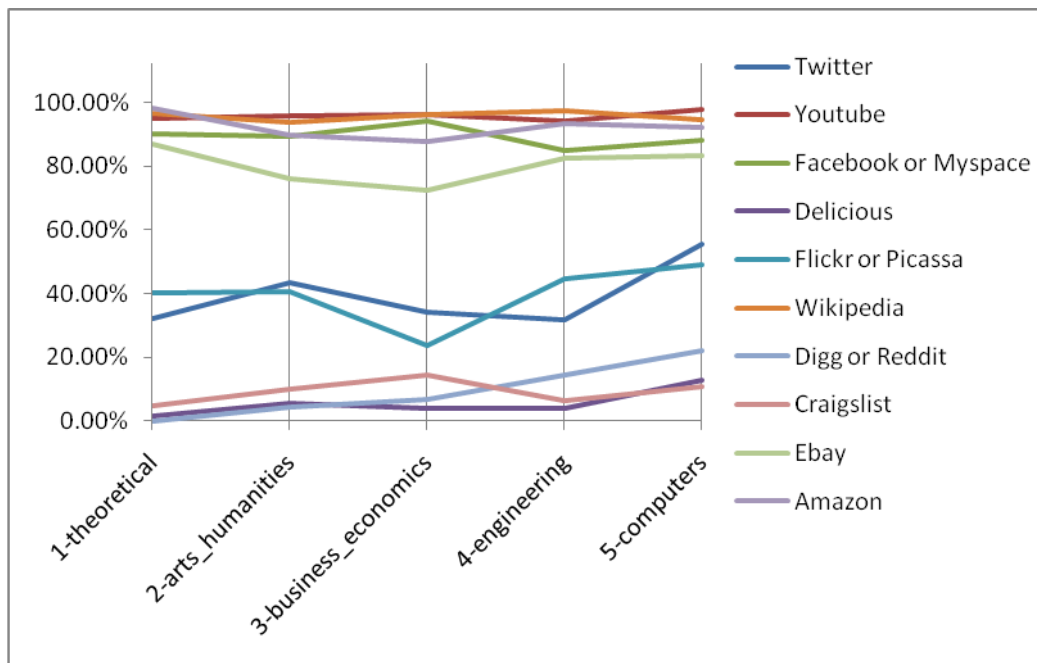


Figure 4.6 – Popularity of web 2.0 applications over different expertise areas (% of respondents confirming use)³³

The figure 4.6, illustrates popularities across various expertise areas, one way of looking at the x-axis is to consider it as ordinal in that from theoretical through arts / humanities, business / economics, engineering to computer science the area of expertise is becoming more IT oriented or technical. From the figure it can be appreciated that especially relatively popular applications and tailing or niche applications are more popular for respondents with IT experience. Interestingly however the most heavily used applications (Youtube, Wikipedia, Amazon, Facebook / Myspace and Ebay) do not show much variability in popularity, except maybe for Ebay. Given IT specific barriers of entry into Ebay auction style shopping, it is considerably less popular with arts / humanities and business / economics minded respondents, despite Craigslist being more popular with business / economics respondents than any other expertise group. The chart illustrates that Wikipedia and Youtube are independent of specialisation, as their popularity doesn't vary substantially with different categories.

4.5.2.4 Popularity of Activities

Now that the most prevalent web 2.0 applications in the sample have been analysed across a number of demographics, it is of much interest to know how these and similar applications are used by those same respondents. In section 2.3.2.2, a number of essential and fundamental activities commonly used to accumulate user generated content on web 2.0 applications were introduced – table 4.7 illustrates the overall ranking of these activities, as based on the sample.

³³ Sample sizes for the different groups are available within table B.8 in the appendix B.

Table 4.7 – Overall ranking of top web 2.0 activities as introduced in chapter 2.3.2.2 (based on all survey responses)

Web 2.0 Activity	Respondents	overall %
<i>uploaded a file</i>	634	87.3%
<i>joined a community</i>	580	79.9%
<i>commented on...</i>	543	74.8%
<i>tagged...</i>	538	74.1%
<i>rated a...</i>	370	51.0%
<i>submitted a blog post</i>	247	34.0%
<i>used RSS</i>	226	31.1%
<i>edited a shared resource</i>	171	23.6%
<i>API or 'Mashup'</i>	93	12.8%
<i>OpenID or DISQUS</i>	66	9.1%

The activity of uploading a file is very common in video, image sharing (e.g. Youtube, Flickr / Picassa) and social sharing (Facebook / Myspace, Twitter) applications, hence it is not surprising that uploading files is the most popular activity in the sample. Many of the web 2.0 applications referred to in the survey allow users to associate with a group of people that share similar values and norms, and in fact interestingly enough 80% of respondents consider “joining a community” to be a type of activity they tend to do or have done in the past on the web. Predictably, commenting, tagging and rating online content are the next top three activities. In chapter 2.3.1.3 a simple model of web 2.0 participation was presented, inspired by this, two types of activities are suggested; **1**–low effort activities and **2**–high effort activities. Rating, Tagging, joining communities for example would present relatively low effort activities since the complexities and time taken to achieving the contribution is minimal. Whereas uploading a file, requires somewhat more effort³⁴, as does submitting a blog post – in fact most likely it is necessary to set-up a blog first in order to begin posting, which in itself isn’t technically challenging but does require some effort. The high number of people (34%) in the sample who claim to have submitted blog post(s) was somewhat surprising. Editing shared resources and commenting on content also requires some amount of effort, maybe not as much as the activities mentioned, yet still more than what could be accomplished in several clicks. It wasn’t expected that as many as 171 respondents would have edited shared resource(s), this is highly encouraging in terms of direct participation in collaborative content creation³⁵. Also about a tenth of the sample was involved in API / Mashup use or creation, and with open ID style online identification providers. The participation in OpenId is still very small but certainly

³⁴ Preparing the file itself, possibly the format and size have to satisfy some requirements, also uploading time, etc.

³⁵ Unfortunately we are unable to determine what kind of shared resources were edited by users. It would seem the likely candidate would be Wikipedia – however there are many other popular collaborative resources.

some awareness does exist. Overall these results indicate that over recent years participation in web 2.0 content creation increased significantly, since Daugherty et al. (2008) found that only 21% of all 325 respondents contributed any UGC (their survey included picture / file uploading, Blog posts, etc.). Although results from both surveys are not directly comparable, since unfortunately Daugherty et al. did not provide useful details of their sampling frame.

Web 2.0 specific activities and how they rank among different demographics of the samples' respondents along age, education level and expertise area are now investigated in the remainder of this sub-section.

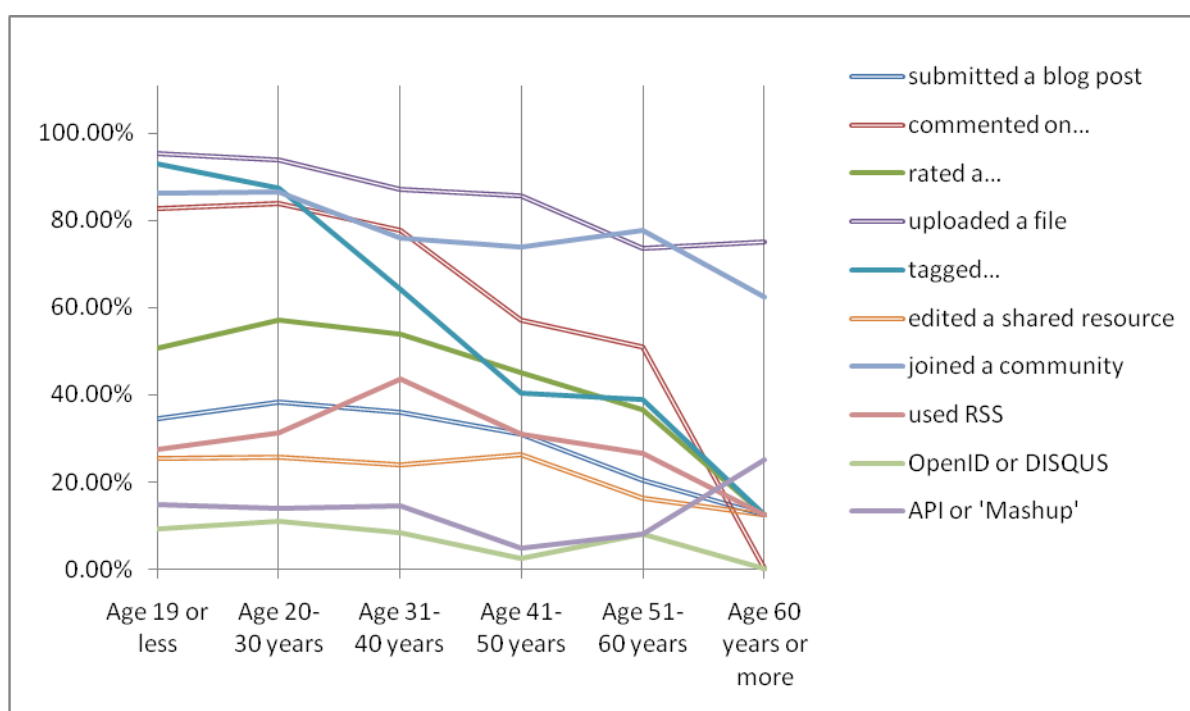


Figure 4.7 – Popularities of web 2.0 related activities broken down by all the age-groups³⁶

It can be clearly observed from figure 4.7 how the popularity of virtually all the activities reduces with age³⁷. Although there is a slight increase (see figure 4.7) in edits of shared resources within the 41-50 age group, and an increased use of RSS in the 31-40 age group. Surprisingly many Blogs posts are submitted by a respectable percentage of respondents throughout the various age-groups.

³⁶ Sample sizes for the different groups are available within table B.9 in appendix B.

³⁷ It is somewhat intriguing why there is a rise in API or Mashup use and uploads of files in the oldest age-group, however as it was mentioned earlier in this chapter, the oldest age-group consists of a small sample that hence likely could contain biased individuals affecting the whole group disproportionately. For example, given popularity of Flickr / Picassa it would seem the file uploading refers to the use of these picture sharing applications – maybe a group of older people who all share a passion for photography has biased this sample. Further to this a spearman's rho correlation of all percentual rankings of age-groups for the 10 different activities, found that all the age-group's rankings were correlated to more than 0.9 correlation coefficient, except the oldest age-group, where the ranking breaks away significantly (at p (two-tailed) > 0.193), and correlation coefficient isn't larger than .449.

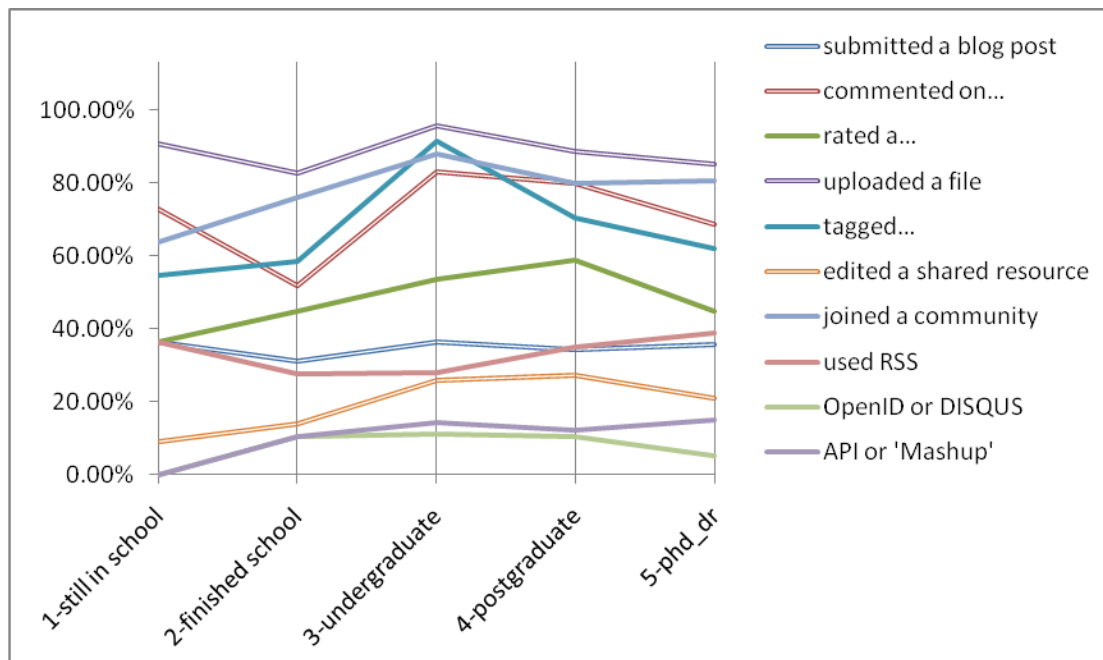


Figure 4.8 – Popularities of web 2.0 related activities broken down by levels of education³⁸

Figure 4.8 presents popularity rankings as they are represented by groups of respondents with different educational backgrounds. Commenting, tagging and uploading files seem to be very popular with undergraduates, this has to do with the use of social media applications which mostly allow and expect one to perform these actions. It is interesting to note the increase in RSS usage by postgraduate students and PhD/Dr, this is most likely due to an increased need to keep on top of research or important news, in contrast to less educated individuals. There is a considerable increase in interest to edit shared resources, beginning with undergraduate students; however, participation of less educated respondents in editing shared resources is indeed much smaller. This would indicate that especially in the case of Wikipedia more highly educated individuals would likely spent their time editing articles. Surprisingly popularity of blog-posts doesn't tend to vary significantly across education levels, however joining online communities is prevalent within undergraduate, postgraduate and PhD/Dr respondents.

Figure 4.9 illustrates convincingly that respondents with computer expertise are on average more accustomed to web 2.0 activities; notice the increase of popularity of activities in the computers expertise group. It also seems that more or less the same activities, without much variation between them seem to be popular amongst other expertise areas.

³⁸ Sample sizes for the different groups are available within table B.10 in appendix B.

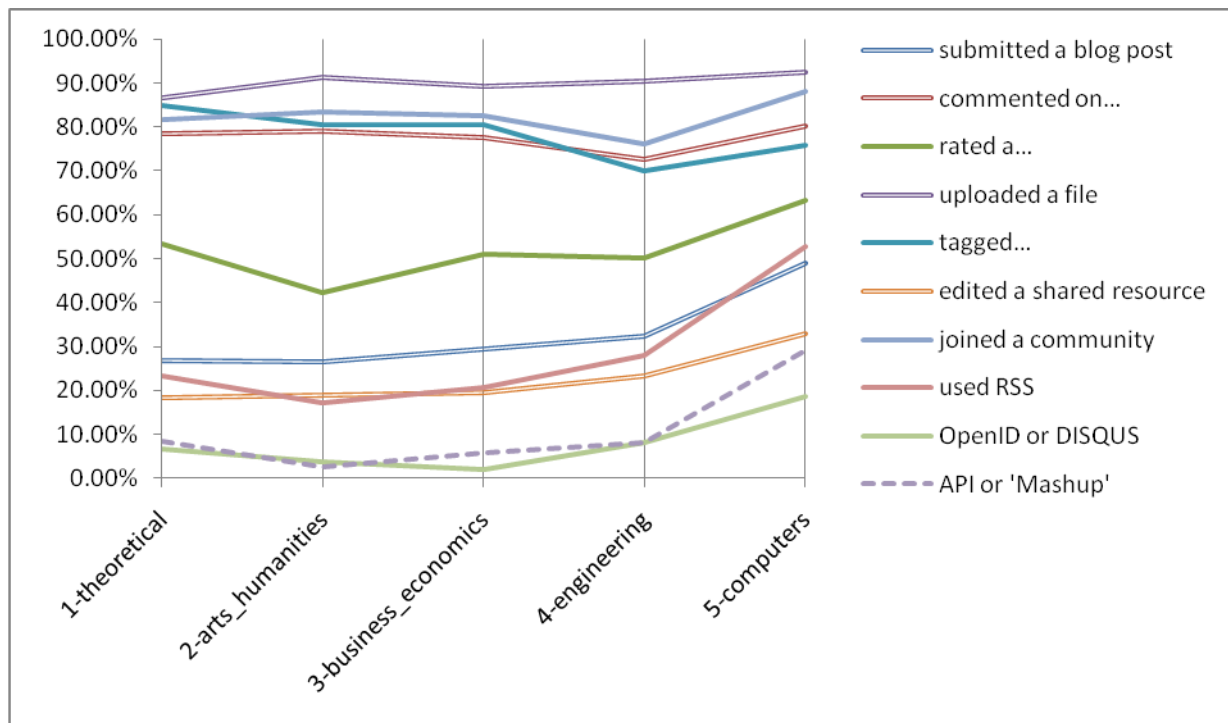


Figure 4.9 – Popularities of web 2.0 related activities broken down by expertise areas³⁹

4.5.3 Business Activity

Survey responses indicate that individuals across all age-groups, education-levels and technical skill-sets are accustomed to online purchasing. As many as 95.9% of individuals from the entire sample indicated they have made purchases online. A decade ago it would have been unthinkable to expect virtually all internet users to have made purchases online. In addition to this (81.5%), 592 individuals claim to have used PayPal or webmoney. Only 55.9% (406 respondents) admitted to following / clicking an add. Not much can be said about variation based on demographics, since any differences amongst age, education level and expert area respondent groups are small and relatively insignificant⁴⁰. Since virtually everyone claims to purchase online there is little use in looking at any patterns that help determine such activity⁴¹; however, an analysis into the profile of group buying respondents is worthwhile⁴² and hence the following section discusses group buying in particular.

³⁹ Sample sizes for the different groups are available within table B.11 in appendix B.

⁴⁰ See tables B.12, B.13 and B.14 in appendix B, for more details, i.e. actual percentages for each demographic group in the sample.

⁴¹ A series of non-parametric median based tests were performed, to see whether the distributions of non-advertising and advertising followers differ along any of the main scores. It was found that respondents who follow adverts tend to have a slightly higher time, trust and web 2.0 awareness score.

⁴² As far as the author is aware, group buying user patterns in terms of web 2.0 application behaviour have not received academic attention. Also see section 2.2.1 for an introduction to the group buying model.

4.5.3.1 Group Buying

Surprisingly, as many as 14.3% of respondents (104 individuals) in the survey have previously used a group buying website, and this highlights the relatively wide spread acceptance of group buying as a new shopping mode. Group buying is a very recent phenomenon (see section 2.2.1) and is generally considered to be a radically different shopping model, contrasted to the more traditional online shops. Due to this, group buying is associated with a certain entry and learning barrier in terms of becoming accustomed to a new way of using the web for shopping. It can hence be expected that those 104 respondents will be well phrased in, and accustomed to web 2.0 applications. In fact, it was found that of the 104 people less than 33.7% used only 5 and less web 2.0 apps, as opposed to 49.5% of the 622 non group buying individuals – pointing to the group buying respondents being heavier web 2.0 users. To confirm this, with statistical significance, the web 2.0 competence scores⁴³ of both groups were compared.

H_0 – there is no difference in the group buying and non group buying web 2.0 competence scores⁴⁴

H_1 – there is a significant and systematic difference in the web 2.0 competence scores

A non parametric, independent samples Mann-Whitney U Test found a significant and systematic difference between web 2.0 competence scores in the two groups, rejecting hypothesis H_0 , *standardized test statistic* = 4.209, p (two-tailed) < .0005. The group buying median was 14 (mean-rank was 443.41), compared to median of the other group 12 (mean rank 350.14), hence the web 2.0 awareness is significantly higher in the group buying respondents. The same statistical test was performed on all three demographic variables in order to check whether being a user of group buying websites is potentially related to particular age, education level or technical IT abilities. It was found that the distributions are the same across both groups (non group buying, and group buying), with significance values being very far off any alternative hypotheses, .726, .515 and .142 respectively. A visual inspection of distribution histograms further shows that the distribution shapes are generally identical to the entire survey dataset. The distribution of time spent, trust awareness, motives variation, perceived Wikipedia usefulness were all also found to have the same distribution across both groups, at p (two-tailed) < .05. Eventhough the significance value for trust-awareness was .052, very closely missing the somewhat arbitrary significance level of .05, followed by significance .062 for time spent online, and .073 for motives variation. Hence, one may conclude that trust awareness, time spent online and to some degree the number of different motives a user has for using web

43 Web 2.0 competence is a simple summation of Q1, Q2 and Q3 scores. See sections 4.4.2 / 4.4.3 for details.

44 Group buying group (N=104), non group buying group (N=622).

2.0 apps, tends to be higher with group buying users.

In summary it was found that surprisingly many people actually use or have used group buying in the past. There doesn't seem to be any systematic age, educational or IT skill-level variation that predisposes users to be more likely to having used group buying websites. At the same time it was discovered that the awareness and intensity of web 2.0 use is strongly related to the likelihood of having had used group buying websites.

4.5.4 Trust

4.5.4.1 Trust in Web 2.0

Quite surprisingly the survey found that only 57% respondents consider themselves able to judge trustworthy from non-trustworthy sites (based on Q6-5). This is rather low, considering that 100% of our respondents have used at least one web 2.0 application, where data sharing is a dominant activity. Breaking down this answer by age-groups reveals that, over the sample, younger people are consistently more confident of their ability to judge trustworthiness, see the within age group trustworthiness percentage decrease along the (increasing) age groups in table 4.8.

Table 4.8 – Trustworthiness (Q6-5) responses broken down by age groups of respondents

	Age group					Total
	≤ 19 years	20 - 30 years	31 - 40 years	41 - 50 years	≥ 51 years*	
Q6-5 (trustworthiness) Count	62	248	62	22	21	415
% within Age group	69.7%	62.6%	50.4%	43.1%	31.3%	

*. Age groups 51-60 and ≥61 were joined into same group (since >61 only contained 5 responses)

Despite the overall low ability of respondents to judge trustworthiness of sites, people do want to be safe (rather than sorry), as only 1.7% (12 individuals) feel comfortable sharing their personal details on web-pages. In contrast as many as 485 individuals (66.8%) only feel comfortable sharing personal details on secure web-applications⁴⁵. Similarly, as many as 648 (89.3%) individuals would only feel comfortable purchasing from trustworthy stores, whereas only 23 (3.2%) don't mind. These are large differences, clearly pointing out, how important trust is, yet an alarmingly large proportion of respondents cannot judge trustworthy from non-trustworthy websites.

The Likert scale question Q5 (i.e. I trust most websites that I use on a regular basis) revealed

⁴⁵ 241 (33.2%) did not answer this question, this could be explained in a number of ways, for example that these people simply did not concern themselves with the issue at all, or they felt it didn't apply to them specifically.

that for each age group the median and mode was 4, which simply indicates that people tend to use websites that they trust. The mean decreased (from 4.02 to 3.64)⁴⁶ with the increasing age-group, which confirms the earlier observation of older users being generally more concerned and careful on the web. The negative and neutral responses to the Likert scale (i.e. strongly disagree, disagree, neutral towards Q5 statement) added up to only 21%, which indicates that people tend to mostly use websites that they trust, even though many might not consider themselves to be good judges of when to trust a new website and when not.

The overall factor score for trust (amalgamation of Q5 and Q6 as described in section 4.4.2) can be used to statistically substantiate the role of trust⁴⁷ in relation to other factors that might be related. Hence the sample is split into two groups; group *1* – where trust score < 6 (low trust awareness group), and group *2* – where trust score > 5 (high trust awareness group)⁴⁸. Since correlation analysis (see section 4.5.1) pointed to potential relationships to *1*-web 2.0 competence, *2*-business activity, *3*-time spent online and *4*-motives heterogeneity, the following hypothesis is tested for each.

H_0 – there is no difference in the low and high trust awareness group for the tested score⁴⁹

H_1 – there is a significant and systematic difference in the tested score

Non parametric, independent samples Mann-Whitney U Test found a significant and systematic difference between all factors; web 2.0 competence scores, business score, time spent online and motive heterogeneity in the two groups, hence hypothesis H_0 is rejected at p (two-tailed) < .0005⁵⁰, for each score. With high trust group having higher scores than the low trust groups, thus confirming the initial finding in the correlation analysis. Trust seems to be an important and dominant feature in most aspects of the web 2.0 area.

4.5.4.2 Web 2.0 Habits across Trust Levels

Since trust was found to be such a significant factor, this section will look at the most popular web 2.0 applications and web 2.0 related activities as they relate to different trust awareness

46 No. of responses in each increasing age group order was 89, 396, 123, 51 and 55. The age group for more than 61 year olds actually had a mean of 4.08; however, a small number, with 11 out of 12 individuals scoring a 4 or 5.

47 Since the score is based on amalgamation of Q5 and Q6, the trust score can be thought of as a measure of trust awareness – i.e. the higher the score, the more trust becomes a priority for a respondent.

48 Low trust awareness group, N=177, high trust awareness group, N=549.

49 As explained, score is either web 2.0 competence, business score, time spent online or motives heterogeneity.

50 *Standardized test statistic* = -7.094 (median high trust = 13, low trust = 11), -6.494 (median high trust = 3, low trust = 2), -5.919 (median high trust = 4, low trust = 3), -6.063 (median high trust = 1, low trust = 1) respectively.

levels. In order to get an idea of popular applications used and web 2.0 activities performed by respondents with certain trust awareness levels, the sample was split into three separate groups based on the trust awareness score⁵¹; **1-Low** trust (score <5), **2-Normal** trust (4 < score < 8) and **3-High** trust (score > 7). Finally the tables 4.9 and 4.10 simply present popularity counts of activities and applications as respondents of specific trust levels selected them from the whole sample⁵². Notice, the within trust level percentage (highlighted in bold) in table 4.9, and the popularity of each application in high trust column, through normal trust down to low trust column. When interpreting the table, column and row totals should also be considered carefully; hence a column with the Chi-Square test statistic and its p value, i.e. significance level, is provided (see appendix B, tables B.19 for the Chi-Square test tables). The highest difference between the highest and lowest within trust level percentage is the largest for the activity of *joining a community*; this indicates that there are proportionally more people with higher trust awareness willing to join an online community, similarly for commenting, rating, tagging, or using OpenID or DISQUS. All these activities require a certain amount of trust awareness and this is corroborated by responses.

Table 4.9 – Popular activities of respondents broken down by trust score based groups (with Chi-Square tests)

		Trust score based trust-level			Total	χ^2 - test.
		Low trust	Normal trust	High trust		
Q3-1bog post	Count	21	110	116	247	10.350 ¹
	% within trust. l.	28.0%	32.3%	41.1%		p < .01 ²
Q3-2comment on	Count	50	254	239	543	25.330
	% within trust. l.	66.7%	74.5%	84.8%		p < .0001
Q3-3rated	Count	34	164	172	370	17.354
	% within trust. l.	45.3%	48.1%	61.0%		p < .0005
Q3-4uploaded file	Count	64	303	267	634	21.763
	% within trust. l.	85.3%	88.9%	94.7%		p < .0005
Q3-5tagged content	Count	46	248	244	538	40.051
	% within trust. l.	61.3%	72.7%	86.5%		p < .0001
Q3-6edited shared resource	Count	19	69	83	171	8.490
	% within trust. l.	25.3%	20.2%	29.4%		p < .05
Q3-7joined community	Count	51	266	263	580	52.893
	% within trust. l.	68.0%	78.0%	93.3%		p < .0001
Q3-8rss	Count	22	96	109	226	11.088
	% within trust. l.	29.3%	27.9%	38.7%		p < .005
Q3-9openid / disqus	Count	8	17	41	66	17.783
	% within trust. l.	10.7%	5.0%	14.5%		p < .0005
Q3-10api / mashup	Count	6	40	47	93	6.697
	% within trust. l.	8.0%	11.7%	16.7%		p < .05
Total	Count	75	341	282	698	-

¹Pearson Chi-Square (df, 2)

²two-tailed test

51 The score ranges from 0 to 9, inclusive. Note: only 75 responses out of 698 have a low trust score.

52 698 respondents provided activity information and all 726 respondents provided applications used information.

Similarly, from table 4.10 one may appreciate that certain applications are more popular among higher trust respondents than among lower trust respondents. Ebay illustrates this convincingly, as the percentage of Ebay users for high trust group is 86.7%, however only 59.5% for the low trust group. This could mean that people who are more confident on the web (i.e. more trust aware) will also more likely use an auction site such as Ebay to sell and buy items. Similarly, other applications with trust awareness bias can be appreciated from the table⁵³.

Table 4.10 – Popular web 2.0 applications broken down by trust score based groups (with Chi-Square tests⁵⁴)

		Trust score based trust-level			Total	χ^2 - test.
		Low trust	Normal trust	High trust		
Q2-1twitter	Count	32	130	147	309	15.683 ¹
	% within trust. l.	38.1%	36.4%	51.6%		p < .0005 ²
Q2-2youtube	Count	80	339	279	698	3.896
	% within trust. l.	95.2%	95.0%	97.9%		p = .142
Q2-3facebook / myspace	Count	65	309	271	645	24.232
	% within trust. l.	77.4%	86.6%	95.1%		p < .0001
Q2-4delicious	Count	7	17	26	50	5.010
	% within trust. l.	8.3%	4.8%	9.1%		p = .082
Q2-5flickr / picassa	Count	30	139	134	303	5.673
	% within trust. l.	35.7%	38.9%	47.0%		p = .069
Q2-6wikipedia	Count	75	340	278	693	10.273
	% within trust. l.	89.3%	95.2%	97.5%		p < .01
Q2-7digg / reddit	Count	8	24	56	88	25.462
	% within trust. l.	9.5%	6.7%	19.6%		p < .0001
Q2-8craigslist	Count	4	32	34	70	4.197
	% within trust. l.	4.8%	9.0%	11.9%		p = .123
Q2-9ebay	Count	50	286	247	583	30.235
	% within trust. l.	59.5%	80.1%	86.7%		p < .0001
Q2-10amazon	Count	66	325	276	667	29.666
	% within trust. l.	78.6%	91.0%	96.8%		p < .0001
Total	Count	84	357	285	726	-

¹ Pearson Chi-Square (df. 2)

² two-tailed test

It can be concluded that respondents prefer secure, reputable and known (to them) web-applications, yet only 57.2% are able to judge trustworthy apart from non-trustworthy sites. This could well be one of the reasons that make it difficult for new web 2.0 platforms to gain a substantial following. Attaining trustworthiness in the user's eyes is a major obstacle, and reaching critical-mass (in terms of participation) for collaborative web 2.0 projects is still a real challenge for web 2.0 system deployments. Some indication of web 2.0 applications

⁵³ The popularity patterns do not seem to hold across trust levels. The provided analysis to some extent is only indicative, given this survey sample, these patterns are indeed recognisable; however, extrapolating statistical validity onto the population cannot be done with much certainty (some groups are relatively small and this presents little challenge to any bias).

⁵⁴ See appendix B, tables B.20 for the Chi-Square test tables.

characterised by more social exposure requirements being more popular among respondents with high trust levels was found. Similarly, activities that require some socially identifiable information have been less popular with respondents of lower trust levels. Overall these findings are indicative of the high importance of trust awareness amongst the web 2.0 user community.

4.5.5 Time

4.5.5.1 Time Assessment: Relative to Peers vs. Self Reflection

Before an analysis of the role that time spent online plays in relation to other factors, an interesting phenomenon observed in the relationship between Q7 and Q8 is discussed. This is important since the time based score (on which further analysis is performed) is an amalgamated score based on Q7 and Q8 Likert responses. Both questions asked the respondents to answer how much time they spend online, however Q7 asked the respondent to evaluate the time spent in relation to their peers⁵⁵, whereas Q8 simply asked the individual to rate themselves⁵⁶. This seemingly insignificant difference in the way the question was posed might result in different tendency of responses, and indeed does. The distributions of both Likert score responses (in both cases N=726) were compared using non parametric, independent samples Mann-Whitney U test, to establish whether, (H_1) the distributions are significantly and systematically different or whether (H_0) there is no systematic and significant change between Likert score distributions for Q7 and Q8.

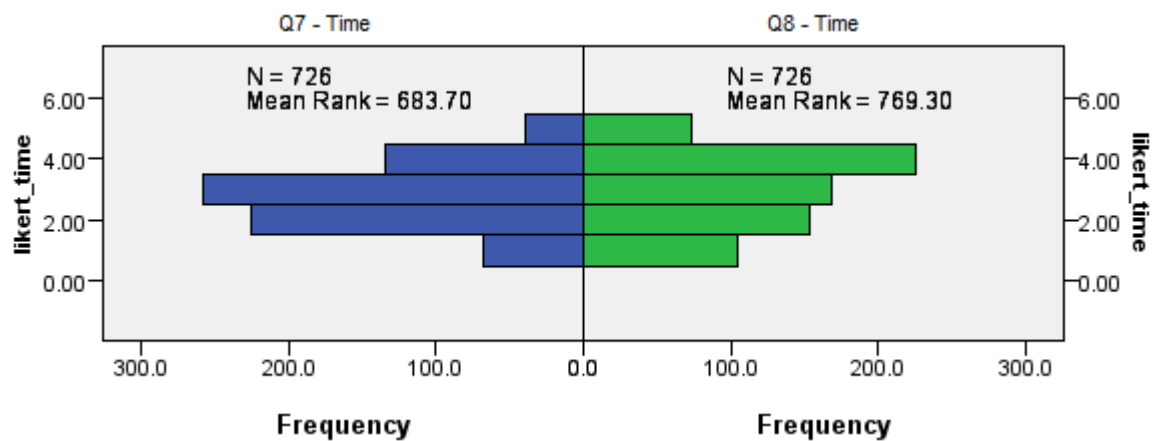


Figure 4.10 –Comparison of Q7 and Q8 likert score distributions, output of Mann Whitney U test

55 **Q:** Compared to your friends, in your free time how much time per day do you spent on sites such as Facebook, Digg, Youtube ... **A:** 5-Probably much more than most of my friends, 4-Probably more than some of my friends, 3-Same as majority of my friends, 2-Less than most of my friends, 1-No time at all, or nearly no time

56 **Q:** I spent too much time on Twitter, Facebook, Youtube, Wikipedia ... **A:** 5-Strongly Agree, 4-Agree, 3-Neutral, 2-Disagree, 1-Strongly Disagree

The hypothesis H_0 , was rejected at p (two-tailed) $< .0005$ ⁵⁷, both distributions are highlighted in figure 4.10, with Q7 on the left and Q8 on the right, with the mean rank of Q8 being 86 higher, indicating that the distribution of Q8 likert score tends to be somewhat higher than Q7. Eventhough the difference isn't substantial⁵⁸, there is some systematic difference between both distributions. The difference among the set of responses seems to be, in that answers to Q7 tend to under-represent the time spent online in relation to the answers in Q8 – i.e. respondents are more conservative when they compare the time spent online against their peers, than when they simply judge whether they spent too much time online. Figure 4.11 illustrates how time spent online decreases with the older age groups along both likert scores. What is interesting is the preference of the two younger generations⁵⁹ to indicate that they spent less time online when assessed relative to their peers. A similar pattern can be observed in figure 4.12 where the less IT-technical individuals consistently underrate time spent online when measured against their peers. The described tendency can be explained in a number of possible ways – the most feasible one is that younger people do indeed have some friends that spent much more time online than they do. This is further supported by respondents with less technical expertise indicating that some of their peers might spent more time online than they do themselves, which is likely less the case with more technical individuals, who probably themselves spent a lot of time online. However, ultimately in further analysis the scores of Q7 and Q8 will be combined into an overall time score, which will take into account the whole picture.

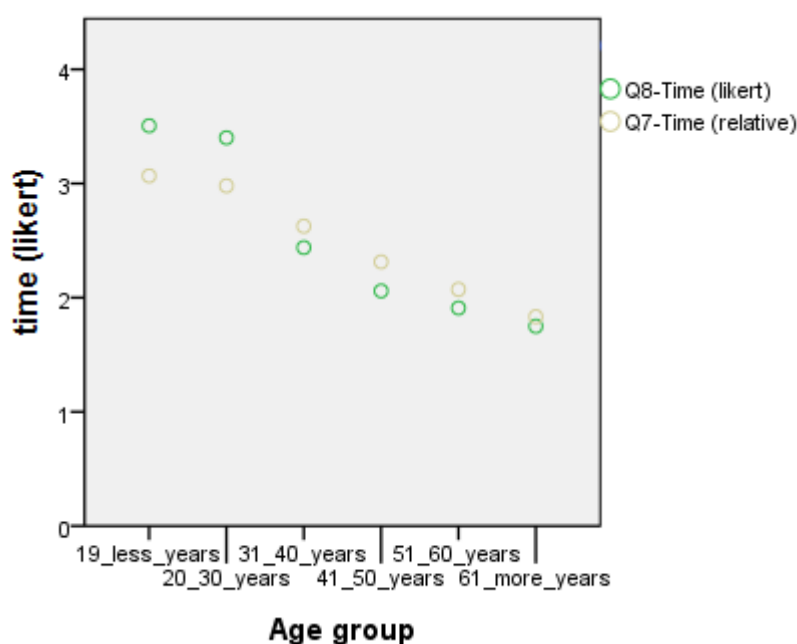


Figure 4.11 – Mean likert scores for Q7 and Q8 plotted against each age group

⁵⁷ Standardized test statistic = 4.013

⁵⁸ Mean rank is not easily interpretable hence we can look at the central tendency measures, median = 3 for both, mode = 3 for Q7, and 4 for Q8 and the mean = 2.8 for Q7 and 3.02 for Q8.

⁵⁹ Both account for 67% of the sample

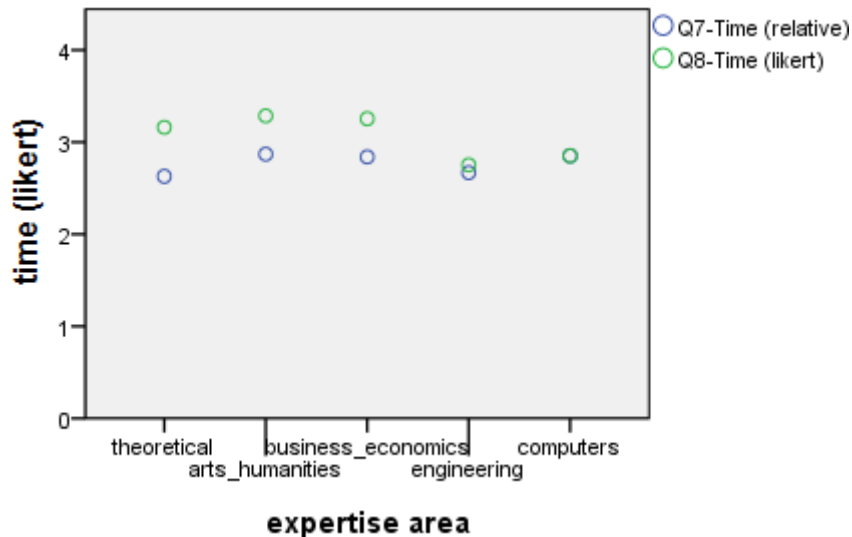


Figure 4.12 – Mean likert scores for Q7 and Q8 plotted against each expertise area group

4.5.5.2 Time Spent Online

The correlations discussed in section 4.5.1 suggest that older respondents spend less time on web 2.0. It was further indicated that the more time a respondent spends online the more on average he or she will be competent with regards to web 2.0 (probably due to the longer exposure and experience on the web). The overall factor score for time (amalgamation of Q7 and Q8) can be used to statistically substantiate the role of time⁶⁰ in relation to the factors that seem correlated. The entire sample is split into two groups; group 1 – where time score < 5 (low time-spent group), and group 2 – where time score > 4 (high time-spent group)⁶¹. Since correlation analysis pointed to potential relationships with 1-web 2.0 competence, 2-business score, 3- online trust awareness and 4-motives heterogeneity, the following hypothesis is tested for each of the four factors.

H_0 – there is no difference in the low and high time spent groups for the tested score⁶²

H_1 – there is a significant and systematic difference in the tested score

Non parametric, independent samples Mann-Whitney U Test found a significant and systematic difference between; web 2.0 competence, trust awareness and motive heterogeneity scores in the two groups, hence for these, hypothesis H_0 is rejected at p (two-tailed) < .0005⁶³. Despite

⁶⁰ Since the score is based on amalgamation of Q7 and Q8, the time score can be thought of as an overall measure of time spent online, as judged by the respondent's relative judgment and their self-reflexion. The higher the score, more time the respondent spends online.

⁶¹ Low time spent group, N=420, high time spent group, N=306.

⁶² Score is either web 2.0 competence, business score, trust awareness or motives heterogeneity.

⁶³ Standardized test statistic = -7.094 (median high time =13, low time =11), -6.494 (median high time =3, low time =2), -5.919 (median high time =4, low time =3), -6.063 (median high time =1, low time =1) respectively.

the business score test being significant at p (two-tailed) = .017, further visual inspection of box-plots, distribution-plots and comparison of means⁶⁴ and medians for both groups, point to very small differences in the distribution, and not to a shift of higher business scores for time spent online. The same inspection for the other factors found that web 2.0 competence had the biggest difference between both time groups, followed closely by motive heterogeneity and trust scores.

One may conclude from the tests above that time spent online is significantly and substantially related to web 2.0 competence, trust awareness and to respondents having more motivations for using web 2.0 applications with increasing time spent online. Causality cannot be determined, web 2.0 competence and trust for example, likely builds up with increased use, yet web 2.0 competent and trust aware individuals will likely spent more time online than average users. Further non-parametric means tests also confirmed strong relationship between time spent online and age-groups of respondents – a negative relationship was found. Hence whether people spent more time online seems to be pre-conditioned on a number of factors including the trend of younger people doing so. However, a connection to business activity wasn't confirmed. Trust awareness (previous section 4.5.4) on the other hand is strongly related to business activity.

4.5.5.3 Web 2.0 Habits across Time Levels

In order to analyse the popularity of individual web 2.0 applications and of web 2.0 related activities over a number of different time spent on web 2.0 sites score levels, the overall time score was split into three groups based on the score⁶⁵; people who according to themselves spent **1**-little time, **2**-normal time and **3**-much time on web 2.0 applications and activities. From table 4.11, one may appreciate that tagging, commenting, uploading files, joining communities, posting to blogs, or rating content are activities that are by far more popular among respondents who indicated high amount of time spent online, than among respondents with lower online time spending tendencies. More respondents who use OpenID / DISQUS also tend to spent more time on web 2.0 applications in general.

⁶⁴ 5% trimmed means were also checked.

⁶⁵ The score ranges from 0 to 9, inclusive.

Table 4.11 – Popular activities of respondents broken down by time spent online score based groups (with Chi-Square tests⁶⁶)

		Time score based time-spent			Total	χ^2 - test.
		Little time	Normal time	Much time		
Q3-1bog post	Count	42	142	63	247	26.388 ¹
	% within trust. I.	23.0%	38.1%	44.4%		p < .0001 ²
Q3-2comment on	Count	104	313	126	543	94.366
	% within trust. I.	56.8%	83.9%	88.7%		p < .0001
Q3-3rated	Count	80	199	91	370	22.847
	% within trust. I.	43.7%	53.4%	64.1%		p < .0001
Q3-4uploaded file	Count	150	347	137	634	59.646
	% within trust. I.	82.0%	93.0%	96.5%		p < .0001
Q3-5tagged content	Count	90	313	135	538	149.336
	% within trust. I.	49.2%	83.9%	95.1%		p < .0001
Q3-6edited shared resource	Count	35	93	43	171	8.907
	% within trust. I.	19.1%	24.9%	30.3%		p < .05
Q3-7joined community	Count	132	320	128	580	48.620
	% within trust. I.	72.1%	85.8%	90.1%		p < .0001
Q3-8rss	Count	55	124	47	226	2.810
	% within trust. I.	30.1%	33.2%	33.1%		p = .245
Q3-9openId / Disqus	Count	11	31	24	66	14.398
	% within trust. I.	6.0%	8.3%	16.9%		p < .001
Q3-10api / mashup	Count	18	54	21	93	4.408
	% within trust. I.	9.8%	14.5%	14.8%		p = .110
Total	Count	183	373	142	698	-

¹Pearson Chi-Square (df, 2)

²two-tailed test

From table 4.12 differences in users' time habits of certain applications can be observed. For instance the respondents who use Twitter or Digg / Reddit are more likely to spend a lot of time on web 2.0 applications in general, than respondents who spent little time. The same cannot be said about Ebay, Amazon, Craigslist or even Flickr / Picassa. That is, there is roughly the same proportion of users who spent a lot, and a little time on web 2.0 applications in general, who indicated that they use the mentioned applications.

66 See appendix B, tables B.21 for the Chi-Square test tables.

Table 4.12 – Popular activities of respondents broken down by time spent online score based groups (with Chi-Square tests⁶⁷)

		Time score based time-spent			Total	χ^2 - test.
		Little time	Normal time	Much time		
Q2-1twitter	Count	55	160	94	309	54.105 ¹
	% within trust. I.	26.6%	42.4%	66.2%		p < .0001 ²
Q2-2youtube	Count	190	366	142	698	17.185
	% within trust. I.	91.8%	97.1%	100.0%		p < .0005
Q2-3facebook / myspace	Count	150	356	139	645	79.615
	% within trust. I.	72.5%	94.4%	97.9%		p < .0001
Q2-4delicious	Count	12	22	16	50	5.283
	% within trust. I.	5.8%	5.8%	11.3%		p = .074
Q2-5flickr / picassa	Count	75	158	70	303	5.921
	% within trust. I.	36.2%	41.9%	49.3%		p = .052
Q2-6wikipedia	Count	191	366	136	693	7.172
	% within trust. I.	92.3%	97.1%	95.8%		p < .05
Q2-7digg / reddit	Count	15	40	33	88	21.905
	% within trust. I.	7.2%	10.6%	23.2%		p < .0001
Q2-8craigslist	Count	15	35	20	70	4.636
	% within trust. I.	7.2%	9.3%	14.1%		p = .098
Q2-9ebay	Count	160	302	121	583	3.356
	% within trust. I.	77.3%	80.1%	85.2%		p = .187
Q2-10amazon	Count	189	347	131	667	0.132
	% within trust. I.	91.3%	92.0%	92.3%		p = .936
Total	Count	207	377	142	726	-

¹Pearson Chi-Square (df, 2)

²two-tailed test

Overall respondents are more conservative when they compare the time spent online against their peers, than when they simply judge whether they spent too much time online. Younger people and less technical people seem to have friends that spent more time online than they do. Time spent online has been found to be significant and important element over the entire sample, with several related factors, such as relationship with higher trust, higher web 2.0 competence and higher number of motivations for UGC contributions. More time spent online was found to be detrimental to increased activity within web 2.0 applications.

4.5.6 Motivation

4.5.6.1 Individuals without Motives

29% of the sample respondents did not indicate a single motivation that drives them to contribute content to web 2.0 type systems. Presumably users without motives are likely

⁶⁷ See appendix B, tables B.22 for the Chi-Square test tables.

characterised as non, or small contributors to web 2.0. A break-down of activities showed that users without motives do contribute far less in terms of writing blogs, tagging, rating, commenting, etc... These differences are significant, where Chi-Square tests were used to confirm the significant differences for each activity (see appendix B, tables B.23 for all the two-way tables used in the Chi-Square tests). The hypothesis H_0, H_1 for this and the next Chi-Square set of tests are, H_0 – There is no significant difference between the two variables, H_1 – There is significant difference between the variables. Table 4.13 presents each activity, with a column for respondents that did not indicate any motivation, for respondents that did indicate at least one motivation, and a column with the Chi-Square test statistic and its p value, i.e. significance level. The respondent set which indicated what motivates them, submits twice as many blog posts and edits twice as many shared resources, for example. An inspection of table 4.13 will help to better appreciate these differences⁶⁸.

Table 4.13 – Popularity of activities amongst motivated and non-motivated users (with Chi-Square tests)

		With / without motivation(s)		χ^2 - test
		No motivation	Motivation indicated	
Q3-1blog post	Count	36	211	30.262 ¹
	%	19.0%	41.5%	$p < .0005$ ²
Q3-2comment on	Count	118	425	35.396
	%	62.4%	83.5%	$p < .0005$
Q3-3rated	Count	67	303	32.079
	%	35.4%	59.5%	$p < .0005$
Q3-4uploaded file	Count	162	472	8.147
	%	85.7%	92.7%	$p < .005$
Q3-5tagged content	Count	128	410	12.831
	%	67.7%	80.6%	$p < .0005$
Q3-6edited shared resource	Count	26	145	16.169
	%	13.8%	28.5%	$p < .0005$
Q3-7joined community	Count	138	442	18.742
	%	73.0%	86.8%	$p < .0005$
Q3-8rss	Count	42	184	12.210
	%	22.2%	36.1%	$p < .0005$
Q3-9openId / Disqus	Count	4	62	16.306
	%	2.1%	12.2%	$p < .0005$
Q3-10api / mashup	Count	9	84	16.452
	%	4.8%	16.5%	$p < .0005$
Total	Count	189	509	-

¹ Pearson Chi-Square (df, 1) ² two-tailed test

⁶⁸ In the entire dataset 4 respondents indicated a motive but did not provide any web 2.0 activity (i.e. answers for Q3 missing). Also 213 did not indicate a motive and 24 did not provide any web 2.0 activity either. This was taken into account within the Chi-Square tests.

Since a visible and robust difference in contribution activity for the two response sets was found, it was further investigated whether there are also significant differences in popularities of various web 2.0 applications. The table 4.14 presents web 2.0 applications used in the response set that have not provided any motives (first column), response set that did provide at least one motivation (second column), followed by Chi-Square test statistic (last column) and its p value (see appendix B, tables B.24 for all the two-way tables). Popularity differences among Amazon, Ebay, Craigslist, Youtube and Delicious between the two response sets was not found to be significant, yet significantly higher popularity for Twitter, Facebook, Flickr / Picassa, Wikipedia, Digg / Reddit in the motivated response set does exist.

Table 4.14 – Popularity of applications amongst motivated and non-motivated users (with Chi-Square tests)

		With / without motivation(s)		χ^2 - test
		No motivation	Motivation indicated	
Q2-1twitter	Count	57	252	30.787 ¹
	%	26.8%	49.1%	$p < .0005$ ²
Q2-2youtube	Count	201	497	2.567
	%	94.4%	96.9%	$p = .109$
Q2-3facebook Myspace	Count	167	478	33.141
	%	78.4%	93.2%	$p < .0005$
Q2-4delicious	Count	10	40	2.259
	%	4.7%	7.8%	$p = .133$
Q2-5flickr / picassa	Count	62	241	19.766
	%	29.1%	47.0%	$p < .0005$
Q2-6wikipedia	Count	198	495	4.331
	%	93.0%	96.5%	$p < .05$
Q2-7digg/reddit	Count	10	78	15.607
	%	4.7%	15.2%	$p < .0005$
Q2-8craigslist	Count	15	55	2.338
	%	7.0%	10.7%	$p = .126$
Q2-9ebay	Count	165	418	1.535
	%	77.5%	81.5%	$p = .215$
Q2-10amazon	Count	190	477	2.881
	%	89.2%	93.0%	$p = .090$
Total	Count	213	513	-

¹ Pearson Chi-Square (df, 1) ² two-tailed test

Not having motives means that respondents are less likely to contribute content to web 2.0 systems, i.e. they *consume* more rather than *produce* more content. Non-parametric means-tests on all factors and demographic variables pointed out never-the less that respondents without motives are less web 2.0 competent and spend less time online. Non parametric, independent samples Mann-Whitney U Test was used to compare distributions of independent samples of

non-motivated respondents and motivated respondents. Hypotheses were as in other experiments, H_0 – there is no difference, H_1 – there is a significant and systematic difference in the distributions of variables. Only the tests for education level and expertise area variables weren't significant at any reasonable p value. From the other variables, after performing visual inspection of box-plots, distribution-plots and comparison of medians, it was found that web 2.0 awareness (*Standardized test statistic* = 9.583) and time spent online (*Standardized test statistic* = 6.579) variables showed the most recognisable difference in central tendency (at significance of p (two-tailed) < .0005) – i.e. time spent online and web 2.0 awareness tends to be consistently lower for respondents who did not indicate a motivation for using web 2.0. Age (*Standardized test statistic* = -4.400, at p (two-tailed) < .0005) was also found to be somewhat higher for such respondents, for example the 12 respondents over 61 year old did not indicate any motivation.

Hence, it could be argued that non-motivated users are relatively new to web 2.0, don't spend much time on it and haven't quite adjusted to it and “decided” what drives their contributions. Also having motivations was found to play a significant role in the use of Twitter, Facebook / MySpace, Flickr / Picassa, Wikipedia and Digg / Reddit, whereas it didn't matter too much in relation to the use of Amazon, Ebay, Craigslist, Delicious and Youtube. This maybe because more people do not feel there needs to be a reason (such as the ones suggested in Q9) to contribute in order to use the latter mentioned applications – i.e. user generated content is considered less important.

4.5.6.2 *Individuals with Motives*

Out of the 513 respondents who selected at least one or more motivations, all these motives can be ranked as follow⁶⁹: **1**–Altruism (23.8%, 235 responses), **2**–Belonging to a community (22.0%, 218 responses), **3**–Knowledge autonomy (20.9%, 207 responses), **4**–Self presentation (18.2%, 180 responses), **5**–Reciprocal altruism (15.1%, 149 responses). These results corroborate with what was reported in Kuznetsov (2006), except that a somewhat larger percentage of respondents identified with self-centred motives. As far as the profiles of respondents with motives, in terms of web 2.0 applications used, and activities performed, are concerned, tables 4.15 and 4.16 present these details.

⁶⁹ Keep in mind that Q9 is a multiple-response set question, where one respondent was indeed allowed to select more than one motivation for their web 2.0 use.

Table 4.15 – Popular web 2.0 applications broken down by respondent’s various motives

		Motivations for web 2.0 use					Total
		Altruism	Reciprocal altruism	Belonging / community	Self presentation	Knowledge autonomy	
Q2-1twitter	Count	129	81	113	116	115	252
	% within Motivation	54.9%	54.4%	51.8%	64.4%	55.6%	
Q2-2youtube	Count	230	145	215	176	200	497
	% within Motivation	97.9%	97.3%	98.6%	97.8%	96.6%	
Q2-3facebook / Myspace	Count	220	138	205	174	195	478
	% within Motivation	93.6%	92.6%	94.0%	96.7%	94.2%	
Q2-4delicious	Count	22	13	19	17	17	40
	% within Motivation	9.4%	8.7%	8.7%	9.4%	8.2%	
Q2-5flickr / picassa	Count	122	80	118	88	97	241
	% within Motivation	51.9%	53.7%	54.1%	48.9%	46.9%	
Q2-6wikipedia	Count	228	143	214	176	201	495
	% within Motivation	97.0%	96.0%	98.2%	97.8%	97.1%	
Q2-7digg / reddit	Count	47	32	38	39	41	78
	% within Motivation	20.0%	21.5%	17.4%	21.7%	19.8%	
Q2-8craigslist	Count	29	18	29	23	28	55
	% within Motivation	12.3%	12.1%	13.3%	12.8%	13.5%	
Q2-9ebay	Count	193	127	176	150	177	418
	% within Motivation	82.1%	85.2%	80.7%	83.3%	85.5%	
Q2-10amazon	Count	219	138	205	170	202	477
	% within Motivation	93.2%	92.6%	94.0%	94.4%	97.6%	
Total	Count	235	149	218	180	207	513

Percentages and totals are based on respondents.

The breakdown of motivations per applications is very useful. One can for example appreciate from the table 4.15 how **self presentation** is the most common motivation for Twitter (Chi-Square test statistic 26.047, significant at p (two-tailed) $< .0005$, for two-way table see appendix B, table B.25), Facebook / Myspace (only marginally however) and even Digg / Reddit (closely followed by **reciprocal altruism**). Interestingly Flickr / Picassa depend on users with strong feelings of **belonging to a community** and **reciprocal altruism**. This is characteristic of picture sharing applications, where the main reasons for sharing often seem to “*boil down to*” wanting to show pictures to family, colleagues, society members (i.e. **community sharing, belonging** based motivation stems from this), and the prevailing attitude of; I show some pictures and I also expect you to eventually show me some of your pictures at some point

(**reciprocal altruism** is related to this). Not quite surprisingly contributions to Wikipedia are most often equally motivated by **belonging to community**, **altruism / reciprocal altruism**, **knowledge and autonomy** or feelings of **self presentation**. Due to the encyclopaedic nature of Wikipedia, there is a exceedingly wide range of users with many motivations who access Wikipedia. Motivations on Wikipedia have been well researched (Kuznetsov 2006, Bishr 2009), as opposed to other applications, see section 2.2.4.2 for a presentation of previous work.

Similar patterns discussed above, for application specific motivations, can be appreciated from table 4.16 for indicative motives of actions, such as submitting Blog posts, which is often done by individuals who are motivated by self presentation and knowledge / autonomy. Interestingly the acts of rating is more often done by respondents motivated by altruism, knowledge / autonomy, whereas commenting content is motivated by self presentation and knowledge / autonomy.

Table 4.16 – Popular activities of respondents broken down by respondent’s various motives

		Motivations for web 2.0 use					Total
		Altruism	Reciprocal altruism	Belonging / community	Self presentation	Knowledge autonomy	
Q3-1blog post	Count	112	64	102	109	112	211
	% within Motivation	48.3%	43.0%	46.8%	60.6%	54.4%	
Q3-2comment on	Count	196	132	180	160	183	425
	% within Motivation	84.5%	88.6%	82.6%	88.9%	88.8%	
Q3-3rated	Count	153	95	136	115	134	303
	% within Motivation	65.9%	63.8%	62.4%	63.9%	65.0%	
Q3-4uploaded file	Count	212	140	209	171	193	472
	% within Motivation	91.4%	94.0%	95.9%	95.0%	93.7%	
Q3-5tagged content	Count	191	130	178	157	173	410
	% within Motivation	82.3%	87.2%	81.7%	87.2%	84.0%	
Q3-6edited shared resource	Count	87	51	80	64	72	145
	% within Motivation	37.5%	34.2%	36.7%	35.6%	35.0%	
Q3-7joined community	Count	207	129	199	162	187	442
	% within Motivation	89.2%	86.6%	91.3%	90.0%	90.8%	
Q3-8rss	Count	99	66	92	79	92	184
	% within Motivation	42.7%	44.3%	42.2%	43.9%	44.7%	
Q3-9openId disqus	Count	44	23	32	34	36	62
	% within Motivation	19.0%	15.4%	14.7%	18.9%	17.5%	
Q3-10api mashup	Count	51	36	38	43	51	84
	% within Motivation	22.0%	24.2%	17.4%	23.9%	24.8%	
Total	Count	232	149	218	180	206	509

Percentages and totals are based on respondents.

4.5.6.3 Web 2.0 Awareness and Motivation

Several interesting aspects about motivation and user behaviour in regards to web 2.0 applications and activities were already presented. However, what about people who simply indicated that they are aware of web 2.0 in question 1 (*i.e. respondents who indicated that they have at least heard the term web 2.0 and also respondents that have a rough understanding of the concept*), are these individuals more likely to be aware of their motivations as well, presumably since they are more educated about the meaning behind web 2.0. A two-way table Chi-Square test was performed to find out whether H_0 – there is no relationship between motivation and web 2.0 term awareness, and H_1 – there is a statistically significant relationship between motivation and web 2.0 term awareness. The two way table is shown below (see table 4.17), and Chi-Squared test statistic was 19.380, which was significant at p (two-tailed) < .0005.

Table 4.17 – Two-way table – Web 2.0 aware vs. Motivation exists (used in Chi-Square test)

		Motivation exists		Total
		No	Yes	
web 2.0 aware	No	121	200	321
	Yes	92	313	405
Total		213	513	726

Hence it can be concluded with some conviction that indeed respondents who are aware of web 2.0 as a concept will more likely have a tendency to have motives. However this is not a causal relationship since it is more likely that motives of individuals are conceptualised from intensive use of and contribution to web 2.0 systems.

In summary it was found that the more motives a respondent has the likelier they are to spent a lot of time online (*time score, non-parametric test*), and they would also tend to use more web 2.0 applications in different ways (*web 2.0 competence score, non-parametric test*). It was further found that the likelihood to contribute content to web 2.0 applications differs significantly for respondents who are motivated by one or more factors, as opposed to respondents who don't have any explicit motivation(s) to contribute content to web 2.0 applications (*a series of Chi-Squared tests and cross-tabulation tables*). Not having motives means that respondents are less likely to contribute content to web 2.0 systems, i.e. they *consume* more rather than *produce* more content. Such users are less, web 2.0 competent and spend less time online. It could hence be argued non-motivated users are relatively new to web 2.0, don't spend much time on it and haven't quite adjusted to it and decided what drives their contributions. Also motivations that play a significant role in the usage of specific applications were identified.

Understanding of motivations as they apply to different UGC (i.e. User Generated Content) based applications is little understood and the breakdown in this section provides an important contribution to knowledge, also as far as we can relate the applications to our proposed web 2.0 taxonomy. Throughout this thesis, support for the proposed web 2.0 taxonomy is emphasised to help us present its validity and to help provide support for it on substantial, quantitative and qualitative insights.

4.5.7 Wikipedia

The final question in the survey asked respondents to indicate whether they find Wikipedia to be a useful encyclopaedic reference text. There has been much debate in the media and academic literature about potential inaccuracies, and other issues relating to Wikipedia as a useful body of encyclopaedic reference. This study discovered that 12 times more respondents fully agreed with the statement (strongly agree vs. strongly disagree) “I consider wikipedia.com to be a useful body of encyclopaedic reference”. The frequency table 4.18 highlights the overall bias in the distribution towards people being prepared to accept Wikipedia as a useful encyclopaedic reference text⁷⁰, the mode and median being 4 (*i.e.* 4 = *agree*) and the 25th percentile = 3, and 75th percentile = 4.

Table 4.18 – Wikipedia usefulness – Frequency table for the 726 survey-responses

	Frequency	Percent	Cumulative Percent
Strongly disagree	12	1.7	1.7
Disagree	48	6.6	8.3
Neutral	154	21.2	29.5
Agree	341	47.0	76.4
Strongly agree	171	23.6	100.0
Total	726	100.0	

Given responses from the survey, Wikipedia is a publicly, well respected encyclopaedic resource (71% of respondents consider it to be the case). Correlations in table B.4 in appendix B, and visual inspection of the distribution box-plots for different demographic groups did not point to any explicit pattern underlying the perceived usefulness of Wikipedia. There were 693 out of 726 individuals who indicated they use Wikipedia, there were 57 respondents out of a total of 60 who said that they do not find it to be a useful encyclopaedic reference yet they use it. The table 4.19 illustrates this, together with the activity of editing shared resources, which is

⁷⁰ It must be noted that, since Wikipedia is considered to be useful it also implies that it must be accurate to a satisfactory degree that is acceptable for people to associate with the statement that Wikipedia is a useful body of encyclopaedic reference. Hence this implies that usefulness – accuracy are synonymous.

associated with Wikipedia collaborative participation.

Table 4.19 – Wikipedia usefulness per Wikipedia (Q2, top table) and edited shared resource (Q3, bottom table)

		Q10-Wikipedia useful (likert)					Total
		strongly disagree	disagree	neutral	agree	strongly agree	
Q2-6Wikipedia	Count	11	46	142	325	169	693
	% within Q10	91.7%	95.8%	92.2%	95.3%	98.8%	
Total	Count	12	48	154	341	171	726

		Q10-Wikipedia useful (likert)					Total
		strongly disagree	disagree	neutral	agree	strongly agree	
Q3-6edited shared resource	Count	3	8	29	86	45	171
	% within Q10	25.3%	17.0%	20.4%	25.9%	27.3%	
Total	Count	12	47	142	332	165	698

Table 4.20 shows the motivations from Q9 grouped by different Q10 levels. The percentage of respondents for the motivation categories; altruism and knowledge autonomy are the largest, with altruism being not present as a motivation for respondents who do not find Wikipedia useful. Wikipedia is probably dominated by Altruism as the most important motivation for user generated content.

Table 4.20 – Wikipedia usefulness broken down by motivations (Q9), 513 survey-responses

		Q10-Wikipedia useful (likert)					Total
		strongly disagree	disagree	neutral	agree	strongly agree	
Q9-Motivation altruism	Count	0	12	39	114	70	235
	% within Q10n	.0%	35.3%	41.1%	47.1%	50.7%	
Q9-Motivation reciprocal altruism	Count	2	8	24	72	43	149
	% within Q10n	50.0%	23.5%	25.3%	29.8%	31.2%	
Q9-Motivation belonging community	Count	2	14	37	106	59	218
	% within Q10n	50.0%	41.2%	38.9%	43.8%	42.8%	
Q9-Motivation self presentation	Count	3	9	24	93	51	180
	% within Q10n	75.0%	26.5%	25.3%	38.4%	37.0%	
Q9-Motivation knowledge autonomy	Count	4	14	39	91	59	207
	% within Q10n	100.0%	41.2%	41.1%	37.6%	42.8%	
Total	Count	4	34	95	242	138	513

Overall Wikipedia is used by 96% of all respondents, which ranks it just second after Youtube in the sample. In question Q10, 60 people altogether indicated that they strongly disagree, or

simply disagree with the statement that Wikipedia is a useful body of encyclopaedic reference, yet only 2 respondents of those 60 did not use Wikipedia at all. Overall 71% of the sample considers Wikipedia to be useful, but it has not been possible to determine convincingly whether this opinion differs significantly over various demographic variables. Correlation analysis hinted to a relationship between the increasing number of motivations and usefulness of Wikipedia. A breakdown of different motivations revealed that altruism was a motivation that generally seemed to be out of favour with respondents who do not find Wikipedia useful; however, it was not possible to deduce much more out of this data. The sample strongly indicates that Wikipedia is indeed well respected among the majority of web users, including academics (i.e. PhD/Dr, 67% find Wikipedia useful).

4.6 Limitations of the Study

Many issues associated with correct survey design were taken into account when designing this survey, in section 4.2. Considering survey studies of web 2.0 from prior literature, the design methodologies employed in this chapter are more complete, and help to avoid numerous issues. However, the survey results were limited in several respects.

The sample contains bias towards an academic population, since the snowball sampling was primarily initiated from a sample of individuals at a British University (Loughborough University). For instance, 93% of all respondents are studying at university or attained higher level degrees, and 55% of the sample is in the 20-30 year old age-group (see section 4.3.5). Hence, in aggregate responses are clearly biased, as various strata of the population of all web users are simply not represented in their proper proportions. It is crucial to keep this bias in consideration when interpreting responses and inferring conclusions from our survey. Nevertheless the bias is somewhat alleviated with snowball sampling which facilitated collection from a wider range of population strata than a simple sampling approach would (see section 4.2.3). Some patterns of trust, time spent online, or motivations of web 2.0 use across various demographics were observed, and although clearly not representative they can with some likelihood be considered indicative.

Since the goal was to collect a large sample, questions had to be kept simple and short. This meant that elements such as checkboxes had to be used where maybe response likert-scale items would have provided a far better response resolution, than a simple yes / no answer. Clearly the short length of the survey was prohibitive in discovering more detailed sample data. Although, as was argued in section 4.2, a longer survey might have decreased the accuracy of

responses and lowered the response rates altogether⁷¹; however, getting the trade-off between length of survey and depth of response data right is a challenge.

Potential problems with subsequent analysis of survey responses were also given due attention in section 4.3. Some of the results that lack statistical significance should only be considered as indicative. A large sample of over 700 respondents does allow to, draw some insightful conclusions nevertheless. Especially the results regarding web 2.0 competence and awareness of social media, and web 2.0 concepts are significant.

4.7 Summary

In chapter 2 and 3 the concept of web 2.0 and social media with its contextual background were established. It was possible to show convincingly that there is strong interest into the phenomenon and also that some web 2.0 terms are becoming widely used. This chapter presented a large survey with 726 responses and the sole aim to substantiate and elaborate the concept of “web 2.0” and “social media”, as it is publicly perceived. A strong indication of how users relate to various issues, such as time spent, motivations, economic participation, or trust awareness in the web 2.0 context was also presented. This type of understanding was missing from current body of literature. The proportion of web 2.0 users who contribute UGC was found to have increased radically over the last few years, considering previous reports. Also the popularity of web 2.0 applications with older respondents was found to be rather high. Money based online transactions, including group buying were found to be quite prevalent. The role of trust and investment of time, as highlighted in chapter 2 were substantiated and confirmed. The significance of motives for contributing UGC and for using certain web 2.0 applications were shown. Below, the bullet points highlight some findings in this chapter.

- Half of the entire sample was found to be aware of both the terms social media and web 2.0, but social media is more prevalent among respondents than web 2.0 – 81% vs. 55%. Both terms, or at least one was known to 88%, indicating that some percentage of users use web 2.0 applications but don't associate the terms with their use (section 4.5.2.1). More educated and technical users were aware of the terms web 2.0 and social media.
- Surprisingly many people were found to use or having used group buying in the past. This group of respondents was analysed in some detail in section 4.5.3.1, and most interestingly it was found that there wasn't any systematic demographic variation that predisposes users to be more likely to having had used group buying websites, other

71 For example, Daugherty et al. in their web 2.0 survey achieved 325 survey responses, although 23% of respondents did not complete the survey, as they abandoned it due to its length.

than the general awareness and intensity of web 2.0 use.

- Considering that 100% of respondents used at least one web 2.0 application, we found that only 57% respondents consider themselves able to judge trustworthy from non-trustworthy sites. Younger people were found to be consistently more confident of their ability to judge trustworthiness. Yet only 1.7% of respondents feel comfortable sharing their personal details on non-secure web-applications.
- Trust was found to be a significant factor, as it relates to other factors on the web, i.e. web 2.0 competence, business activity, time spent online and motives (section 4.5.4.1). This is consistent with and lends support to discussions from section 2.2.
- Insightful details are provided for web 2.0 activities, specific applications and the relative importance of trust. Trust was found to be most important for the web 2.0 activity of joining a community, and Amazon, Ebay, Facebook or Twitter were more likely to be used by users with higher trust awareness levels (section 4.5.4.2).
- A breakdown of web 2.0 applications and activities by how much time users spent online is provided in section 4.5.5.3. Facebook / Myspace, Twitter, Digg / Reddit, and tagging content, commenting, uploading files, are some of the applications and activities with a higher number of users who spend more time online. Time spent on web 2.0 decreases with increasing age groups.
- A detailed breakdown of motives in use of individual web 2.0 applications and web 2.0 activities is provided within section 4.5.6.1 and 4.5.6.2. For instance motives were found to be significant in applications such as Twitter and Facebook, but unimportant in the use of Ebay, Amazon, or Craigslist.
- It was found (see section 4.5.6.3) that respondents who are aware of web 2.0 as a concept will more likely have a tendency to have well defined motives. This is not a causal relationship since it is more likely that motives of individuals are conceptualised from intensive use of and contribution to web 2.0 systems.
- Wikipedia was confirmed as a well respected resource, in particular 67% of academics found it to be a useful encyclopaedic reference text.

The survey presents interesting results and is a valuable contribution to prior literature, since it substantiates several important claims. The survey was designed and validated, following survey design recommendations from literature.

In the next section the historical evolution of the web towards web 2.0 will be investigated in an empirical manner, in order to substantiate the concept of web 2.0, as it was introduced in earlier chapters.

5 Historical Evolution towards Web 2.0



5.1 Background

Despite initial and ongoing criticism of web 2.0 (Zimmer 2008; see section 2.2.4), the term and concept behind it provides a valuable insight into the way in which web has evolved over time (Millard and Ross 2006). As was suggested by Alexander (2006); “*the term Web 2.0 assumes a certain interpretation of Web history, including enough progress in certain direction to trigger a succession [i.e. Web 1.0 → Web 2.0]*”. The most recent generation of web sites have been considered by some to be fundamentally different from the ones found on the early web (see section 2.2.1), these have been grouped together under the term web 2.0. The name is arguably misleading, since it implies a designed version and a discrete evolution, although presumably the emergence of web 2.0 applications did not occur as a sharp break with the old but, rather, the gradual emergence of a new type of practice. However, the concepts behind it provide a valuable insight to the simple observation that the web has evolved. Within this chapter an original retrospective and historical study of seven major websites is undertaken in order to investigate, what it is that has changed on actual websites over time. This is the first kind of its study, as far as the author is aware.

In order to understand the historical evolution towards web 2.0, it would be necessary to

investigate the evolution of websites throughout time. This is impossible unless a reliable, independent, and trustworthy historical dataset exists. In this study such a dataset was identified and an empirical analysis of the Wayback Machine¹, an internet archive, was undertaken. The Wayback Machine (from now on referred to as WM) is a time-capsule library of cached websites with over 3 petabytes of stored web-page content (as of 2009), with about 20 terabytes (10¹² bytes) of new digital content being added each month to the archive (Murphy et. al. 2008). Web pages are usually cached repeatedly at various multiple month intervals². The library is a non-commercial project supported by Alexa Internet Inc and its significant uses included as evidence in court cases (Gelman 2004, Howell 2006), and also a significant portion of the archive was donated to the US congressional library as early as 1998 (2 years into the project's existence). Since the system started to aggregate web-page snapshots in 1996, many popular websites are represented over their entire lifespan. Hence the WM is a unique dataset, literally a series of time-capsules of WWW at various points in its evolution, and suggests itself as a suitable tool for a historical web-page analysis³. Indeed earlier studies using the WM dataset also exist. However, these have in particular investigated the evolution of design techniques and web accessibility⁴ using historical snapshots from the WM (Fukuda et al. 2005, Hackett and Parmanto 2005, respectively). These studies used various complexity and accessibility scores derived from historical page features over statistically large samples to successfully infer interesting patterns. The reliability of data on WM has been scrutinised and explicitly validated in a separate study, and was found to be of reliable quality (Murphy et al. 2008)⁵. Given previous successful studies using the WM (Fukuda et al. 2005, Hackett and Parmanto 2005, Murphy et al. 2008), it was decided to undertake an extensive, first of its kind, empirical analysis looking at web 2.0 related developments using the WM archive.

5.2 Evolution of Web 2.0 and Web-design Considerations

Web design and presentation habits have changed markedly over time (Zeldman 2007). Some HTML syntax has become obsolete while new development patterns such as AJAX or CSS based design, *i.e. design from content separation* have emerged (Murphy and Persson 2008, Budd et al. 2009). In this study some 40 features per page, for every website were extracted to

1 Accessible at <http://web.archive.org>

2 Some web-pages are archived more frequently than others however there is a lag of 6-12 months between the caching and the time the cached content is made publicly available on the Wayback Machine.

3 The site itself argues that they “*seek to collect and preserve the digital artefacts of our culture for the benefit of future researchers and generations*”

4 In numerous countries web accessibility standards became a regulatory requirement by law and have therefore received some attention (Zeldman 2007).

5 The study analysed the accuracy of WM on a sample of travel industry websites. See Murphy et al. (2008) for a discussion of numerous issues and the full analysis of Wayback Machine reliability.

be analysed. These features were split into 4 streams of features in order to help organise the various aspects related to evolution of the web. The main emphasis was on the temporal adoption of web 2.0 elements, activities and design patterns. Most of the features presented in table 5.1 can be directly related to the adoption of web 2.0 (4th stream of features, right-most column in the table), where the web 2.0 elements and web 2.0 activities are largely based on discussion within chapter 2, specifically section 2.3.2. The 3rd stream of features (2nd column from the right, table 5.1) is represented by JavaScript, AJAX, and CSS Stylesheet adoption, as these were identified to be conducive to web 2.0 adoption, in acting as a catalyst to the process of change. Outdated tags (in the 2nd stream, table 5.1) should illustrate the gradual disappearance of problematic design elements that might have acted as barriers to wider web application standardisation and adoption.

The websites for analysis were chosen on the basis of being considered to be representative of web 2.0 systems. In particular, Youtube, Amazon, Flickr, Twitter, Craigslist, Digg and Yahoo; a breakdown of reasons for the choice is provided in table 5.2. As mentioned earlier, several historical internet studies were undertaken by researchers in the past; however, this study is the first to investigate web 2.0 adoption related trends, rather than accessibility or generic design issues.

Table 5.1 – Wayback Machine study; overview of features extracted from every page analysed

General	Outdated HTML	Script and Styling	Web 2.0 activities & elements
1- Page Title 2- Page length (characters) 3- Unique Tags (count) 4- Lexical Density 5- Unique Alpha 6- Unique Digits 7- In-site Links 8- Out Links 9- Form elements (count)	1- table (tr, td) 2- font formatting tags (<i>b, i, u, big, small, font</i>) 3- center tag 4- menu tag 5- layer tag 6- blink tag 7- marquee tag 8- attributes (<i>align, bgcolor, background, text, link, vlink, alink</i>)	1- All JavaScript tags 2- JavaScript source tags (<i>external JS code</i>) 3- Script libraries used ⁶ 4- AJAX use 5- CSS Styling (in-page) 6- CSS Styling (source)	Web 2.0 Elements 1- RSS 2- API 3- Mashups 4- Podcast 5- Blog 6- Tagcloud 7- Wiki 8- Permalink Web 2.0 Activities 1- Share 2- Comment 3- Submit / Upload 4- Rate 5- Tag 6- Like/Favourite/Unlike/Report This 7- Edit

Since the investigation is a temporal study in understanding the evolution of web 2.0, it is appropriate to look back at some factors behind WWW evolution⁷, and to relate these to the

6 JQuery, Prototype, Joose, Dojo, GWT, Processing, Scriptaculous, Midori, Pyjamas, Rico YUI, Qooxdoo, Mootools, Mochikit were the libraries that our script checked for (http://en.wikipedia.org/wiki/List_of_JavaScript_libraries). If a top level site or a referenced script made a reference to one of the libraries then our script would detect this as well.

7 See section 2.2.1 on the Historical Perspective

features from table 5.1.

1. **Development became easier** – the appearance of JavaScript Libraries, improved XHTML (*generally much better cross-browser standardisation*)
2. **Architecture / Integration** – RSS, Mashups and APIs. These three features are understandably rather rough approximation to this factor, yet to some extent indicative.
3. **Standardisation** – Outdated HTML; the phase-out of old mark-up / tag attributes, and the arrival of new tags and emphasis on *design – content* separation
4. **Viability of online business models** – cannot be investigated by this study; however, references to literature dealing with this question were provided in section 2.2.1, and the survey in chapter 4 deals extensively with questions related to business models.
5. **Trust** – has not been measured in this study in any direct manner. However, on a more indirect basis it can be argued that significant increase in sharing of personal data on the web is an effect attributable to web users having more trust in sharing content online. Therefore the features related to forms and Web 2.0 activities of sharing, commenting, submitting may be slightly indicative.

Qualitative studies where websites are manually reviewed are useful; however, quantitative analyses are more feasible over larger datasets and generally tend to be more objective (Moore and McCabe 2001). A qualitative inspection by a human can; however, help to validate experimental results, and one might also observe things that weren't planned for / detectable automatically. Let us consider the main page of Youtube from the 1st January 2006⁸. It contained a tag-cloud⁹, a last five users online widget, and video listings used practically no client-side scripting yet. The Youtube Blog was just launched recently – in other words, it seemed that Youtube was just beginning to use various web 2.0 collaborative elements more intensively, but with some reservations. In the next section it will be shown how this assessment correlates with our quantitative analysis. The second issue which is closely related to what was said above is concerned with the interpretability of such results. Any inference from the dataset was approached with caution as other contributing factors might have been unmonitored or omitted; however, in certain cases inference was relatively self-evident. Consider for example that from within the earliest entries for the title of Youtube (html based header / title tag¹⁰), the slogan in the title changed from “*YouTube – Your Digital Video Repository*” into “*YouTube – Broadcast Yourself*”, this represents the symbolic shift in the site's purpose to a more direct “*Broadcast Yourself*”. This was also noted and described by Burgess and Green (2009). In the

8 Accessible at <http://web.archive.org/web/20060101075658/http://youtube.com/>; page snapshots from over 900 other dates are also available for Youtube over the years 2005, 2006, 2007, 2008, 2009 and 2010 in the WM archive.

9 Tag cloud or word cloud is a visual depiction of textual content within a graphic, where words are scaled based on their occurrence frequencies, see http://en.wikipedia.org/wiki/Tag_cloud.

10 This is an element that makes a title appear as the title in the browser bar (or an active browser tab, on newer browsers). As it became apparent from the study – over time this title does not tend to change much, the only times it has changed was when a website wanted to communicate a different perception about itself.

next section (5.3) methodology details and limitations of the study will be described, followed by section 5.4, which will report results of the analysis with an evaluation. Discussion of the study outcomes are reported in section 5.5.

5.3 Methodological Details and Limitations

Data from the WM was downloaded in multiple phases, and collated into 7,482 (*after initial cleaning*) instances, where each instance describes a historical web-page using 45 features. The overall time-window covered by the dataset ranges from 17/10/1996 to 16/04/2010¹¹. Reasons and justification behind the choice of web applications for analysis are provided in table 5.2.

There is a latency of several months until data crawled by the Wayback crawler appears on the archive's servers, hence our dataset has a cut off date in April 2010. The frequency of entries in the archive also tends to be irregular. This is due to at least two reasons related to how the WM crawls the web¹². Historically, some web-pages were also down for maintenance or technical issues, such as when Amazon experienced disruption to its service during October 2000. Other pages were corrupted or temporally inaccessible from the actual WM servers. In the first download phase over 3,000 pages were corrupted, and hence a second phase of downloads (distributed over an entire week) was necessary. In this phase 90% of the initially corrupted dataset was successfully re-download. A third data download phase to download all the script files associated with pages to additionally improve the accuracy of the dataset, took place.

The overall download methodology is summarised in figure 5.1, and in fact the related scripts and datasets are available at <http://www.newsmental.com/thesis/wayback.html> for inspection.

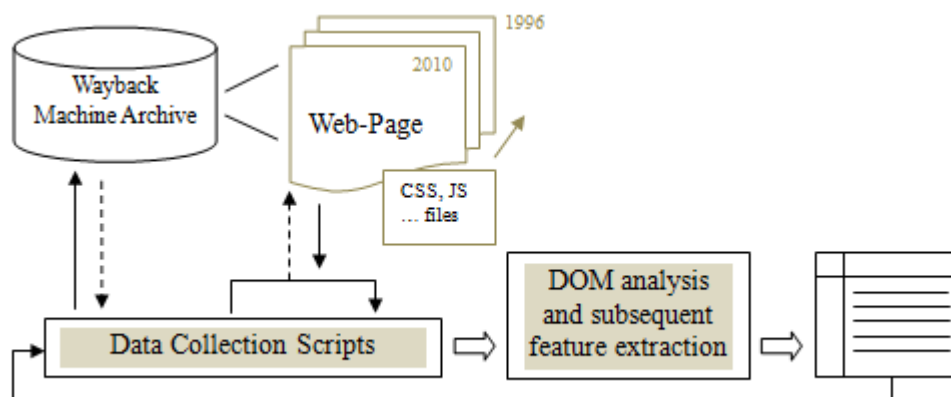


Figure 5.1 – Diagram of Wayback Machine archive extraction methodology

¹¹ Of-course it was possible to access the most recent versions of the websites directly for comparable feature extraction.

¹² See Wayback FAQ - <http://www.archive.org/about/faqs.php>

Essentially the methodology is split into four phases. *Phase 1* – A custom script gets all historical records available from the WM archive (only cases indicated to contain substantial changes by the archive, are considered). *Phase 2* – For each archive entry in the list, the HTML code of the page is requested, in addition with all the *.js*¹³ files linked to it. *Phase 3* – A DOM tree¹⁴ of the HTML code is generated, which allows a custom script to then extract the required tags, and allows the automated content analysis to look for keywords. *Phase 4* – Saves all the extracted features of each HTML page and related *.js* scripts as a record (i.e. row) into a *.csv*¹⁵ file. This can easily be imported into SPSS, WEKA or Excel for further analysis.

A total of 45 features are extracted for each HTML page, and from related *.js* files (57, when Old Tags and Old Attributes are also considered). The following bullet points provide a detailed overview.

1st Stream (General)

- Informational, (8)
 - 1-Wayback ID, 2-website, 3-url, 4-date, 5-month, 6-year, 7-unique digits, 8-unique alpha numeric characters
 - Where feature 5 and 6 above are used for ease of analysis, 7 and 8 for detection of possibly erroneous pages. The Wayback ID represents a unique archive-entry *url* and can be used to access the archived version of the page directly.
- Basic Characteristics, (6)
 - 1-Page Title (in HTML-HEAD-TITLE), 2-Page Length (in characters, all characters), 3-unique Tags (all HTML tags), 4-lexical density %, i.e. ((set(allTags)/allTags)*100), 5-insite links (relative links), 6-out links (contains http)
 - Checks for insite / out links is simplistic, no check against server domain takes place.

2nd Stream (Outdated HTML)

- Input Form Fields, (1)
 - 1-<input ...> tags were searched in the page DOM (Document Object Model)
- Outdated tags, (7, or 9 including *tr* and *td* tags)
 - 1-table (*tr*, *td*), 2-b, 3-i, 4-u, 5-big, 6-small, 7-font
- Old Tags, (5 possible tags, aggregated into one feature)
 - 1-center, 2-menu, 3-layer, 4-blink, 5-marquee
- Old Attributes, (7 possible attributes, aggregated into one feature)
 - 1-align, 2-bgcolor, 3-background, 4-text, 5-link, 6-vlink, 7-alink

3rd Stream (Script and Styling)

- Styling Tags, (2)

13 *.js* is the file extension usually associated with JavaScript files, and is the de-facto standard client-side language on the Web (Crockford 2008).

14 A DOM tree is a standard representation of an (X)HTML page, which allows to programmatically manipulate or access its elements (XPath supported), see <http://www.w3.org/TR/DOM-Level-2-Core/introduction.html>

15 *.csv* is the file extension which stands for a “comma-separated values” file. It is used to store tabular data, and supported by most spreadsheet software or data analysis packages.

- 1-css Style, 2-css Stylesheet (external Stylesheet, i.e. *src=true*)
- Script Tags, (4)
 - 1-Script Tags, 2-Script *src=True* Tags (external JavaScript files referenced from the page), 3-Client-side libraries used; checks for 14 different libraries, in addition all “src” attribute *.js file headers (up to 200 characters) were checked too¹⁶, 4-AJAX use has been checked by searching for *msxml2.xmlhttp*, *microsoft.xmlhttp*, *xmlhttprequest* JavaScript objects. Also all related “src” attribute *.js files for any AJAX were searched. HTML frame based implementations were not checked for; however, these are not common at all.

4th Stream (web 2.0 atomic activities and elements)

- Web 2.0 Elements, (8)
 - 1-RSS, 2-API, 3-Mashup, 4-Podcast, 5-Blog, 6-Tagcloud, 7-Wiki, 8-Permalink
- Web 2.0 Activities, (9)
 - 1-Share, 2-Comment, 3-Submit, 4-Rate, 5-Tag, 6-Like, 7-Unlike, 8-Edit, 9-Report this

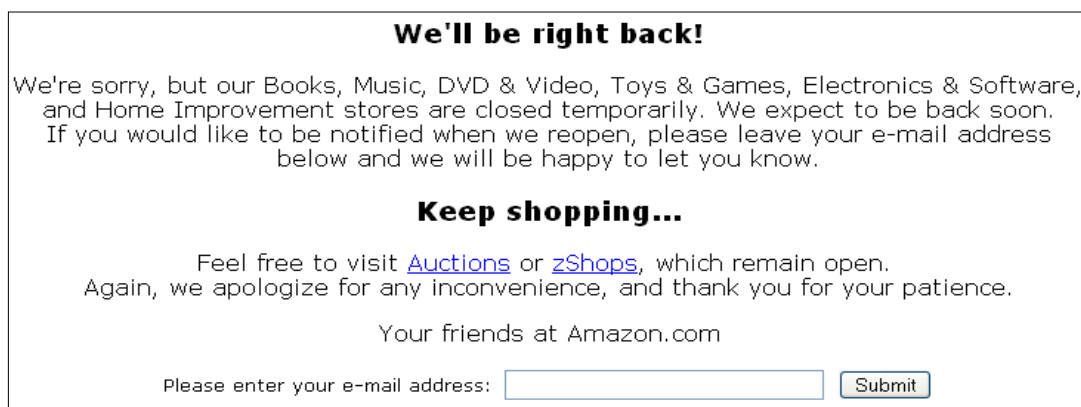


Figure 5.2 – Amazon.com during October of 2000 (page down for technical reasons)

As was mentioned some pages either did not download correctly, had to be re-downloaded, or were completely inaccessible, corrupted, and unusable for meaningful analysis, see Figure 5.2 for an example of an unusable page. Clearly web-pages, such as in figure 5.2 are not to be included in the actual analysis, hence a thorough dataset cleaning procedure was necessary, to ensure a representative dataset quality of valid pages. To this end, most features in the 1st stream were used in a set of data-cleaning rules. To illustrate how this may be done, using at least a couple of such heuristic rules, consider the following. First, the data would be checked for any unavailable pages. This is easily recognised by the *title* of the page containing “Internet Wayback Machine Archive...” reporting an error message, instead of the actual title, such as “Twitter: A Whole World in Your Hands” for Twitter. Secondly the number of unique alphanumeric characters would highlight any pages that contain an unexpectedly low set of such characters (such a check would have detected the page in figure 5.2). Several pages with

16 JQuery, Prototype, Joose, Dojo, GWT, Processing, Scriptaculous, Midori, Pyjamas, Rico YUI, Qooxdoo, Mootools, and Mochikit were the libraries that the script was actively looking for (http://en.wikipedia.org/wiki/List_of_JavaScript_libraries).

extreme values for some of the features were also checked by a human expert, to ensure data quality.

Table 5.2 – Reasons for the choice of websites, analysed in this study

Website	Reason
Youtube	Youtube was started as a video sharing website in February 2005 and has evolved into one of the most heavily used websites on the entire WWW (based on traffic). It allows viewing, updating, commenting, tagging, annotating and rating videos, along with creating user specific social profiles, playlists, and some further user driven features. It seems to be a <i>text-book</i> example of a web 2.0 application.
Flickr	Flickr was launched in February 2004, and it became one of the main destinations for uploading personal photos which can be shared and annotated in several ways. Given Flickr’s ubiquity and significance as a picture sharing application, it was believed to be appropriate to include it in this study.
Digg	Digg is one of the most popular and widely used text-sharing web-sites. It allows to create, share, and peer review news stories. It has employed many web 2.0 elements from the beginning of its existence in late 2004. Digg is an exemplary website of a web 2.0 application and we’d like to see how it has evolved with time.
Craigslist	Craigslist is one of the oldest websites in our selection, which is the main reason for its inclusion in the analysis. It had an online presence since 1996 and the WM data for Craigslist begin in 1998. Craigslist is a collaborative application which started off as a mailing list and has evolved from a relatively centrally managed system to a fully user driven website. It has been ranked at the highest level on O’Reilly’s web 2.0-ness scale (see section 2.3.1.1) and is well known for having kept a simple design over the years, which might provide for an interesting analysis.
Twitter	Twitter was launched in 2006 and has grown at a fast rate in popularity. It has been ranked at the highest web 2.0-ness by O’Reilly. The site represents a whole group of web 2.0 applications, widely known as micro-blogging websites. Understanding how Twitter has changed is crucial to understanding how web 2.0 systems evolved over time, and improved user engagement.
Amazon	Amazon has been ranked at the highest level of web 2.0-ness despite not being really a classical web 2.0 system example. Amazon is a popular shopping website, widely attributed to having pioneered product reviews and numerous other social features, such as shared wishlists or direct recommendations. Amazon is an interesting website for the study for three reasons. First, it has been around since 1995; Secondly it has been a pioneer in web-applications and new design elements; Thirdly, it grew out from a simple online bookstore to a fully fledged web 2.0 system. Observing this gradual transformation seems therefore worthwhile.
Yahoo	Yahoo permits to look far into the past, since it has been online since 1995. It started out as a website directory (as opposed to a classical search engine), and evolved into a web portal. Yahoo stands out from the selection of websites in that it has traditionally been perceived as a web-portal. We would like to understand how a website that was traditionally administered has adapted to the era of social-media and UGC.

There are a number of limiting factors, or drawbacks to this study which must be considered. First of all, only the main page (homepage) for each date was analysed. This is a limitation to the representativeness of the given web-site, yet it provides a glimpse into the site, and weighs against the factor of running time. In other words, experiments that would analyse a set of related pages on the same web-application domain would take prohibitively more time to run,

and secondly the WM does not usually store many related pages, and if it does, then the relative links tend to be often broken. Nevertheless all related JavaScript files to the main page were analysed. A more significant limitation was that pages themselves were not segmented by content, other than based on the HTML mark-up. For example, on a page like Craigslist, some of the user contributed content would also be scanned for words referring to “Wiki” or “Blog”. This can skew the results for web 2.0 activities and web 2.0 elements upwards, since this text isn’t part of the Craigslist application. The assumption was that despite some inaccurate readings, overall the accuracy will be sufficient. Results are averaged out and any intense variability in the readings was detected relatively easily, and removed from the dataset.

Finally it must be taken into account that the dataset retrieval and preparation was a major factor in this study. Separate download phases had to be run over multiple weeks. Altogether over 350 Megabytes of web-page data was extracted, parsed and analysed for the required features. The page download (*from the WM archives*) and feature extraction (*from the actual HTML files*) routines were written in Python¹⁷, and are available at <http://www.newsmental.com/thesis/wayback.html>.

5.4 Results

This section provides an overview of the collected datasets, followed by an analysis of substantive website-design changes, as observed throughout the available sample time range. Since the historical study focused on a narrow set of web 2.0 applications, a second study of a much larger set but current snapshot only, of applications is performed (section 5.5.1). Finally a discussion on the methodology’s potential uses follows, and suggestions are made for its use in similar historical World Wide Web studies (section 5.5.2).

5.4.1 Dataset Overview

The full cleaned dataset contained 7,482 records, with the features as described in sections 5.2 and 5.3. The dataset was broken down by individual websites into seven separate groups; Craigslist (N=304), Twitter (N=312), Youtube (N=922), Flickr (N=1,141), Digg (N=1,141), Amazon (N=1,420), and Yahoo (N=2,247). Some time-periods were more heavily represented than others. Table 3.1 provides a summary of the dataset for the seven websites.

¹⁷ The BeautifulSoup open-source library (<http://www.crummy.com/software/BeautifulSoup/>) was used to parse the DOM tree and extract HTML elements and contents. The built-in urllib library was used to process HTTP requests.

Table 5.3 – Summary of the dataset (with some features from stream 2 – see section 5.3)

		Basic Details					
		Available Dates	table tag	b tag	i tag	font tag	old tags*
Craigslist	<i>Min</i>	11/11/1998	4	1	0	4	0
	<i>Max</i>	20/06/2006	25	45	1	259	2
Twitter	<i>Min</i>	30/09/2006	0				
	<i>Max</i>	31/07/2008	5				
Youtube	<i>Min</i>	28/04/2005	0	0	0		0
	<i>Max</i>	22/08/2008	103	101	5		10
Flickr	<i>Min</i>	26/02/2004	2	0	0	0	
	<i>Max</i>	16/04/2010	5	11	0	1	
Digg	<i>Min</i>	09/12/2004					0
	<i>Max</i>	27/10/2009					1
Amazon	<i>Min</i>	12/12/1998	12	1	0	0	0
	<i>Max</i>	27/10/2009	65	71	39	86	7
Yahoo	<i>Min</i>	17/10/1996	1	5	0	1	1
	<i>Max</i>	24/03/2010	39	105	14	151	6

* Old tags are center, menu, layer, blink and marquee. Their usage is very strongly discouraged (e.g. Zeldman 2007).

Note: **u** and **big** tags were only used by yahoo (min = 0, max = 6; min = 0, max = 4 respectively). The **small** tag was only used by yahoo (min = 0, max = 47) and twitter (min = 0, max = 3).

From the table 5.3 it is evident that Yahoo, Craigslist, Amazon, and Flickr rely on depreciated html tags. This was somewhat unexpected and indeed illustrates that in some areas even leading websites still haven't modernised to more up-to-date W3C specifications. Very rarely their usage can be justified, such as in certain client-scripting scenarios or when dealing with compatibility issues, although this is now rare (Zeldman 2007).

The box-plot in figure 5.3 illustrates the distributions of downloaded pages over a range of dates, Yahoo is represented in the dataset for the longest time-period, over 13 years (17/10/1996 – 24/03/2010). This is followed by Amazon, Craigslist and other applications¹⁸.

¹⁸ It ought to be noted that most of the downloaded pages are distributed over a small period, since the WM web-crawler did not scan websites as frequently, earlier in its existence. The archive's web-crawling engine also tends to crawl the web irregularly in some time-periods.

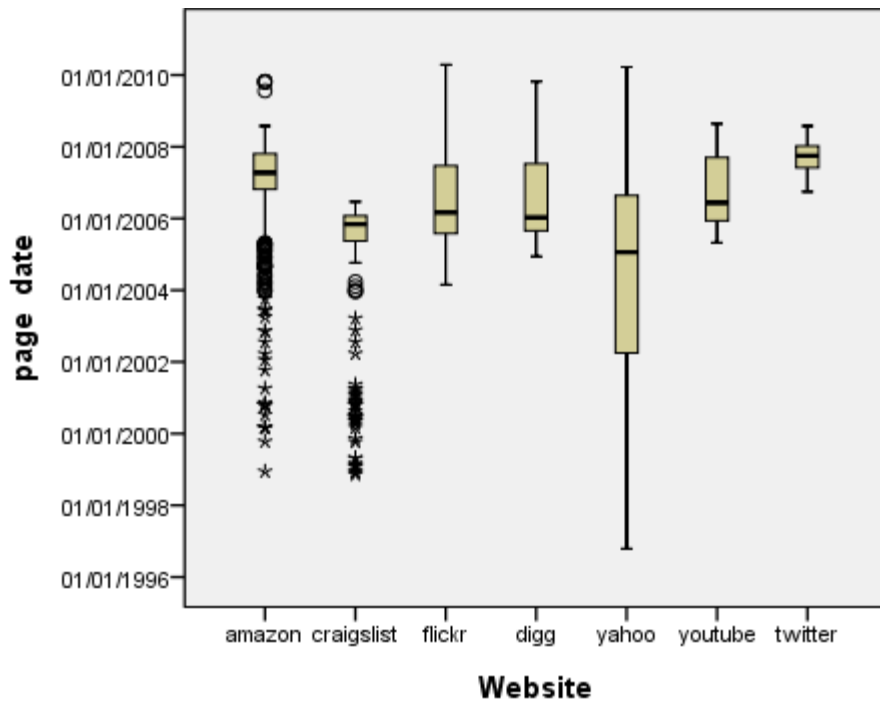


Figure 5.3 – Box-plots of downloaded page frequencies for available dates (all seven websites)

Table 5.4 shows that client side scripts are prevalent on all websites; however, external scripts are used more sparingly. Craigslist surprisingly never made use of external scripts in the sampled period, even though page level scripts were used¹⁹. In terms of advanced scripting, AJAX or well known client Javascript libraries were identified on all websites, with the exception of Yahoo, Amazon and Craigslist. CSS styles are prevalent on all websites, but yet again external Stylesheets are never used by Youtube.

The 4th feature stream contains features related to web 2.0 activities (such as like, share, rate, comment...)²⁰ and web 2.0 elements (such as Blog, RSS, Podcast...). It can be appreciated from table 5.5 that each website contained web 2.0 elements and activities at some point during the sampled period. Further investigation reveals that most elements emerged around 2005, and activities began emerging as early as 1999 and 2000 (see figures C.1 and C.2 in appendix C).

¹⁹ Usually external script files are desired and preferred over page level scripts, since the former speeds up performance thanks to file caching.

²⁰ An inherent problem with this feature in the experimental set-up is that features were extracted by matching keywords in the pages content text and alt attributes (i.e. alternative texts often used on graphics). This is not the safest method of detecting these features, since some keywords can and will appear by chance (section 5.3).

Table 5.4 – Summary of websites – features from stream 3 (see section 5.3)

		Basic Details					
		script tag	script source tag	script libs used	AJAX use	CSS Style	CSS Stylesheet
Craigslist	Min	1				0	0
	Max	3				1	1
Twitter	Min	4	1	0	0	0	0
	Max	9	6	2	3	3	2
Youtube	Min	1	0		0	0	
	Max	23	10		3	3	
Flickr	Min	1	0		0	0	0
	Max	5	1		3	1	2
Digg	Min	4	2	0	0	0	0
	Max	38	19	3	3	3	2
Amazon	Min	1	0			0	0
	Max	84	4			17	3
Yahoo	Min	1	0			0	1
	Max	30	2			4	3

Table 5.5 – Basic summary of Web 2.0 elements and Web 2.0 activities from stream 4 (see section 5.3)

		Websites						
		Craigslist	Twitter	Youtube	Flickr	Digg	Amazon	Yahoo
Web 2.0	Min	0	1	0	0	0	0	0
Elements	Max	2	3	4	2	5	4	2
Web 2.0	Min	0	0	1	0	2	1	0
Activities	Max	3	3	8	4	8	6	6

5.4.2 Analysis of Web-design Changes

Major design changes in respect to page layout and use of old, or depreciated tags, CSS, and the introduction of JavaScript, AJAX, and web 2.0 features, are discussed in detail. Although this study was largely automated through the use of programming scripts (section 5.3), some qualitative analysis, i.e. careful manual inspection, was also undertaken to help make sense out of the various observations throughout the archive. In a way this study is like a historical puzzle piece, and the following sub-sections will aid in putting this puzzle together.

Youtube, Digg and Amazon are discussed first, and results for all three share the same characteristic, which is an emphasis on web 2.0 elements. Yahoo is presented next, followed by Twitter, Flickr and Craigslist. The analysis of Twitter and Flickr was especially challenging, as

both applications were hidden behind a login wall for some time-periods.

5.4.2.1 Youtube

The historical analysis of Youtube is based on archive pages for a period of over three years (*April 2005 to August 2008*). The number of tables used for page layout throughout the three years has been constant until the *10th April 2008*, when the use of tables was scaled down from 15 to only two tables. Initially there was a table for each video (picture, title, description...), and this was reduced down to only one table used for login form input, and the other table for the footer menu. By *19th June 2008* table use was reduced down to only one table, and as of *1st July 2011* no tables are used. The site's layout is now entirely based on CSS. In fact throughout the analysed period Youtube landing page has used external Stylesheets. As for old and depreciated tags, from *April 2006* Youtube began to use b tags instead of inline CSS, to set characters bold. Since September 2006 their use has been decreasing from 101 (at its peak), yet even today Youtube still uses 12 b tags in their menu items – see figure C.5 in appendix C. This is a rather bad practice but it was not possible to discover the reasons behind such design. More predictably, Javascript became more intensely used in recent times, although it was used throughout the entire sample period, with AJAX found on the main Youtube landing page during *Oct. 2006 – Feb. 2008*²¹.

Several web 2.0 elements were identified; Youtube had an API, RSS-feed, and a Blog, and it was possible to identify the exact dates when these were introduced. The RSS feed was introduced on *17th June 2005*²², an API as early as *4th August 2005*²³, and the Blog followed on the *15th December 2005*²⁴. Web 2.0 activities were detected on at least 50% of the 923 web-page entries, and included the activities of sharing, commenting, submitting, rating, tagging and liking. One of Youtube's earliest features was commenting on videos, introduced as early as the *17th June 2005*²⁵. A further manual investigation of the earliest cases containing keywords relating to these activities revealed them to be part of the Youtube website, rather than accidental content within arbitrary Video descriptions.

21 As of *1st July 2011* AJAX was being used within a referenced JavaScript library, which is accessible through the main Youtube landing page at <http://s.ytimg.com/yt/jsbin/www-core-vflporvst.js>

22 <http://web.archive.org/web/20050617015149/http://www.youtube.com/> (bottom of the page)

23 <http://web.archive.org/web/20050804232629/http://www.youtube.com/> (bottom of the page)

24 <http://web.archive.org/web/20051215085406/http://www.youtube.com/> (bottom of the page)

25 Method: essentially we navigated to the first case / record where the feature relating to commenting was set to true, and then checked the actual webpage and the previous days (or more generally any nearest previous record) webpage to investigate the difference relating to commenting. Hence our methodology allows us to mix automatic analysis with a qualitative check to discover reliably (*constrained by WM archive, of course and some factors mentioned in section 5.1*) and effectively (*the process is largely automated*) content or feature based changes of historical webpage evolution.

In summary – The first video uploaded to Youtube was on 23rd April 2005, and since data from WM was available from 28th April 2005, the entire first and most interesting years of Youtube’s historical existence were analysed. It was possible to observe when exactly certain features such as commenting, rating or favourite lists were added to Youtube. Youtube has a heavy web 2.0 footprint, in terms of the number of activities such as uploading / posting videos, rating videos, liking videos into favourite lists and commenting. The entry points for UGC are indeed more varied than Twitter, for example. Youtube was also early to adopt RSS feeds and to provide an API for developers. Expected patterns were observed for features relating to CSS, Javascript (AJAX), and depreciated tags. CSS and Javascript were used from the outset and AJAX was added during *Oct. 2006*. Table based layouts gave way to a more standardised CSS approach.

5.4.2.2 Digg

From the outset Digg has been developed as a social news website. A core activity of it being the idea of average users vetting news-stories, by either voting them up or down, or in other words digging or burying stories, respectively. The interesting thing about Digg is that as a site it has adopted a very clean and standards based design, i.e. no tables and depreciated tags were found throughout the available time period. The site was continuously improved with four major re-designs of user interface, on July 2005²⁶, June 2006²⁷, November 2006²⁸, and June 2007²⁹. The figure C.7 in appendix C highlights the increasing dependence of Digg on client side scripting, with AJAX introduced to the site in January 2005.

As opposed to Flickr and Twitter for example, most of Digg application functionality can be accessed on the main landing page, i.e. reading stories, voting for stories, etc. A manual investigation of changes highlighted by substantial and consistent changes in page features was used to point out when web 2.0 elements might have been introduced. Figure C.8 in appendix C highlights the web 2.0 elements and activities on the Digg landing page as detected through the automated content analysis over time. It was found for example that January 2005 commenting on links was introduced, or the “blog this” feature (allowing to directly post articles on Blogs) was added in April 2005, in August 2005 Digg Podcasts, and on April 2007 an extensive API platform were introduced to Digg.

In summary – It was found that Digg shows a great rate of adoption of client-side scripts, and at the same time CSS is relied upon for the entire surveyed time-period, with virtually no use of

26 <http://web.archive.org/web/20050711082245/http://www.digg.com/>

27 <http://web.archive.org/web/20060628060022/http://www.digg.com/>

28 <http://web.archive.org/web/20061109004131/http://digg.com/>

29 <http://web.archive.org/web/20070606083729/http://digg.com/>

outdated or non-standards HTML. Since Digg relies upon user-participation web 2.0 based elements that facilitate and help trigger social interaction were introduced relatively early to Digg.

5.4.2.3 Amazon

Amazon differs from Yahoo (discussed next) in that it is widely regarded as an innovator in web being used as an interactive media (O'Reilly 2006), and was in fact an early adopter of some UGC facilitating features, such as rating and commenting within its product pages (Spector 2002).

The website layout on 13th October 1999 (earliest available entry from the WM³⁰) is predictably based on a simple set of tables with some *bgcolor* attributes. During this time it also seems that image-maps³¹ were rather popular (*Yahoo was also using one in the 90s*). Amazon was using an image map to give the impression of rounded corners in its main menu, as achieving the same effect back then was complicated using other means. As early as 2002³², Amazon was already using basic elements of CSS and some Javascript in order to detect a wide set of browsers and execute browser specific code where necessary. Figure C.6 in appendix C highlights the decreasing trend of now depreciated tags. The most recent Amazon page now uses an AJAX based element on the main page – however AJAX hasn't been detected in earlier pages, and must have been introduced recently. In 2001 product wishlists which are shareable between customers were introduced³³. In August 2006 Amazon podcasts became available, and an API for developers to build on top of the Amazon platform was provided by Amazon at the time. As of 2007 a corporate Amazon Blog went online. Around March 2008 a tag cloud on the main page³⁴ was introduced for a short time period, and was subsequently removed. A similar tag cloud, for navigation, was also observed on Youtube.

In summary – Amazon has shown a pattern similar to other, already discussed websites. CSS or Javascript usage emerges and increases over the years, as the use of older tags becomes less relied upon. As for the collection of UGC, Amazon it seems has indeed embraced social and collaborative elements of the web, in a web 2.0 exemplar way. Many keywords relating to sharing / openness and web 2.0 specific elements such as Blogs, Podcasts and Tag Clouds were detected. Quite early on Amazon was experimenting with how to engage its wide user-base into

30 <http://web.archive.org/web/19991013091817/http://amazon.com/>

31 A crude technique used to split a picture into different clickable areas – i.e. <map><area> tags.

32 <http://web.archive.org/web/20020123011349/http://www.amazon.com/exec/obidos/subst/home/home.html>

33 <http://web.archive.org/web/20010515030108/http://www.amazon.com/exec/obidos/subst/home/home.html>

34 <http://web.archive.org/web/20080306222747/http://www.amazon.com/>

online social networks of shoppers that might help make online shopping a more social activity. During 2005 the “Friends & Favorites”³⁵ social network was introduced within Amazon and users were able to build profile pages and connect in a number of interesting ways – “a facebook for shoppers”, see figure 5.4.

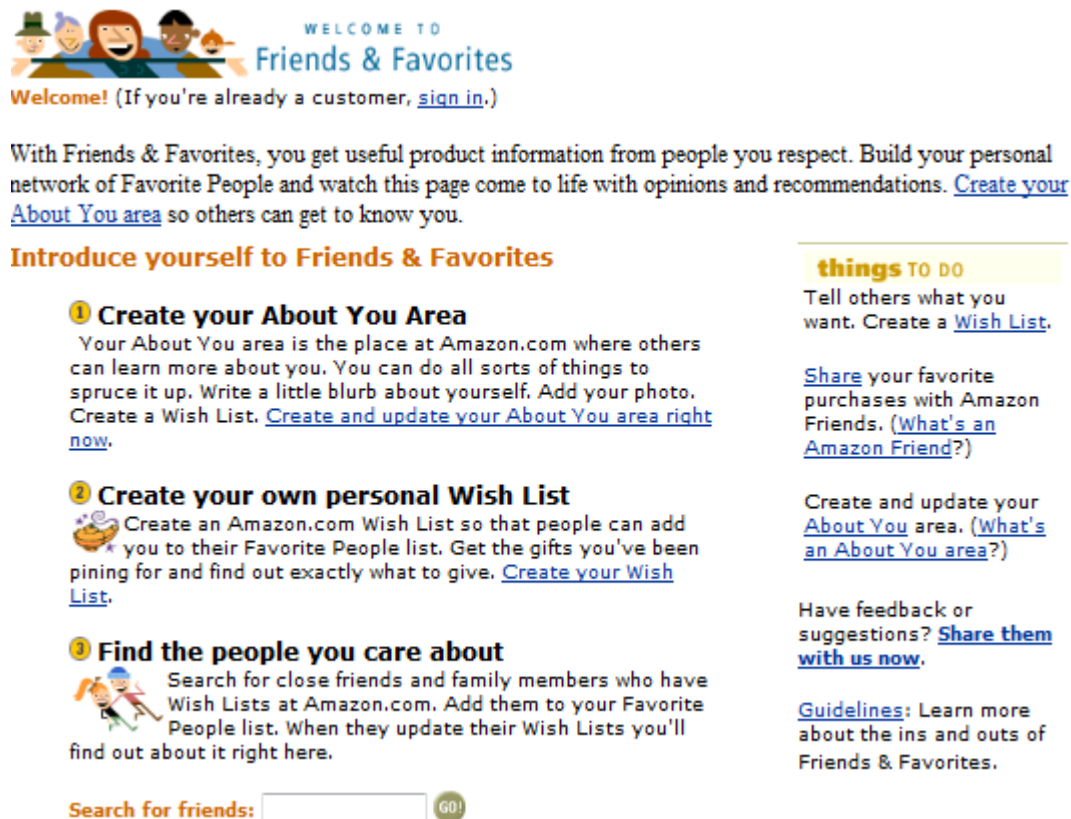


Figure 5.4 – The social network “Friends & Favorites”, built on Amazon (during 2005)

5.4.2.4 Yahoo

The range of dates represented by the historical sample for Yahoo is the widest, and covers a period of over 10 years. The data facilitates an analysis of the late 90s and the pre web 2.0 era. Yahoo is generally perceived to be an administered web-portal³⁶, rather than a social media or web 2.0 website, and hence it was expected that a web 2.0 *footprint* as measured over time would be relatively small, possibly but not necessarily increasing somewhat in more recent years, as well as some decrease in depreciated HTML elements was expected as Yahoo strived for HTML standardisation. Unfortunately after an extensive manual inspection, 591 page records from the period between *August 2006* and *February 2010* had to be excluded from

³⁵ <http://web.archive.org/web/20050601150539/http://amazon.com/exec/obidos/subst/community/community-home.html>

³⁶ Yahoo is best known for its web portal, from which one can access their search engine, Yahoo directory, Yahoo Mail, Yahoo News, Yahoo Maps, Yahoo Videos, etc... Yahoo also operates Flickr (a photo sharing application), however this website is analysed separately. The focus of interest is the design of Yahoo’s main page.

analysis, due to an issue with WM, further details in section 5.5.1.2. There is hence no page representation for years 2008 and 2009.

During the early years of Yahoo, in the months of October, November and December 1996 Yahoo wasn't using any table tags and page design was extremely basic³⁷. These were still very much the early days of the web (*IE 3 and Netscape Navigator 3 both came out in August 1996*), with simple browsers not facilitating much in the way of more complex designs³⁸. The first major redesign discovered by our analysis took place in *January 1997* with Yahoo introducing tables to lay out the still very basic HTML content into a more compact visual display. Two years later (*February 1999*) Yahoo introduced the use of *bgcolor* attributes and table based layout now received a more colourful design. During *March 2001*, basic in-page cookie checking scripts were introduced. It wasn't until *July 2002*, and *November 2004* when Yahoo's visual design was polished considerably; however, the design was still heavily based on tables and with entirely no use of CSS. Only towards the end of 2005 CSS was introduced. As of *February 2010*³⁹ Yahoo layout was finally based on CSS, with old tags now completely removed, and JavaScript was used to some degree⁴⁰. Web 2.0 elements, other than arbitrary mentions in the content of articles, haven't been detected throughout the content analysis. Only, on the most recent page (*1st July 2011*) a corporate Blog is finally available, and there also doesn't seem to be much in terms of keywords pointing to possible web 2.0 activities on the main page.

In summary – It is well known that Yahoo provides extensive abilities to contribute UGC within its child projects such as Flickr, Delicious, Yahoo! Answers, etc. and supports a strong API. However, the main landing page itself has kept most of its classical portal image, with practically no web 2.0 style content sharing. Yahoo, provides an API to much of its data, and has opened up considerably as a platform, yet little of this is observable on Yahoo's main landing page, with no direct encouragement for contributing UGC. Yahoo was also slow to adopt a CSS based design and table based layouts seemed to have been preferred by Yahoo designers. During the year 1996, design was extremely simple as there were numerous compatibility and rendering issues associated with many HTML elements, and CSS was mostly

37 The HTML code of the page contained only some , <a>, <p>, <hr>, , , , <center> and tags. A page example from this period is available for inspection under this url: <http://web.archive.org/web/19961022175643/http://www10.yahoo.com/>

38 One of the main reasons why Yahoo wouldn't be using tables in the early days was that 1-pages took longer to load with the overhead code and 2- some browser versions weren't able to render more involved table layouts.

39 Unfortunately the period between these years is missing, due to the problem described in section 5.5.2.2

40 Yahoo uses their YUI custom open-source Javascript library <http://developer.yahoo.com/yui/>

an unsupported standard⁴¹.

5.4.2.5 Twitter

A look at the 21 months of data for twitter, from September 2006 to July 2008 reveals a number of redesigns to the main Twitter page. As expected Twitter has phased out tables and never looked back. Between 28th and 31st October 2007 Twitter removed tables completely⁴² from its page design. Before then tables were used to lay-out the content on the main page, where table layouts themselves were slightly re-designed several times over the sample period⁴³. The removal of tables during 31st October 2007 was accompanied with other changes in which old and depreciated HTML tags were removed from the pages' design and as for visual changes, Twitter updates were now fully hidden behind a login wall. There were also some changes to the number of referenced CSS files, as well as CSS related tags, but some form of CSS was used throughout the entire sampled period, also during the earlier periods of table based layout (see figure C.3 in appendix C). AJAX was detected to be present⁴⁴ on the main Twitter page, within the earliest available entry; however, after 4 months (from 2nd February 2007) AJAX was moved away from the main landing page, most likely into a deeper level page which was out of scope for the automated scan. AJAX was later re-introduced via the JavaScript libraries; however, it wasn't possible to retrieve this information, since Twitter has changed its robots.txt rules to stop any crawlers from requesting their .js files, hence the WM wasn't archiving Twitter anymore. Nevertheless, client side scripts were prevalent throughout the full sampled period, see figure C.4, in appendix C. As for the content analysis relating to web 2.0; over three times more web 2.0 elements than web 2.0 activities (636 vs. 168 counts) were found within all the historical Twitter entries, and out of the 312 historical entries, 310 referred to an API⁴⁵, and all 312 to a Blog. There were a number of other web 2.0 elements and activities; however, after a more extensive manual investigation these were attributed to the content, which was clearly unrelated to Twitter's own functionality or features⁴⁶.

41 See <http://www.w3.org/> for much more on the history of standards and specific information on CSS and (X)HTML

42 This table-free version can be accessed on (note: the CSS external spreadsheet file does not always seem to load within Wayback Machine properly) <http://web.archive.org/web/20071031185621/http://twitter.com/>

43 Dates of some major re-designs for twitter and the relevant WM Archive links, include – September 2006 (<http://web.archive.org/web/20060930214639/http://twitter.com/>), November 2006 (<http://web.archive.org/web/20061109073453/http://twitter.com/>), January 2007 (<http://web.archive.org/web/20070129052251/http://twitter.com/>), March 2007 (<http://web.archive.org/web/20070304191014/http://twitter.com/>)

44 AJAX was detected within the prototype JavaScript library (<http://www.prototypejs.org/>) used on Twitter.

45 Twitter had an API available relatively early. It was introduced throughout October / November 2006.

46 For reference these other features were: RSS, Podcast and Wiki, in four, nine and one page entries, respectively. Web 2.0 activities, Share(2), Comment(10), Submit(4), Rate(9), Tag(10), Like(124) and Edit(9).

In summary - throughout the surveyed, early period of its existence (*between September 2006 to July 2008*), it is evident that Twitter displayed a number of elements usually associated with web 2.0. API⁴⁷ and a Blog feature were found to exist on Twitter's main landing page throughout the entire sampled period. Client side scripting, CSS use, and more importantly early use of AJAX with a significant redesign towards the end of 2007 were observed. In this redesign table use and many old html tags were phased out, which highlights some degree of evolution towards standards. Unfortunately a sample over a larger time-period, following July 2008 is not available through the archive. Larger design changes tend not to go unnoticed though and with some effort can be tracked down in the blogosphere, on forums, and other web resources. For example the major overhaul of Twitter during late September 2010 is well documented in the blogosphere, where various blog posts are found to describe the changes in detail⁴⁸.

5.4.2.6 Flickr

An analysis of Flickr turns out to be problematic simply because the website is hidden behind the main landing page. Some indication of the website can be gained from a landing page; however, is limited, since Flickr requires one to register to use the site's features and browse content. Therefore a systematic sampling approach was taken by selecting one URL from each month (usually if the date showed some features change) and inspected by a human subject. This allowed to find for example that during *April 2004*, Flickr allowed to tag or comment on pictures, create / join groups in the Flickr community, easily publish Flickr pictures on Blogs and even a Flash based chat application was available (presumably to facilitate discussion about the Flickr pictures)⁴⁹, and RSS was supported. From *July to August 2004*, *February 2006*, and on *12th June 2007*⁵⁰ the landing page was redesigned. It is interesting that tables were used throughout the landing page until *April 2010*, it is only now (*1st July 2011*) that design has been entirely migrated to a CSS based layout.

47 The API shows Twitters intention to offer its website as an online application platform via the API, from early on.

48 See <http://twitter.com/newtwitter> for the official announcement of the re-design. Some blog-posts describing the changes include <http://www.stevesouders.com/blog/2010/09/22/newtwitter-performance-analysis/>, <http://techcrunch.com/2010/09/14/new-twitter-tips/>, <http://mashable.com/2010/09/14/new-twitter-web-interface/> with much discussion in the comments sections. Blogpulse.com only allows to search past 6 months, yet interestingly a search via this service reveals a huge surge of blog posts around the 5th April 2011 (<http://blogpulse.com/trend?query1=newtwitter&label1=&query2=&label2=&query3=&label3=&days=180&x=16&y=16>) and conversation tracker shows more detailed information (http://blogpulse.com/conversation?query=newtwitter&link=&max_results=25&start_date=20110401&Submit.x=22&Submit.y=14), when a major technical glitch occurred on Twitter.

49 <http://web.archive.org/web/20040401182410/http://flickr.com/tour.gne>

50 It seems that the Flickr blog was introduced on the 12th June 2007 with the re-design, as well.

In summary – Flickr had much to offer in terms of social engagement, from its earliest version in 2004. Users were able to organise themselves into sub-communities, share pictures, comments and opinions. Organise pictures using tags, swap and discuss over live (flash based) chat⁵¹. Unfortunately a large degree of manual investigation of the archived pages was necessary since the Flickr website is mostly hidden behind a login / register landing page. Although the design isn't necessarily the cleanest and despite the fact our conclusions are limited due to limits of access to pages, it seems Flickr has mostly concentrated its efforts at allowing high levels of social engagement from its beginnings.

5.4.2.7 Craigslist

Craigslist is an inherently collaborative application which started off as a mailing list and has evolved from a relatively centrally managed system, with Craig Newman submitting a lot of the content, to a fully user generated, and well categorised list of wanted and for-sale items marketplace. During the end of 1998 and until November 1999 the landing page for craigslist was kept more or less unchanged. A minor re-design in November 1999 was followed by a substantial one during March of 2000, in which HTML tables were introduced to play a larger role in page layout. A seemingly cleaner looking interface was introduced during June 2000⁵², which is when some very basic JavaScript was introduced for the first time and new ads categories were also introduced⁵³. During July 2000 a discussion forum was added and later that year in October, community based vetting of ads was introduced through the “flag for review” link-buttons. In January 2001, a wishlist feature⁵⁴ was introduced, and the Craigslist Blog in November 2003. By June 2004 Craigslist offered localised ads for the UK (London) and Canada (Montreal, Toronto and Vancouver), in addition to the US. The visual design of Craigslist; however, has not changed much since 2004. Today the design looks surprisingly

51 Flickr supports many more interesting features. For example *sets*, mutually non-exclusive collection of pictures, and sets of sets, which can all be geo-tagged, or in-picture annotations. The following sources provide more information on some of these Flickr features: <http://blog.flickr.net/en/2009/03/04/setting-sets-free/>, <http://www.dopiaza.org/flickr/setmgr/v2/index.php>, <http://www.gossinteractive.com/blog/flickr-creating-sets>, <http://web.archive.org/web/20090923081400/http://www.flickr.com/photos/junku-newcleus/417646359/>.

52 One can compare these interfaces by going to: (newer) <http://web.archive.org/web/20000612232944/http://craigslist.com/>, (older) <http://web.archive.org/web/20000303014907/http://www.craigslist.com/>

53 While investigating web-design related features on Craigslist we noticed that the ads categories evolved over time, sometime in quite interesting ways. For example in year 1998/99 there were mostly practical job and house related ads (<http://web.archive.org/web/20000303014907/http://www.craigslist.com/>, <http://web.archive.org/web/20000303183807/http://www.craigslist.com/category.help.html>), which of course has changed considerably over time, with the notorious personals ads and the many international / localised categories introduced (e.g. see <http://web.archive.org/web/20000612232944/http://craigslist.com/>). One would imagine that a historical study of Craigslist could be interesting for some social sciences research. Unfortunately the scope of this thesis does not allow us to explore this idea further.

54 Specifically for schools and non-profits, but essentially the idea was to involve the entire community.

similar, even tables are still used to layout content, although old and depreciated tags were entirely removed from the page-design.

In summary – Craigslist grew organically from a mailing list, into a website, where user contributions played a central role from the beginning. Users contributed mainly in the form of ads and ad vetting. In many ways; however, Craigslist simply resembles an advanced, moderated forum, where participation is driven by the goal of exchanging goods. There isn't an API that would open up Craigslist as an application platform either. It would also seem that much of Craigslist functionality could be accomplished via a more advanced forum design. An extremely simple table-based visual design was employed. JavaScript was introduced quite early to the page, but the pages on Craigslist have never evolved much in complexity. Eventually the usage of depreciated tags has disappeared completely over time.

5.4.2.8 Title Variations over Time (all pages)

The title of various pages (html header / title tag) was found not to change much over time, except to communicate a different perception about a site to its users. Yahoo for example hasn't changed its title from "Yahoo!" one single time during 1996-2010, and as of July 6th 2011 Yahoo still uses the same title. This may be attributed to the strong corporate identity of a web-portal business that Yahoo has had over the years.

As was already mentioned, Youtube had its title slogan changed from "*YouTube – Your Digital Video Repository*" into "*YouTube – Broadcast Yourself*", which represents the symbolic shift in the sites purpose to a more direct "*Broadcast Yourself*". Essentially Youtube desired to present itself as a social application where videos are not only stored but where videos, profiles and experiences are shared⁵⁵. Interestingly this was also mentioned by Burgess and Green (2009), so we are not first to observe and comment on this. Launched in February 2004, Flickr used the simple title "*Flickr.com*". Then throughout March to May 2004, this was changed to "*Welcome to flickr.com!*", however since June 2004 until this day Flickr has been using "*Welcome to Flickr - Photo Sharing*", as to emphasise the goal of sharing on Flickr. Twitter was using the slogan "*Twitter: A Whole World in Your Hands*", during Sept 2006 to 30th December 2006, when it was replaced with the more well known "*Twitter: What are you doing?*" to encourage people to tweet about daily common-place activities. This was used at least until 31st July 2008, with an exception of a week during August for which the title was changed to simply "*Twitter*", a more recent check, as of July 6th 2011, "*Twitter*" is being used as the title page. Amazon, just

⁵⁵ Youtube has many social elements such as profile pages reminiscent of MySpace, an internal messaging system, commenting, rating, video-replies, watch-lists to share and other elements.

as Yahoo, was quite consistent in their title slogan. During the period of late 1998 to mid-August 2006 “*Amazon.com--Earth's Biggest Selection*” and “*Amazon.com: Online shopping for electronics apparel music books DVDs & more*” were interchangeably used. Amazon has emphasised its position as an online shop with the widest selection of products from the beginning (Spector 2002⁵⁶). Since August 2006 till today Amazon uses “*Amazon.com: Online Shopping for Electronics Apparel Computers Books DVDs & more*”; however, on some of its sub-pages the slogan “*Earth's Biggest Selection*” can still be found. From its launch Digg has been simply using “*Digg*” as its page title, but this later changed to “*Digg / News*” (Dec. 2006), “*Digg / All News & Videos*” (Aug. 2007), “*Digg / All News Videos & Images*” (Dec. 2007) and finally to “*Digg – the Latest News Headlines Videos and Images*”. There is no indication of Digg trying to reposition its public image over time, other than increasingly stressing the multimedia content to be found on their website. Since the autumn of 1998, Craigslist was in the process of rebranding itself into ListFoundation, the title of the page reflects this as well (Nov. and Dec. 1998), however due to legal reasons the name could not be adopted and the site had to return to using Craigslist as its business name. Since then “*craigslist: San Francisco bay area online community*” and other localised variations have appeared on Craigslist’s website.

Given the wide temporal coverage of the datasets, it can be clearly observed how these page titles have changed over time. This discussion has been provided for the purpose of completeness, to enable information from our dataset to be used by other researchers, if necessary.

5.5 Discussion

Clearly there was an overall trend within all the investigated websites, to re-design old HTML based code. Old tags were phased out in every case, although some websites seemed to have relatively clean designs from the outset, e.g. Digg. In other cases, such as Yahoo, rather bad and old design elements prevailed for longer than expected. As for web 2.0, Tim Berners Lee did suggest that there is no fundamental change in the technology (see section 2.2); however, this concerns the W3C introduced standards rather than actual adoption of standards by browsers and websites. It wasn’t until 2006 that some of the main browsers became compliant with a substantial set of CSS and HTML W3C standards (Zeldman 2007, Murphy and Persson 2008, Budd et al. 2009). The problems associated with earlier incompatibilities and inabilities of

⁵⁶ Amazon first used the slogan “*Earth's Biggest Bookstore*” which was eventually changed to “*Earth's Biggest Selection*”. The mentioned book provides further background details to the history of the shopping website.

browsers to render much content simply meant that user-interfaces and rendering across different browsers was buggy and inconsistent (Zeldman 2007). Hence Tim Berners Lee's argument when put into context transforms into the argument of the issue of technical feasibility of standards into real world implementations, which did not occur more extensively, until the mid 2000s. This is much more important than it would seem. People didn't take the web seriously at this time and trust was difficult to build, if correct page rendering would be inconsistent. Since full CSS and HTML standards were often not supported, websites had to usually compromise on the web-application design. Ability to manipulate the HTML DOM-tree from client side scripts with proper CSS support, and to integrate this tightly with server side logic via AJAX would have been a major development; however, unlikely to be used if the browsers did not support such implementations consistently.

It was found that interactivity was increasingly being added (in terms of JavaScript). Although Javascript adoption was probably understated, since it was measured on the main landing page, and deeper level pages, where most functionality for sharing and collecting user contributions is located would have been ignored. AJAX was not found over all websites, it was detected on five websites (*as of 1st July 2011*). Maybe AJAX hasn't been as significant for web 2.0 as some might suggest.

Finally it was found that sites have generally opened up during the later 2000s, with incorporating APIs, RSS feeds or opening up corporate Blogs. An API fulfills an important role for a web-application as it allows for the dataset and application logic to be extended by third parties, and effectively allows for more complex and potentially useful applications to be built. Since data on web 2.0 websites is generally user contributed, the access and sharing of the data via APIs is to some extent justified. Not every website however has been found to provide an API, i.e. Craigslist. RSS feeds also open up the website as a platform, since Mashups can be built and information consumed in a number of ways. Blogs are a more symbolic way of opening up communication and allowing for a more informal communication channel. It was found that all of the websites discussed, introduced Blogs at some point, most have done so during the mid 2000s.

5.5.1 Criticism: Breadth of the Study

Despite having been carefully chosen (reasons were given in table 5.2), a major criticism of work within this chapter is that only seven applications were analysed. Out of these, two were behind a login wall for most of the time-period, which reduced results even more. Although the results of this study may be interesting and insightful, the coverage of applications relative to

the universe of existing web 2.0 systems is rather limited. In order to address this criticism of the WM study being too narrow in focus, and partly by a need for baseline features characteristic of web 2.0 applications; a second study was performed, in which a much wider set of web 2.0 applications was analysed⁵⁷. To keep this study simple, the same features extracted from websites were used (as described in 5.2-5.3).

5.5.1.1 Wide Dataset Extraction

Two directories of web 2.0 applications were considered. **1** – In Lindmark (2009) a list of web 2.0 sources is suggested, it is this list, mentioned in Lindmark, which provided the first dataset of 99 top web 2.0 websites, as ranked by the Alexa Inc. traffic ranking (<http://movers20.esnips.com/>). In the top 99 list nearly a third of the URL-resources could not be accessed, hence all in all, 59 unique website landing pages were retrieved from this list. **2** – The other dataset of web 2.0 applications came from an older, larger and seemingly well organised directory of web 2.0 websites (<http://www.web20searchengine.com/web20/web-2.0-list.htm>). At first this index seemed usable, however after removal of duplicate and dead links this source provided only 633 supposedly “web 2.0” websites from an initial 1021 URLs. Further inspection showed that the list contained many spam sites and similar⁵⁸. The main issue being that web 2.0 websites were contributed by users to this list and were clearly not vetted strictly.

Due to the low quality of this list, the 59 top ranked web 2.0 sites (by traffic) were analysed, instead.

5.5.1.2 Dataset Evaluation and Conclusions

HTML pages of all items in the list were downloaded on the 5th January 2011. It was found that 64% (38 out of the 59) websites did not use tables at all, and 49% did not use old depreciated HTML tags either (*such as b, i, u, marquee...*). This is compatible with the historical study. Also, only on 2 websites it was found that no JavaScript, and on 1 website that no CSS were being used. Hence, interactivity and standards based layout is completely prevalent within the web 2.0 applications space. It was further found that well known Javascript libraries were referenced within 36 websites (61%), and as many as 40 websites (68%) contained or

⁵⁷ Performing a historical analysis on many more applications would be highly demanding, as over 350MB of data had to be retrieved and analysed only for seven websites. Also many websites may not be tracked within the WM archives.

⁵⁸ Sites that are mostly irrelevant to the topic, or claim functionality, but provide very primitive capabilities. Often such sites are loaded with Google keyword ads, and are largely unusable.

referenced code with AJAX. Nearly 82% of surveyed pages seemed to contain Blogs, followed by pages containing APIs (37%), RSS (29%), Wiki (19%), Mashups (3%), and one page referred to Podcasts. The following web 2.0 activities were mentioned across the given websites; tag (72%), like (64%), rate (61%), submit (56%), edit (49%), share (46%), and comment (31%).

Evidently most websites use client-side Javascript and CSS, despite a relatively wide set of pages relying on outdated HTML, most interestingly, 36% of the websites still rely on tables. This could have been expected, and does not contradict the results from the historical study. It is interesting to see that Client-side Javascript libraries are so widely spread amongst large websites. This implies that well known, high-traffic, websites leverage the advantages and efficiency in application development gained through using libraries. In line with the historical study, most websites now have a Blog, and even though potentially not accurate, the content analysis points to some widespread UGC activities. Especially lighter or easier activities, in terms of time and effort needed for generating UGC, such as tagging, like-ing, or rating tend to be more widespread.

5.5.2 Reflections on the Methodology

5.5.2.1 General Considerations

In a more complete historical web study, historical entries from the Blogosphere could be searched and used as an explanatory tool in addition to the automatic WM analysis⁵⁹. A combination of both approaches would potentially provide more detailed insights, while being able to automate some elements of the study with the WM based approach. For example, a well formalised and systematic content-analysis of related Blog-posts in conjunction with an automated (or indeed also a qualitative content analysis) of resources from the WM could provide a more complete picture. In this particular case it was felt that such an extension to the study would not be appropriate or necessary, since a narrow set of features to be observed over time was well defined. Some online Blogs were used to double check major re-designs of sites. For studies of a similar nature to this, where more qualitative insight is necessary, a mixed methodology where the WM is analysed in conjunction with Blog posts from Blogosphere to provide corroborating evidence seems sensible for relatively recent events. An investigation of the earlier years of the web, early 2000s and late 90s one would have to restrict a study to the

⁵⁹ Unfortunately some sources, such as <http://www.blogpulse.com> limit blogosphere searches to six most recent months. However, for example Google's <http://blogsearch.google.com/> does not impose a limit on retrospective searches.

WM based methodology. Despite some restrictions and issues imposed by the WM archives, this data-source is largely unexplored and may provide for a number of potentially interesting studies. Others have also argued for the potential benefits from using WM archive for social science research, specifically Arms et al. (2006) from Cornell University. Their website provides more details <http://weblab.infosci.cornell.edu/>; however, the tools are only accessible to Cornell University researchers, although the WM now provides a browser styled historical surfing tool on <http://wayback.archive.org/web/>, which might be useful to the qualitative researcher. The lack of applied studies has meant that WM as a research resource has been largely overlooked.

5.5.2.2 WM Archive Considerations

It has been shown that the archives are relatively accurate, yet not much information on the system seems to exist in academic circles. More research is needed to better understand the capabilities of the archive, for example on the proportion of cached or still active external website links within the archive. Veronin (2002) found that only 24% of links (specifically links on medical / health topics) could still be found after an approximate three years of revisiting the same links. This clearly points to high link attrition on the web⁶⁰. Understanding how reliable and complete the archive tends to be within different areas would be useful for historical link and other analyses. The WM needs some further investigation in order to establish support for any such methodology, since without better understanding, it will be challenging to apply and automate WM studies within academic research.

In this study, it was found that the archive exhibited large amount of variance for covered periods and frequency of page archiving. Some sub-pages were found to contain corrupted, unusable data, and exclusions of entire websites from the archives were not uncommon. This usually happens since WM crawler respects the courtesy robots.txt exclusion standard⁶¹, or when website owners specifically ask for their websites to be removed from the archives. Availability of archived entries is sometimes discontinued from specific dates or parts of the websites made unavailable, such as JavaScript files due to the already mentioned exclusion issues⁶².

The WM crawler can sometime download and archive perfectly valid page server responses

60 At least within the medical domain, even though it is highly unlikely that this effect does not extrapolate to other parts of the web.

61 Robot Exclusion Standard is a convention to prevent web crawlers and robots from accessing all or part of an otherwise public website, see here <http://www.robotstxt.org/orig.html> for more details.

62 This was also an issue for some websites in this study, where .js files, linked to, from the main landing page were made unavailable.

that; however, turn out to be unusable for the purposes of a specific study. In particular output of a page from a website's web server can sometime vary depending on the geographical / time-zone location and user-agent identity of the requestor (i.e. WM crawler). This may generate biased responses that show (undesired) localised content or content specific to the user-agent's identity which is not compatible with the study. Such tailored responses can as for example in the case of our Yahoo analysis during the years 2007, 2008 and 2009 render responses unusable for historical evolution analysis. On 27th June 2006, Yahoo applied a redesign to their landing page which did not display correctly on older / incompatible user-agents. Unfortunately WM crawler was using an incompatible user-agent id at the time, which resulted in Yahoo replying to the crawler with a relatively (useless) scaled down version of the landing page. Figure 5.5 illustrates the unusable response that was archived for a period of over 3 years by the WM archive.

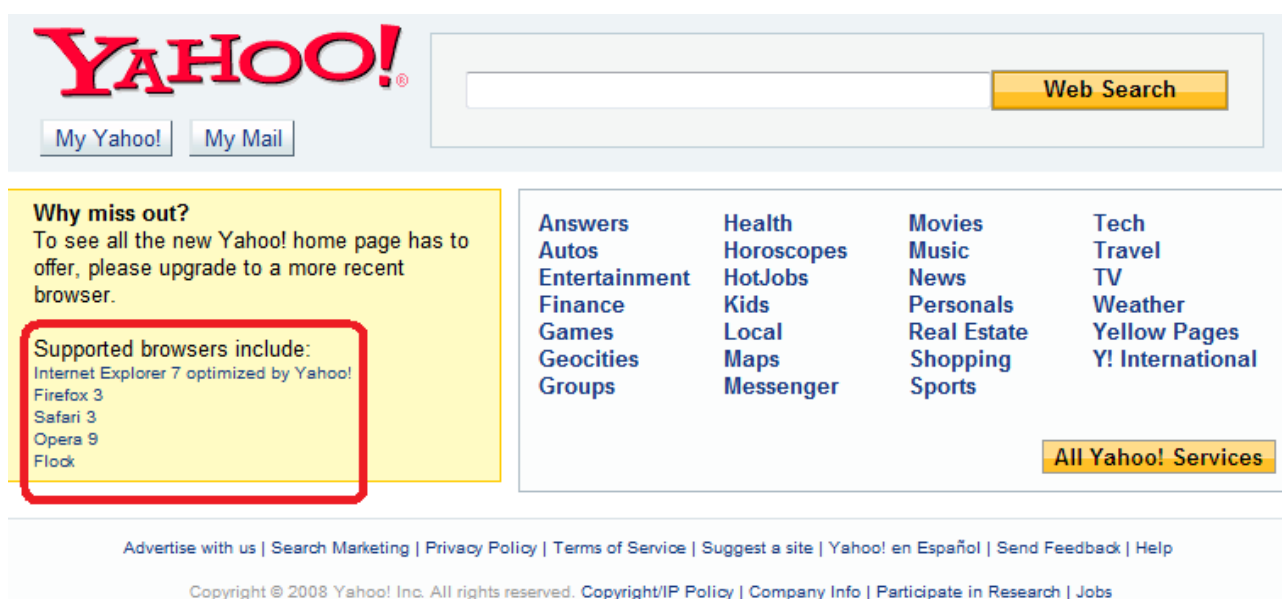


Figure 5.5 – Unusable Yahoo response as archived over 3 years by the WM⁶³

A paper by Howell (2006) mentions, in passing a number of other issues with the WM, as potential uses of the archives in the court of law are discussed. The consistency and accuracy of the archives are of utmost importance in the field of law, maybe even more than in many other areas.

The Wayback Machine provides an admirable data-source of historical webpages. It can be used in an automated or qualitative content analysis to track evolution of websites over time, and depending on the study, insightful patterns may be detected over a large set of websites and historical date / time ranges. The described approach; however, requires caution. Any automated analysis must be carefully checked for outliers and potentially unusable archive

63 For example, accessible under <http://web.archive.org/web/20091016215646/http://www.yahoo.com/>

pages must be dealt with appropriately. In an ideal scenario a study should be accompanied with an element of qualitative investigation, maybe using the Blogosphere or similar timely resources where relevant, to confirm validity of the data, and findings.

5.6 Summary

This chapter contains a historical study of web 2.0 applications, using a relatively novel methodology, to accomplish this. Given the recent hype over web 2.0, it is worthwhile to provide supporting evidence for the (mostly) hypothesised developments on the web, leading up to web 2.0, by means of an empirical, systematic and qualitative study. As far as the author is aware, the presented Wayback Machine study and results are a valid contribution to existing literature. Results from the study, support the web 2.0 phenomenon, in that tangible, i.e. measurable, changes occurred. It was shown that increasing standardisation and adoption of proper web design elements was a significant trend over the years, more so than AJAX. It was further observed that websites opened up their communication and datasets via Blogs and API platforms, which in turn helped built more trust with users, and allowed deeper integration with related websites and web services to effectively built more complex web applications, rather than just *stand-alone* websites. All of this happened gradually, and significant differences in the timings of such trends exist between the analysed pages. Since the retrospective sample only included seven carefully selected websites, a second study was performed on current websites, which analysed over 50, high traffic web 2.0 applications. Results from this broader study corroborate with results from the historical study. In addition to the study's outcomes, a set of guiding rules for leveraging the Wayback Machine archives for historical studies of the web, were presented. A number of issues with the methodology were critically evaluated. It is hoped that further research studies using the sizeable and freely available historical web archives will continue to emerge in the social and information sciences.

Considering that the last two chapters analysed the use and the historical evolution of web 2.0 applications, with the understanding that these chapters have provided, the next chapter will focus on the significance of web 2.0 applications and how web 2.0 applications may be leveraged in useful, practical applications; specifically within the field of finance.

6 Collective Intelligence Data Sources



Section 2.4 presented various practical applications of web 2.0 within medicine, libraries, education and other areas; the vast overall current range of applications and potential of web 2.0 were highlighted in chapter 2. Although in many vocational fields web 2.0 tools are still not used to their full capacity, or only in rudimentary ways. It was suggested that it may take some time for a more complete adoption of web 2.0, and grassroots efforts in individual fields are ongoing and necessary. However, the simple use of web 2.0 creates another level, or type of information that has far reaching implications, and presents an area of research in itself; this is the so called collective intelligence, generated by web 2.0 application use.

O'Reilly (2005) discussed collective intelligence (abbreviated to CI from now on) and its potential uses in applications, as a major element of web 2.0. Although CI should be seen as a consequence of web 2.0 use. It was also, already used in chapter 3, to identify neologisms. A simple definition of CI follows; “*Collective intelligence is a shared or group intelligence that emerges from the collaboration, competition, or sharing of many individuals and is essentially pattern based decision making based on collective knowledge, where collective knowledge can be effectively collected via web 2.0 systems.*” Chapter 8 will discuss CI in much more depth, for now the above definition will suffice. Arguably, real benefits of web 2.0 stem from the possible applications of web 2.0 facilitated CI. In order to understand how CI can be leveraged, a good appreciation of the relevant web 2.0 UGC data is necessary. Some encouraging and interesting

prior uses of CI exist and have been reported in literature.

Beginning with this chapter, and followed by chapters 7 and 8, this part of the thesis is dedicated to a treatise of CI use from web 2.0 applications. In this chapter various well known and public web 2.0 sites will be analysed and discussed, in terms of their CI potential. Given the presented examples in this chapter (and literature), chapter 8 will attempt to, then define a rough framework, based on communalities between web 2.0 applications on how to leverage and integrate web 2.0 for CI uses, effectively providing a systematic guide for its use.

It is the goal of this chapter to illustrate the variety of sources that are available, and to show how relatively simple techniques can generate useful insights from web 2.0 data-sources, and the sources themselves are discussed from a CI perspective. The primary focus of this chapter is the domain of financial markets and the case of the recent financial crisis. It was felt that this was an area of much interest due to its complexity, and recent events. There are other advantages of employing financial markets as a case-study. Market-indices can act as a useful *proxy* for efficient information transmission, which will be explained in sub-section 6.2.4.

A number of original contributions to existing academic work are presented within this chapter. First of all, some web 2.0 applications have not been investigated in the context of the recent financial crisis yet. Secondly the chapter is significant since it helps to build an understanding of UGC from the various data-sources. As will become apparent, there are also web 2.0 systems that seemingly do not relate to a given domain; however, it is shown that looking at UGC data should be carefully considered in these cases.

The chapter first presents a number of generic considerations, followed by a detailed analysis of the video sharing, web 2.0 application Youtube, in section 6.2. This section deals not only with an analysis of Youtube within financial markets, but also attempts to make a valid and highly important point about information transmission on web 2.0 systems. Delicious uses a comparable and relatively simple technique for extracting collective intelligence; however, a number of factors related to Delicious as a CI data-source are critically evaluated in section 6.3. In the chapter dealing with Amazon UGC it is illustrated how CI from a source that may seem unlikely to provide much benefit, can sometimes be leveraged in a relevant manner (section 6.4). In addition Amazon has not been investigated in this context before. The in-depth study of Youtube, Delicious, and Amazon are summarised and limitations are discussed at the end of each of the three sections. Due to its significance, Wikipedia is also introduced as a potential CI source in section 6.5, although an in-depth analysis was avoided. Finally, a number of relatively recent financial web 2.0 applications, specifically community finance and trend prediction websites are introduced in section 6.6, for completeness. The next section, 6.1, presents some of the existing literature in the assessment of CI using Blogs, and other applications.

6.1 Overview and Previous Literature

Before Blogs, social bookmarking, media sharing and other types of web 2.0 applications will be analysed within the context of finance in this chapter, first some relevant literature in CI within finance and other fields ought to be briefly introduced. Interesting and valuable work exists, usually raw UGC data has to be aggregated or pre-processed to facilitate CI. Adamic and Glance (2005) for example looked at Blog posts, one of the most common UGC. Posts of many individuals were analysed for URL links, and aggregated based on a-priori known political orientations of the Bloggers, and links between liberal and conservative blogs were investigated during the 2004 US presidential elections. In aggregate this revealed valuable information about political discourse and was compatible to prevailing political debates. There is strong evidence that political opinion is represented in online media sentiment (Mullen and Malouf 2006, Malouf and Mullen 2007, Johnson et al. 2007, Farrell and Drezner 2008), and being able to aggregate and pre-process the raw UGC to understand collective opinions seems useful. This could have number of applications, such as commentary and analysis of political opinion, i.e. better understanding, decision support, or even uses in forecasting (e.g. Tumasjan et al. 2010, used CI from microblogging to forecast German federal elections more accurately than pre-election polls). More generally, Mishne and de Rijke attempted to identify overall topic independent mood sentiments, represented in Blog posts, and classify them into mood categories such as tired, cheerful, happy, calm, angry, etc. (Mishne and de Rijke, 2006a/2006b; Balog et al., 2006; Mishne et al., 2007). This is possible via semantic text analysis and explicit mood tags. It has been shown that mood is intrinsically present in informal text posts and previous authors suggested that this information if extracted, may be of value. This chapter will not limit itself to a discussion of text processing for those purposes; however, extracting mood, sentiment, and understanding informal natural language text is a large problem domain with many difficult problems (see Pang and Lee 2008, Liu 2010).

Of most interest to us, is sentiment analysis and CI work within the financial-markets domain. Choudhury et al. (2008) investigated correlations of tech-companies on the Stockmarket with aggregated activity on Blogs. Using a Support Vector regression model¹, they found very encouraging associations with stock magnitude and price direction advances. As early as 1999, (Wysocki, 1999) and then Jones (2006) analysed the impact of posted discussion board messages on stock moves. They found that after online forums were introduced to World Wide Web, trading volume and volatility have significantly increased and daily absolute returns on

¹ Support Vector Machine (SVM) is a prediction model from the field of machine learning and data mining. Various prediction (also known as classification or regression) models will be mentioned briefly throughout this chapter; however only where relevant. A good introduction is the book by Han and Kamber (2006).

average decreased. This is an interesting observation as it highlights strong effect online information seems to have with the markets. Tumarkin and Whitelaw (2001) found correlations between abnormal activity on popular online forum Raging Bull (www.ragingbull.com), and abnormal share returns. Thomas and Sycara (2000) implemented a simple text processing (bag-of-words frequency)² GA based trading system, using the very same discussion board and reported successful trading performance. Antweiler and Frank (2004) applied text analysis techniques to capture the meaning of forum posts, and also found significant correlations with markets. Das et al. (2005) investigated message board posting and news correlations with stock returns, but at same time considered the disagreement between news and message board postings. Gloor et al. (2008) looked at correlations between Blogs and Stockmarkets, taking into account the social network structure of participants of the online discussion. Only recently first work appeared, which explored the linkages between micro-blogging (i.e. Twitter) and financial-markets (Bollen et al. 2011). Also several others (Fung et al., 2005; Clarkson et al., 2006; Sabherwal et al., 2008) have looked at whether features from free-form text can be extracted and correlated to financial market activity.

The literature overview above shows that most work in the financial context focused on online discussion-forums and only more recently the analysis of Blog and Twitter based UGC has emerged. There is strong indication from the literature that online sentiment and UGC does relate to Stockmarkets, and since web 2.0 applications have received little or no attention at all in the financial context, naturally more web 2.0 based UGC for CI purposes needs to be investigated. There are clear motivations from the financial domain as well, as this kind of research can be used to help monitor public interests, opinions or sentiment on a large set of different assets which may be relevant to particular investment or trading strategies, or techniques. Decision support and forecasting systems in the financial industry might benefit from the opportunities web 2.0 based CI could have to offer. This will be discussed in several sections of this chapter. In fact at the time of writing this *thesis* it has become public knowledge that the first hedge-fund has began to use micro-blogging aggregated features within their algorithmic trading models³.

Overall it is expected that CI from web 2.0 based UGC, in this chapter will reveal some relationships with financial-markets, and at least help provide clarity and decision support with regards to the financial crisis.

² Bag-of-words is a vector based text representation model, often used in text-processing, see section 7.2.1.3

³ <http://www.theatlantic.com/business/archive/2011/05/the-worlds-first-twitter-based-hedge-fund-is-finally-open-for-business/239097/>, last accessed; 1st June 2011. The news was first reported on 17th May 2011.

6.2 Youtube

Previous published work, discussed in section 6.1 (above), was performed on Blogs, Twitter, news websites, or discussion boards; however, media sharing websites and other types of web 2.0 applications were largely ignored. This section investigates the large media sharing community, Youtube. A significant association between video meta-data and textual data using a content driven sentiment text mining approach is found, and it is shown that efficient information transfer on online media sharing communities exists (Sykora and Panek 2009, Sykora 2009). This latter finding adds further support to the practical use of CI, to which chapter 8 is dedicated.

6.2.1 Youtube Background Information

Youtube.com was established in February 2005, as an *online video and the premier destination (...) to watch and share original videos worldwide through a Web experience*⁴. It is a free community-driven website through which registered users can upload unlimited number of videos and share them with other users. Each video must be given a title and be assigned to a specified category (e.g. News, Music). A publisher can optionally provide further details. Essentially everybody with an internet connection, whether with or without a camera, can contribute to Youtube, hence the content is dynamic with high update frequencies. There is; however, the issue of a higher barrier of entry. In order to contribute to Youtube, submitting videos generally requires the effort of recording, editing, formatting and uploading, which all represent certain costs in time. According to alexa.com, web traffic has been constantly growing since its founding, earning Youtube a ranking in top three most frequently visited websites in the world. It reaches about 5% of internet users in a day and generates 20% of all http based pageviews on Internet. These figures make Youtube most popular community based website.

6.2.1.1 Youtube in the Literature

Havley and Keane (2007b) in “Exploring Social Dynamics in Online Media Sharing” found that Youtube users prefer browsing rather than uploading their own videos. On average, they view 966 clips against 11 submitted files. Moreover, community facilities available after signing-in are not widely exploited. Most users are anonymous and do not participate in various

4 According to Youtube http://www.youtube.com/t/about?hl=en_GB (consulted on 2nd April 2009)

web 2.0 activities like commenting, video responding or rating. Furthermore, only small percentage of users has subscriptions for favourite clips or channels. In the paper “I tube, You Tube, Everyone Tubes: Analyzing the World’s Largest User Generated Content Video System” (Cha et al. 2007), it was found that Youtube on average experienced 65,000 new video submissions per day. Similarly to Havley and Keane (2007b), this research concludes that there is very little web 2.0 activity, most users stay anonymous and even if they register, the level of their participation is low. Authors also draw attention to a very important aspect, i.e. content aliasing. As anyone can upload nearly any content on Youtube, there exist multiple copies of videos relating to a single event. Hence there are many identical videos submitted by different users and this probably dilutes popularity of a corresponding video. “Analysis of Online Video Search and Sharing” written also by Havley and Keane (2007a), reveals that in general all of the videos that receive greater number of hits, have more descriptive meta information. It means that more textual information in the form of tags, title, and description makes these pages more popular than others. This is probably a consequence of the fact that the internal search algorithm picks these videos up with more likelihood and matches them with search terms. Freeman and Chapman (2007) investigated whether Youtube videos promote smoking. It was found that search term “smoking” had returned 29,325 results on Youtube. After in depth analysis of the content of top 50 clips, it turned out that Youtube is used as a channel for advertising tobacco.

It is clear that with 65,000 new video submission every day (Cha et. al 2007), and as will be shown in section 6.2.3, still rising, although Halvey and Keane (2007b), Cha et al. (2007) point out low web 2.0 activity in relation to the entire user base, Youtube is clearly a powerful web 2.0 phenomenon. Freeman and Chapman (2007) present an interesting aspect of one of many ways in which advertisers make use of Youtube.

6.2.1.2 User Generated Content, or what to get?

We are interested in as much Youtube video related data as possible. Every uploaded video on Youtube is in the form of a video file and a set of related meta-data describing the file. Such meta-data contains video title, description, category, date of submission, view count, duration and author. Since Youtube is a social website it also allows users to comment, rate (1 out of 5) and submit response videos. Videos can also be tagged with arbitrary tags that might help identify a video better. The meta-data attached to a file is shown in figure 6.1.

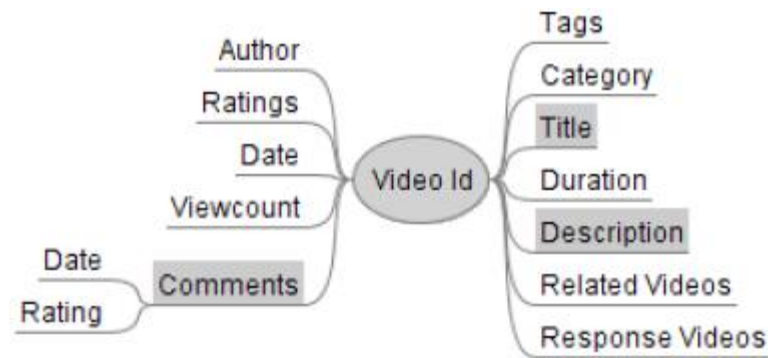


Figure 6.1 – Video meta-data associated with a Youtube file. Highlighted fields represent the three main streams of textual data

In figure 6.1 – Title, description, category and tags provide basic information as to the content of a video clip. Author, date, communicate who and when submitted the file. Ratings and duration tell a little bit more about the video. Viewcount and comments can be quite important for an analysis, the former can be useful in judging the popularity of a video and the latter also provides collective opinion about a video contribution in textual form. Related videos are video recommendations that might be of similar content to the target video. This is done by an algorithm that is based on the cosine similarity of two word-vectors, but the exact details of the algorithm aren't available⁵. Response videos are actual file responses to the original clip, and are usually used to create a so called "video-debate". For a typical example of a Youtube video page please see figure D.1 in appendix D.

6.2.2 Methodology

6.2.2.1 Data Acquisition

In November 2006 Youtube was acquired by Google, Inc. After a few months Google implemented their API called GDATA, enabling developers to integrate systems with Youtube platform. This API was used to extract as many financial market related videos as possible for the entire available time period. Videos based on search keywords FTSE, DOW JONES, NASDAQ, NIKKEI, CAC, DAX⁶, and also related and response videos were retrieved (see section 6.2.1.2). Since each of these videos has a lot of meta-data associated, altogether about 90,400 videos, 89,000 tags and 3,749,000 comments on submissions related to finance news were extracted.

A number of issues with the API were encountered, some errors were discovered and some

⁵ According to Youtube support section, <http://help.youtube.com/support/youtube/bin/answer.py?hl=iw&answer=95612> (Last accessed on 2nd April 2009), the exact form of this algorithm is kept secret.

⁶ These represent UK, USA, Japanese, French, and German stock indices respectively.

limitations imposed by the Youtube terms and conditions⁷. It was ensured that terms and conditions were complied with by an appropriate implementation of our scripts. One of the major limitations was the number of search results returned for a keyword video search. This was overcome by retrieving all related videos for each video-item in the search results. Since related videos are based on a similarity recommendation system, this was a reasonable and justifiable approach. Another challenge was to ensure that videos contain the target content, this required some filtering. It was found that best way to constrain a search was to impose restrictions on the tags that could be associated with a video submission. It was found that filtering content by a combination of tags was very effective. Queries that were only filtered by keywords or topic (i.e. News) often returned too much unrelated content. The main bulk of data extraction process took over seven days. After this extraction scripts were run daily, to ensure database was kept up to date with recent video submissions. Since the system required manipulation of large amounts of data, a powerful server set up was used with a MySQL database backend.

6.2.2.2 Pre-processing Textual Content

For the sentiment analysis in section 6.2.4 textual data had to be pre-processed using standard text processing techniques. Tokenising, stemming (*finding the root form of words, Porter's stemming algorithms was used*), stop word removal had to be applied to the text. Comments are full of difficult expressions and jargon, such as emoticons “:-)”, forum talk “gr8”, rude language, negations “not good”, etc... these had to be handled appropriately (*a dictionary of such expressions was manually constructed within the code*). For example emoticons were quantified as they express sentiment, and rude language was filtered.

6.2.2.3 Pre-processing Stockmarket Prices

Stock prices tend to be very noisy, especially at high frequencies, i.e. hourly or daily price data. In such data, the prevailing short to medium term trend can get lost within the data. Therefore the noise was smoothed away using time-series segmentation. A windowing based time-series

⁷ For example, after extraction of 90,000 videos, there were about 130 videos with a blank date (0000-00-00 00:00:00) or empty title and description. The reason for this can be the fact that Youtube staff takes down videos which violate terms and conditions. According to the discussion group, http://groups.google.com/group/youtube-api-gdata/browse_thread/thread/8334c8e8b6daf30/def011e8c2716129 (Consulted on 15th March 2009), API has latency in updating feeds and video details in comparison to <http://youtube.com> website. Therefore when accessing taken down video, some corrupted responses may occur.

segmentation algorithm was used, that is price trends of at least 5% moves in magnitude must occur in order to be detected, figure 6.2 illustrates this step. The first chart in figure 6.2 shows an original stock index with daily fluctuations. As stock time series have rising and falling trends, some daily fluctuations can have an opposite direction to the direction of the overall segment. The second chart represents post processed, segmented and smoothed data, which eliminated noise. Start date and end date of segment 1, for example are t_1 and t_2 respectively, in figure 6.2 second chart. Price movement m of a segment is computed as follows:

$$m = \frac{p_i - p_{i-1}}{p_{i-1}}$$

, where p is a price at time i ($t_1 \leq i \leq t_n$, n being total number of prices in the time-series)

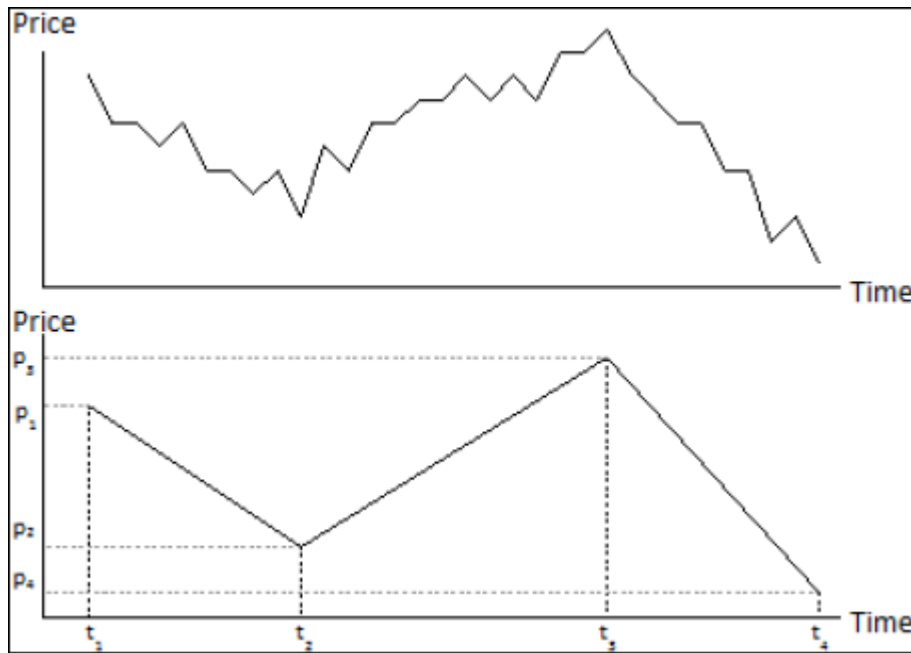


Figure 6.2 – Segmentation of Stock Index (original and segmented price data)

The full algorithm of the time-series segmentation can be downloaded as c#.net code from <http://www.newsmental.com/thesis>.

6.2.2.4 Sentiment Analysis of Textual Features

Three sentiment classification dictionaries of positive and negative words were implemented using a semi-automated process based on most common occurring stemmed words in each stream of text. Two human subjects reviewed these top stemmed word-roots and selected terms that were perceived as most positive and negative into each of the three dictionaries. This was

necessary as each stream of texts tends to use different vocabulary and phrasing to express a message. The dictionaries are provided in the appendix D, table D.1 for reference. A simple score based technique for good and bad words, was used with all three dictionaries. The idea behind the scoring function is to provide quantitative indication of sentiment for textual information of a video-clip. The scoring function is of the form,

$$s = \frac{p - n}{p + n}$$

, where s is the sentiment score, p is number of positive words and n is number of negative words in the text, $p \geq 0$, $n \geq 0$, $-1 \leq s \leq 1$. This scoring function is relatively intuitive and self explanatory. Even though quite simple, it is robust in capturing word bias in a piece of text, and has been used by other researchers in the past (Tetlock 2007). It ought to be noted that a number of other techniques for retrieving sentiment from unstructured and informal texts that use different approaches exist, for example more recently SentiStrength (Thelwall et. al 2010).

6.2.3 Youtube, the Financial Crisis and Authors of Content

First and foremost, the question of how much market news is really submitted on Youtube arises. That is, it is important to acknowledge that maybe there is too few video submissions on financial news within Youtube. Contrary to this; however, it was found that Youtube contains a large number of file submissions. A sample of video files has been inspected manually, most of the videos are of relatively high quality, often reporting on financial events throughout the day or analysing possible strategies for the next day or week(s). As can be appreciated from figure 6.3, especially in the period after September 2008 there was a rapid increase in video submissions. This could be attributed to the financial crisis. During this time, awareness of crisis and risk of recession became widespread (for example, bankruptcy of Lehman Brothers at this time dragged attention of many reporters to financial collapse and economic instability). However, it could just be due to the rapid increase in overall popularity of Youtube. Therefore benchmark data was needed to associate this trend with one or the other reason. Hence similar quantity of videos were retrieved (as described in section 6.2.2.1) from three independent categories, namely; music, entertainment and sport. A comparison of monthly time series data for each category showed that indeed only financial video submissions experienced a rapid increase (statistically significant) in fourth quarter of 2008, see figure 6.3 below.

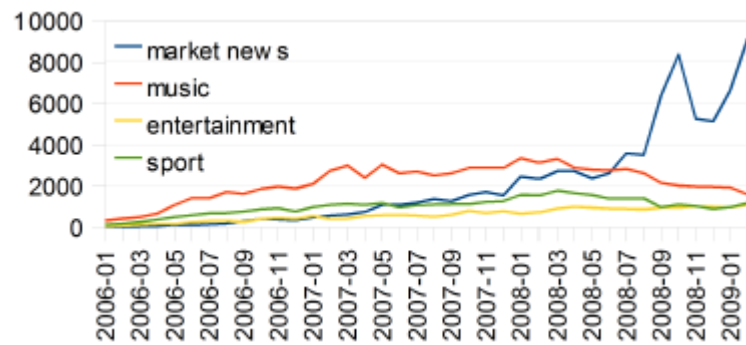


Figure 6.3 – Comparison of video submissions over various categories

Given this rapid increase in financial news video submissions, we were interested into who actually uploads videos and how these video uploads are distributed over time.

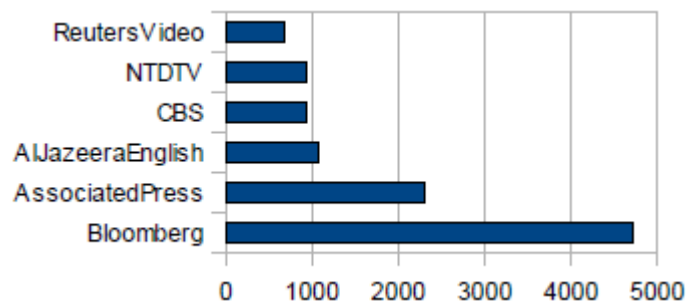


Figure 6.4 – Ranking of authors (top 5 and Reuters)

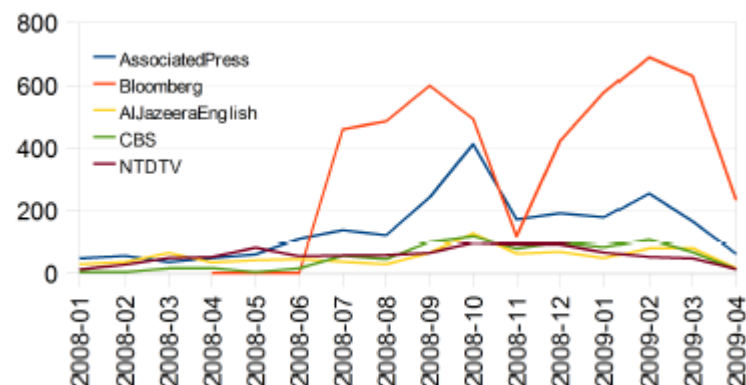


Figure 6.5 – Video submissions of selected popular authors

Figures 6.4 and 6.5 show that not only passionate users upload videos but apparently, most active users turned out to be worldwide press agencies like Bloomberg, Associated Press, Al Jazeera or CBS. They became much more active in second half of 2008 and maintain this trend in 2009. Together the top 100 authors account for 32% of videos with the remaining videos distributed amongst 27,998 unique authors.

Table 6.1 – Author statistics

Author	Video submission properties					
	Avg. Title Length	Avg. Desc. Length	Avg. View Count	Avg. Rating	Avg. no. of Raters	Avg. Duration (secs)
Bloom-berg	51	126	308	1.81	1	268
Assoc-iated Press	39	289	10'082	3.79	27	85
AlJazee-ra English	50	284	13'090	4.62	41	416
CBS	26	172	34'363	3.98	100	192
NTDTV	36	1'065	2'776	4.39	8	122
Reuters Video	28	272	5146	3.51	8	115

Table 6.1 shows quite interesting data about these authors, i.e. different attributes of their videos and how users perceive these clips. For example, Bloomberg tends to submit videos with very short descriptions of 126 characters and duration of 268 seconds on average. Associated Press's videos average description length is more than twice greater (289 characters) however videos themselves have shorter duration (84 seconds on average). The community appreciates Associated Press videos more and manifests that by much higher rating (3.8 against 1.8) and intensity of rating (27 against 1 rates per video). AP videos have average view count of 10,089, whereas the same figure for Bloomberg equals 308. It would confirm the findings of (Halvey and Keane 2007a) mentioned in section 6.2.2.1. As AP has longer descriptions, it therefore has more keywords to match by search engine. This may be the reason of such disproportion of view counts in contrast to Bloomberg.

Inspecting these attributes over a total of 1,000 publishers the following statistically significant ($p < .05$, two-tailed) correlations were found:

- Avg. View Count – Avg. no. of Raters 0.852
- Avg. Title Length – Avg. Desc. Length 0.315
- Avg. Title Length – Avg. Rating 0.187
- Avg. View Count – Avg. Rating 0.133
- Avg. Rating – Avg no. of Raters 0.174
- Avg Rating – Avg Duration 0.149

Most of these are self explanatory, such as the relationship between number of people who saw a clip and the number of people who also rated a clip is clearly correlated. An interesting insight is provided by the second correlation that is that most publishers who use descriptive (longer)

titles also tend to use more descriptive video descriptions or summaries. The correlation of 0.133 between view count and increasing average rating, indicates that users generally appreciate videos on Youtube. This does not go against intuition; that since there is a recession, ratings should be negatively correlated. This is because ratings do not usually address the content of the video messages themselves. Instead, as was discovered, ratings are generally only relevant towards quality, accuracy and stylistic factors of a video.

A number of regression models were built to describe some of the attributes in table 6.1. Models for different attributes as dependent variables were optimised on 1,000 instances of top publishers. No significant or interesting model other than the bivariate correlation relationships described above was found using simple linear regression. As an example a linear regression model for the number of raters (y), is of the form, see equation 1.

$$y = 7.79 - 0.957 * t + 19.59 * r + 0.003 * v + \mathcal{E} \quad (1)$$

, where t , r and v stand for title length, rating and view count, respectively. This is not quite so interesting and can be summarised as; number of raters depend on increasing number of viewers (standardised beta = 0.842), to some degree on average rating and a smaller title length.

Numerous financial news publishers that actively report on the Youtube platform were looked at; however, what is it that they report on? It turns out an accurate way to categorise videos by sub-topics is to filter them over associated tags (see subsection 6.2.2.1). A number of recent topics that received tremendous attention over last few months were selected and aggregated over months.

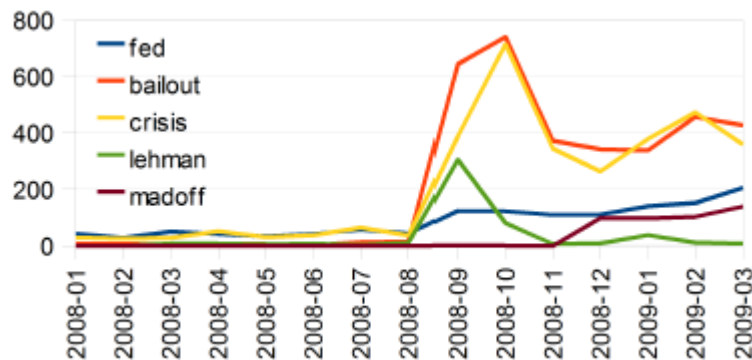


Figure 6.6 – Video submissions related to significant events / subjects

In figure 6.6 video submissions related to significant events in the financial world, are presented. When Lehman Brothers collapsed, governments of different countries tried to rescue banks from bankruptcies through bailouts. It was one of many signs that global economic

slowdown and a recession was approaching. Much attention was devoted to this problem. There is hence strong increase of videos containing tags ‘bailout’ and ‘crisis’ in September 2008 to above 700 videos in the month. Likewise, fall of Lehman brothers was reflected on Youtube by about 300 uploaded videos in the same month. The fraud case of Madoff’s financial pyramid was revealed in December 2008 and resulted in loud, public arresting of Bernard Madoff. More than 100 videos in December and in consecutive months refer to the issue.

6.2.3.1 Publishers in other Topic Categories

Financial news publishers on Youtube were inspected in quite some detail. However, what about other categories of video topics on Youtube, i.e. who is responsible for their submissions. Table 6.2 presents top 10 publishers of content in music, entertainment and the sports categories.

Table 6.2 – Top 10 publishers in different categories

Rank	Market news	Music	Entertainment	Sports
1	Bloomberg	SonyBMG	Machinima	NBA
2	Associated Press	Hoover4000	Booklvr1256	Expertvillage
3	AlJazeera English	Andromeda 881	Ginaya	NHLVideo
4	CBS	Jwcfree	CBS	RandomPokr
5	NTDTV	Rpoland	Uluvshane	TNAwrestlmg
6	Thermal1	Somedia	Ichglotzutube	TeamFlight
7	Expertvillage	Parlophone	JPizzle1122	Pennycw
8	Ehowfinance	Kinagrannis	QuickUploadr	Jon747
9	Reuters Video	Sori1004jy	LaaDida3	Ialousse24
10	FXbootcamp	SilvaGunner	Newscribe	TrueWrestling

As can be appreciated from top 10 publishers in table 6.2, most content within Music and Entertainment seems to be submitted by individual users and / or artists, except the odd big label company, such as “SonyBMG” in Music. Within sports and market news on contrary, professional publishers seem to provide bulk of the content. NBA or NHL, are the official channels (publishers) for the U.S. Basketball and Ice-Hockey leagues respectively.

In summary this sub-section looked at who publishes content and in what quantities about the financial crisis. Various properties and community feedback was analysed, and a number of

interesting relationships discovered. An investigation of publisher dynamics over time showed that numerous reputable financial news agencies, now actively submit video content. This highlights the significance that Web 2.0 communities have gained. In fact it seems that for Bloomberg and others, Youtube is becoming an information publishing channel of importance. The financial crisis has pushed financial news reporting on Youtube to before unprecedented levels. There is no reason why this should change and in fact, media sharing services are set to continue in this trend.

6.2.4 Information Efficiency on Youtube

6.2.4.1 Efficiency Expectations

The purpose of the experiments presented in this sub-section is an investigation into the degree of efficiency in propagation of financial news within Youtube. This can be measured by the in-time correlation with stockmarket price data. Price movements in financial markets are consequences of decisions taken by both stockholders and stock buyers based on how they perceive a market, sector, company or asset. Actions taken by them are not only influenced by the rational information on market but also what actions other investors took, what somebody said or wrote, and simply sentiment and emotion. According to the recently emerged field of behavioural finance (Siegel 2002), feelings of anger, fear, uncertainty or confidence and subjective perceptions of financial perspectives of economic agents have real impact on entire markets and therefore price movements. In its simplified form, Efficient Market Hypothesis, originally proposed in the 60s (Fama 1965), essentially states that market participants have equal access to information, and as new information affecting a market emerges, this information is counted-in into the market almost instantly. An offshoot of this hypothesis is the Adaptive Market Hypothesis (Lo 2004). AMH takes behavioural finance into account, and it is within this framework that it is acceptable to expect some short to medium term predictability, based on the information extracted from Youtube. This is possible; however, only if assuming information propagates into Youtube *quickly enough*, and can be *filtered* well from non relevant information. Since these assumptions cannot be guaranteed, the main question is whether a relationship between Youtube and Stockmarkets is present, and if so, of what strength, and in what form. The findings reported in this section point to the hypothesis that news data in fact must propagate through Youtube quite efficiently, see the next subsection (6.2.4.2) for results.

6.2.4.2 Information Transfer, i.e. Correlation to the Financial Market

The goal of this section's experiments is to show whether there is a detectable correlation between changes in Stockmarket prices and community submitted information⁸ on the Youtube platform. The hypothesis is that since most popular news videos receive attention, information gets propagated through Youtube quick enough to satisfy efficiency expectations (see previous section 6.2.4.1), and importantly this information represents value in terms of capturing financial news events that can be correlated to the markets. There is; however, the issue of a higher barrier of entry for video contributions which is central to this study. In order to contribute to Youtube, submitting videos generally requires the effort of recording, editing, formatting and uploading, which all represent certain costs in time, with little guarantee of any extrinsic rewards for the effort. Hence, experiments in this section had the aim to answer two questions. **First**, whether it is possible to relate intensity of content submissions with market volatility. **Secondly**, whether it is possible to quantify sentiments of videos and relate them to directional market moves. Assuming the second aim is shown to be true one may conclude that web 2.0 systems are indeed efficient.

To find whether there is a standing connection between stock price volatility and Youtube, monthly time series from video submissions and total posted comments per month were prepared, for the period between January 2007 to April 2009 (months before January 2007 contained too few video submissions). Figure 6.7 compares intensity of video submissions against absolute value of price movements of the Dow Jones index for the highlighted time-period.

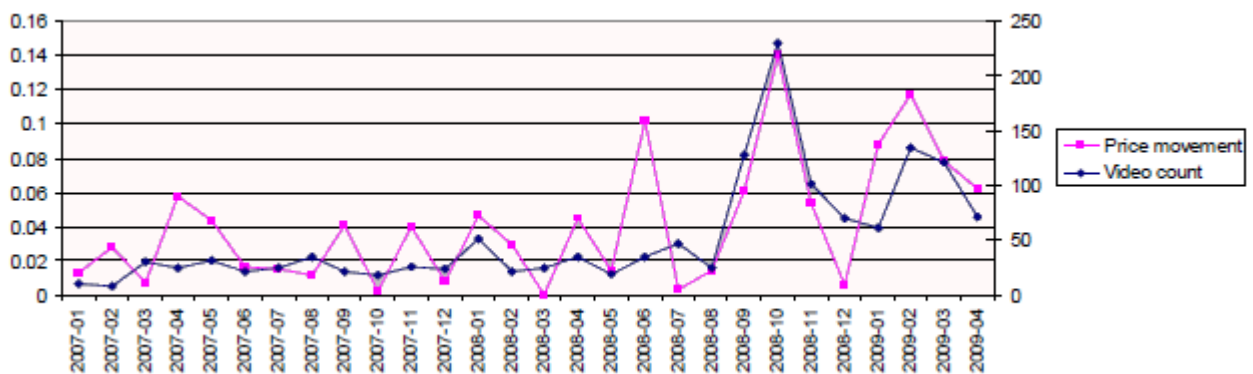


Figure 6.7 – Video submissions related to significant events / subjects (x-axis – year / months)

As can be appreciated from figure 6.7 in the second half of 2008 there is a strict relationship between intensity of video and the market. The Pearson correlation coefficient for the whole

⁸ The main UGC investigated are video-submissions.

period is 0.745 for video and 0.697 for comments submissions. Both values are statistically significant, at p (two-tailed) $< .05$, and point towards a relationship between stockmarket and Youtube. Of course causality of relationship presented above cannot be deduced from Pearson's correlation alone. In the data presented, there is a rapid increase of financial (Dow Jones) related Youtube activity beginning September 2008, which can be safely attributed to the financial crisis and risk of recession, as was established in 6.2.3.

To examine the correlation between sentiment and directional price movement of Stockmarket data, three models were built as described in sub-section 6.2.2.4. When sentiment scores were aligned against stock index returns⁹, correlations of 0.423, 0.387 and 0.033 were measured for title, description and comment models respectively, where the first two correlations are substantial and statistically significant, p (two-tailed) $< .05$. These varying strengths of correlations are due to the fact that there are noticeable differences between the three streams of text. Title often expresses the main content message of the video in a concise manner, e.g. *"Dow Closes Below 10,000, a four-year low"*. It often represents facts, as in the former example (Dow fell to the 10,000 level). The description gives more insight as to the video content, and words such as downtrend, suffer, hope or opportunity would occur. Comments on the other hand are filled with subjective opinions of users as to their interpretation of videos. The problem that was faced with comments was that sometime users would comment on the quality of video rather than the message conveyed (e.g. *"the video was well done"*, *"the guy has amazing presentation skills"*, etc ...). It was tried to take this into consideration when constructing the model vocabulary; however, filtering comments from noisy contributions can be difficult. Improved results were achieved when the scores were combined into a single indicator by averaging the individual scores. See figure 6.8 for this statistically significant 0.543 correlation. As one can see, the resulting score tends to correlate in local turning points to the market very consistently (*note; sentiment correlations were performed against the segmented relative returns, as per section 6.2.2.3*).

⁹ Stock index returns of the US Dow Jones were used. These were pre-processed as detailed in the relevant methodology section 6.2.2.3, i.e. the raw prices were segmented (percentage parameter = 5%) into 33 segments.

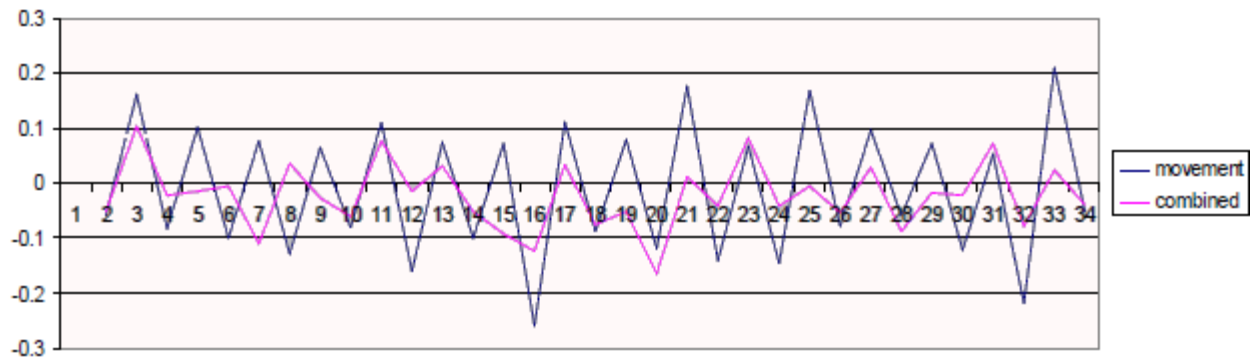


Figure 6.8 – Combined (averaged) scores (pink line) and segmented price data (blue line)

Finally we employed thresholds to the combined score, in order to change soft classification to proper classification. Since scores are distributed normally with a mean (μ) of -0.021 and variance (σ^2) of 0.0037, this was used to eliminate some of the more frequent values close to the mean. Table 6.3 illustrates the rather good (76% up to 89%) model accuracies of directional move forecasts.

Table 6.3 – Combined (averaged) scores (pink line) and segmented price data (blue line)

Scenario	Lower and upper limit	Matches / Hits	Hit rate
No limits	None	25/33	76.00%
$\mu \pm 1/4\sigma$	-0.0367 to -0.0059	21/25	84.00%
$\mu \pm 1/2\sigma$	-0.0521 to 0.0094	16/18	89.00%
$\mu \pm \sigma$	-0.0829 to 0.0402	6/7	86.00%

6.2.5 Summary and Limitations

Users of Youtube upload all kinds of videos, ranging from personal videos, mobile phone videos of US soldiers in Iraq, post election riots in Iran (coverage of the riots has been exclusively from mobile phones) to daily market analyses of the world's economies (Tapscott and Williams, 2008). In regards to the financial markets it was found that there are considerably more videos contributed during the financial crisis, than on other topics. This can to a large degree be explained by press agencies and TV stations like BBC or Bloomberg duplicating some of their video content from traditional distribution channels, yet there is a very sizeable group of amateur, hobby and professional analysts uploading daily market related videos or trading day predictions onto Youtube. A random sample of 100 videos on the Dow Jones was manually reviewed and the quality found to be good. The dates of a number of important events during the financial crisis have been found to coincide with trends of Youtube video content submission activity. The same holds for community comments, as they were found to highly correlate with absolute market volatility. The second part of this study analysed sentiment of

videos. It was found that the directional agreement between aggregate sentiments over a segment of 5% directional market volatility¹⁰ was high and significant. It must be noted that no prediction attempt on the markets was made (*since this wasn't the aim of the study and involves a number of separate issues*) instead a substantial, significant and consistent relationship of Youtube financial video content was discovered to exist in relation with major market moves, over a test period of just over 2 years.

Video titles and video descriptions were found to be significantly correlated to the Dow Jones market, except for comments, which leads to the first limitation of this study. Extracting sentiment from highly informal textual data is a difficult task and the sentiment scoring technique employed (see 6.2.2.4) in this study may not be the most suitable, after all there are a number of different and more advanced approaches available (Liu and Hu 2004, Pang and Lee 2008, Liu 2010, Thelwall et al. 2010). Unfortunately the comment analysis is complicated by another level of complexity. One must be able to detect the target of the comments sentiment (i.e. context of the comment – what is the comment about), to be able to more accurately relate the sentiment. This is further complicated by the very nature of the comments, which tend to be overly aggressive, offensive, positive, unrelated and often spammy. Another limitation of this study was the inability to process the user contributed video itself. That is, in addition to the description of a given video, it would be of much interest to understand the actual video content itself, and it was found that a useful technique to achieve further understanding would be a voice / audio analysis of Youtube videos, similar to the one employed by Dowman et al. (2005). The feasibility of implementing such a system was investigated, and it was found that Youtube recently launched automatic transcription and user submitted subtitles¹¹ for some of its videos. Unfortunately after further effort, involving direct communication with Google's Youtube representatives, we were unable to conduct this study due to limited freely available video-data of such nature. More generally, Youtube was found to be a useful, efficient and interesting resource¹².

10 During a highly volatile market, such as the financial crisis of 2008/09, the time duration of a segment would be shorter than in non volatile markets, as larger price moves were likelier.

11 A special time-stamped transcription format allows for this type of data to be submitted, with some effort.

12 Many more social and cultural phenomena of interest can be easily studied on Youtube which have been omitted in this chapter. However, we did find for example what people on Youtube liked the most in the Sports, Music and Entertainment categories, respectively: [NBA, football, basketball, sports, soccer, world, Jordan, Dunk, Michael, hockey], [music, live, rock, guitar, pop, cover, piano, love, acoustic, song] and [episode, show, movie, TV, video, funny, game, comedy, trailer, Xbox].

6.3 Delicious

Delicious has to a large extent helped to popularise the process of tagging (Golder and Huberman 2006). Although recently the website experienced some problems and competing services (e.g. www.google.com/bookmarks) have taken hold, it is still according to alexa.com in the top 500 websites, as ranked by unique visitor traffic¹³. This and a number of important reasons, already discussed in chapter 3, make Delicious a valuable web 2.0 resource for aggregate opinion, over collectively agreed concepts, and a useful source of CI. Chapter 3 used the system successfully in support of experiments concerned with web 2.0 related neologisms, and a valid new methodology for term emergence was presented. To establish that Delicious bookmarking patterns can reflect real world events, section 3.2.1, presented an example of Madoff's Ponzi scheme and how the Delicious UGC correlates with these real-world events.

To the author's knowledge there is no prior published work linking UGC from a social bookmarking platform (e.g. Delicious) explicitly to the financial markets. In this section the significance of collective action on Delicious is related to the financial markets, specifically the recent financial crisis. Several peripheral issues are considered and pointed out for further work.

6.3.1 Methodology

Delicious has a clean and simple design (which probably contributed to its success). The user interface is consistent, and the URL folder structure is transparent, with RSS feeds available for any combination of search tags. Although an API is available¹⁴, it does not support retrieval of public datasets. Hence a custom script in C# with a MySQL database was built for extracting and storing bookmarks. Further details of the methodology and motivation for using delicious are available in sections 3.1 and 3.2.2.

Search on delicious is performed using tags for a given resource; for instance, results from the search shown in figure 6.9 would return all the links, where those three tags were found. Each result-item can then be drilled down into, for a list of individual users who bookmarked the link with the date, comments (if any), and tags used.

¹³ Recently Delicious changed ownership and was redesigned and re-launched in September 2011. This section refers to the old design, although this change does not affect the analysis and conclusions drawn.

¹⁴ <http://www.delicious.com/help/api>, however only dataset belonging to a given user can be retrieved directly

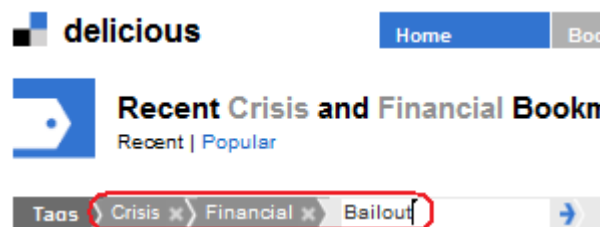


Figure 6.9 – Delicious tag based querying (returns a list of resources tagged by the search tags)

As many links and comments as possible were extracted based on tag search for a number of well known companies, trading related keywords, currencies, and commodities – specifically:

- **Financial Crisis:** *Crisis+financial+bailout, Lehman, Madoff*
- **Currencies:** *USD, GBP, GBP+USD*
- **Trading:** *Daytrading, Trend+Trading, Swing+Trading, HedgeFunds*
- **Companies:** *FTSE-index, Vodafone, HSBC, AstraZeneca, Barclays, Tesco, RioTinto*
- **Other assets:** *Commodities, Derivatives*

The plus sign refers to a conjunction search of those tags. The major limiting factor in the data-retrieval for this study was the limitation of unique links returned from each search. Although the limit was more relaxed with retrieving the users who have bookmarked a given site, delicious further enforces a strict pre-emptive policy on automated data-access by blocking requests that access resources frequently. Retrieval code had to be written to comply with official request limits, yet it was necessary to compose a proxy IP switching mechanism to avoid being unjustly blocked¹⁵. A possible extension would be to use query splitting as suggested in Thelwall (2009). Since Delicious supports logical expression in its search, this process is viable. Some of the data retrieved, based on the above term-tags, are discussed in more detail and overall results corroborated with the financial crisis in section 6.3.3. The next section presents a summary of the financial crisis.

6.3.2 A brief Summary of the Financial Crisis

A brief recapitulation of the real world financial events leading to the financial crisis is necessary. There are a number of useful *time-lines* summaries of the financial crisis available on the Web¹⁶ (the following brief description of events is based on these sources). Essentially the first warning signs came in February 2007 when the *Federal Home Loan Mortgage Corporation (Freddie Mac)* announced that it would no longer buy the most risky subprime

¹⁵ The author found despite complying to the limitations to be blocked more frequently, which can occur for a number of reasons. Delicious did not get back after initial request by the author and a proxy switching procedure was used.

¹⁶ See <http://news.bbc.co.uk/1/hi/7521250.stm>, <http://timeline.stlouisfed.org/index.cfm?p=timeline#> or http://www.ny.frb.org/research/global_economy/IRCTimelinePublic.pdf directly.

mortgages and mortgage-related securities, and in April that same year *New Century Financial Corporation* which was an expert subprime mortgage lender filed for bankruptcy. A number of events followed in which rating agencies such as *Standard and Poor's* and *Moody's* downgraded several mortgage related products. During August and September 2007 interest rates are cut in the *UK*, *US* and *EU* and on 13th September 2007 *Northern Rock* (a British Bank) asks the government for emergency financial support. During October *UBS* and *Merill Lynch* are the first among bigger banks to report significant losses, and in December *US* government unveils a plan to help homeowners facing *foreclosure*. During February 2008 the *British government* announces, that *Northern Rock* is to be nationalised. During the summer of 2008 a number of struggling companies announce plans of *rights issues* and similar steps in order to raise capital to cover their losses on bad loans however reviewed sources agree that September 2008 was the worst month in the crisis, in fact BBC dubbed it "*The Eye of the Storm*"¹⁷. Mortgage lenders *Fannie Mae* and *Freddie Mac*, which accounted for nearly half the outstanding mortgages in *US* are rescued by the *US* government in one of the largest bailouts in *US* history. During late August some speculation emerged that state owned Korea Development Bank was considering buying the investment house *Lehman Brothers* however on September 9th it was reported that takeover talks were put on hold. The next day *Lehman* reported a large loss of nearly \$4 billion for a three month period, and on the 15th September (after a weekend of frantic negotiations) shortly before 1 am (EST-time) *Lehman* filed for bankruptcy, citing debt of over \$613 billion as one of the reasons¹⁸. On the 17th September *New York Stock Exchange* delisted *Lehman Brothers*¹⁹. Towards the end of September *US lawmakers* announce the largest rescue plan since the great depression. In Britain, the mortgage lender *Bradford & Bingley* is nationalised on the 29th September, and a number of other European banks have to be bailed out, including *Dexia* and *Fortis*. During October of 2008, a number of rescue packages and legislation is passed to help the struggling economy, most importantly on the 3rd October the house of representatives approves a the *US rescue plan*. The *US recession* is officially declared by the National Bureau of Economic Research on the 1st December 2008. In January 2009, *US* and *UK interest rates* are at their historically lowest point, *US Jobless rate* is the highest for over 16 years and *China's exports* register the biggest decline in a decade. On 2nd March 2009 the insurance giant *AIG* reports the *largest quarterly loss* in *US* corporate history. The *US Federal Reserve* announced on the 13th March that it will buy \$1.2 trillion worth of debt to help boost economic recovery.

This suffices as a reminder of some of the significant events, for the rest of this chapter and

17 <http://news.bbc.co.uk/1/hi/7521250.stm>, last accessed on 1st March 2011

18 http://www.lehman.com/press/pdf_2008/091508_lbhi_chapter11_announce.pdf, last accessed on 1st March 2011

19 <http://www.nyse.com/press/1221647871334.html>, last accessed on 1st March 2011

where necessary in sub-sections, relevant events will be highlighted.

6.3.3 Analysis and Results

6.3.3.1 The Financial Crisis

The appeal of using a web 2.0 bookmarking systems (such as Delicious) as opposed to Youtube and other web 2.0 systems, is the minimal effort required for contributing UGC. Pages on the World Wide Web emerge much quicker in response to events than they would with many other media. Keeping track of a vast number of ever growing online resources is a major challenge and the participation of implicitly motivated humans in intelligently annotating and indexing these resources might allow for speedy identification of resources relating to particular events, companies or technologies. The process on a bookmaking page such as *delicious* is simple, where a page P_i gets detected by a user and this user is the first to detect the URL $U_i^{P_i}$, then other users are likely to follow in a short time period given that; **1**-the resource P_i is useful, **2**-the resource P_i is timely / relevant right now. The number of users $n=|U_j^{P_i}|$ who bookmarked the resource is therefore an indication of the resource's overall significance. A resource P_i has been bookmarked n times where each bookmark can be more formally written as $B_j^{P_i} \langle d, c, t \rangle$, where d is the date-time stamp, c is the comment and t the set of tags used for the bookmark.

It can be safely said that certain topics are underrepresented on Delicious, due to a small subset of users interested in a given topic, as for example, far less people would care about online resources concerning the mining company RioTinto than HSBCs new offers on credit cards and travel insurance. At least on Delicious this was found to be the case; however, on another bookmarking application where the community shares different demographics and cultural / community norms this may be different²⁰.

How then, should these user generated opinions on web resources be analysed over time? One approach is to simply aggregate the number of resources on a topic over regular time-intervals to produce a time series, in which the counts can be weighted by the number of users who bookmarked the resource. In this study such an approach wouldn't be practical with niche topics for which there are far fewer resources on Delicious. Therefore the individual bookmark counts (rather than resource counts) are aggregated into time series. This approach measures the interest in a topic over time. No explicit weighting is required since individual bookmark counts are included in the actual time series, and only resources that were bookmarked by at least 3

²⁰ There are numerous social bookmarking web 2.0 applications on the Web with different focus, yet even delicious is used by a number of groups other than the general public, for example teachers share resources and even medical professionals were found to be using delicious to track and share topical resources easily.

unique users are considered. Although bookmarks could be weighted differently based on user-profiles of the users who made them, why this might sometime make sense is discussed in section 6.3.4.

According to the description given above, the following monthly, weekly, and daily time-series charts of bookmarks were constructed, see figures 6.10, 6.11 and 6.12 respectively. From the monthly figure 6.10 it is clearly visible how Lehman Brothers, Financial Crisis and Bailout related resources became bookmarked significantly more frequently from early September 2008 onwards. In fact the weekly chart in figure 6.11 shows more clearly how the bankruptcy of *Lehman Brothers* simultaneously caused a rise in interest into *Crisis+Finance+Bailout* tagged resources on Delicious. This interest prevailed for the rest of 2008 and the first half of 2009. From the daily chart in figure 6.12 the reaction can be seen at a day's granularity.

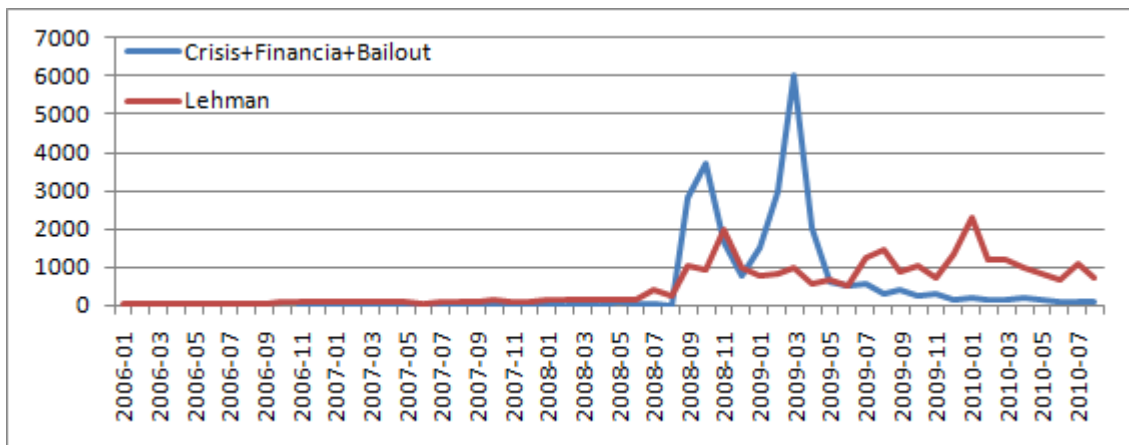


Figure 6.10 – *Financial Crisis* and *Lehman* related bookmark counts (monthly)

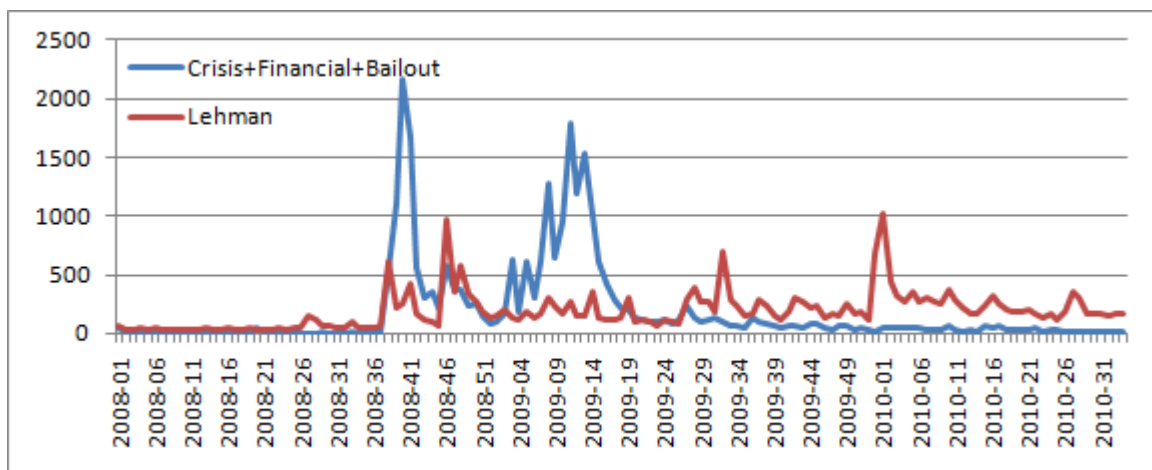


Figure 6.11 – *Financial Crisis* and *Lehman* related bookmark counts (weekly)

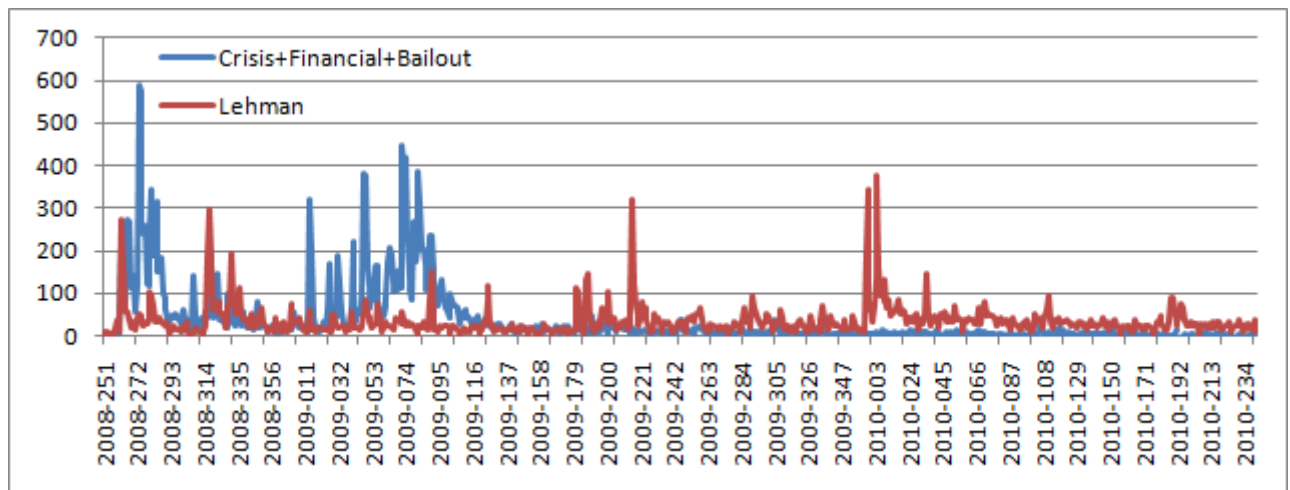


Figure 6.12 – *Financial Crisis* and *Lehman* related bookmark counts (daily)

Given that the bankruptcy of *Lehman Brothers* was announced in early hours of 15th September, several interesting observations about Delicious can be made from the daily data (not quite visible on the chart). Before the bankruptcy on 15th September, *Lehman Brothers* received virtually no attention on Delicious. It was found there were few bookmarks from the 9th September linking to an article criticising *Lehman Brothers* accounting practices. Over twice as many (compared to the last seven day's average), or 19 bookmarks relating to *Lehman* were submitted to delicious on the day of the bankruptcy (15th September), indicating some response from the user-base. The response was small considering the importance of the event and was followed by an increase to 36 bookmarks on the 16th. It wasn't until the 18th September that the bookmark count shot up to over 270 for the day, which is the first clearly visible spike in figure 6.12. This delayed reaction from the Delicious user-base is attributed in part to the unavailability of relevant and fully informative articles on the *Lehman* bankruptcy, and partly to the fact that only resources bookmarked by at least 3 unique users were retrieved. It wasn't until a few days later, when better written and digested resources appeared on the topic, that the Delicious community considered these useful enough to be bookmarked.

Interestingly the count of bookmarked resources for *Crisis+Financial+Bailout* tags seems to have spiked with the failure of *Lehman Brothers*, despite the fact that the subprime mortgage crisis and a number of somewhat higher profile bank failures and losses had already taken place (see 6.3.2). During August 2008, before *Lehman Brothers* Bankruptcy, there were only 15 bookmarked websites with such a tag combination (i.e. *Crisis+Financial+Bailout*), however during September as many as 2,831 bookmarks accumulated within one single month. This was followed by a further 3,716 bookmarks during October. Indeed strongly indicating that people became substantially more worried and alarmed about the ongoing economic issues. The trigger seems to have been the failure of *Lehman Brothers*, as it caused more, wide spread worries

about a financial crisis and bailout expectations. In fact the high number of new bookmarks for *Crisis+Financial+Bailout* tags did persist for much longer than Lehman Brothers related bookmarks, increasing to a new peak of just less than 6000 bookmarked resources during March 2009. Much *buzz* and debate at the time was concerning the largest US bailout package (see 6.3.2). Delicious bookmarking activity for the mentioned tags dropped-off eventually; however, as of August 2010 the count of bookmarks was still higher than the pre-*Lehman Brothers* time-period.

These findings suggest that efficient information transfer may not be as high on Delicious as it was found to be on Youtube. However, it is interesting to see when interest into a topic is triggered, as above the single event of *Lehman Brothers* failure has done. A bookmarking system such as Delicious is predominantly used to detect resources, the table 6.4 presents five top links during the period of most bookmarking activity related to financial crisis and bailouts. Clearly Delicious is useful in pin-pointing information hubs, or significant resources on the web, since these bookmarks are all crisis specific information portals, which in turn may take some time to appear on the web, hence the information transfer delays, described above.

Table 6.4 – Top 5 URLs bookmarked during August 2008-September 2009 (highest bookmarking activity)

Resource URL	#resource bookmarked
http://www.recovery.gov/	1966
http://flowingdata.com/2009/03/13/27-visualizations-and-infographics-to-understand-the-financial-crisis	1232
http://www.pagetutor.com/trillion/index.html	1016
http://baselinescenario.com/	888
http://www.themoneymeltdown.com/	821

6.3.3.2 Companies, Commodities, and Currencies

Several entities from the financial markets are examined in this section, beginning with one of the most popular currency pairs. For instance the actual GBP/USD monthly averaged²¹ exchange rates and *GBP+USD* tags associated bookmark counts can be compared. Figure 6.13 shows both monthly time-series over the 2005 to 2010 time-period. During the early months of 2009 the British Pound hit some of its weakest exchange rate against the US dollar. Gordon Brown's government, at the time, presented an economic aid package; however, this had a negative effect on investor confidence, causing much uncertainty on the foreign exchanges. Around this time a significant spike in currency related bookmarks can be appreciated from figure 6.13. The bookmarking activity shows predominantly links to a mix of currency

²¹ Source of currency data <http://www.oanda.com/currency/average>, last accessed 2nd March 2010

information resources and articles on GBP currency issues, for example the well known and highly influential currency investor Jim Rogers²² at the time expressed some strong negative sentiments about the British Pound.

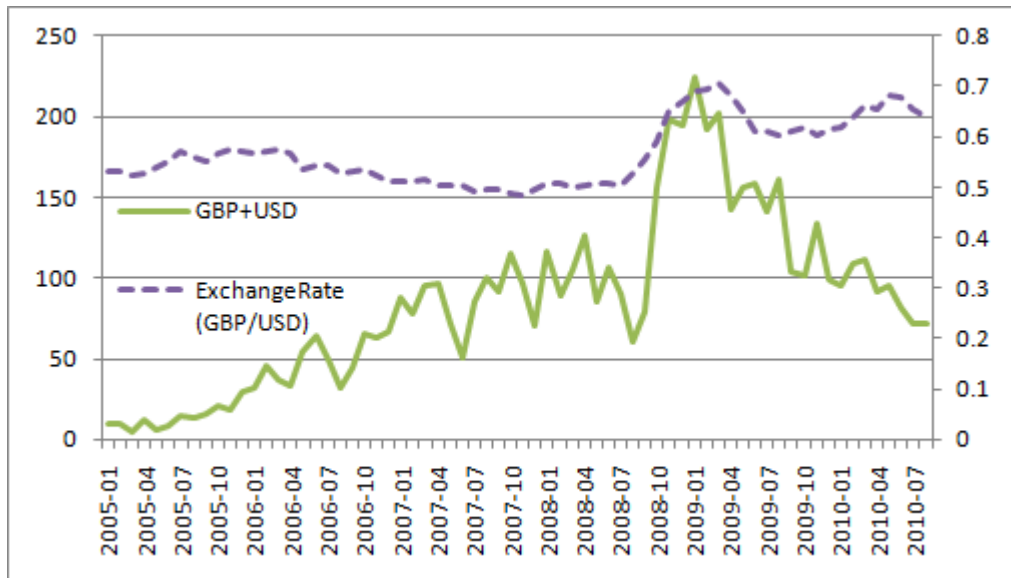


Figure 6.13 – GBP/USD exchange rate vs. GBP+USD tagged bookmark counts (monthly)

The bookmarks associated with Hedge funds and Daytrading tags distributed over time, are presented in figure 6.14. These exemplify a slightly different aspect of the financial crisis. Day trading has been growing in popularity for numerous years before the financial crisis, and in fact there exists a strong community of amateur and semi-professional day traders (Turner 2007, Burns 2011). It is interesting to observe some decline in interest into day-trading throughout 2009 and 2010, yet the bookmarks per month are still high. As for Hedge Funds, these were widely criticised for being responsible for a lot of the bad debt and overly risky trading techniques that exacerbated the financial crisis (Friedman 2010). Hence, it is understandable that for a while, Hedge Funds were a point of interest on Delicious.

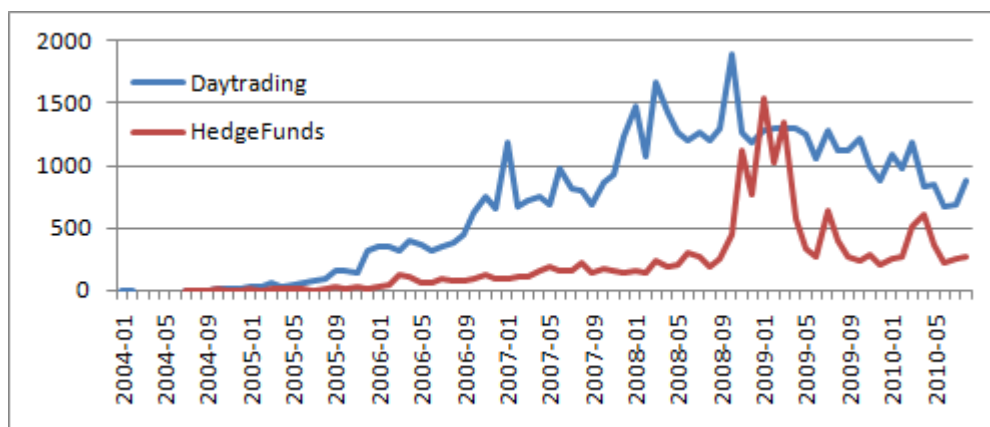


Figure 6.14 – *Daytrading* and *HedgeFunds* related bookmark counts (monthly)

²² Jim Rogers began the famous Quantum fund together with George Soros (Soros 1995).

Figure 6.15 illustrates the overall popularity of commodities and derivatives over the years. Interest into both increased somewhat after the Lehman Brothers crisis. The noticeable peak of derivatives bookmark counts in April 2010 is related to the accusations and criticism of Goldman Sachs derivatives trading desk and the role it played in worsening the effects of the crisis²³. During the period from 2007 to mid-2008 commodities bookmark counts were substantially higher in relation to derivatives as this was a period of a major bull/up-market in the commodities market²⁴.

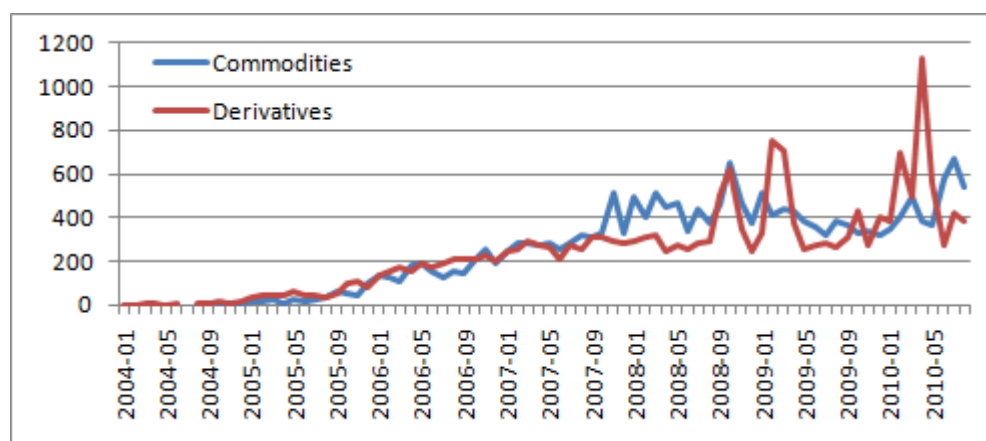


Figure 6.15 – *Commodities* and *Derivatives* related bookmark counts (monthly)

Individual stocks and the FTSE-index related bookmark counts from Delicious are presented in figure 6.16. A number of major peaks were found to correlate with actual stock price moves, such as April 2009 peak for Barclays. Establishing a consistent correlation for most of the local minima and maxima points based solely on bookmarking frequencies to Stockmarket volatility is limited by the quantity and quality of bookmarking data. Valuable company specific resources are discovered by the Delicious community; however, a statistically significant relationship with Stock-prices could not be proven.

²³ <http://finance.yahoo.com/news/SEC-accuses-Goldman-Sachs-of-apf-1523020722.html?x=0>

²⁴ <http://finance.yahoo.com/echarts?s=DBC>

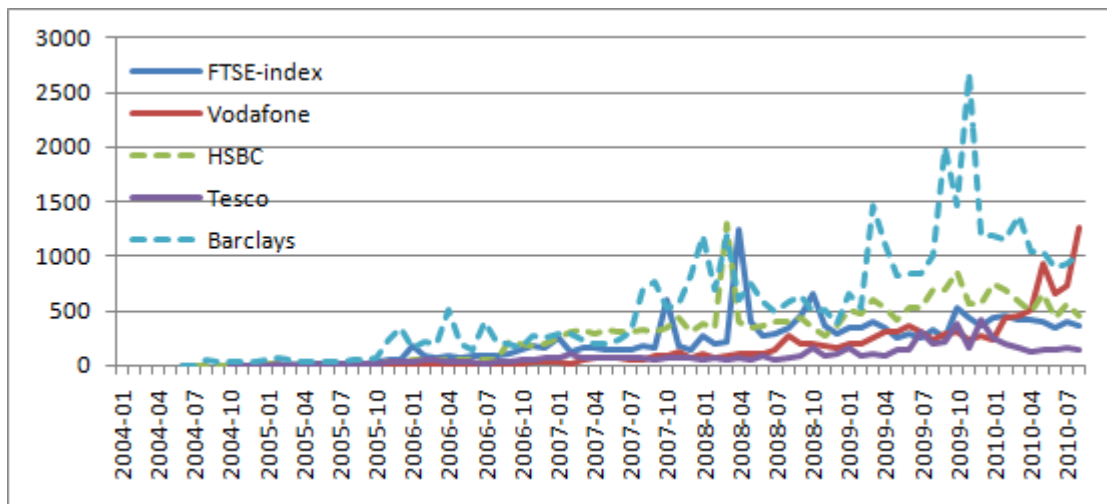


Figure 6.16 – FTSE-index and Vodafone, HSBC, Tesco and Barclays related bookmark counts (monthly)

6.3.4 Summary and Limitations

Given that the Delicious user-base is probably somewhat more technically minded than the average population, and likely to have higher education (see section 3.4), it might not be the best community for financial news-analysis. This is well illustrated on the Rio Tinto (a large FTSE-100 mining company) example where the tag *RioTinto* was used to retrieve relevant bookmarks. Only 455 bookmarks, over 24 unique links, where at least 3 individuals bookmarked one resource, were found. Many of these bookmarks were made during a few days following 23rd April 2007, when Rio Tinto discovered a new crystal in Siberia²⁵, with a similar composition as the fictional Kryptonite crystal from the Superman novel. Wider and probably more leisurely interests seem to be more dominated on Delicious, when it comes to individual stocks. In another example, during April 2009²⁶ and the following months there was renewed discussions about AstraZeneca's (the drug manufacturer) controversial drug Seroquel, and a spike in bookmarking during April 2009 was detected. Although our study found a correlation of widely public financial crisis events to exist with bookmarking habits, the nature of the bookmarks seems to be less representative of the financial analysis / news domain, unless the news has a broader appeal.

Top users on delicious can be identified by the number of links bookmarked, or alternatively by the sum of the length (*in characters*) of comments made, if one of the goals is to analyse the optional free-form comments. It may well make sense in tracking a subset of users on delicious, rather than tracking topics purely based on content. For example the 5th top user (*Jason*) as ranked by the sum of characters, claims to be a *Banker* and has supplied comments extensively

²⁵ <http://news.bbc.co.uk/2/hi/science/nature/6584229.stm>, last accessed on 1st March 2011

²⁶ <http://www.fda.gov/AdvisoryCommittees/Calendar/ucm136250.htm>, last accessed on 1st March 2011

to note down his own opinions about financial crisis and other mostly financial related resources. Although comments are used only on about 20% of all bookmarks, and on average 22% of all bookmark comments per link are repeat comments. However, the first few users in the ranking use relatively in-depth comments with each bookmarked resource; see table 6.5 for a list of top users. In the future the value of comments, of such active users could be investigated.

Table 6.5 – Top 10 users on the delicious dataset, as ranked by volume of comment data submitted (*highlighted in gray are also within top 10 based on tag volume of data submitted*)

Delicious User	Sum (chr) Comments	Sum (chr) Tags	Comment count	Avg comment length	Avg tag length
Adamcrowe	40009	4731	68	588.3676	69.5735
PlanMaestro	37829	4590	75	504.3867	61.2
Jkstyle	21874	3899	65	336.5231	59.9846
Getpost	21493	773	48	447.7708	16.1042
Jason	20629	8741	174	118.5575	50.2356
Alex Boden	18199	2498	67	271.6269	37.2836
asterisk2a	17074	5566	59	289.3898	94.339
Michel Bauwens	16653	1202	58	287.1207	20.7241
Kiffmeister	16397	474	41	399.9268	11.561
C. Maoxian	16273	1525	133	122.3534	11.4662

In this study bookmarks were retrieved based on a combination (*conjunction*) of tags, and also based on simple one-word tags, which yielded considerably more unrelated links, in terms of topical relevance. It was found that results for *Crisis+Financial+Bailout* were highly relevant to economic crisis, for example the highest spike in figure 6.12 was a result of many bookmarks on economic crisis and bailout related links, and this was consistent for the remaining time-period observed. Unfortunately, bookmarks retrieved using a search with a single tag²⁷, were prone to introducing noise into the dataset. Delicious search is based on at least one bookmarker (even if it is out of 500 or more) using the search-tag, this can cause many unrelated bookmarks to be included in the results list. A presentation on slide-share completely unrelated to Lehman Brothers or banking by *Lynn Lehman*, resulted in noise being introduced into the Lehman Brothers related bookmarks, as a few individuals used Lynn's surname as their bookmark tag. For example the spike in August 2009 for Lehman Brothers was caused mostly by a surge of bookmarks from this unrelated resource. Another example relates to *GBP* currency search where a site dedicated to ginger beer making was bookmarked by over one hundred individuals, however one of the users selected GBP (standing for Ginger Beer Plant) as a tag. This is a

²⁷ Unless the single tag itself is tag composed of multiple words, such as *HedgeFund*, or *TrendTrading*. This is a common tag shorthand convention (Orchard 2006).

limitation and the data needs to be either carefully cleansed, retrieved using a selection of multiple tags, or a minimum user count of a given tag checked for. The latter two options seem most appropriate, although the search may result in less results, hence it may be advisable to vary the minimum bookmarkers required, for a resource to be retrieved.

In summary Delicious was found to be a lively community as previously reported by others (Orchard 2006). Social bookmarking data on a number of carefully chosen tags was extracted, and the nature of bookmarked resources was briefly discussed. Currency problems during 2009, daytrading, hedge funds, derivatives, commodities and British stocks were looked at. An indicative link to financial crisis events and the financial markets could be established, with this study being one of the first to look at such an explicit connection to the markets (to the author's knowledge). It can also be concluded that efficient information transfer is not as high on Delicious as it was found to be on Youtube.

6.4 Amazon

In sections 2.3 and 5.4.2.3 it was highlighted that Amazon is one of the oldest websites that has successfully innovated web 2.0 style social engagement with its users. Despite being usually perceived as an online store (Spector 2002), Amazon incorporates a number of social features, such as public / private wishlists, tagging, submitting product images or manuals²⁸, user profiles, and even social networking (*i.e. see section 5.4.2.3*). However, the feature that stands out the most are the product reviews, and the focus of this section is an investigation of the collective intelligence locked within Amazon reviews. Each Amazon product review breaks down into two elements; *1*– a rating on a 5 point ordinal scale (*the minimum and maximum values on the scale are perceived as negative and positive sentiments about a product, respectively*) and *2*– a free-form text review with no real word limit.

A number of papers from the field of economics have shown there to be some effect of word of mouth on sales figures (Coleman et. al 1966, Foster and Rosenzweig 1995). Much work has been done on free-text form datasets of reviews, for example; opinion extraction from product ratings (Dave et. al 2003, Pang and Lee 2005, Pang and Lee 2008), or product specific feature extraction from reviews (Liu and Hu 2004, Liu et. al 2005). However, the author has not come across any study investigating possible connection of Amazon product reviews, or ratings with the financial markets. Although Amazon may not seem most relevant, Amazon contains a great

²⁸ Amazon also uses collective intelligence within its website, based on implicit user generated data, such as 1-“people who have viewed this and similar books end up buying this, or 2-“customers who bought this also bought that”. However it must be pointed out that collective intelligence within Amazon itself is used sparingly, considering the large amounts of explicit user generated data.

amount of valuable UGC content. Two different types of products will be investigated within this section, and as will be demonstrated, there is some value in the UGC related to finance and financial-markets.

The study in this chapter can be broken down into two parts (*part i, part ii*) based on the Amazon review data investigated in each case. The first part of this study investigates reviews for a number of products that are the lead-products of competing global companies. It can be argued that the volume of reviews for products on Amazon, closely correlates with popularity and hence sales of products, and that those Amazon sales (see Chevalier and Mayzlin 2006) are in turn representative of the real online sales figures of the company's products²⁹. Connection between sales and reputation exists, assuming that more feedback to a product is due to the products popularity and not due to its problem features, for example. Also, according to social psychology literature (Cialdini 2000), social validation has an influence on consumers' attitudes. The aggregated reviews for a product represent the consensus opinion, which when carefully analysed might provide useful insights on good, bad points, or other issues with a given product, and this is valuable information to manufacturers³⁰. It could be hypothesized that reviews have some predictive element to them, since if they are generally positive they may influence potential customers to make similar purchases or vice-versa, this can be referred to as *aggregate review influence*. In the first **part (i)** of this study the review frequency for competing products will be compared and investigated for a relative relationship with financial markets. The second **part (ii)** of this study is based on somewhat different assumptions. Book reviews relating to particular topics on the financial markets are investigated for possible relationships with market performance. This is based on two arguments, **1**-an increased interest into particular areas of finance will likely cause an increase in the number of published books on the topic within a given time-period, this was shown to be the case in many areas of social and cultural life (Michel et al. 2010). Recently a new tool was introduced by Michel et al. (2010) in cooperation with Google, which allows study of publishing history over time and has been dubbed as some sort of cultural DNA-code, for social-scientists to explore and study (Bohannon 2010). **2**-There is a large self-improvement and part-time, amateur, but also professional trader community, which actively recommends reading material between its members, with Amazon being one of the venues for such individuals (authors subjective

29 Amazon represents a large share of online shopping, e.g. see for example <http://www.thebookseller.com/news/amazon-has-80-online-share-claims-new-survey.html>.

30 In fact this is often a major motivation for research in this emerging area, and a number of commercial ventures that analyse such review data are active in this field, e.g. <http://www.buzznumbershq.com/>, <http://www.visibletechnologies.com/>, <http://www.radian6.com/>, <http://www.collectiveintellect.com/>, etc.

observations at financial conferences; e.g. Traders Expo 2006/09; Turner 2007; Burns 2011). Based on this observation many reviews of particularly practical trading books may be representative of real world traders to some extent (i.e. people who are likely to be actively trading). If enough reviews are available in aggregate, it may be possible to measure the interest in particular trading topics from simple review frequencies. Further to this, the opinion on a given book may be reviewed for certain keywords and sentiment. Arguably the sentiment refers to the book; however, from qualitative observation, this is not always the case. A summary of the two parts to the proposed study is provided in table 6.6, below.

Table 6.6 – Overview of the two part study of the Amazon reviews and financial market linkage

Amazon review study, part i	Amazon review study, part ii
<p><u>Reviews:</u> Leading Products (no-books) Reviews</p> <p><u>Hypothesis:</u> frequencies of (<i>selected</i>) competing product reviews, when compared to each other represent the same real world scenario, possibly exhibited on an averaged stock-price or similar proxy of company performance.</p> <p><u>Other:</u> The investigation of presence of certain business specific keywords or sentiment in reviews may provide useful explanatory insights.</p>	<p><u>Reviews:</u> Trading Books Reviews</p> <p><u>Hypothesis:</u> frequencies and sentiment for aggregated book category reviews may exhibit a relation to the financial-market, or at least certain financial events.</p> <p><u>Other:</u> It is interesting to understand the content of actual reviews, whether contextual information is prevalent or at least whether it exists in such Amazon reviews.</p>

This study explores and attempts to establish a connection to financial markets based solely on UGC on product and book reviews from Amazon.

6.4.1 Methodology

6.4.1.1 General Methodology

Amazon is open about its data, and using its API allows tight account integration; software can be build on top of the website, which is an important feature of web 2.0 as a platform (see section 2.2.1). The API did not support functionality to extract certain information from its pages³¹, therefore a custom web crawler to crawl the Amazon search results and subsequently extract book / product reviews and information from each page, had to be developed. The crawler script was written entirely in C# and all data was stored to an online MySql database

31 See <https://forums.aws.amazon.com/thread.jspa?threadID=39172&tstart=0> for some details about the APIs limitations

server for further analysis, around *600MB* of review and product data. The crawler also collected most occurring and significant phrases for books, which were irrelevant for other products.

6.4.1.2 Leading Product Reviews (part i) Data Collection

Since a large number of product reviews would require a lot of storage space, only certain products, i.e. a subset of products, were selected for retrieval. During this selection there was an effort to adhere to certain rules where possible. First manufacturer's that would be considered to be competitors in some well known product domains were selected, *e.g. Acer, Asus, Apple, Microsoft or Kindle, Ectaco, Bookeen*. Secondly for some companies, product ranges that were representative of the entire company were picked, *e.g. Microsoft (OS, Office, Xbox console, Zune MP3, peripherals), Apple (Macbooks, Desktop Macs, Ipod, Ipad)*, and finally leading products within a company's product range were preferred for selection, *e.g. Microsoft Xbox, Nintendo Wii, Sony PS3, for the games console market*. Product ranges were also picked based on the number of total reviews. Manufacturers with an insignificant review count on its products would be avoided in favour of manufactures for which there were more product reviews. In order to increase the number of retrieved products and reviews, a set of products that appeared as related products on an Amazon's product range would also be extracted (*for the same manufacturer*), *e.g. for Microsoft Windows XP Home; the XP Professional and XP Enterprise versions would be extracted as well*. The break-down of the product-ranges selected for download are presented in table 6.7.

Table 6.7 – Overview of Companies and Products / Product Groups that were extracted from Amazon (legend in bottom left corner)

bottom left corner)

Apple	Microsoft	Acer	Asus	HP	Samsung
Notebooks	Zune	Netbooks	Notebooks	Desktop	Netbook
Desktops	Business & Office soft.	Notebooks	Netbooks	Notebook	LCD
iPod touch	Operating Systems				Netbook
iPod classics	Games				
iPod nanos	Mouse/keyboards				
iPod shuffles	Xbox consoles				
Soft.					
iPad					
Toshiba	Sony	Nintendo	Amazon	Bookeen	Ectaco
Notebooks	ps3 consoles	Wii console	Kindle 6" Display	Cybook	jetBook
	Vaio notebooks		Kindle 9,7" Display		
	Reader				
Legend	Product count				
computers	12				
mp3 players	5				
software	4				
consoles	3				
Book Readers	5				
Misc	2				

The guiding principle for selection of product ranges and manufacturers were set-out in such a way as to build a product review sample that would represent some real world concerns. These concerns are briefly illustrated in the following bullet points.

- Competitors and well known rivals were chosen intentionally with a certain representative product range that characterised the company. A relative comparison of reviews / ratings between rival companies would hence be possible on the dataset.
- A representative set of products was downloaded for manufacturers, or alternatively the lead product range(s). This lead product would often be representative for a large share of sales for a company (i.e. this is based on the simple assumption of Pareto Principle, 80/20 rule). The reputation of the product, to an extent can be expected to be a reflection of its sales success; Chevalier and Mayzlin (2006) found that an improvement in reviews, lead to an increase in sales within Amazon and Barnes & Noble (another online book-store). They further revealed that the impact of negative (1-star) reviews is generally greater on sales than the impact of positive (5-star) reviews.

Analysis of lead-product reviews (part i) is based on the assumption that ratings and reviews reflect the mood about a product, and when these are aggregated, this reflects the overall popularity. However, the popularity reflected in reviews is only representative of the website

community in question. This community, depending on the product range, is only a fraction of overall customers, who, in addition, tend to be more web / IT aware and possibly younger, although this does not necessarily hold as was shown in the survey results chapter 4. From a sampling point of view, research bias is minimal, since users do contribute reviews voluntarily and in their own time, which other (*more traditional*) sampling methods often cannot guarantee.

6.4.1.3 Trading Books Reviews (part ii) Data Collection

Several issues emerged during data-extraction task for book reviews. The most important issue was the consideration of what particular books would be most relevant to financial-markets. Amazon essentially provides two means of accessing its book catalogue; either using its search feature, based on a keyword relevance search, or based on Amazon's book directory, which is a hierarchical set of topical book categories. A list of candidate books for possible download was put together, during a qualitative review of the catalogue. This list is presented in tables 6.8, 6.9, 6.10 and 6.11, and the decision for inclusion was made based on whether there were enough books returned for a search or topical category, but also consideration was given for selecting Stockmarket relevant topics. These topics would appeal to amateur, hobby and other financial market enthusiasts. For example, technical analysis and fundamental analysis are probably the two most popular streams of thought on how to value and analyse share-prices, and they have wide acceptance within amateur and semi-professional market trading circles (Francis 1988, Murphy 1999, O'Neil 2002).

In the mentioned tables the column "*Amazon Categories*" contains the specific categories used to access books, on the topic highlighted in the first column "*Topic*", and the middle column "*Search Keywords*" contains the keywords used in the catalogue search³².

³² During the catalogue category request; results lists were sorted in *bestselling* order for category returned result-sets, as this was shown to be the most optimal ordering, using a trial extraction based on all possible sorting methods for a test topic. During the search feature request; results were sorted based on search keyword relevance order.

Table 6.8 – Download Group 1 (*Stockmarkets, Economics, Beginning and Advanced Investing*)

Topic	Specific Keywords	Amazon Categories
Stockmarkets	Stocks, Stock-market, Markets, Financial Markets, Finance	Books › Business & Investing › [Finance Investing Personal Finance] › Printed Books
Economics	Economics, Economy, Economic theory, economics textbook, Banking, Macroeconomics, Microeconomics	Books › Business & Investing › [Popular Economics Economics] › Printed Books,
Beginning Investing	Beginner Investing	
Advanced Investing	Advanced Investing	

Table 6.8 represents *group 1*, i.e. broad and basic Stockmarket and Economics related literature. Sometimes very similar keywords were used, such as “*Economics*” and “*Economy*”. Both keywords would; however, return slightly different result-sets and hence more books and reviews could be retrieved. One major limitation imposed by Amazon, is that search results only display top 1,200 book-items (*Amazon only allows to view the first 100 pages of 12 items per page*), yet the actual number of books is even smaller since some of the results, are fan pages, wishlists, posters, audio-books and other irrelevant items. Also, since there is no need for books without reviews, it was a requirement for a book to have at least one review. *Group 2* (table 6.9) relates to more specific market and trading topics, than *group 1*.

Table 6.9 – Download Group 2 (*Investing, Trading, Trading Term, Strategies, Technical and Fundamental Analysis*)

Topic	Search Keywords
Investing	Investing
Trading	Trading
Trading Term	Day Trading, Swing Trading, Long term Investing
Strategies	Trend Trading, Trading Strategies, Algorithmic Trading
Technical Analysis	Technical Analysis
Fundamental Analysis	Fundamental Analysis

Book topics in *group 3* (table 6.10) represent further specific assets, such as; derivatives, commodities, real estate and other. Job hunting and careers related books can potentially exhibit some relationship to the unemployment rate³³ which is why they are considered.

³³ Recently a strong connection between Job searches on search engines and real unemployment rates has been shown (Askitas and Zimmermann 2009). On the same basis, a similar relationship might exist with book reviews.

Table 6.10 – Download Group 3 (*Currencies, Commodities, Real Estate, Derivatives, Bonds, Funds, Job Hunting / Careers*)

Topic	Keywords	Amazon Categories
Currencies / FOREX	Currency, Forex, Forex Trading, Currency Trading, Euro, Eurozone , Dollar	
Commodities	Commodities, Gold, Oil, Grains	
Real Estate		Books › Business & Investing › [Real Estate] › Printed Books
Derivatives	Derivatives, derivatives markets, derivative trading, futures, options	
Bonds	bonds investing, municipal bonds, bond market	
Funds	Mutual Funds, Hedge Funds	
Job Hunting / Careers		Books › Business & Investing › [Job Hunting & Careers] › Printed Books

Group 4 essentially represents the control group, using book topics that have little or nothing at all, to do with the financial markets, such as Science Fiction, or World War II, see table 6.11.

Table 6.11 – Download Group 4 (*Control group: Business, Science Fiction, World War II*)

Topic	Amazon Categories
Science Fiction	Books › Science Fiction & Fantasy › Science Fiction › Printed Books
World War II	Books › History › Military › World War II › Printed Books

Amazon provides several versions of a book, for example, in video or audio formats. For the purpose of this study printed books were only (Hardcover / Paperbacks) extracted. An example screenshot of a typical books results page is shown in figure 6.17.

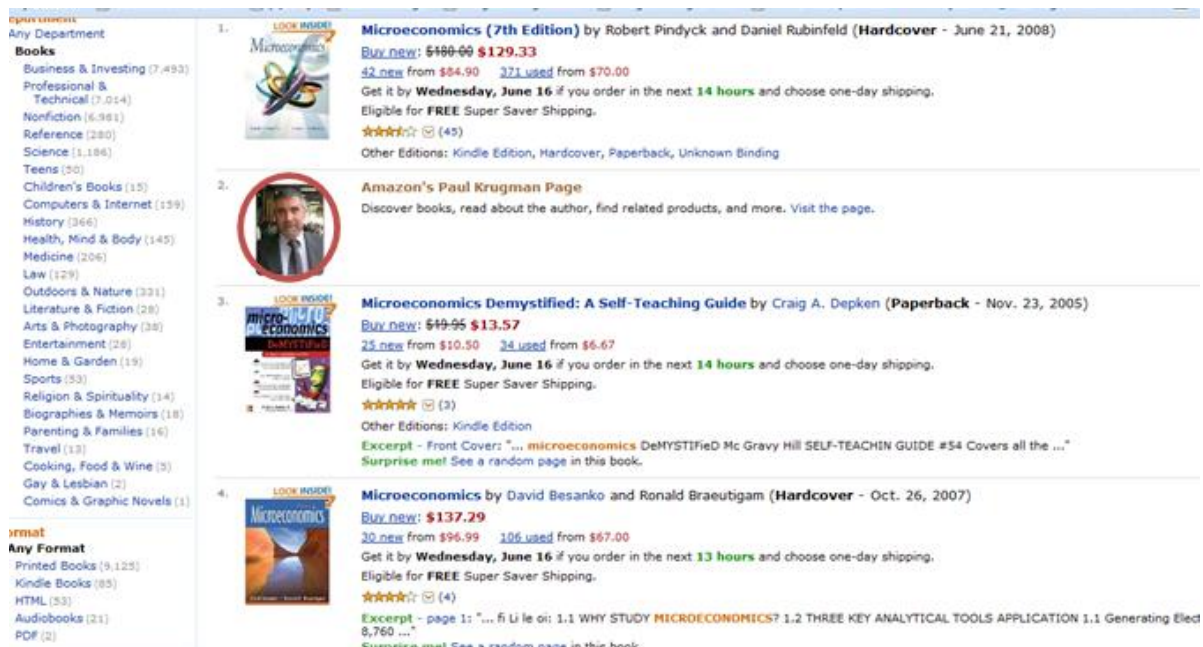


Figure 6.17 – A typical Amazon search / category results page (circled, a fan page – ignored by the crawler)

For each book a number of details are usually available, such as price, publisher, etc. More recently Amazon introduced the *Key Phrases* widget, which provides an insight into the books content via *Statistically Improbable Phrases* (some form of TF*IDF n-gram based method is used, but it is not clear what is used exactly) and *Capitalised Phrases* (possibly noun identification coupled with some form of Gazetteer list lookups, but again it is not clear what Amazon uses exactly), from the book's actual content. These were also extracted into the database repository, where available³⁴.

In summary, book categories to download, were selected based mainly on a-priori anecdotal knowledge about the field of trading and investing culture. Only books that have appeared in print were considered, and the Amazon crawler was extracting date of book publication (and the *Capitalised* and *Improbable Phrases*) and reviews for each book, this includes the 5 star ordinal rating, the usefulness of a rating (*i.e. people who found the review useful, or not*). A book with less than one review would be ignored. In order to, restrict Amazon book search and to return relevant books, search based retrieval was limited to the Business section, which is a top level category for books.

34 View this page <http://goo.gl/tfsFd> for an example. Even more recently Amazon introduced a set of further book contents based statistics, including readability indices, see <http://goo.gl/DGFrc> for an example.

6.4.2 Results: Competing Product Reviews (part i)

6.4.2.1 Overview of Product Reviews

Figure 6.18 illustrates a full monthly summary of all the reviews, with the most frequently reviewed products highlighted. A clear increase in reviews from late 2005 onwards is observable. Most reviews come from last few years, but the dataset goes as far back as 1998.

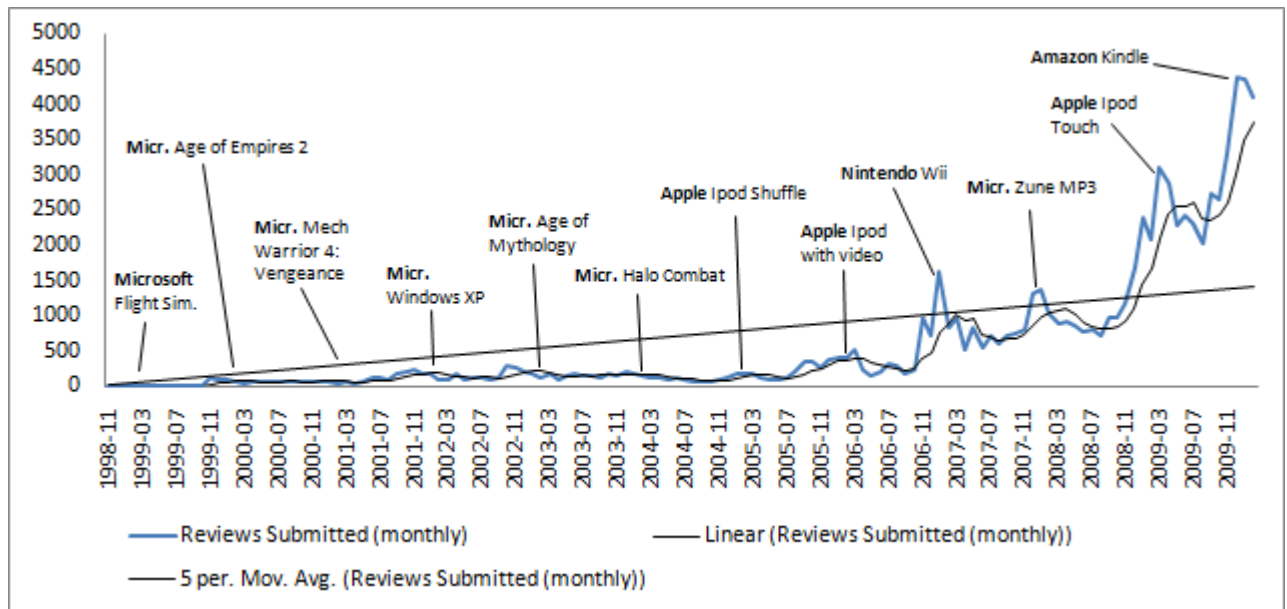


Figure 6.18 – All product reviews aggregated by month and plotted as a time-series, with dominant products highlighted

The early years in the dataset are represented by Microsoft products, several product launches, such as the launch of Nintendo Wii or the Amazon Kindle are visible in the time series as spikes in reviews in later months.

Table 6.12 presents product reviews aggregated for each of the 12 manufacturers from Amazon. Most products were downloaded for Microsoft, followed by HP and Samsung. The review count is not strictly associated with the product count (*i.e. more products downloaded do not mean more reviews*). This is due to a few products being responsible for majority of reviews. Nintendo and Apple are such examples, where Nintendo with only 2 products has approximately same number of reviews as Acer, with 67 products.

Table 6.12 – Basic details of products and reviews, grouped by manufacturers³⁵

Manufacturer	Products	Reviews	Avg. stars	StdDev. stars	Review usefulness	Max reviews rated
Microsoft	407	29362	3.7503	1.4706	0.6554	1289
HP	248	2533	3.7035	1.4917	0.7695	270
Samsung	119	3081	4.2801	1.1376	0.7767	948
Sony	92	4271	4.1765	1.2816	0.6042	2664
Apple	82	16680	4.2245	1.251	0.644	2819
Acer	67	2607	4.2221	1.2083	0.795	832
Asus	58	3618	4.249	1.1852	0.8019	1358
Toshiba	53	742	3.9987	1.2595	0.8084	728
Ectaco	5	35	3.9429	1.145	0.8189	83
Amazon	2	15230	4.2216	1.2643	0.6907	22584
Bookeen	2	10	4.3	0.9	0.925	20
Nintendo	2	2045	4.6465	0.8231	0.5693	3631

When manufacturers are ranked based on how “hotly” reviewed their products were (i.e. companies ranked by the average number of reviews per product), then the ranking is as follows; Amazon, Nintendo, Apple, Microsoft, Asus, and Sony are all, some of the more hotly rated companies' products than the rest³⁶. The reviews for these 6 manufacturers are predominantly positive, except for Microsoft's average rating of only 3.75-star average over all of its 407 products. The average star rating for Nintendo is very high (4.65), given the number of reviews and very low standard deviation, there is clearly an overall high satisfaction with Nintendo's products. At the same time not too much value seems to be placed on these reviews since the usefulness of the reviews in aggregate is only 0.57, the lowest. This means that the reviews on average were found not to be useful by nearly as much as half of the readers who cared to rate a review. Chen et al. (2008) found that more influential reviewers have a stronger effect on buying decisions. They ask why consumers would trust the information provided by strangers they may have never met and how trust is formed among consumers themselves? Credibility is a critical issue in effective information sharing, which involves information reliability and consumer trust. Amazon.com identifies individual reviewers based on a ranking system where reviewers who post more reviews and have a higher number of helpful votes are singled out to other community members.

As far as overall ratings are concerned, 56% of all reviews are 5 stars rated, and the distribution expresses overall positive opinion of reviewers over the dataset, see figure 6.19. Only 16% of all reviews are negative (1 or 2 stars).

³⁵ Each review can be rated as useful or not-useful, the last two columns in the table relate to this review rating.

³⁶ On average for all products in the experiment there are 760 reviews per product, or 43, if the median is used as measure of central tendency. The distribution of reviews per product is highly skewed by Amazon and Nintendo which have received an extraordinary number of reviews. The high review count for Amazon is due to Amazon Kindle which can be attributed to the frequent advertising on Amazon webpages.

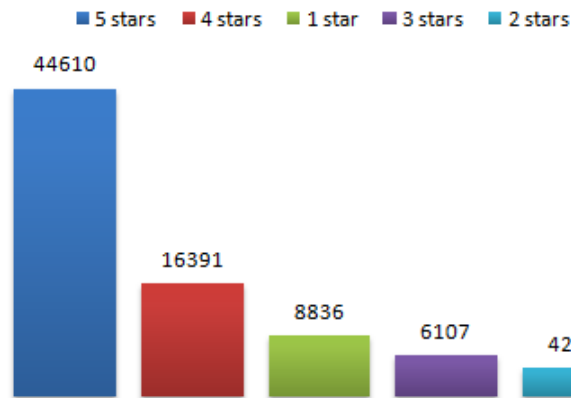


Figure 6.19 – Review star ratings distribution – over all manufacturers (1,137 products, and 80,214 reviews)

The tendency of overall reviews being positive is an observation compatible with the distribution of star ratings, as reported by Chevalier and Mayzlin (2006), who also conclude that ratings tend to be on average positive, see table 6.13 for their results.

Table 6.13 – Review star ratings distribution, two online shops (Chevalier and Mayzlin 2006) – dataset of ~6,000 books

Amazon		Barnes and Noble.com	
Star Rating	Percentage	Star Rating	Percentage
1 star	8.97	1 star	3.44
2 stars	7.53	2 stars	4.07
3 stars	10.56	3 stars	6.00
4 stars	19.89	4 stars	19.27
5 stars	53.05	5 stars	67.22
Average Rating		Average Rating	
4.01 stars		4.45 stars	

Interestingly, this tendency extrapolates to other online communities, Resnick and Zeckhauser (2003) find that reviews on Ebay have a lot less variance, compared to the reviews presented within our study, where on average only about **0.5%** provide negative or neutral feedback. That is 48.3% of buyers on Ebay provide no feedback on transactions and 51.2% provide positive feedback. This is partly explained by the fact that on Ebay both sellers and buyers rate each other, which results in incentive to post positive reviews by the buyer that are in turn reciprocated by the seller. On Amazon a user who hasn't bought the product via Amazon can still review it, and in fact with some work-arounds it might be possible for a user to submit numerous reviews for a product. Chevalier and Mayzlin (2006) briefly address this issue, but dismiss it using pre-order figures of a popular book.

As was already mentioned, each review on Amazon can also be marked as *useful* or *not useful*. This is a feature that tends to be used by Amazon users frequently, as only 127 out of all 80,214

reviews have not been assigned a rating of usefulness – this means that nearly all reviews at some point were read, and hence reviews are actually used. Only about 9% of all 80,087 reviews were not considered helpful, which points to a relative usefulness of reviews. An interesting observation highlighted by the rating feature is, that there is a number of “*super popular*” reviews, which get rated as useful by readers in the thousands. The “*super popular*” reviews tend to be more in the form of well written articles than average, simple few line summaries about products. Pearson Correlation coefficient between text length and a review being marked as helpful is 0.196, statistically significant ($p < 0.01$, two-tailed). For example the most frequently liked review in the dataset was for “*Amazon Kindle 6" Display*”, titled “*BEWARE of the SIGNIFICANT DIFFERENCES between Kindle 1 and Kindle 2!*”, the reviewer took great care in thoroughly comparing and describing technical intricacies and issues with the product, and in fact this review was a negative (1 star review), yet 95% of all known readers found it to be useful (out of 22,584). On average, a review was found helpful by 434 readers, the mean is biased by the very popular reviews and therefore the median is more appropriate measure of central tendency which still shows 183 readers found individual reviews helpful.

Table 6.14 presents the most frequently used titles in positive (4 and 5 - star) and negative (1 and 2 - star) reviews. Positive reviews occur more often, and the phrases in review titles tend to clearly express the sentiment of a review e.g. “*I Love it*”, “*Excellent*”, “*Amazing*”, “*Don't waste your money*”, “*junk*”, “*very disappointed*”. Among top neutral ratings (ratings where stars=3) review titles such as “*average*”, “*It's OK*”, “*mixed feelings*”, “*not bad*”, “*could be better*” are common. Among less frequent review-titles we find titles which tend to be descriptive and usually more creative, e.g. “*Two years of lovin' my imac*”, “*two weeks of ecstasy with this new imac*”, “*Fast, beautiful graphics, easy to add memory, couldn't be happier*”, “*Better than 500, worse than 505; light and touch is my preference*”, “*The fastest computer? Absolute rubbish...*”, “*meh - glad it was a gift, and I didn't waste my money*”, “*Very disappointed, not worth the price*” are common.

Table 6.14 – Top review titles used for positive and negative reviews

Review title (<3 stars)	Occurrence count	Review title (>3 stars)	Occurrence count
<i>Disappointed</i>	75	<i>Kindle</i>	416
<i>Disappointing</i>	30	<i>Great product</i>	389
<i>buyer beware</i>	28	<i>Love it!</i>	383
<i>Don't waste your money</i>	26	<i>Kindle 2</i>	338
<i>terrible</i>	24	<i>Love It</i>	269
<i>very disappointed</i>	20	<i>Awesome</i>	225
<i>junk</i>	20	<i>love my kindle</i>	195
<i>horrible</i>	18	<i>Excellent</i>	177
<i>Defective by design</i>	18	<i>GREAT</i>	170
<i>Kindle</i>	15	<i>Ipod</i>	139
<i>Disappointment</i>	14	<i>Awesome!</i>	137
<i>Not ready for prime time</i>	14	<i>I love it!</i>	127
<i>Too Expensive</i>	14	<i>Kindle review</i>	120
<i>waste of MONEY</i>	13	<i>Amazing</i>	118
<i>Not worth it</i>	13	<i>Great Mouse</i>	116

The dataset in this section confirms results already reported by Chen et. al (2008), who found that neutral reviews tend to be longer since they explain positives and negatives to defend the neutral view, (see relevant mean values in appendix table D.1).

6.4.2.2 Product Groups: Consoles

Within the consoles market the gesture / motion controlled gaming console Nintendo Wii, Microsoft's entry product into the lucrative consoles market – Xbox, and the newest generation of Play Station consoles – PS3 from Sony were selected for a comparative analysis. There are various editions and bundles (14 for Playstation, 15 for Xbox, 2 for Wii) as sold on Amazon which were downloaded. All consoles together contained 5,477 reviews. The table 6.15 indicates that Nintendo Wii is perceived as the best console on Amazon, followed by Sony PS 3, and then Microsoft Xbox, as based on the share of 5 star reviews against 1 star reviews. It has been empirically shown by Chevalier and Mayzlin (2006), that 1-star reviews matter the most, and that ratings in between 1 and 5 stars don't matter too much (Hu et. al 2009). To compute the overall ranking one can assume r_j^c is the star rating of star c (i.e. $1 \leq c \leq 5$) for a product group j (i.e. Sony PS3, Nintendo Wii,...), then the ranking is simply based on $d=r_j^5-r_j^1$.

Table 6.15 – Extreme ratings for game consoles

	5 star share	1 star share
Nintendo Wii	79.1%	2.0%
Sony PS3	70.6%	6.3%
Playstation Xbox	62.9%	13.5%

The review frequency for Nintendo Wii, observed over time in figure 6.20, exhibits yearly seasonality, just after the December holiday season, i.e. January³⁷. Figure 6.20 shows monthly and figure 6.21 daily charts, from which various product related events can be appreciated.

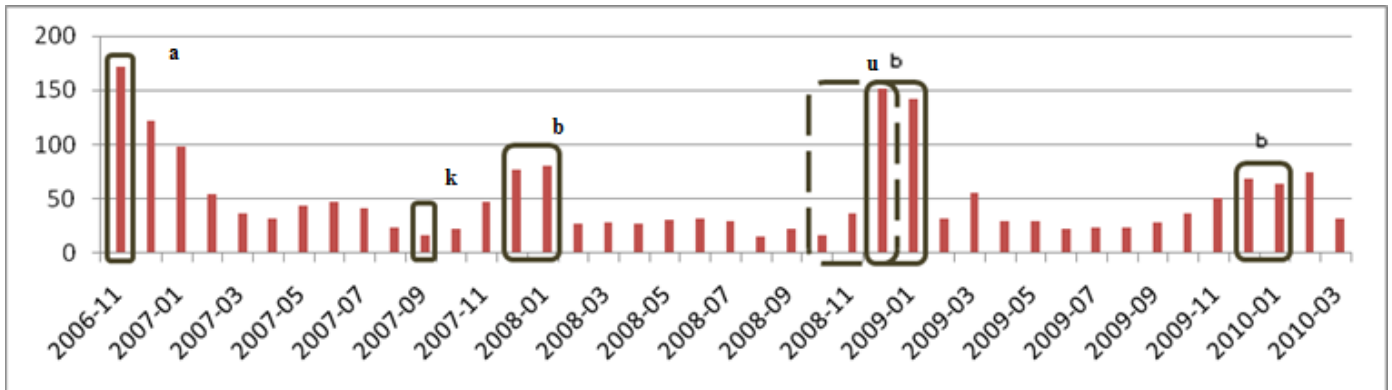


Figure 6.20 – Review frequency chart for Nintendo Wii (monthly aggregated reviews)

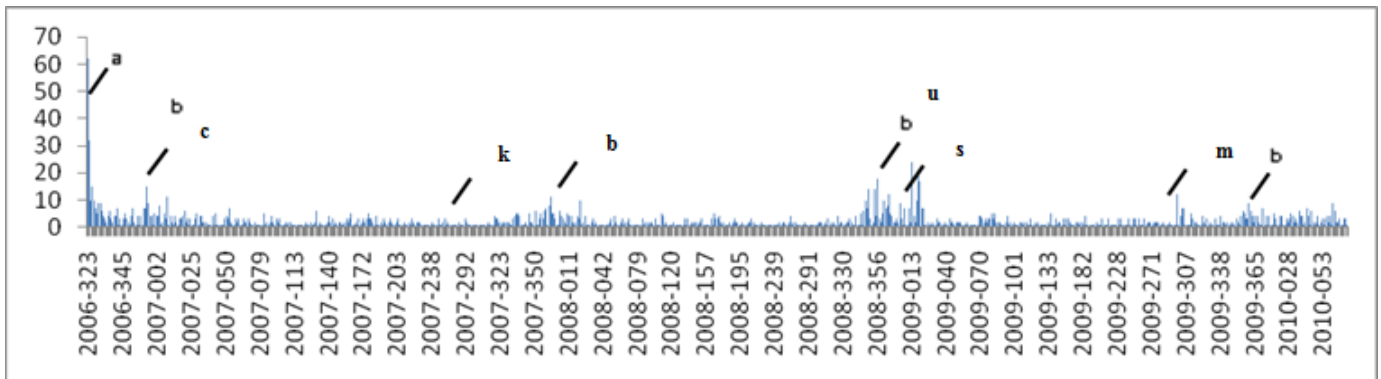


Figure 6.21 – Review frequency chart for Nintendo Wii (daily aggregated reviews)

The Nintendo Wii was launched on November 19th 2006. The first spike in figure 6.20/21 represents the launch day and indeed a large proportion of the months' reviews occurred on the actual day of the 19th November. Over the subsequent months of December, January and February the inflow of reviews was high but steadily decreasing, as the initial excitement faded. The highlighted months of December and January (*b*) represent the yearly cycle of calendar based holiday reviews, for example (*c*) in figure 6.21 points to the 27th December, when an influx of reviews occurred just after Christmas and the boxing days. Further example events are listed in the bullet points below:

- **a** – 19th Nov. 2006, Nintendo Wii launched
- **b** – Dec./Jan., Seasonal Holiday Sales

³⁷ The high frequency of reviews in January, likely points to the lag of review *readiness* by new users of products purchased during the December holiday season.

- **c** – 27th Dec. 2006, influx of reviews after Christmas 2006
- **k** – 12th Sep. 2007, Wii surpassed the sales of Xbox and became market leader³⁸
- **u** – October 2008, Nintendo announced that between October and December 2008 the Wii would have its North American supplies increased considerably from 2007's levels, producing 2.4 million Wii units a month worldwide, compared to 1.6 million per month in 2007³⁹.
- **s** – 3rd Jan. 2009, In the UK, the Wii leads in home consoles sales with 4.9 million units sold⁴⁰
- **m** – 20th Nov. 2009, Nintendo's limited-edition black Wii launched in UK⁴¹

Console reviews for three competing manufacturers were explored in this sub-section. A clear sentiment ranking based on extreme negative and positive reviews, corroborates with the real world interpretation of sales of these three manufacturers. Further analysis of Nintendo Wii reviews showed a number of patterns observable over time. Review submissions were also found to be efficient in terms of reviews being submitted on a given product launch for example. Even though originally it was expected that correlation of events with date-stamped reviews would be low, basing this on the assumption that users submit reviews whenever this might be convenient to them, rather when some auxiliary event takes place. From data analysed it seems that this is a mix between the two, since for example, during the Christmas period, (*c*) in figure 6.21, reviews were submitted after the holidays, on the 27th December, suggesting that reviews were made when this was convenient⁴².

6.4.2.3 Product Groups: E-Books Readers

The general advances in computing technology have created a new and highly competitive market of E-Book readers, with the ambitious goal of taking over newspaper and book print into the electronic domain. The highly successful Amazon product Kindle (its two marketed editions account for an impressive 15,230 reviews⁴³), Sony's serious attempt to enter the market with Sony Reader (718 reviews), the only just freshly released Apple Ipad (61 reviews, since this is a new product with reviews to accumulate as more are sold), and the seemingly less popular but serious competitors Bookeen Cybook (10 reviews, 2 product editions) and Ectaco jetBook (35 reviews, 5 product bundles) were downloaded.

38 Sanchanta M., 2007. Nintendo's Wii takes console lead, Financial Times, Last accessed on 2010.05.26 at <http://www.ft.com/cms/s/0/51df0c84-6154-11dc-bf25-0000779fd2ac.html>

39 Ashley P., 2007. Can't Find a Wii? Take a Rain Check, ABC News. Last accessed on 2010.11.23 at <http://abcnews.go.com/Technology/GadgetGuide/story?id=4001054&page=1>

40 Ryan K., 2008. E3: Nintendo Wii pulls ahead of Xbox 360 in console sales. San Francisco Chronicle, Last accessed on Retrieved 2010-11-22 at http://www.sfgate.com/cgi-bin/blogs/techchron/detail?entry_id=28286

41 Cnet UK, Crave, 20 October 2009, Nate Lanxon, <http://crave.cnet.co.uk/gamesgear/0,39029441,49303993,00.htm>

42 Note that this pattern doesn't hold for other holiday seasons.

43 Other products are at a disadvantage since Amazon has marketed its own product heavily and the competition has received little attention in comparison.

Table 6.16 – Extreme ratings for e-book readers

	1 star share	5 star share
Amazon Kindle	9.2%	62.5%
Bookeen Cybook	0%	50%
Ectaco jetBook	5.7%	40%
Sony Reader	11.1%	43.0%
Apple Ipad	24.6%	37.3%

From table 6.16 one can appreciate the order of sentiment based ranking from top to bottom, as perceived by the reviewers on Amazon. The Bookeen Cybook ranking is based on only 10 reviews and hence its' ranking is not significant. The reviews for apple Ipad were on average 3,533 characters long, compared to 840 characters mean review length for all E-books. This can be explained by innovative products, such as the apple Ipad was getting detailed and exhaustive coverage within reviews, during a short period of time; however, its rating is the most negative one too.

6.4.2.4 Product Groups: Netbook Computers

The popularity of mobile computing brought with it light and portable low spec style computers that have become extremely popular in recent years, with a maturing market of competitors. Reviews for Asus, one of the pioneers of Netbook computers (30 netbooks, 3,173 reviews), and a number of other well known computer manufacturers, Acer (26 netbooks, 2,064 reviews), Samsung (21 netbooks, 923 reviews), and HP (18 netbooks, 365 reviews) were downloaded. The table 6.17 highlights the sentiment rankings as based on star-ratings.

Table 6.17 – Extreme ratings for e-book readers

	1 star share	5 star share
Samsung Netbooks	2.8%	72.6%
Asus Netbooks	7.6%	61.2%
Acer Netbooks	7.9%	59.1%
HP Netbooks	14.2%	37.5%

Given the low percentage of 1-star ratings and the very high proportion of 5-star ratings for Samsung Netbooks it would indeed seem that this manufacturers produces netbooks that receive the best online word-of-mouth recommendations. Clearly this is not the case for HP Netbooks.

6.4.2.5 Microsoft vs. Apple

Apple and Microsoft are two long standing rivals, and to compare their performance in terms of Amazon review sentiment is interesting. The product ranges under analysis for both companies were outlined in section 6.4.1. From the two tables that follow (table 6.18 and table 6.19) one may appreciate the relative review sentiment performance. Microsoft has a small 5 star share (only 45.7%) and a large 1 star share (15.1%). Microsoft's operating systems, numerous peripheral products and Zune multimedia player especially have systematically received very low ratings. Apple has a much better distribution of positive ratings with 62.3%, 5 star ratings. Apple received very positive ratings for its "Ipod Touch", PC's and Laptops in contrast to a lower rated "Ipod Shuffles".

Table 6.18 – All star ratings for Apple and Microsoft reviews

	1 star share	2 star share	3 star share	4 star share	5 star share
Apple	8.5%	4.1%	5.9%	19.1%	62.3%
Microsoft	15.1%	7.7%	10.0%	21.5%	45.7%

Table 6.19 shows a breakdown of sentiment as it evolved over 4 distinct time periods of over 10 years. Apple has steadily kept a better sentiment than Microsoft with lower standard deviations for the star ratings. It is interesting to note that prior to 2002 Microsoft products were very actively rated. Apple reviews on Amazon jumped by 834%, for the 2006-2008 time-period. Judging from the reviews alone, Microsoft rated products are the ones that are more frequently rated, which is a reflection of more widely used products (as of 2010).

Table 6.19 – Review summaries for Microsoft and Apple through the years 1998-02, 2003-05, 2006-08 and 2009-10

	Microsoft			Apple		
	<i>Star Avg.</i>	<i>Star. StdDev.</i>	<i>Count</i>	<i>Star Avg.</i>	<i>Star. StdDev.</i>	<i>Count</i>
1998-2002	3.72	1.48	4'111	4.39	1.30	41
2003-2005	3.64	1.48	4'090	4.03	1.32	1'058
2006-2008	3.64	1.50	12'551	4.20	1.26	8'826
2009-2010	3.86	1.42	8'563	4.11	1.27	6'516

The review sentiment and review counts in table 6.19 can be compared and contrasted with figure 6.22. This figure presents the stock-prices for *Microsoft* and *Apple* on a log scale, i.e. up / down moves are unit independent. Effectively, one could use total market capitalization (*i.e. outstanding shares * avg. share price*) during these time periods to assess the value of both companies in absolute terms, however, share price changes compared to each other on log-scale

show the rate of growth of the companies more appropriately (*the prices are stock-split adjusted*). The stock-prices roughly approximate the overall better rankings of Apple's product reviews throughout Amazon, its shares growing at a much more significant pace, as review frequencies increased over time. Clearly this connection is only indicative and one may only postulate that the increase in reviews and their overall sentiment highlights how good products are, which in turn affects real or at least potential sales, and then the sales and sentiment in aggregate are projected through the share-price into the Stockmarket up and down moves (i.e. rate of growth of the individual companies). It is likely that, if any such connection does exist, that it works over a longer-term, rather than short term period. In the short-term there are too many random / noise and other factors affecting a price; however, a strong profile on a review site for a manufacturer may be indicative overall of a well performing stock.



Figure 6.22 – Apple Inc. (blue line) and Microsoft Inc. (red line), stock-prices plotted on a log scale (base 0, 1999-2011)

6.4.2.6 Other Review Related Considerations

The presented product review dataset allows investigation of further useful and relevant topics. In particular, it has been shown by Chen et. al (2008) for books and confirmed by us for products as well, that neutral reviews (3-stars) tend to be longer on average – this is probably due to a natural tendency of expressing mixed, positive and negative points in neutral reviews. It would be interesting to explore whether there is, on average, any variation in complexity of the free-form reviews text. To the authors knowledge this has not been commented upon in previous literature, and it could be useful in understanding any possible difference in such free form text reviews.

The methodology used was straight-forward, a random sample of 500 reviews⁴⁴ was selected from 5-star rated reviews and as much data as possible for 1-star, 2-star...4-star ratings, from all the Nintendo Wii reviews. The Nintendo Wii reviews were chosen on purpose since given overall review sentiment being highly positive; it was of interest to see whether negative reviews differed in writing style. For all of the five review groups, Lexical density, Gunning-Fox index, avg. syllables per word, avg. sentence length, and sentences per review were computed. Where **Lexical density**, represents a measure of content per functional (grammatical) and lexical units (lexemes)⁴⁵. This is a simple ratio $L_d = \frac{N_L}{N}$ where L_d is the lexical density, N_L is the number of lexical word tokens (*i.e. nouns, adjectives, verbs, adverbs*) and N is the number of all tokens (*total number of words in the analysed text*). **The Gunning-Fog index** (Gunning 1952) is a measure of readability of a sample of English text. The resulting number is a rough estimate of the number of years of formal education needed to understand the text on a first reading⁴⁶, and its computation is defined as,

$$G_F = 0.4 \left(\frac{W_N}{S_N} + 100 \frac{C_N}{W_N} \right)$$

where G_F is the Gunning-Fox index, W_N is the number of words in a sample full passage of text (*about 100 words, no broken sentences*), S_N is the number of sentences in the sample text, C_N stands for the count of complex words (*that is words with three or more syllables, not including proper nouns, familiar jargon, compound words, or common suffixes*). The remaining features don't need an explanation. Table 6.20 shows that except for the 3 star reviews being marginally higher the Gunning-Fox isn't significantly higher for other ratings, and in fact one may conclude that Amazon reviews are overall easy to understand across all rating levels, with no substantial differences.

44 The software used for the evaluation of these features imposed a strict word limit.

45 This measure is also known as a text complexity measure, and represents a descriptive statistic for text where text with lower lexical density measure is generally considered to be more easily understood. The measures' value will vary depending on the source and style of writing (Ure 1971).

46 A Fog-index of 10 has the reading level of a UK A-level student. A piece of writing that is desired to be understood by a wide audience generally requires having Fog-index of 10 or less. The index was developed by Robert Gunning – see Gunning (1952) for more details. There are similar readability statistics; “Flesch Kincaid Reading Ease”, “Flesch Kincaid Grade Level”, “SMOG Index”, “Coleman Liau Index” and “Automated Readability Index”. “Coleman Liau” and “Automated Readability Index” rely on counting characters, words and sentences, whereas the other indices are based on counts of complex words (*i.e. polysyllabic words*).

Table 6.20 – Features of complexity of language, as they vary over ratings

NINTENDO WII	<i>Lexical Density</i>	<i>Gunning- Fox</i>	<i>Avg. Syllables per word</i>	<i>Avg. Sentence length</i>	<i>Sentences per review</i>
5 Stars	34.20%	7.40	1.54	15.64	12.83
4 Stars	31.50%	7.90	1.57	16.53	14.48
3 Stars	33.30%	8.50	1.56	17.5	14.68
2 Stars	31.60%	7.5	1.52	15.75	16.60
1 Star	37.10%	7	1.53	15.80	12.83

6.4.2.7 Summary

It was shown on an example of two major technology companies, Microsoft Inc. and Apple Inc. that a connection between longer term market performance and Amazon relative frequency of reviews and sentiment exists. These two companies were compared to the markets since most of their leading product line reviews were available for analysis, which wasn't the case for other manufacturers in this dataset. In future work it will be useful to confirm the indicative result obtained by this study, on a larger set of examples. In order to collect enough reviews for analysis other product opinion websites such as Eopinion.com could be used to extract additional reviews for analysis. A detailed analysis of the main games-consoles market, especially the Nintendo Wii console was undertaken. The ranking of top consoles was shown to mirror the real world relative success. Information propagation via reviews was found to be efficient and a number of previous results from literature were also confirmed. Accurate sales figures for the Netbook and E-Book market weren't available for this study, hence Amazon based review results were only presented and comments were made. The question of text complexity was studied and it was discovered that no substantial differences in the language used in differently rated reviews exist, except some detectable difference in neutral 3-star rated reviews. Some of the results present an original contribution, even though the study has certain limitations.

6.4.3 Results: Trading Books Reviews (part ii)

Altogether over eighteen thousand unique books and nearly half a million reviews were retrieved on the topics highlighted in section 6.4.1.3, which works out overall to about 26 reviews on average, per book. A look at book publishing frequencies distributed over time shows that most books in our sample have been published during the last 20 years (~95%). Due to the nature of publishing and the many steps involved, the publishing industry might be sluggish in responding to emerging topics of interest, as it can take numerous months for books

to get to print. In this part of the experiment it is hypothesised that UGC contributed in the form of book reviews can carry contextual information relating to current events in the financial markets. In other words, the reviews might be timely with the review putting a book into the current context, i.e. financial world events, especially with major financial events taking place. Hence such reviews would hold informative value, relevant to the books but possibly a much wider scope of insight, than purely covering the book.

As was mentioned in the introduction to section 6.4, a recent research project in co-operation with Google Books has resulted in a new social science research tool, which allows performing within books keyword searches over the history of several centuries, of book publishing. This is possible due to Google having digitized many books over the last few years for its Google Books program⁴⁷; however, legal copyright issues in making this data widely available, had to be avoided by storing and processing the text as n-grams⁴⁸. The exact details of the technical implications and issues to consider when interpreting results from this tool are discussed at length in Michel et al. (2010). The figure 6.23 generated by this tool can confirm that there is some obvious connection with historical book content and specific prevailing temporal interests, within the financial markets. It illustrates the occurrence of the term “*short selling*” within books being much more common just after the 1929 Stockmarket – *Great Depression* crisis, and then during the early 2000s short-selling becomes a topic of interest again, and it was widely seen as the most common strategy within hedge funds (see Jaeger 2002, for example), and becoming popular amongst traders, again.



Figure 6.23 – “*Short selling*” as mentioned throughout books over time, the two peaks highlight the 1929 recession, and the modern day trading period⁴⁹

47 See <http://books.google.com/intl/en/googlebooks/about.html> for much more information on the project.

48 N-grams are a standard text-processing representation method, essentially a adjacent sequence of n items from a given sequence of text.

49 In order to avoid skew of results (with many more publications) towards modern days, the book count is normalised by the number of books published in each year (*y-axis*), and only books with at least 40 mentions of an n-gram are considered. Hence, the reason for the 1929 peak being so high is due to there being fewer books

The book data from Amazon also confirms there being an increase during the 2000s in “*short selling*” within contents of books (within contents book search is possible as Statistically Improbable Phrases and Capitalised Phrases were downloaded for many books; see section 6.4.1.3). The question can then be extended to, whether certain topics of temporal interest to the financial markets (i.e. such as the sub-prime mortgage crisis, banking collapse, Lehman Brothers...) also prevail within user generated reviews, to what extent, and what information do they provide. That is, do time-stamped user contributed reviews reflect current issues to some extent? Clearly one may expect reviews to be correlated with the number of books brought onto the market. This section (part *ii*) of the Amazon study looks at; **1**-Simple review frequencies within various topical groups of books, **2**-Reviews themselves are analysed for associated text that may relate to recent financial events. The frequency of reviews may also point to a topic of emerging interest, before the book publishing industry has a chance to respond. Hence, it seems sensible that a measure of interest into a given topic could be approximately represented by reviews.

6.4.3.1 Book Review Frequencies and Ratings

Reviews submitted on topics outlined within section 6.4.1.3 were aggregated by month and their review frequencies and average star-rating based sentiments calculated. Month was chosen as the most adequate aggregation time-unit since number of books published per month is more meaningful than shorter time-periods, and hence can be readily compared with reviews for each month as well. Figure 6.24 reveals new monthly review submissions for the baseline topics “*science fiction*” and “*World War II*”.

published overall; the peak appears much higher than the more recent 2000s peak.

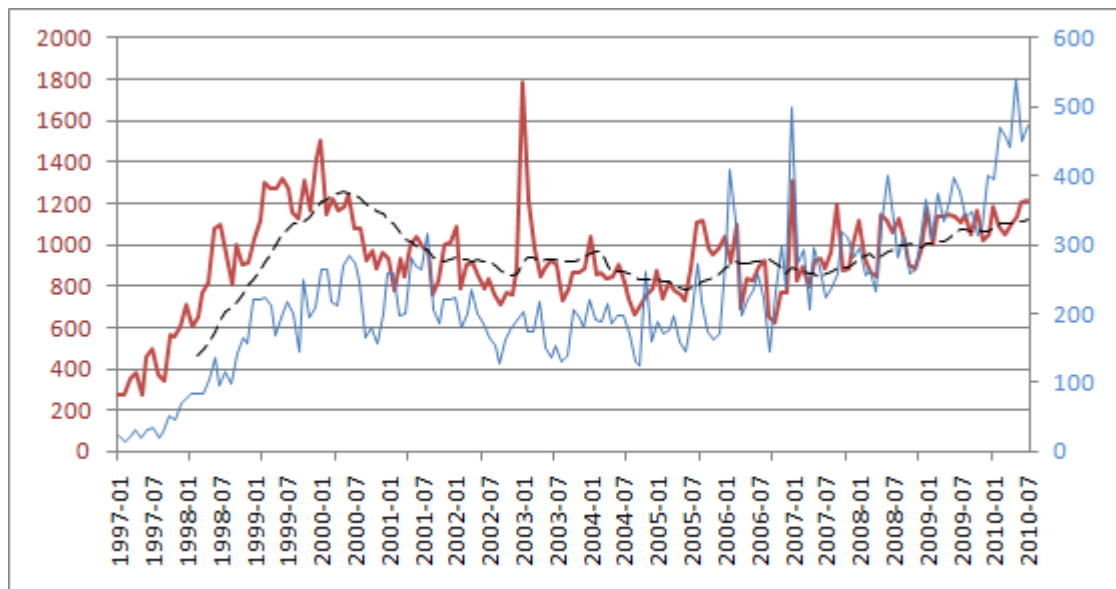


Figure 6.24 – Science Fiction (in red) and World War II (in blue) related book reviews (black dashed line is a 20 day moving average)

It can be observed from the initial years between 1997⁵⁰ and 2000 how the rate of submitted reviews steadily increased, this likely due to the underlying higher awareness and contributions of online reviews on *Amazon*. It is encouraging that during the period from around early 2001 to 2009, the reviews for both baseline topics have been relatively stable. The spikes visible from the chart represent bursts of reviews in response to best-selling books. For example the spike of reviews for Science Fiction during January 2003, is directly attributable to a Science Fiction book entitled “*Crossroads of Twilight (Wheel of Time, Book 10)*”. Written by Robert Jordan (Jordan 2003), this book came out 10th in a series of related stories, which has a strong cult following; however, this particular book has been very badly perceived by the community. The book was the 6th most rated book in the entire dataset, but it has received the lowest average star-rating as well. This is quite pronounced and observable in figure 6.25 – averaged review ratings per month show that this book has affected negatively the overall average.

⁵⁰ The dataset in this study contains sporadic reviews since mid 1996 (not shown on the chart).

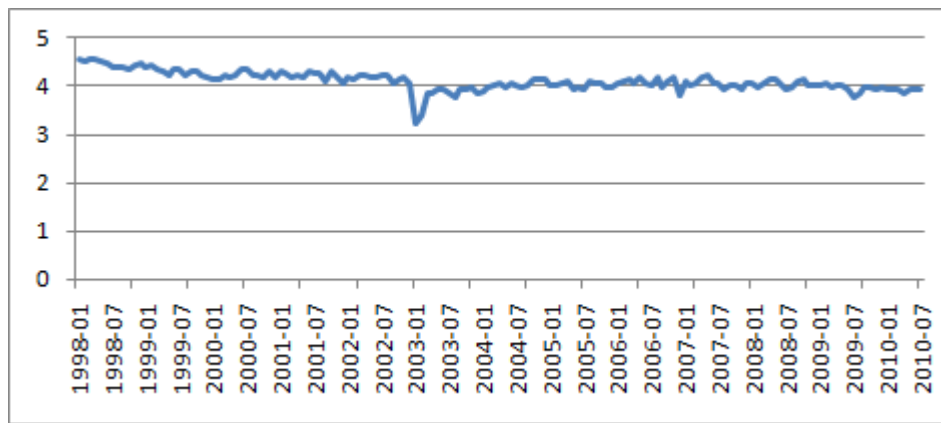


Figure 6.25 – Science Fiction review sentiment (*based on star-rating*) averaged per month

Clearly not every book spurs such an intense interest and decisive agreement on polarity concerning a books' quality, and indeed other topic groups (as identified in 6.4.1.3) have not shown such pronounced reaction to single individual books. Pearson's correlation coefficient for *World War II* and *Science Fiction* reviews was found to be positive and significant, .627 (N=163), p (two-tailed) < .01. The baseline topics provide a good idea of the underlying popularity of Amazon's review feature, and various financial topics can now be examined taking into account this baseline.

Overall, it was found that review sentiment (as expressed by star-rating⁵¹), over time is very stable and tends to be positive. This has been found to be the case for the rating distribution and was also, already suggested in Chevalier and Mayzlin (2006), and confirmed by results on technology products within section 6.4.2. Figure 6.26 illustrates how the rating average over time is indeed quite stable, and never goes below 3 stars, in any month, for the topics investigated.

A-priory it was expected that as different books are rated they might be rated in relation to current events or how the book fits into context of recent events. This does not seem to be the case though. Since the sentiment rating is this stable, it seems of little use for purposes of analysis, and in fact sentiments mostly relate to the books.

⁵¹ Note; there was no need for polarity classification of the free-form text, since the polarity or sentiment for reviews is provided via star-ratings.

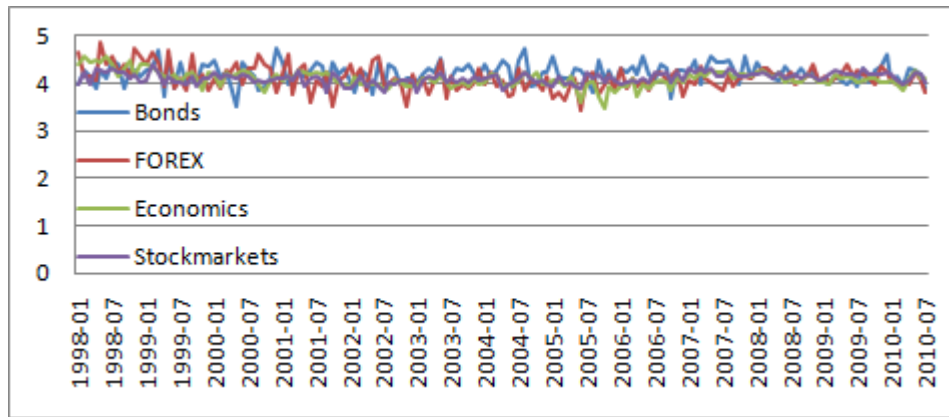


Figure 6.26 – Review sentiment (*based on star-rating*) averaged per month for different topics (*i.e. Bonds, FOREX, Economics, and Stockmarkets*)

As expected, there is a clear connection between books published and reviews submitted each month. Figures 6.27 and 6.28 highlight how for the topic of Stockmarkets, they tend to increase together.

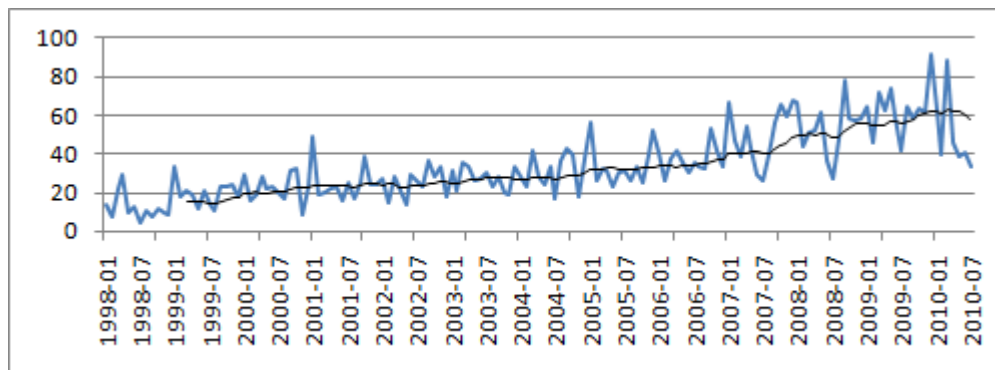


Figure 6.27 – Stockmarket related books published (*black line is the 15 day Moving Average*)

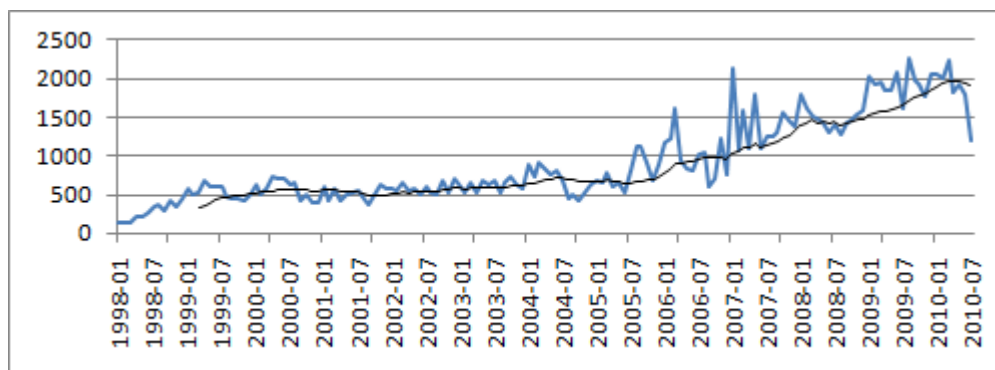


Figure 6.28 – Stockmarket related book reviews (*black line is the 15 day Moving Average*)

Real Estate is another book topic that has been on the rise over the years. Especially during the pre-mortgage subprime crisis period, real estate was a sought after asset and much was published on the topic in the years preceding the 2007-09 crisis. Following the years after the subprime crisis, new publication of literature on real estate dropped significantly. Although

number of reviews each month has not decreased as substantially and there is still strong inflow of reviews each month on the topic of real estate, see figure 6.29 for the books published, and figure 6.30 for the monthly reviews.

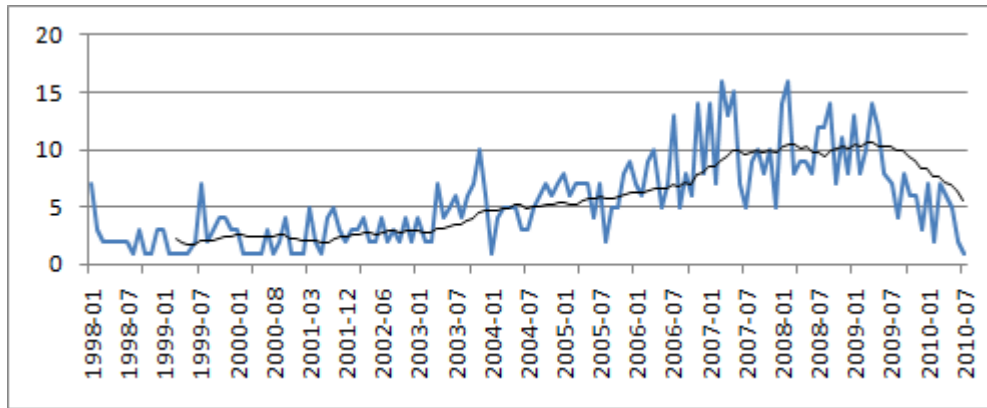


Figure 6.29 – Real Estate books published each month (*black line is the 15 day Moving Average*)

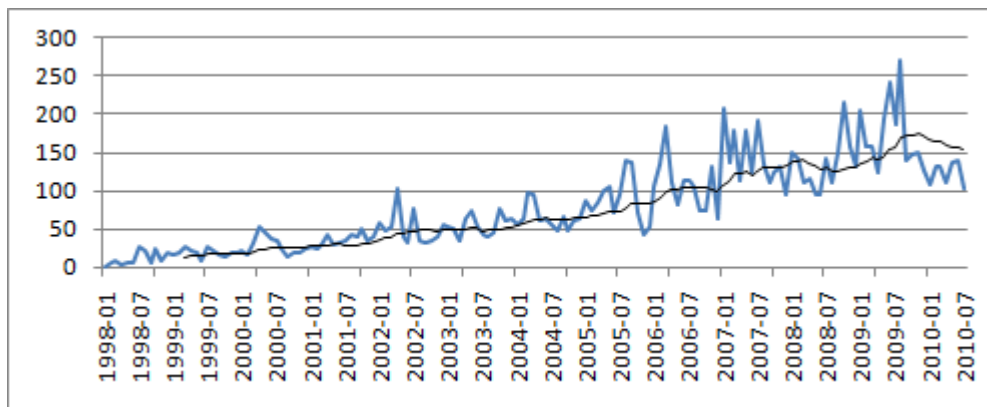


Figure 6.30 – Real Estate related book reviews (*black line is the 15 day Moving Average*)

Some book topics were found to exhibit increasing reviews in spite of relatively steady, constant and even decreasing supply of new books each month. This may point towards a stronger interest into a given topic than what is accounted for, by the number of released books. Arguably contextual information affects reviews rather than the books themselves affecting reviews. Figure 6.31 illustrates the relatively constant number of new books on *derivatives*. Compared to *derivatives* reviews in see figure 6.32; since the year 2007, interest into derivatives has been constantly growing. This could quite likely have to do with people's increased interest to understand the financial crisis, learn about it and ultimately share opinions and reviews.

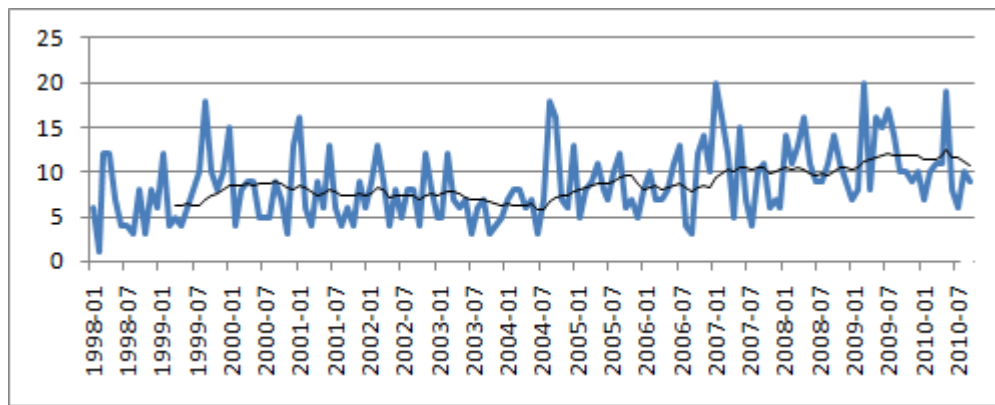


Figure 6.31 – Derivatives books published each month (black line is the 15 day Moving Average)

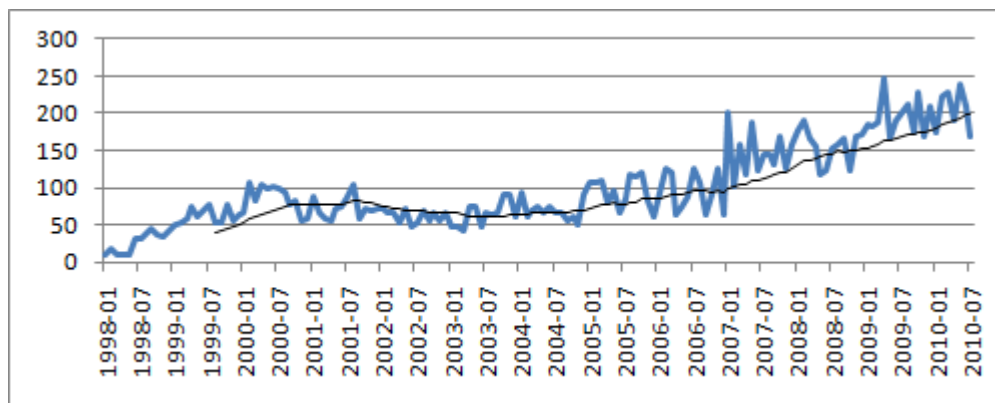


Figure 6.32 – Derivatives related book reviews (black line is the 20 day Moving Average)

Algorithmic trading is a relatively recent field that has emerged in computational finance⁵². The first book on the topic of *Algorithmic trading*, is from 1994 and the first review from 1998, with very few reviews until about 2000. The number of published books per month was small over time. For numerous months there wasn't one single new book published on the topic and the highest numbers of books published a month, were two books. Yet, the reviews per month were slowly, but steadily increasing over the years, which indicates a higher interest into the area of Algorithmic trading.

⁵² Algorithmic trading is increasingly receiving more attention. As of July 2009, 50% of all NYSE stock-exchange trades are attributed to automated algorithmic trades. See *Economist* article "Rise of the Machines", 30th July 2009. <http://www.economist.com/node/14133802>

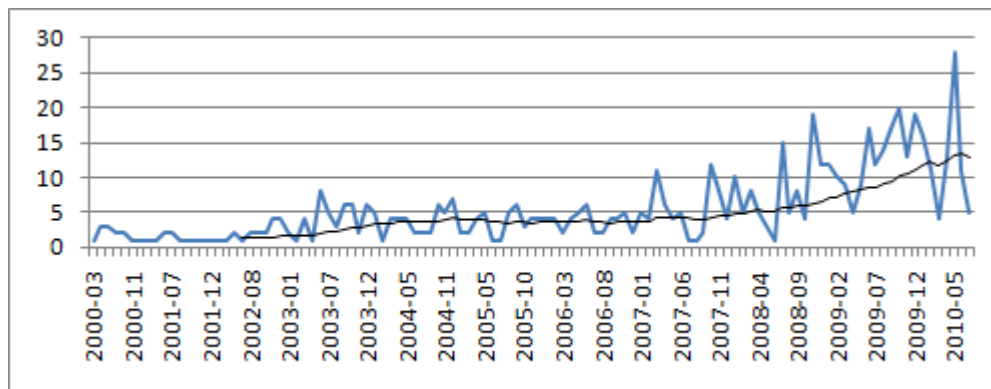


Figure 6.33 – “Algorithmic trading” book related reviews (monthly)

All book groups discussed in the earlier section were briefly analysed, however since no new interesting review frequency patterns were found these shall not be discussed here further, although some charts illustrating topics of book reviews are provided in the appendix *D*.

Given, that general interest into particular topics may be reflected in review frequency, but since most of these reviews directly relate to books it seems quite unlikely that much real world (i.e. financial) events would be mirrored in the book reviews available. Nevertheless a textual analysis of the reviews may reveal some connection to current events and hence in the next section two sets of text analysis experiments are performed to help understand what book reviews are generally concerned with.

6.4.3.2 Text based Book Review Analysis

Since an aggregate direct connection of user generated reviews and the financial markets or recent events could not be convincingly established, a textual analysis of the reviews was undertaken to investigate the actual content of reviews. Each review contains in addition to the review *text* and *star-rating* also a review *title*. This is a concise summary of the review that might be used by the reviewer to focus a readers’ attention to the main points or opinions shared. The years from 2005 to 2010 were analysed for most frequent noun terms within these review titles. Presumably titles of reviews are generally similar for the same polarities and books, with the idea being that any significant changes in the use of nouns within the title can be further qualitatively inspected by a human researcher against the database of reviews, so that it can be established whether the review refers specifically to the book, or whether the review is put into the context of current, financial events. In order to keep this investigation manageable, two financial crisis related topics were selected, specifically *Derivatives* and *Real Estate*. Both represent assets that were directly related to the subprime mortgage crisis and the aftermath that followed. The now relatively widely known OTC (*over the counter*) Collateral Debt

Obligations were the derivatives hiding away unsafe and highly risky mortgaged real estate assets (Brownell 2008).

Table 6.21 presents the top 10 noun-terms for every year since 2005. This is generated from over 9000 review titles over the 6 years. The methodology used to generate this table is similar to the one described in section 3.2.3. Essentially title text was tokenized, tokens lemmatised (i.e. normalised to root words) and frequency of each word over a time period of a year established. In fact the original output returns around 40 top terms from which a human subject can then choose the top 10 terms, since certain tokens, such as “book” occur too often in review titles (e.g. “Great book, explains a lot about Debt Obligations...”) to have any useful meaning.

Table 6.21 – Top 10, *Derivatives* review title terms, as retrieved for each year in the sample (words are normalised, lemmatised)

2005	2006	2007	2008	2009	2010
1.Future	Option	Option	Option	Option	Option
2.Option	Future	Future	Trading	Future	Trading
3.Trading	Finance	Trading	Future	Money	Future
4.Finance	Trading	Money	Guide	Trader	Crisis
5.Investor	Business	Guide	Trader	Guide	Business
6.Guide	World	Trader	China	Review	Information
7.Commodity	Money	Reference	Review	Derivative Market	Money
8.Introduction	Guide	Trader	Market	Strategy	Commodity
9.Money	Work	Business	Business	Business	Wall Street
10.Strategy	Analysis	China	Strategy	Information	Energy

Now that top terms from review titles with a temporal dimension, have been identified (and will be discussed shortly) a second and closely related text analysis is used to extract word associations from the main body of the reviews text.

Word association, also known as word co-occurrence, essentially produces estimates of the degree of word association based on word frequencies in close proximity with the word (Bird et al. 2009). The following methodology was used to detect word associations, where first of all text was tokenized⁵³, then a trained POS⁵⁴ tagger found all the nouns in the review-text, and for each noun token all the following noun tokens in the text sequence that lie within a fixed window (*of 5 words in this case, not counting stop-words*) were added to a conditional

⁵³ In an ideal scenario for word association text would be segmented by sentences, so that neighbourhood search can be constrained to sentences, however with the nature of the review text being highly informal, using dubious punctuation if any, this intermediary step had to be omitted.

⁵⁴ POS – Part of Speech, e.g.: Verb, Adjective, Noun, Pronoun...

frequency distribution for the token noun in question⁵⁵. This can be useful in identifying related terms within actual reviews, which can be based on the top terms already identified from review titles. This helps to gain a quick overview over the relatively large body of 9000 reviews, and is a technique in addition to the already mentioned qualitative investigation of the reviews by a human subject.

From table 6.21, it can be appreciated that Futures and Options are within the top three terms, since they are both the basic derivatives in finance, hence users who submit reviews would probably use them ubiquitously to describe Derivatives related books. Terms such as *Guide*, *Resource*, *Review*, *Reference*, or *Information* occur to some extent independently of book topic (i.e. in both tables 6.21 and 6.22). Further specific terms such as *Trading*, *Finance*, and *Strategy* are common to the derivatives review, and the real estate reviews often employ terms such as *Home*, *Landlord*, *Investing*⁵⁶ and terms relating to both, i.e. *Money*, are shared. More interesting; however, are the new terms – in 2007, **China** has appeared among the top 10 terms of a rather large sample of title terms. In fact China did rise even higher in 2008, and then suddenly disappeared from the top 40 terms as well. Similarly in 2010, the term **Crisis** and **Wall Street** appeared within top terms. This hints to some manifestation of financial real world events, yet the question whether these reviews simply describe new books on these topics or whether the reviewers go as far as introducing these issues into the reviews, is a question to be answered. A qualitative review reveals that 81 reviews mention China in connection to 10 books on the emergence of China as a major economic power, and 58 reviews mention the term Crisis in some manner for 30 different books. Indeed reviews seem to be strongly related with books published on these topics, than anything else. However, this is not always the case, as for example a book on LEAPS⁵⁷ option derivatives, entitled “*The Alpha Hunter: Profiting from Option LEAPS*”, was essentially unrelated to China; however, one of the reviewers mentions that a case study in the final chapter of the book presents an excellent introduction to Chinese economy. The author of the review considered this important enough to mention it in the title of their review. In the case of Crisis related reviews, a number of reviews actually refer to the financial Crisis of the later 90s and early 00s, and all investigated reviews essentially referred to books that explicitly dealt with *Crises*. It was quite interesting to discover for example a book from 2005, “*Running on Empty*” which clearly and quite accurately anticipated the deficit issues leading to a major financial crisis during recent years. The more recent books are explicitly dedicated to the recent financial crisis and reviews generally reflect and relate to the

55 The Python code responsible for achieving both text analyses is available on <http://www.newsmental.com/thesis>

56 Note that a derivative trading is generally considered a shorter term activity hence the reference to *trading* in derivatives related reviews, whereas reference to *investing* within the longer term and less liquid real estate.

57 LEAPS stand for *Long term Equity AnticiPation Security*, which are essentially options with typically at least 2 year maturity terms, as opposed to standard derivatives options of 3, 6 or 9 months maturities.

books content; however, what was found through reading individual reviews was that often reviewers go to great lengths to incorporate the books arguments into a current context, which is clearly more than would be expected for simple type of review⁵⁸.

An analysis of word(*noun*)-associations within the actual reviews, reveals a number of commonly co-occurring words, such as; with “oil” these were 'gas', 'coal', 'reserves', 'prices', 'production', 'gold', 'energy' co-occur or with “futures” → 'options', 'trading', 'markets', 'contracts', 'market', 'derivatives', 'swaps', with “position” → 'market', 'money', 'size', 'return' co-occur which isn't interesting, except that it simply shows the comments tend to be to the point and consistent with the reviewed topic.

Table 6.22 – Top 10, Real Estate review title terms, as retrieved for each year in the sample (*words are normalised, stemmed*)

2005	2006	2007	2008	2009	2010
1.Real Estate	Real Estate	Real Estate	Real Estate	Real Estate	Real Estate
2.Investor	Guide	Guide	Home	Guide	Home
3.Information	Home	Information	Information	Home	Guide
4.Money	Information	Home	Guide	Information	Information
5.Property	Investing	Investor	Investor	Landlord	Mortgage
6.Guide	Time	Agent	Time	Investing	House
7.Home	Info	Money	Mortgage	Resource	Landlord
8.Advice	Informative	Mortgage	Resource	Mortgage	Review
9.Resource	Investment	Resource	Investing	Market	Advice
10.Time	Money	Advice	Foreclosure	Housing Crisis	Investing

From table 6.22 one can appreciate how **Mortgage** emerged as a frequent term in the year 2007, when a wider sense of sub-prime mortgage related worries have indeed surfaced into mainstream debate. During 2008 and 2009 the financial crisis terms **Foreclosure** and **Housing Crisis** became used more frequently. There were 76 reviews which mention *Foreclosure* for 26 unique books, a number of these books were written as guides on how to profitably deal within the real estate Foreclosure market. A qualitative look at the reviews found that indeed there is some reference to current events. For example a Foreclosure book from late 2007 was reviewed on January 2009 and the title of the review went by “No longer relavant to the foreclosure market we are in.”⁵⁹, it is also observed that the new book titles have refocused from helping people begin a real estate business to how one may avoid Foreclosure and save their home. In fact it was found that a number of reviewers shared some of their and their friends' stories with dealing with Foreclosures during the recent crisis in actual reviews. The term *Mortgage* occurs

58 See, for example, <http://goo.gl/ytLGr> for the review by EWC which is a great example of highly elaborate reviews.

59 A number of these reviews are available on <http://www.newsmental.com/thesis> as excel styled spreadsheets.

more frequently within reviews from 2007 onwards, in fact most reviews from after 2007 were about the mortgage crisis, reviewers often give additional details and information. Such as in one case the author of a book on the mortgage crisis being criticized for being an employee of Moody's rating agency, which the reviewer blamed for the crisis in a number of insightful ways; see appendix D, figure D.6.

Throughout the co-occurrence text analysis for real estate, an exploratory review of the most frequent reviews showed for example that the most co-occurring nouns with "*housing*" were '*market*', '*boom*', '*bust*', '*crisis*', '*prices*', '*collapse*', '*markets*', '*crash*', '*mortgage*', '*policy*', with "*crisis*" these were '*mortgage*', '*housing*', '*causes*', '*economy*', '*nation*', with "*property*" these were '*management*', '*manager*', '*managers*', '*owners*', '*investor*', '*rights*', '*value*', '*business*', '*owner*' or with "*landlord*" these were '*law*', '*tax*', '*liability*', '*forms*', '*business*', '*property*', '*tenant*', '*deduction*'. Hence it is clear that much of the reviews' textual content is indeed to the point, which confirms to some extent the qualitative analysis of a number of reviews. For both, derivatives and real estate, these word co-occurrences can be relatively easily reproduced using the *Python* scripts and sample reviews available from the accompanying website for this thesis.

6.4.3.3 Summary

This section was devoted to an analysis of Amazon book reviews in order to investigate any possible relationship of large amount of user generated content with the financial markets, i.e. possibly the recent financial crisis. Frequencies of books and reviews on particular topics were investigated, in a similar approach as with Delicious, where publishing frequencies aggregated over regular time intervals were scrutinised. It was found that some books are responsible for much of the reviews, but it was also found that the frequencies of reviews can be used to measure the interest into a topic, over time. It was observed that review topics evolve over time, this is indeed largely due to new book titles on the market that deal with current issues; however, a qualitative investigation of individual reviews showed that in many cases reviewers provide contextual and additional information, often to a greater extent than anticipated. For example books that have been published on certain financial topics, before the crisis, contain more recent reviews that often put the book into current context and provide additional crisis specific information. Being able to detect such reviews might be beneficial. Sentiment of reviews seems to be largely irrelevant and stable over time, with a few exceptions there is generally a positive consensus, and sentiment relates to the books quality / content, directly.

6.4.4 Summary and Limitations

Potentially Amazon represents a valuable source of collective intelligence, given the large repository of UGC. Due to substantial review activity, especially as in recent years the uptake in online shopping meant an uptake in use of sites such as Amazon. Previous work discovered review sentiment played a significant role in sales ranking of products on Amazon and similar online retailers. Hence, the first part (i) of this study investigated review activity and the relative ranking that may result for manufacturers. This relative ranking was confirmed to some extent; however, one limitation imposed on the study was the lack of detailed sales data.

The distribution of star-ratings is skewed towards positive review sentiment, this was confirmed in previous studies and it was shown that it does not vary significantly over time either. It was further found that information travels rather efficiently, with a few exceptions (for example during the holidays). Various other points of interest were discussed, such as the readability and style differences in textual reviews of different star-rating polarities.

It was discovered in the second part (ii) of the study that some reviews tend to refer not just to the book itself but also to the wider topical context, often adding additional or new information that may be of some use. Detecting such reviews automatically, maybe using a machine-learning classifier model may be a useful future extension to the work. The study in this sub-chapter also had a qualitative aspect aided by automated text analysis, in which many reviews were manually investigated. In aggregate some connection to the financial environment were observed over the time, with related book reviews; however, much of the evidence is to a degree anecdotal.

In both experiments each reviewer was also treated equally to provide an unbiased examination of Amazon's UGC data. Although influential users could be considered, and a separate study, where Amazon users are weighted according to their social significance might be worthwhile in future. Amazon implements a customer profile, which is usually associated with a review based on the name on the customer's credit card itself (*i.e. the "real nameTM" badge⁶⁰*). This simplifies the task of identifying individuals behind particular user accounts, as real individuals (see section 8.3). Amazon also uses the badge "VineTM Voice" to highlight pre-selected reviewers (*i.e. these are pre-selected reviewers who review items before the product becomes widely available*), and the badges "Top n reviewer", "The" and "Artist" badges point to important reviewers as well. Importance of reviews could hence be weighted, based on the badge, and reviews from "Author", "Manufacturer" badge holders interpreted with more care, especially

⁶⁰ See http://www.amazon.com/gp/help/customer/display.html/ref=cm_rn_bdg_help?ie=UTF8&nodeId=14279681&pop-up=1#RN for all the badges Amazon uses.

for competing products. There are; however, several problems with tracking individual users such as; customers can choose to use pseudonyms, instead of real names, which complicates matters. Users are also allowed to change their pseudonyms at any time⁶¹, and there is no guarantee that pseudonyms and real names are actually unique within Amazon, although the latter issue can be overcome by inspecting public user profiles and employing Amazon URL-IDs as unique user identifiers. The number of reviews overall was large, product count relatively small, and our aim was to investigate all possible reviews rather than a small subset, especially since the given topic hasn't been studied in literature before.

To the author's knowledge no previous study investigated an explicit connection between Amazon reviews and financial markets before. Some evidence in support was found and text analysis revealed useful longer term insights in relation to finance. A number of results from previous work were confirmed, which in turn lent some support to this study.

6.5 Wikipedia

Up to this point UGC data from three web 2.0 applications was analysed in depth. In this section Wikipedia is briefly described in terms of potential use as a collective intelligence data source. UGC data from Wikipedia represents a notable entity of human-sourced knowledge. Wikipedia receives visits totaling more than 380 million a month as the fifth most popular website in the world⁶². To illustrate its lively community of contributors and how it may be used, the next sub-section provides an overview of the Wikipedia culture, and the subsequent sub-section presents four main article related data-sources that might be useful for tapping into the collective intelligence of Wikipedia.

6.5.1 The Wikipedia Culture

The commitment and enthusiasm of *Wikipedians*⁶³ (*users who write, edit or contribute to Wikipedia articles*) are based on altruism, reciprocity and a sense of community. In fact there is a significant body of interesting literature investigating the wider motivations behind Web 2.0 participation, especially for Wikipedia (see section 4.5.7). *“Wikipedians enjoy a sense of accomplishment, collectivism and benevolence, while working with exceptional freedom and*

61 <http://www.amazon.com/gp/help/customer/display.html?ie=UTF8&nodeId=14279641>

62 *As of April 2011 – for a complete set of statistic see* <http://stats.wikimedia.org/reportcard/>. On 25th Feb. 2011, Wikimedia foundation also published its 5 year plan for Wikipedia with aim of achieving 1 Billion monthly visitors by the year 2015 <http://blog.wikimedia.org/2011/02/25/wikimedia-presents-its-five-year-strategic-plan/>

63 <http://en.wikipedia.org/wiki/Wikipedia:Wikipedians>, last accessed on 2nd July 2011.

ease...effectively creating a near-utopian society in which individuals voluntarily collaborate and learn together.” (Kunznetsov 2006) Its popularity amongst users has grown so that an annual international conference known as *Wikimania*⁶⁴ is held every year to show presentations and discuss topics relating to sub projects, open source software and the different social and technical aspects that surround Wikipedia. There are 13.5 million registered contributors to Wikipedia¹³, not including a large number of unregistered contributors. This has risen from 1.25 million in 2006 (Kunznetsov). Fewer than 2% of Wikipedia users ever contribute to Wikipedia (O'Sullivan 2009), which is a characteristic shared by Youtube, for example. Wikipedia operates by giving every user power to contribute. They are each given their own user / profile pages where they may track articles they will edit in the future and present themselves to other community members. Wikipedians are able to recognise each other in their subsets of collaboration on topics. It has been suggested by O'Sullivan (2009) that subsets of users with similar interests are almost like groups that may not necessarily interact with other Wikipedians. They can further nominate each other for awards to honour distinguished work and potentially increase their ranking to a moderator with extra responsibilities. This actually highlights that despite Wikipedia's apparent anonymity – i.e. “*anybody is free to edit a page*” reputation is somewhat misleading as there seems to be a sense of individuality and authorship that can be associated with individual users. Honoured articles can be nominated as candidates for “Featured articles”⁶⁵ that appear on the front page of Wikipedia. This workflow is the same for portals like Wikiversity and other subprojects under the Wiki branching. Portals implement communal task lists that allow people to organise collaboration pages for articles and set targets for their completion.

Accuracy is maintained on Wikipedia as a direct result of the Wikipedians who act as moderators for articles. As a result of collaboration on talk pages, any new information that is added can be discussed a-priori and monitored with ease. Wikipedia implements an edit page history which allows users to see two article snapshots in a side by side format, with highlighting to show differences between the two dated snapshots. Wikipedia maintains its credibility by using five main principles⁶⁶ that it expects all users to follow while contributing. The most fiercely debated of the five principles is “Wikipedia has a neutral point of view”; NPOV mandates that contributors refrain from judgement or opinions being written into the articles. The article should comprise predominantly of facts. There should not be decisions of truth but instead there should be decisions about reputable sources. The discussion and talk pages for controversial articles are where NPOV is disputed, and debated. The article should

64 <http://en.wikipedia.org/wiki/Wikimania>, last accessed on 2nd July 2011

65 http://en.wikipedia.org/wiki/Wikipedia:Featured_articles, last accessed on 2nd July 2011

66 http://en.wikipedia.org/wiki/Wikipedia:Five_pillars, last accessed on 2nd July 2011

represent all viewpoints, but this is incredibly difficult to achieve. With multiple editors and collaborators on these articles, it is almost ideological to think that neutrality can be maintained without locking the page itself and there being an ongoing discussion. O’Sullivan (2009) points out that it is highly unlikely for editors with radically different views on controversial topics to reach an understanding, and secondly no consensus is required. The aim of NPOV policy is to allow both viewpoints on a page, perhaps, under two separate paragraphs without a judgement or conclusion forming.

6.5.2 Useful Wikipedia Features

Given the available collaborative content, one may consider three sources of UGC datasets. The article-page itself, the edit history and the discussion pages associated with the article. In addition since Wikipedia is run as a free and openly licensed foundation, anonymised traffic logs are available. Below we discuss each source individually.

The article-page represents a concept that is uniquely identifiable with a URL, although sometime there are multiple URLs which redirect to a particular page or disambiguation pages. Usually the article itself is a multi-author collaboratively constructed *semi-structured* free-text. Semi-structured because there is some representational structure such as headings, citation-lists, tables, info-boxes, and other elements that are described in much detail within O’Sullivan (2009). Researchers in the past have used out-links from an article to provide context to the concept represented by an article (Gabrilovich 2006, Hu et al. 2008). Citation page count can also be used as an approximate measure of reliability, and other features extracted from the article itself.

Edit history of an article represents valuable information on the changes made to an article, who and when performed them. It has been suggested that the number of times an article has been edited points to an interesting issue (Kittur et al. 2007). Reasons for a significant increase in editing activity can often be attributed to emergence of new information on the topic or vandalism, whereas highly controversial topics would tend to have a relatively high but constant editing activity.

Discussion Pages can be used to highlight points of frequent dispute (i.e. difference of opinion), vandalism (i.e. hatred against someone / something), make recommendations (i.e. suggest changes / improvements) and other types of discussion related to the overall improvement of the article (Ehmann 2008). They are available for every article, and can often be used to gain some background information about an article. Viegas (2007) was one of the

first to analyse talk pages, and found they grew at a 11x growth rate against a 9x growth rate for articles, and it was found that coordination related issues were mostly debated on the discussion pages. Schneider et al. (2010) provide a very complete and more recent analysis of discussion pages.

Traffic History or anonymised traffic logs are available for many time-periods. One inherent issue that has to be handled with care is the management of very large datasets that result from logging every single web-page request. Traffic log analysis is well known from fields such as Webometrics (Thelwall 2009). In the case of Wikipedia not much is different; traffic logs simply represent public interest into pages, although pages are focused on narrow and well defined concepts, which can make a traffic analysis quite revealing.

A number of features can be extracted from the mentioned sources. For example, number of major and minor page edits over different time periods can be extracted, or whether discussion pages contain certain discussion specific elements (as identified in Schneider et al. 2010), out-link counts from articles or one can also consider inspecting **user profiles** of editors to add weight to an article's credibility based on past edits, and for traffic history, one would be able to directly use time-aggregated traffic log counts for analysis. Some of these data sources can be directly visualized with open-source tools <http://stats.wikimedia.org/#fragment-13>, <http://www.trendingtopics.org>, or <http://stats.grok.se>.

6.6 Community Finance and Prediction Websites

In the last two years a number of web 2.0 websites, where individuals can share their personal and financial matters in a social environment have become popular. Two types of applications emerged. First, social websites which allow users to share their financial net-worth publicly, that is their incoming / outgoing cash-flows, assets owned, and assets in debt. These applications allow its users to be ranked in relation to other users, and financial comparisons can be made and discussed online. A number of such sites now exist, such as <https://www.networthiq.com/>, or <http://www.investimonials.com>. An article in the wall street journal has recently discussed this new phenomenon, “*Managing your money in public view*”⁶⁷. Second type of financial social applications that recently emerged is a public trading and investment set of social websites. Websites such as <http://www.etoro.com>, <http://www.currensee.com> or <https://www.zecco.com/communitydashboard.aspx> allow its users to share their trades publicly, for discussion and comments and further social interactions

67 <http://online.wsj.com/article/SB118177906703834565.html>, last accessed on 20th May 2011

between users, when executing real trades. Currensee, for example also selects its top performing traders, and allows other users to “*follow*” their trades, which essentially facilitates automatic mirroring of trades and subsequent trade execution. This allows individuals to trade even without any financial knowledge, and the legal implications of such processes are yet to be seen⁶⁸.

The uptake of social media applications that allow the sharing of such sensitive information, as individual trades, account balances or cash-flows, is intriguing. It can be attributed to the growing readiness of people to contribute and share personal information through the web more readily; that is, a general increased level of trust has made sharing financial details on the web a reality.

6.6.1 Trend Prediction Websites

So called trend prediction websites are introduced in this sub-section. These websites allow its users to vote in favor of, or against a certain prediction or event, in the hope that the aggregate vote of many users will allow the community to arrive at a relatively accurate prediction. Trend prediction websites are augmented with many web 2.0 features that allow rich social interactions. There is evidence that systems like this have been used by companies internally (Kambil and Van Heck 2002, Kambil 2003).

Most often trend prediction websites are essentially prediction markets in a web 2.0 disguise. One of the earliest prediction markets is the Iowa Electronic Exchange Market⁶⁹. The underlying behavior of prediction markets is on some level synonymous with the evolution process of User Generated Data on a web 2.0 system, towards an implicit fitness function (Sykora 2009). Much of the enthusiasm for prediction markets derives from the efficient market hypothesis or EMH (Wolfers and Zitzewitz 2004), as discussed in section 6.2.4 this hypothesis states that in the market mechanisms information will efficiently transfer into the most optimal price, given all known, available information. Sometimes also referred to as “*information market*” or also known as “*event futures*” they allow individuals to trade on future events, where in aggregate these votes represent the probability of the given event. Prediction markets are speculative markets which have the sole purpose of making predictions. They can be used to predict a variety of events such as sports, politics, movies, films or the stock market. The aim of a prediction market is that through a large collective user base there will be a greater accuracy

68 Automated trade mirroring essentially replaces the need for a trading, or investment fund, without the overhead of the entity of a fund. Clearly the fact that short trades as well as long trades, and in fact nearly any trades are allowed, may have some wider implications for financial regulators.

69 Iowa Electronic Exchange Market – <http://tippie.uiowa.edu/iem/>

in predictions which will benefit those who give correct predictions through financial rewards but will detriment those who give incorrect predictions through financial or other losses. For a prediction market to be efficient however, it was found that it is not required that all individuals in a market be rational, as long as the marginal trade in the market is motivated by rational traders (Wolfers and Zitzewitz 2004). Servan-Schreiber et al. (2004) have shown there not to be any substantial difference between real or play money markets. prediction markets were found to perform considerably better than individual human forecasters (Servan-Schreiber et al. 2004). In a study by Luckner et. al (2007) the advantages of prediction markets were highlighted where FIFA 2006 World Cup matches were predicted with 59.4% accuracy against Fifa ranking accuracy of only 46.9%, over 16 matches. Finally a good overview of the main different types of prediction markets and further literature is provided in Wolfers and Zitzewitz (2004).

Even though classical prediction markets (Kambil and Van Heck 2002) come from a separate field that has little to do with web 2.0, it was found that recently there are more elements consistent with web 2.0 websites on such systems. Collective knowledge generated and collected via atomic elements (*i.e. points of collaboration*), may in aggregate present an interesting additional source for any collective intelligence based decision support system. A detailed study of these was unfortunately outside of scope for this thesis. Nevertheless a list of possible candidates for use in the framework from chapter 8 is presented here. It ought to be noted that a trend prediction website can be based on virtually any concept abstraction and topic of interest. Prediction websites often range in complexity from simple voting systems to more complex prediction market mechanisms. In the following paragraphs a number of trend prediction websites are presented.

A well known trend prediction system is the Hollywood Stock Exchange (HSX)⁷⁰. It allows users to buy and sell prediction shares of movies, actors, directors, and of other film-related events. HSX user driven prices in Oscar, Emmy, and Grammy awards were found to correlate well with actual award outcome frequencies, and prices of movie stocks accurately predict real box office results (Pennock et al. 2001).

Intrade⁷¹ is a system where users can vote on many matters such as, Arts, Global entertainment, Financial, or Legal issues. The main featured financial markets on this site are, Dow Jones to close on or above a value on a particular date and whether the US economy will go into recession in 2011, etc. The website focuses more on prediction markets based on future political events. This website has approximately 66,000 predictions on 1,000 prediction markets.

⁷⁰ HSX – <http://www.hsx.com/>

⁷¹ Intrade – <http://www.intrade.com>

In PredictWallStreet⁷² a user has to cast a vote in order to see the predictions that have been made by the community. The site also uses contests instead of real money to motivate the users to predict accurately, the cash prizes range from \$25, \$50 to \$100 for the top three contestants; however, there is a lack of users, which may be an issue for the website manifested in less accurate predictions. Tradesmarter⁷³ is another website where users can predict on up or down moves. The website is based on a winner takes all prediction market.

Stockpair⁷⁴ is a website in which the users vote which stock will do better between a specified pair, such as the stocks of Apple vs. Microsoft, which somewhat differentiates this website from others. The system has a sizeable user following.

6.7 Summary

In the next chapter we will present a novel news-analysis web 2.0 application, called Newsmental. This application was built specifically for the thesis, as an example of a custom collective intelligence data-source. Despite their advantages, existing web 2.0 data sources are sometime insufficient or inappropriate for an application of collective intelligence. Especially automated news analysis has been a hard problem, due to the inherent difficulties of natural language and human subjectivity in judgment.

72 PredictWallStreet – <http://predictwallstreet.com/>

73 Tradesmarter – <http://www.tradesmarter.com>

74 Stockpair – <http://www.stockpair.com>

7 Newsmental; A Custom Source of Collective Intelligence



Many web 2.0 applications have emerged over the recent years as web platforms for sharing information, some very generic ones, and others more tailored to specific communities of users (*see chapter 2*). In the previous chapter a variety of interesting *collective intelligence* sources were presented. Several contributions to the body of existing literature and insightful observations relating to the *financial markets* and *UGC data* were made. In this chapter we built on the concept of collective intelligence sources and present a custom built web 2.0 system that was developed to fill the gap for sharing specific type of knowledge. It is argued that in some situations it may be feasible and indeed highly desirable to introduce a custom web 2.0 platform. The work in this chapter was a significant undertaking and provides an interesting practical aspect to the idea of system design for collective intelligence; see chapter 6 for a more generic discussion.

This chapter deals with issues relating to the identification of the need for custom intelligence and how to approach web 2.0 system design where the aim is to satisfy a collective intelligence acquisition need for specific domain use, in this case the financial markets. The design of the system is generic and applicable to diverse domains, and may be used as a guiding process (design case-study) for developing other custom web 2.0 applications. Issues encountered along the way and future work are discussed. Finally the chapter explores the user generated content and conclusions are drawn. One of the main contributions is also in terms of the custom built

web 2.0 system itself, which is the first of its kind, as far as the author is aware.

7.1. Motivation

Given the vastness of the web 2.0 application space, it is often sufficient to leverage existing sources for collective intelligence, and up till now most studies in this research area were concerned with analysing existing web 2.0 applications. However, in some cases custom sources of collective intelligence (henceforth referred to as CI) are needed to fill a gap in the web 2.0 application space. Deciding to do so may be a relatively costly initiative (in terms of time, complexity and resources), therefore it must be carefully considered whether existing web 2.0 sources might provide the necessary CI. Using custom sources may provide finer grained user generated contributions, at a level that can be custom-built for the CI task. Since all the data is available / “owned”, complicated information extraction and HTML-page parsing can be completely avoided and of course the potential flexibility to accommodate a given problem domain is generally incomparably more substantial than with existing systems. There are unfortunately also major challenges in constructing custom-built CI sources in development of the software, its maintenance / fault free operation, provisions for architecture / hardware, and most importantly attracting and socially engaging user-participation. The positive and negative implications associated with custom CI sources are summarised throughout the following bullet point lists.

Advantages of custom CI sources

- Any set of atomic activities can be incorporated into a custom design (see section 2.3.2.2 for a discussion of atomic activities). This provides potentially a great degree of flexibility and as needs evolve with time, new atomic activities allowing further sharing of user content can be added.
- The freedom to develop a web 2.0 application tailored towards a problem domain is restricted only with the appeal to and uptake of a potential user-base. Many existing web 2.0 systems are relatively generic, used by various groups of interest hence customization may make a lot of sense.
- The system and database(s) responsible for implementing the web 2.0 system are accessible hence all user based contributions are directly available. This is as a rule, almost never the case with existing web 2.0 applications – the flexibility allowed by direct access is notable.

Problems with custom CI sources

- The main advantage of using existing web 2.0 systems is that these tend to have a strong and established user-base. Attracting users to a completely new system may be challenging unless an existing tangible community already exists (i.e. company employees, club members, etc...).
- Developing a new web 2.0 system from the ground-up requires very careful design considerations as successful design choices that encourage sharing and participation must be made. Actual development and testing may be prohibitively time and resource expensive in many situations, on the other hand there are a number of off-the shelf deployable software packages and libraries¹.
- Maintenance of a web 2.0 system will incur further costs and requires constant supervision and service support. Bugs in the software, feature requests or hardware failures might mean unexpected interruption that can be expensive to deal with.

If the design and implementation of the web 2.0 system is done well, then the quality, quantity and richness of user contributions will be of great value in a CI system (also discussed in chapters 6 and 8). In the previous chapter a number of valuable CI sources were presented, Youtube submissions were found to provide surprisingly insightful sentiment indication, Delicious bookmarking corroborated with financial event based resources, Amazon reviews were found to be of some limited use in being insightful in retrospect for financial events, and trend prediction and other web 2.0 based applications were briefly introduced. These are useful sources of CI with much user generated data relating to the financial markets domain (and as discussed in chapters 6 and 8, to numerous other fields); however, these web 2.0 sources fall short of providing fine grained insights into what really matters in finance. A more consistent and explicit indication on sentiment, impact, scope and relevance of news-events would be strongly desired from a CI source that is to be used in practical finance applications.

In computational finance much research has been done in forecasting and interpreting the financial markets and for a better understanding of financial events (e.g. Taylor et al. 2002, Zemke 2003). As was already explained throughout chapter 6 on Youtube, perceived wider sentiment is detrimental to the price of assets and it was also shown that a substantial effect on the price creation process in finance is explained by news-events sentiment (Ederington and Lee 1993, Barberis et al. 1998, Chan 2003). Researchers have highlighted (Fung et al. 2005) that in the existing literature on forecasting and trading models there is an overwhelming tendency to focus on quantitative (macro-economic and price based) data, with very little work

¹ Many companies have reported internal use of trend prediction, Wiki's and other collaborative web 2.0 applications (<http://twiki.org/>, <https://traitwise.com/>, <http://telligent.com/>, <http://sharepoint.microsoft.com/>, etc.). Much of the deployment can be speeded up by using open-source or a variety of commercial solutions, in the case of Newsmental in this chapter, several libraries were leveraged however the system was custom-developed.

investigating the use of qualitative datasets in such models. Given that there is an enormous quantity of qualitative news data in the form of unstructured text, there have been numerous efforts to automatically annotate sentiment in financial news (Mittermayer and Knolmayer 2006). Understanding or analysing news is inherently a very difficult natural language processing task, since even human experts often fail, or disagree on what a particular news actual means and how it applies to various entities. Depending on the perspective, situation and background one same news-item may appear to have different polarity and impact to various individuals (Koppel and Shtrimberg 2006) hence a collective agreement on news is in fact highly desirable. The task of news analysis lends itself well to human based processing. A web 2.0 system that explicitly facilitates participants' collection of their opinion on financial news events in a productive, social and streamlined manner could be a useful source of collective intelligence. The motivations for the system in this chapter can be summarised in a few bullet points.

- Motivation from the field of finance is strong in the context of existing web 2.0 applications and this chapter illustrates that sometimes a need for a custom CI source is justified, and how this relates to other existing CI sources is important. Further discussed in chapter 6 and 8.
- A system that is generic and applicable to any news analysis domain is presented and can be applied to other fields. The process described in designing the custom CI sources is also of value to the body of literature.
- A solution similar to the one presented does not exist. There are live *Reuters* or *Dow-Jones* commercial sentiment annotated news feeds²; however, this is significantly different from collective news based annotation. The closest solution to the one presented was found to be <http://appinions.com/> however this system is still relatively different in a number of respects³. Existing web 2.0 news websites, such as Digg facilitate the aggregation of news stories, but the system is limited to judging significance of news by up-voting or down-voting news (Lerman 2007).

In the next section the design of the news-judgment process will be discussed, followed by an architectural explanation and algorithms involved in creating the news rating web 2.0 system. The final system itself is subsequently presented, the launch of the website and its management discussed and evaluated. The final part of this chapter analyzes the collective intelligence and it is shown how it relates to and adds interpretative value to financial market events.

² http://thomsonreuters.com/products_services/financial/financial_products/a-z/newsscope_application_license/, <http://www.dowjones.com/product-news-analytics.asp>

³ Appinions collects commentary from several websites and analyses those for sentiment. The most interesting product from the company is the Appinions Lens, which is based on a bookmarklet to highlight content of interest. <http://appinions.com/applications/appinions-lens/>, last accessed on 27th February 2011.

7.2 Design

The system discussed in this chapter is entitled *Newsmental*⁴ (accessible on the following domain <http://www.newsmental.com>). The application was designed, built, tested and launched over a four month period during the PhD. A news rating competition was also initiated over a two week period with prizes, in the form of book vouchers, to motivate regular user participation.

The basic idea was to build a system that would **1-Extract** news-articles from a number of (mostly British) financial news-sources. **2-News-articles** would be automatically analysed, pre-processed and entity recognition applied to the unstructured text to extract entities and some basic relationships between certain types of entities. **3-Cluster** similar news (using a clustering algorithm based on the extracted entities). **4-Top** news-items would be presented on the website with a break-down of the news and charts (based on the extracted entities and relations). **5-Each** news-item would be available for a quick non-obtrusive evaluation by the readers / visitors and the ratings would be shared amongst the entire community. **6-News** reading history would be tracked automatically and made available to all registered users with historical views, charts, and other comparisons against community ratings.

Generally it is of utmost importance that a web 2.0 application is designed so that users would be implicitly incentivised to make contributions and participate in at least the use of the main features. The following design choices highlight some benefits for *Newsmental* users:

1. News articles are automatically re-checked at regular intervals from all sources, which means that a single place for reading all the news may be more convenient for visitors. The news articles are clustered, so that the same news from various sources doesn't repeat unnecessarily. This is essentially what many news aggregators such as *news.google.com* also do.
2. Entities, Facts and relationships between some entities are automatically annotated, and presented as a break-down analysis with each news-item. This provides for a useful breakdown of the main actors, out-takes from a lengthy article and greatly increases the speed and efficiency at which news can be read. It also makes it easier for a reader to comprehend the news and help speed up news-analysis in general (Zwaan et al. 1993⁵).
3. News reading and understanding is augmented via collective news analysis, in that all previous news-ratings are summarised / averaged out and presented to all subsequent visitors (i.e. shared within the community of readers). This effectively facilitates reading news "*in a collaborative*" manner, since chances are other users have read and analysed

⁴ A play on the words; *news*, *mental* (crazy about), *smental* – *sentimental* about news

⁵ In Zwaan et al. (1993) four levels of cognitive text comprehension model were presented where readers built incremental, mental representations of text in their mind as they read text. The presentation of entities, facts and relationships extracted from the news, present a higher level of text-comprehension, hence naturally speeding up text comprehension.

the news already, which may in turn drastically speed-up news reading and the news opinion forming process.

4. News reading becomes tractable. Submitting judgments for news essentially creates a footprint of all articles read and opinions felt at specific times about specific topics. The news reading process has been generally intractable – even with most large news portals. However, *Newsmental* provides features for retrieving news-reading history by time, topic and other views (more on this in section 7.2.3).

All in all *Newsmental* should make news-reading less time consuming, more tractable and ultimately engaging enough for the users to come back and hopefully use the system on a regular basis.

7.2.1 Architecture

7.2.1.1 News-item Retrieval

News extraction was limited to eight news sources⁶ and financial news category⁷ only in order to limit the scope of the project and to make data processing manageable. In order to achieve the most complete text-analysis, it was necessary to extract full length articles for each news-item. Identifying actual article text on a page (*i.e. how and where it separates from advertising content, navigation menus and other irrelevant page elements*) is a non-trivial task, as is discussed in section 8.3.1.5⁸. Since there was a limited number of unique websites to consider, it made perfect sense to use the most accurate and relatively robust technique of page information extraction from HTML pages, based on XML querying using XPATH⁹. First, HTML pages were corrected for any malformed tag structure, subsequently they were loaded into a tree based DOM representation¹⁰. Since the resulting HTML was compatible with standard XML, XPATH based expressions were used to extract exact article portions from the HTML pages for each of the seven news websites. An XPATH expression, such as; `//div[@class='story-body']/*`¹¹, retrieves all the contents of a div tag with the *story-body* CSS class attribute. This is a robust method since any page re-design will unlikely break the XPATH expression, unless CSS class / id names are changed specifically. For the actual page retrieval,

6 The news sources were; ONS (UK Office of National Statistics), Bloomberg, Daily Telegraph, BBC, Reuters, Yahoo finance, Guardian and the Independent.

7 Although the system can easily be adapted to any type of news, *i.e.* political, world-events, domestic, etc. In fact extraction code to support many other news categories has been developed and tested. The choice to limit news category was purely a research design choice.

8 Google news for example uses a special annotation format that it asks news-publishers to submit to them, given Google's scale this is clearly feasible (http://www.google.com/support/news_pub/bin/topic.py?topic=11666)

9 Proposed by the W3C web standards organisation, XPATH is a standards compatible querying language for XML documents. See <http://www.w3.org/TR/xpath/> for more details on specification.

10 The open source c#.net library HtmlAgilityHelper was used for some of this processing - <http://htmlagilitypack.codeplex.com/>.

11 Much more complicated XPATH expressions had to be employed in some other cases, such as for Yahoo; `//div[@class='mod related-companies']//ul[@class='symbols']//h3/text() = 'Companies:'`

RSS was used with most news-sources to access the URLs of the latest news-articles, however *Bloomberg* was scraped directly and the *Guardian* has a powerful API platform¹², hence their http-rest API protocol was used. OOP design techniques such as the *strategy* software development pattern were used extensively throughout the code design. Design, testing and software development principles were followed as described in Dawson (2009) and partly in Freeman et al. (2004). Addition of new news-sources is straight forward and therefore only requires some template code and interface inheritance with the custom *XPATH* query expressions. In addition to the article-body, the title, author(s), and for *Guardian*, *Independent* and *BBC* comments associated with specific articles were retrieved. Comment counts (but not individual comments) were retrieved at regular intervals from *Yahoo*, *Reuters*, and *Telegraph* (*ONS* and *Bloomberg* did not support comments). All unnecessary HTML was stripped, except for meaningful sub-headlines. Finally, in order to provide users of *Newsmental* with a breadth of news summaries as extensive as news.google.com (over 1000s of news-sources are supported, whereas *Newsmental* supports seven major news sources), top news-clusters from news.google.com were retrieved and presented on the main page. This has the added advantage of minimising user likelihood of leaving *Newsmental* for other news-aggregator websites.

7.2.1.2 Entity Extraction

Named Entity Extraction (NE) is the process in which definite noun phrases that refer to specific types of entities, such as organisations, companies, countries, cities, persons or dates are identified and their occurrences in the text detected (Bird et al. 2009). The raw text / article requires sentence segmentation, tokenisation, POS tagging and the actual NE extraction, which is based on gazetteers of entity names and types, with entries to handle synonyms and similar linguistic issues (Ye 2003: pp. 482)¹³. In the building of the NE extraction module for *Newsmental* the gazetteers provided by a third party API service *Open Calais* from Clear Forest Inc.¹⁴ were employed. The libraries leveraged for Named Entity Extraction are mostly based on work by Ronen Feldman. The so called Declarative Information Analysis Language (DIAL) was used to manually annotate a vast amount of Entity Extraction rules (over 18,600 rules) by experienced linguistics experts over a number of years (Feldman et al. 2001). The Entity Extraction rules together with gazetteers are at the core of this system. Soft matching (i.e. resolution of abbreviations, resolution of formal and informal company names, etc.) and

¹² See <http://www.guardian.co.uk/open-platform> for Guardian's web 2.0 initiative and their API

¹³ Much valuable work has been done on NE extraction, resolution of co-references, recognition of textual entailment and other related advanced fields in computational linguistics as part of the MUC (*Message Understanding Conference*) competitions organised regularly by DARPA (Grishman and Sundheim 1996).

¹⁴ Clear Forest Inc. is now owned by Thompson Reuters, however the technology behind it was mostly developed and coordinated by Prof. Ronen Feldman, see <http://www.clearforest.com/> and (Ye 2003: pp. 482).

anaphora resolution (i.e. resolving co-references, such as the “he” relating to “Martin”, in “Martin went to the shop. He bought some food.”) were also supported in these NE libraries, as described in Ronen Feldman’s chapter in (Ye 2003: pp. 491-508). Several other tools for Named Entity recognition were also considered, such as *GATE*, *GExp* or *Zemanta*¹⁵, however given the completeness of *Clear Forests*’ implementation (Butuc 2009), and the added support for *RDF* based semantic web annotation with *Linked-data* (discussed further in future work section) as compared to other tools, made this library the best choice overall for *Newsmental*’s NE extraction module. The bullet list below illustrates the types of entities detected within news articles.

- **Labour Department:** [Shortly after the meeting started, the]<entity type=organisation>Labour Department</entity> [said second-quarter productivity slipped at a]
- **Natural Feature:** [for the people living on Phnom Penh's] <entity type=natural_feature>Boeung Kak lake</entity>[. The World Bank's last loan to]

Relatively basic relationships between entities, were also extracted.

- **Acquisition – AOL:** [grew 5 percent to \$319 million, helped by] <relation type=acquisition>AOL's acquisition of Huffington Post</relation> [.Subscription revenue fell 23 percent to \$201.3]
- **Action – Purchase:** [President and CEO Joe Clayton in a statement.]<relation type=action>Dish Network recently purchased Blockbuster Inc.'s assets</relation>[out of bankruptcy.]
- **Natural Disaster – Singapore:** [which could lead to a large number of its shares]<relation type=natural_disaster>flooding the market if cornerstone investors including Singapore</relation>[’s Temasek TEM.UL and the Qatar Investment]

In the last example a natural disaster in Singapore was incorrectly identified in-lieu of a specific market event, described in the given piece of text. The NE extraction is not always accurate but its overall performance seems acceptable and satisfactory. Overall entities can be categorised into temporal entities (Anniversary, Holiday, Political Event...), geographic entities (Country, Continent, City, Natural Feature...), actor / animate entities (Company, Organisation, Person, Facility, Published Medium, TV Station, Radio Station...), object / inanimate entities (Currency, Industry Term, Market Index, Product, Technology...) and there are a number of simple relations (Acquisition, Arrest, Bankruptcy, Product Release, Generic Action...) – for an exhaustive list of these, see Calais (2011).

In the next section the technique used to cluster similar articles from same and different news-

15 GATE (General Architecture for Text Engineering) – <http://gate.ac.uk/>, GExp – <http://code.google.com/p/graph-expression/>, Zemanta – <http://www.zemanta.com/>

sources will be presented, followed by a description of overall system architecture integration.

7.2.1.3 Clustering Algorithm

There is a need to group similar news together in order to aggregate same events and allow non-repeating news display on the front page of Newsmental. Techniques in text-clustering are generally based on *bag of words* (BOW), also known as *word vector space* in which classical data mining clustering algorithms can usually be employed, after the textual data has been transformed into word vector space. A useful summary of available clustering algorithms in the field of data-mining is provided in Han and Kamber's (2006) chapter 7. However, the bag of words representation model has been criticised in literature for its high dimensionality of feature space¹⁶, the inherent data sparsity of the vector space (Beil et al. 2002, Liu et al. 2003), and for the fact that the discovered means of the clusters do not provide an understandable and ready description of the documents grouped in these clusters (Beil et al. 2002). Hence, alternative methods have been proposed, notably Beil et al. suggest association rule mining (using the well known Apriori algorithm, Agrawal and Srikant, 1994) in order to detect frequent term sets which are then used as document representations. The key idea is not to cluster the high dimensionality vector space, but to consider only low dimensional frequent term sets¹⁷. A well selected subset of the set of all frequent term sets can be considered as a clustering¹⁸. Our clustering follows a similar approach to Beil et al., but instead of using terms extracted with *association rule mining*, Named Entities representative of news articles are used as term sets. This is convenient since NE extraction on all news has to be performed anyway. This technique also provides an understandable and easy to interpret description of the discovered clusters (e.g. the first three NEs are used).

In order to determine how similar two article representations are to each other, several alternative measures were considered. The cosine measure is often used in word vector models, however in this case set overlap measures were more appropriate. The formula 7.1 shows what is usually known as the *Jaccard similarity coefficient*.

$$J(a_i, a_j) = \frac{|a_i \cap a_j|}{|a_i \cup a_j|} \quad (7.1)$$

, where a_i and a_j represent two news article derived Named Entity term sets or cluster(s), and $0 \leq J(a_i, a_j) \leq 1$. The more terms in a_i and a_j overlap the higher $J(a_i, a_j)$ and more similar the two articles are. There is however an issue, since in its current form formula 7.1 disadvantages

¹⁶ Starting with a set of d documents and a set of t terms, we can model each document as a vector v in the t dimensional space R^t , which is why this method is often referred to as the vector-space model and which is why the dimensionality of the feature space can be very large given that t tends to be large (Han and Kamber 2006).

¹⁷ Detected using *association rule mining*

¹⁸ Strictly speaking, a frequent term set is not a cluster but only the description of a cluster.

shorter articles by making it harder for them to be joined into a cluster, hence the formula 7.2 with a tweaked denominator is used instead.

$$sim(a_i, a_j) = \frac{|a_i \cap a_j|}{min(|a_i|, |a_j|)} \quad (7.2)$$

Most articles analysed tend to be of similar length, however when shorter articles are encountered (i.e. articles with less NEs), it is preferable for these to be assigned to a cluster. In order to decide whether two articles should belong to the same cluster, formula 7.3 can be used, which allows easy interpretation, since $0 \leq p \leq 1$ expresses the minimum *percentage* of the smaller articles entity term-set that must match for a_i and a_j to be similar enough in order for them be joined.

$$isMemberofClust(a_i, a_j) = \begin{cases} 1, & \text{if } 0 \leq \left[\frac{|a_i \cap a_j|}{min(|a_i|, |a_j|)} - p \right] \\ 0, & \text{otherwise} \end{cases} \quad (7.3)$$

Essentially the clustering algorithm is implemented as greedy clustering, where the first cluster is created by the very first pair of news articles found to be similar enough (formula 7.3). After this every article (not already contained within some other cluster) is checked whether it can be added to an existing cluster, if it cannot, then the algorithm attempts to join up articles into a new cluster. This means that clusters are exclusive (i.e. a single article can only belong to one cluster). After some tuning of the algorithm it was found that the final algorithm performed optimally when $p=0.4$ ¹⁹ and the absolute intersection required for a cluster was equal to at least three NEs²⁰. Thanks to the expressiveness and specificity of Named Entities this method of clustering worked surprisingly well and consistently.

The final Newsmental system, c#.net code implementation was made available under the MIT open source licence on <http://www.newsmental.com/thesis/>. The actual implementation of the clustering algorithm can be inspected in the *SimeEntityAgreementClustering.cs* file of the project.

7.2.1.3 Overall Architecture

A relatively involved architecture had to be used in order to satisfy functionality requirements and minimise latency of the time-demanding long-running article extraction and text-analysis processes. RSS parsing, Restful-API, HTML processing and article extraction using *XPATH*, entity recognition, clustering, programmatic caching and an AJAX and JQuery based interface

¹⁹ Using a small value would build one or very few “junk” clusters, as it becomes too easy for an article to join a cluster.

²⁰ The clustering algorithm was evaluated on a range of p and number of *common NEs* values, to find the most appropriate parameters.

were integrated within *Newsmental*. In order to ensure that news articles are up to date, a background process on a separate and (page-serve) independent thread is run. This thread is responsible for processing news articles, updating the database and ensuring that the memory cache represents most recent state, mirroring database data. Caching was an important consideration for *Newsmental* since the system had to achieve good response times despite working with memory heavy data (i.e. large chunks of article text and text summaries). At a page request, article data from the memory cache would be returned rather than from database; hence, greatly increasing response times. Although running a background thread introduced some complexities into the application, performance of the page-serving thread-pool was improved (consequently, none of the news-updating process took a tangible toll on page-request latency).

7.2.2 News Presentation

Given the large amounts of article text and the need for an additional user-interface facilitating rating of news articles, it was a challenge to display these on limited page-space. Arrangement of collaborative elements in a web 2.0 system deserves careful attention. It was important that *1*-as many news-items are shown on a page and *2*-as much of the news article is shown for a headline as possible, yet without overburdening a regular user. Aspects of news comprehension have been studied by researchers in the field of cognitive psychology. It has been shown that newspaper headlines are effective conveyors of news (Dor 2003). However, Andrew (2007) illustrated on a political example that topics received considerably different treatment in headlines than they did in the full-text stories. It is, therefore, important that the full news story is presented and also that a break-down of extracted entities and community ratings are shown, as these present a higher level of text comprehension (speeding up the news-reading) according to the model by Zwaan et al. (1993).

The final design showed news in reverse chronological order in consistent news-boxes, where each box is composed of three main elements. From left to right, these are; *1*-the article headline and the body of the news article, *2*-community ratings of overall sentiment and impact, *3*-break down of extracted entities and simple relationships between the entities. A “RATE NEWS” button slides out the user-interface for judging the news-article (discussed in section 7.2.3). Full instructions on how to perform most activities on *Newsmental* are provided within the *Newsmental* tutorial on <http://www.newsmental.com/tutorial.aspx>, and therefore only some features will be highlighted here.

By clicking on an article’s headline a new window with the news-article from source opens up. Since readers may prefer to view articles on *Newsmental*, font-sizes can be changed from

within the page²¹ and a full-screen view for the article’s body text is available. A full-screen view of an article can also be brought up and an arbitrary piece of text from an article can be selected / highlighted using the mouse cursor, and as soon as the mouse (highlighting the text) is released the piece of text is stored within database under the logged-in user’s account. This is a handy feature allowing news-readers to store specific excerpts from articles without interruption to their reading-flow (the DB-storage script is run via asynchronous page-postback / AJAX). Features such as these, which encourage UGC contributions in an intuitive manner, without disruption to use-case flow, should be carefully placed throughout a web 2.0 system’s interface.

Numerous entities and entity relations are automatically identified in news articles, as described in section 7.2.1.2. By clicking on an entity / relation item a semi-transparent information box slides out over the article with further information on that entity / relation. This allows users to drill down to more detail on items in the story. Some users found the visualisations of entities and their types more useful to get a quick overview of the article’s content, as shown in figure 7.1 (see section 7.3.3 for user feedback).



Figure 7.1 – Visual display of the entity composition of a news-article (available for each article)

7.2.3 Rating News and Related Issues

With any web 2.0 system the atomic collaborative elements must be as non-intrusive, implicit and as common sense as possible, in order to encourage sharing. This is why a light-weight JavaScript and AJAX based slider panel was introduced – shown in figure 7.2. The slider panel is brought up by clicking on the blue “RATE NEWS” button and automatically slides back after some time of inactivity. Each panel is composed of a set of horizontal and vertical slider

²¹ Font-sizes can be changed using the “+”, “-“ and “=” buttons in top right corner of the page.

controls, a comment text box with 120 character limit, a text-box with autosuggestions²² for tagging the news-article, and a submit button. The submit functionality is implemented entirely as an AJAX partial page postback script²³.

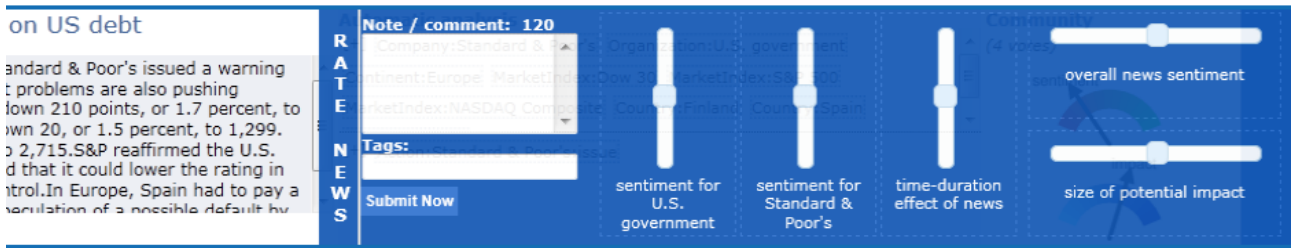


Figure 7.2 – Slide-out rating panel for news article (from left to right: comment [max, 120 char], Tags [autosuggest], sentiment for entity 1, sentiment for entity 2, time-duration impact, overall news sentiment and size of potential impact)

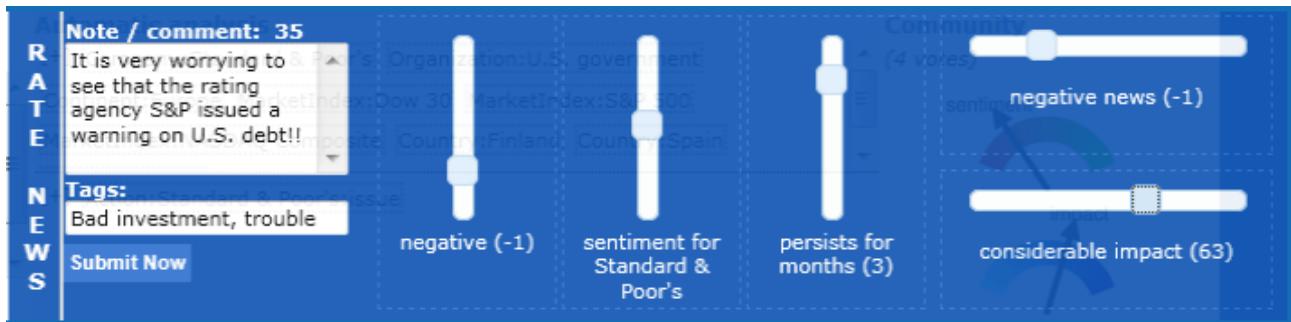


Figure 7.3 – News-rating panel, ready for submission (“Submit Now” button click)

Figure 7.2 shows the slider panel in its full view, and figure 7.3 shows a judged news item, ready for submission. The comment from the 120 character comment text-box is shared with other users of Newsmental and tags are used in search. It can be observed from the two figures that the first two vertical sliders relate to “sentiment for U.S. government” and “sentiment for Standard & Poor’s”, and that only the former was rated negatively (sentiments are rated on a 5 point ordinal scale, i.e.: -2|very bad, -1|bad, 0|alright/no-opinion, +1|good, +2|very good news). The announcement of the news that S&P has issued a debt warning for the U.S. is clearly bad news for the U.S., however S&P is more or less unaffected by the news, hence it is noteworthy the interface allows for such a complicated sentiment / opinion to be expressed.

The third vertical slider is always present in the rating panel and allows a choice of five values, relating to the time-duration effect of news (minutes/hours, days, weeks, months, years). In other words, given the news, what temporal impact in terms of duration of the effect the single news-item will likely or could potentially have on the financial market / financial ecosystem. In the example from figure 7.3, the judgment of impact is likely to be in terms of months, since the debt warning might be an indication of further troubles for the U.S. economy that could take a

22 Autosuggestion text-box is a regular text-box, except that tag-word suggestions appear as drop-down items.
23 It was mentioned throughout chapter 2, that AJAX has allowed for relatively complex implementations of atomic collaborative activities, i.e. page interactions can be streamlined as the page doesn’t have to reload.

few months to materialise. The two horizontal sliders relate to the “*overall news sentiment*” (5 point ordinal sentiment scale) and “*size of potential impact*” (a percentage 0-100, with 5 ordinal bins; no impact [0], very little impact [1-25], some impact [26-50], considerable impact [51-75], very high impact [76-100]). In the given judgment example the overall news sentiment (-1) is of-course quite bad for the U.S. and in fact most other economies that depend on the U.S. The impact rating is also high, *considerable impact* (63), as the impact of the debt warning news will probably move the markets.

However, since it is quite likely that people from different regions and especially various backgrounds will interpret certain news quite differently, each registered user was asked for their demographic details²⁴, i.e.: age group, location, level of education (university level, pre-university level), interest (finance, politics, technology, world events), financial experience (none at all, interested, knowledgeable, expert) and news reading frequency (only sometime, once every few days, every day, every few hours). It was hoped that in aggregate, with several participants reading and rating the same news, consensus opinion will emerge.

7.2.3.1 Streamlining the News Reading Process

The central idea behind Newsmental is to employ a non-intrusive manner of collecting news judgement opinions. There are several advantages of sharing news-analyses on a web 2.0 system, one which will be briefly discussed in next section (7.2.4) and relates to the historical records of all read / rated news, which can be compared to the *community*.

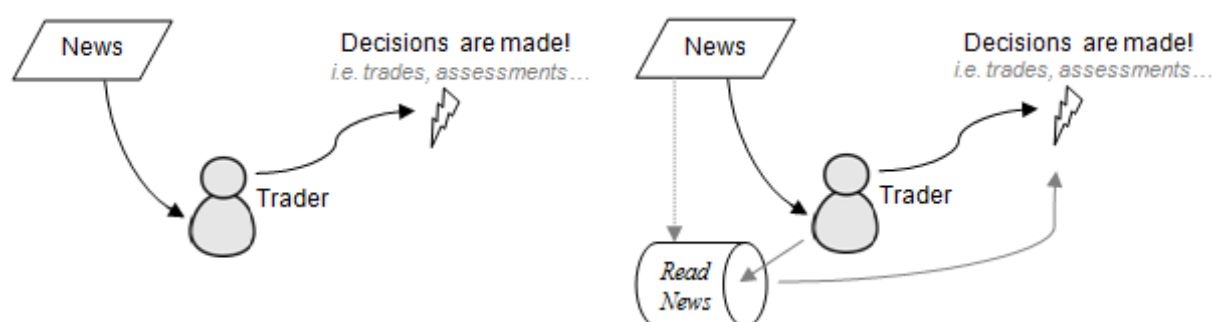


Figure 7.4 – User-case of a Trader using Newsmental
(left figure – usual work-process, right figure – Newsmental based work-process)

Newsmental website’s aim has been to streamline the news reading and rating process by taking an existing need and employing a web 2.0 system to aid in delivering a working solution towards this need. Figure 7.4 presents a realistic scenario involving a low frequency trader²⁵

²⁴ It is important to build an understanding off the individual contributors to a web 2.0 system (this is further discussed in section 7.4.1 and in chapter 8).

²⁵ High frequency traders are individuals who perform many transactions within a day, usually in the time-frame of minutes rather than hours, whereas low frequency traders tend to trade significantly less.

who makes regular trading-related decisions based on a news-feed. The left figure in 7.4, illustrates a trader taking-in news information which is eventually actioned into trading decisions, with many news articles being read over time. Unfortunately, due to the effect of selective memory, it is very difficult to reconstruct the thinking process behind historical trading decisions without explicit note-taking. Traders are known to keep logs (i.e. diaries) of trades (Schwager 1993), however having to take a note of each news-article that was important (e.g. into a spreadsheet) would break down the natural news reading workflow of a trader, to the point where it becomes infeasible. Instead Newsmental allows for all read news to be automatically tracked over time, with no disruptions to the workflow, as illustrated in right-hand side of figure 7.4. The trader rates, comments and highlights text excerpts from news articles in a streamlined way (using the light-weight web 2.0 style UI, described earlier), which gets stored into database (*Read News* in figure 7.4). This data is an accurate representation of a trader's opinions, perceptions of significance and sentiment over time, and can be reviewed by the trader and compared to the rest of the community. Ultimately the trader can use this information in making trading decisions.

7.2.4 Miscellaneous Features

One of the most useful features provided to a Newsmental user is the ability to track topics and particular news-items and news sentiments as they were read and perceived at a particular time in the past.

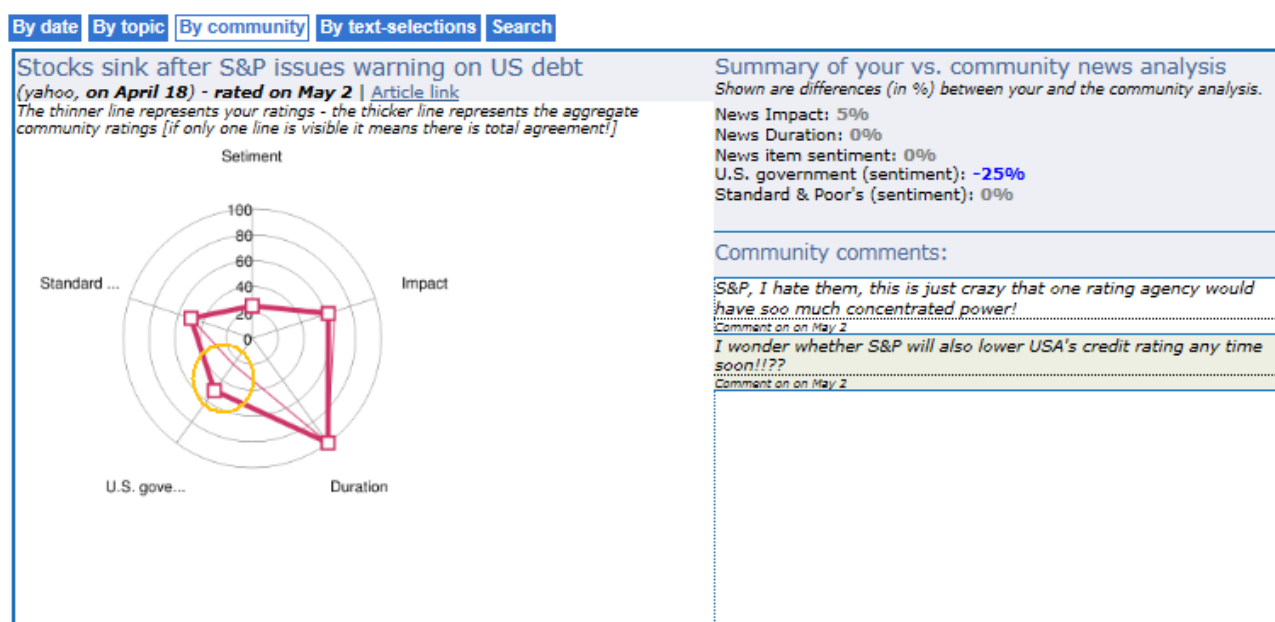


Figure 7.5 – Browse by community view (radar chart showing the agreement amongst ratings with the community)

Each registered user can inspect news as they were rated, in reverse chronological order, with

direct links to the original sources – effectively a kind of augmented bookmarking service for news articles. Other views include; view by most recent topic or a view that compares ratings with the community, see figure 7.5. For each news article a radar-chart is generated, which simply shows a thin line (user opinion) and a thick line (community opinion) for each of the news-rating dimensions. A yellow circle on the chart in figure 7.5 highlights that user sentiment for the "U.S. Government" is visibly lower than the community's, otherwise the ratings are pretty much the same. To the right of figure 7.5 are all the community comments that other users made concerning a particular news item. Figure 7.6 illustrates the browsing interface for all saved excerpts from news articles. A search by keyword can also be performed which will return matches on tags, headlines, comments or notes, in the user's historical news-judgment entries.

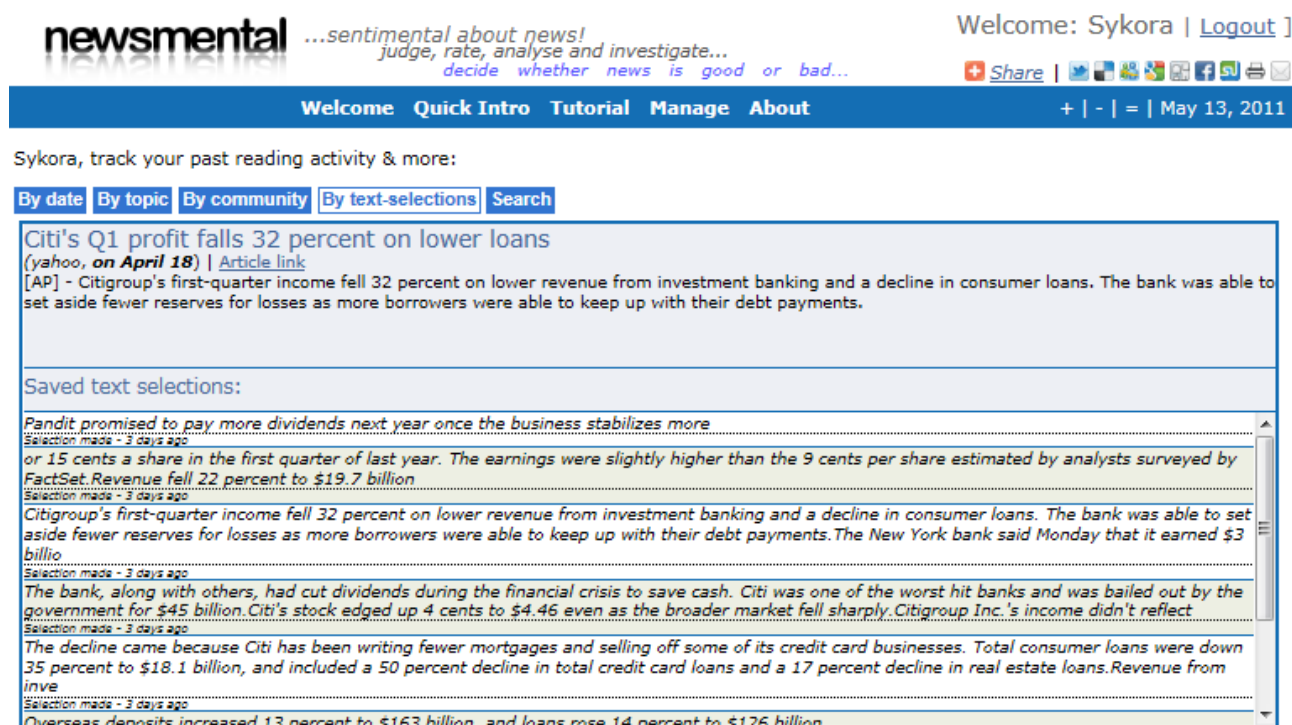


Figure 7.6 – Browse of news article excerpts (as saved by highlighting article text with the mouse in fullscreen view)

Finally, trending news and news.google.com clustered news are displayed as drop-down panels at the top of the main page, with a help bar that brings up easy to follow step-by-step interactive instructions on how to use Newsmental efficiently. Many articles also have comment counts from the original web-pages associated with them. Complete instructions and a functional description are available on the Newsmental tutorial pages.

7.3 Launch and Maintenance

Newsmental was launched during the last semester of the academic term at Loughborough University – it was live and ready on the 30th May 2011. In order to tap into existing communities of interest the website was advertised with eleven student based university finance societies across the UK. Committee members of the relevant societies were approached and most agreed to inform their members internally in addition to social media based (i.e. facebook) society pages. Some word of mouth spread through twitter and facebook as the initial news judgments were starting to pour in. With the kind support of departmental administrators, undergraduate and postgraduate students across various departments (including economics and business) were also informed via departmental mailing lists at Loughborough. A mailing list of around 200 individuals from the survey (chapter 4), who agreed to be informed of the launch of Newsmental were also notified about its launch. In addition more traditional advertising for Newsmental was done, using fliers and posters placed at frequented locations in the Loughborough campus, including the department of Economics / Business and the library. The overall time-frame of the study span from 30th May 2011 to 11th July 2011, with a few days of outage as described in next subsection, 7.3.1.

7.3.1 Issues and Maintenance

Based on initial feedback in the early days of Newsmental being live, some changes to the user-interface were made. An RSS feed, an interactive how-to-tutorial bar and several other minor design tweaks were added. Before the launch of the Amazon book voucher competition, a top-raters feature was introduced, in order to make the news-rating more social. In addition to implicit benefits of use (section 7.2), there is some evidence to suggest that, badges, prizes or other forms of incentives within an online social environment can dramatically increase initial user engagement (see section 2.2.4.2, and Malinen 2009).

Most changes were performed over night to minimise any disruption to the application²⁶. Unfortunately over the lifetime of the project two episodes of server disruption occurred, which meant that unexpected and unscheduled maintenance was necessary to resolve the issues as fast as possible. The first episode of down-time was during 9th–12th June²⁷, which was due to a server software update fault by the hosting provider. The second episode of down-time was during 17th–20th June²⁸, caused by the database server – this time major optimisation of the

26 The user-base was mostly from the UK, hence BST (GMT+1) time is referred to here.

27 The first indication of issues with updated software was on the 9th; however the site was operational except for periods during the 12th June, but nine news ratings were still submitted on the day.

28 During this period the server was fully unavailable, whereas in the previous downtime, the fault wasn't as

database and numerous code-redesigns were completed which improved performance several-fold. Since the 20th June the application ran completely problem-free, until the end of the project's lifetime. The project's lifetime during which news data was collected, was six weeks, as illustrated in table 7.1.

Week 1 – 30 th /May/2011...05 th /June/2011	Site is launched and widely advertised
Week 2 – 06 th /June/2011...12 th /June/2011	RSS feeds and requested features introduced
12 th /June/2011	Disruption to server, website down for maintenance
Week 3 – 13 th /June/2011...19 th /June/2011	Continued usage
17 th /June/2011 – 20 th /June/2011	Disruption to server, website down for maintenance.
Week 4 – 20 th /June/2011...26 th /June/2011	Application speeded up considerably, after redesign
Week 5 – 27 th /June/2011...03 rd /July/2011	Competition for Amazon vouchers begins
Week 6 – 04 th /July/2011...11 th /July/2011	Competition ends, data collection finished

Table 7.1 – Time period covered by the six week pilot run of Newsmental

7.3.2 Traffic

In order to monitor visitor traffic, Google traffic analytics²⁹ was employed. In figure 7.7 the analytics dashboard is illustrated, with the chart showing weekly unique visitors on Newsmental. The Google analytics software made it possible to monitor the traffic sources, user behaviour, user origin and platforms used by visitors accessing Newsmental. One rather unfortunate issue with Newsmental's launch has inherently been the bad timing of the launch, as it clashed with the exam period at Loughborough and the end of term across many other UK universities. Retrospectively it seems that a launch if better timed would have been able to generate a considerably higher rate of participation amongst the student community. Fortunately Newsmental was also advertised with several sizeable non-student financial market groups on LinkedIn³⁰. In order to motivate regular participants, a two week "competition" for £40 Amazon vouchers was launched on the 27th June 2011 (and ran until the 11th July 2011), with simple but carefully phrased conditions, these are available for reference here – http://www.newsmental.com/amazon_competition.aspx. As is the case with any social reward system, the form of its implementation must be given due attention.

significant.

²⁹ A free traffic monitoring and profiling service available to developers, see <https://www.google.com/analytics/>.

³⁰ LinkedIn connects professionals within a social network, and many interest groups of professionals exist on LinkedIn, including professional traders, investors, and market enthusiasts.

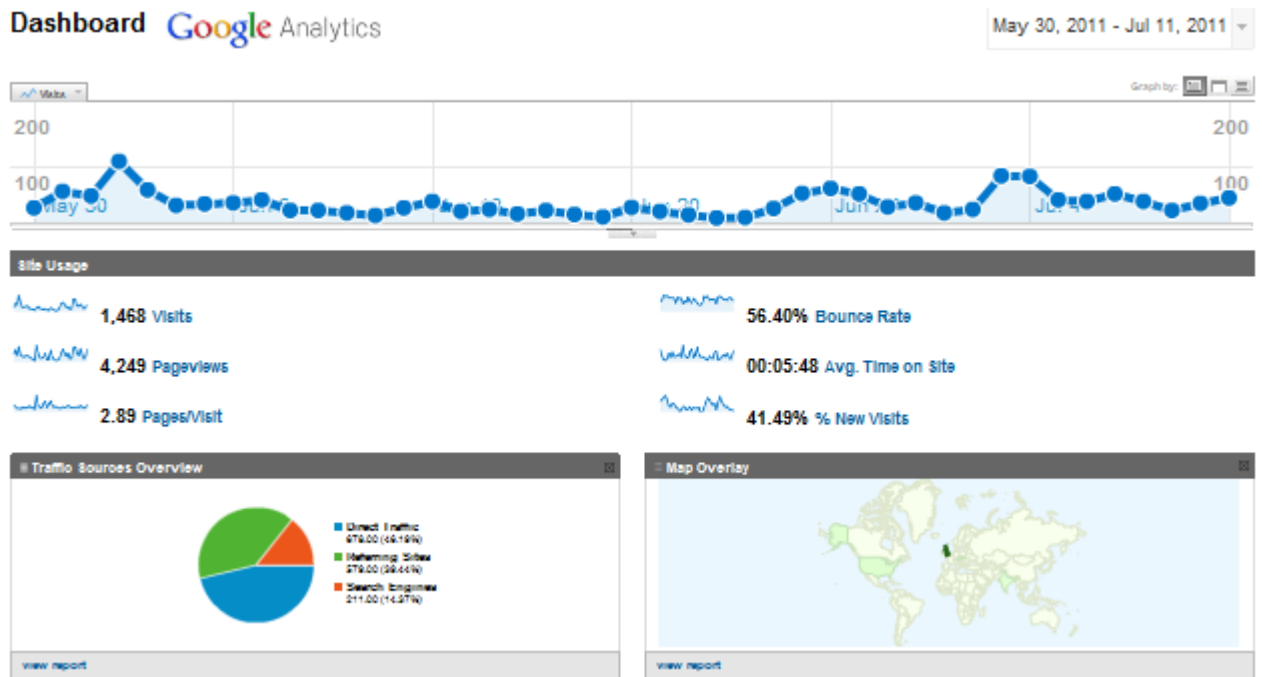


Figure 7.7 – Analytics dashboard for Newsmental, 30th May 2011 – 11th July 2011 (chart shows daily traffic data)

7.3.3 Feedback from Users

Provision for user feedback, especially on the collaborative atomic and social elements of a web 2.0 system, is important in assessing and improving the application. As far as Newsmental is concerned, feedback from users was received ad-hoc through social networks and emails; however, an informal focus group was also held with six users of Newsmental, three weeks into the project. After the Newsmental study was concluded on the 11th July, a message to all registered users was sent out asking for feedback and evaluation. The main points from the collected feedback are briefly summarised in the bullet point list below.

- Feedback indicated that many users enjoyed news-reading on Newsmental, except in some cases where users were not quite clear on how to make best use of all the system's features – despite two tutorials available on the website. As a response to this concern, a quick-intro bar on how to use Newsmental was added to the top of the home-page³¹, which had a positive effect.
- The overall design of the page was otherwise positively received, an example user comment being; *"The website was very simple to use, and really well structured, and by keeping it very light and lots of white space, it made for very easy viewing"*.
- The functionality of being able to keep track of all rated news was very well received, and there was some indication that users appreciated the collaborative news-analysis, as this particular feedback from a user explains; *"I'm an avid online news reader, and love keeping myself updated, but by adding an opinion/rating to an article it gives you an idea as to what the overwhelming response to an article is. The real bonus of this type of rating means that you can*

³¹ This would only show to new / un-registered users.

say more than other news websites, where all you seem to be able to do these days is 'like!', and 'Tweet', but this process gives the reader an opinion on the article”.

- There was one instance in which a user complained to us about suspicion of another user’s news-ratings being overly negative. This is an inevitable effect of a social web 2.0 system with transparent user contributions. Disagreement between users will sometime arise and as maintainers of the system, we may be approached to help resolve the situation, unless a social or automatic problem-resolution is not in place. Since it was rather interesting to investigate this complaint further, the user was subsequently contacted for their reasoning in their news judgment, but explained that they felt news were generally quite negative and a complete satisfactory explanation was given by the user.
- Initially several people felt slightly overwhelmed with the Automatic Analysis (i.e. presentation of entities / relations) however in all followed-up cases (once subjects got to use the system) they deemed the presentation highly useful and conducive to reading articles. A possible improvement for the future would be to present extracted entities in a visual manner and simplify user-interface further (see section 7.2).

7.4 Results

During the project’s lifetime 55 different individuals rated news, 48 users registered an account with Newsmental, of which 19 user accounts actively rated news. Since it wasn’t a requirement to be logged-in to rate news it is likely that some users forget to log-into the system on occasion, despite reminders to do so on the main page. Users who registered were encouraged to provide optional demographic details, the frequency tables of these are provided within appendix E, tables E.1–E.6. In summary, users were predominantly male, between the age of 20-39, 72% claimed to read news daily or more frequently, over 70% were interested or knowledgeable in finance and 91% of users came from the UK, US and rest of Europe.

All in all 2,138 ratings were submitted during the 40 days of the study being online (averaging to 54 ratings per day), however 650 ratings were submitted by anonymous users, i.e. not logged into the system. Out of all 2,138 there were 199 ratings where a shared / public user-note (i.e. comment) was left behind. All the ratings covered a total of 1,070 individual news articles³². The ten most rated news-items are shown in table 7.2 (*also see appendix E, figure E.2 for a dot-plot of 35 top headlines*), from which it can be appreciated that four news stories were perceived positively (highlighted items in column “Avg Sentiment”), and that in relation to each

³² During the same period there were 4,429 news articles collected in total, although due to news clustering only the first article belonging to a cluster would be available for rating in order to avoid showing duplicate stories. Hence the 1,070 articles and their 2,138 ratings will be considered exclusively for further analysis (even though more articles during the period were available).

other, the average duration and impact of the ten news stories makes sense and can be explained well. For example, “Kate and William give UK wine a boost” was perceived to be good news (std.dev. – 0.45), but with a low impact (std. dev., 19.91) and time-duration (std. dev., 0.73), on the other hand “Obama: Still differences on debt, new talks Sunday” was strongly perceived at the time as bad news with the lowest standard deviation for sentiment and impact (among the 10 news), however a higher standard deviation for duration. Standard deviation in this context highlights the disagreement or uncertainty associated with an average news-item judgement. Interestingly the highest average sentiment disagreement (among the 10 news) was for “Pope’s finance back in the black”, even though this seemed like good news, some people disagreed, which seems to hint to a possibly unpopular perception of the pope among the readers. A news-item that carries a lot of significance (highest average duration and impact) is “Trichet says debt is global, not European problem”, and the news-story “George Osborne needs a bolder plan for growth” was rated with high duration; however, the impact is lower, which seems to be a logical interpretation of the news and makes relatively good sense.

News Title	Avg Sentiment	Avg Duration	Avg Impact	#No of Readers
JD Sports could have JJB chain in its sights	0.778	1.778	54.333	9
Obama: Still differences on debt, new talks Sunday	-0.875	2.250	67.125	8
Black economic gains reversed in Great Recession	-0.875	2.250	34.500	8
Trichet says debt is global, not European problem	-0.875	2.875	66.250	8
George Osborne needs a bolder plan for growth	-1	2.429	41.429	7
Home Depot accused of violating Buy American Act	-1	1.571	35.286	7
Kate and William give UK wine a boost	1.286	1.571	26	7
Pope's finances back in the black	0.143	2	30.286	7
Waitrose goes little to grow big	1	2.143	38.143	7
Beko fridge fires started in 2007	-1.286	2.571	51.714	7

Table 7.2 – Ten most rated news-items

A dot-plot chart of further news-items, highlighting sentiment, impact and duration is available in the appendix E, figure E.2. Unfortunately there were only 536 news-items rated by more than one news-reader, this presents a major limitation to the study and is discussed further in section 7.5.1. In order to have as many news-item ratings available as possible, all ratings, including from non logged-in users are considered throughout the analysis in this section, unless otherwise stated. Figures 7.8 to 7.12 highlight the distributions of all ratings for sentiment, impact, duration of news effect, sentiment for entity 1 and sentiment for entity 2.

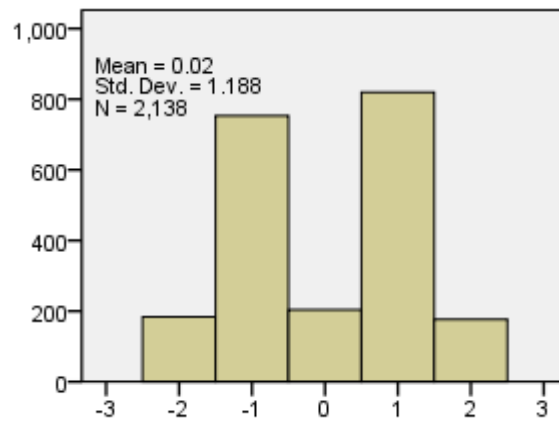


Figure 7.8 – News sentiment distribution, all ratings

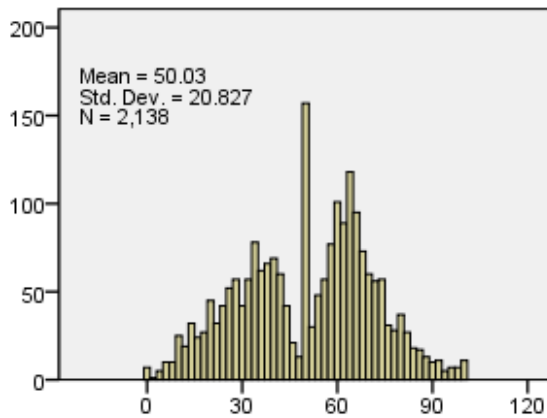


Figure 7.9 – News impact distribution, all ratings

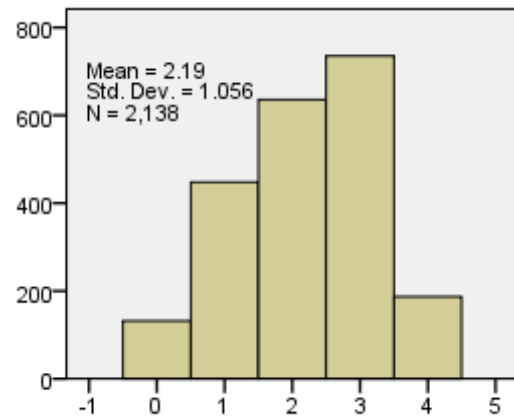


Figure 7.10 – News duration distribution, all ratings

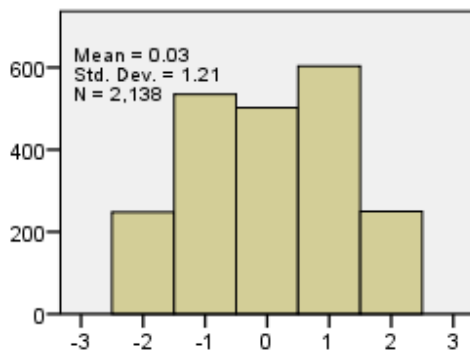


Figure 7.11 – News sentiment distribution for entity 1, all ratings

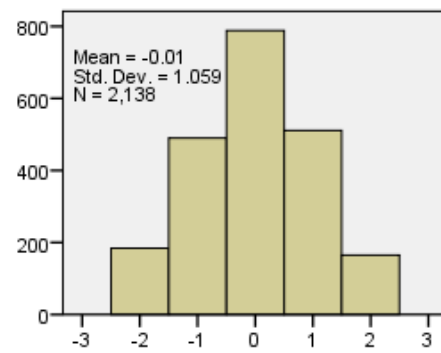


Figure 7.12 – News sentiment distribution for entity 2, all ratings

It is interesting to note how the sentiment distribution in figure 7.8 is bi-modal, with most ratings at -1 and +1, and how the sentiment distributions for entity 1 and entity 2 increasingly contain more zero valued ratings. A manual investigation of relevant news-items revealed that the top entity (most talked about in an article) tended to play a more important role than the second entity for which in numerous cases the news-articles did not necessarily imply any polarity. From the focus group (mentioned in section 7.3.3) it was discovered that when readers weren't certain about a rating, i.e. didn't have an opinion, they would leave the rating-slider in

its default position. This explains the increased number of zero-valued sentiment in figures 7.11 and 7.12. A similar phenomenon can be observed in figure 7.9, which also exhibits a bi-modal distribution that clusters around 30 (news have a lower importance) and 70 (news are more important) impact values. To some extent it was expected that entity 1 and entity 2 sentiment will be related to the overall news-item sentiment, a spearman's rho correlation on ratings confirms this with a .660 and .533, significant, p (two-tailed) $< .001$ correlation between overall sentiment and entity 1 sentiment and entity 2 sentiment, respectively³³. It was hoped that the entity sentiments provide a finer grained insight into the sentiment of a news story.

7.4.1 Agreement among Readers

Given that users rated news-items independently³⁴, it begs the question of the degree of agreement between them. Inter-rater agreement or reliability measures assess the agreement between two or more observers who describe the units of analysis separately from each other. These statistic measures are in frequent use in social sciences, especially in human driven content analysis and similar methods (Krippendorff 2004). They help to answer the question of whether ratings are the result of irreproducible human idiosyncrasies or whether they reflect properties of the phenomena of interest on which others could agree as well. Over the years a number of measures were proposed, such as Cohen's kappa (Cohen 1960; κ), Fleiss's kappa (Fleiss 1971; κ) or Cronbach's alpha (Cronbach 1951; α), however the most appropriate type of inter-rater reliability statistic for the data in this study (multiple raters, missing values, scale of measurement) is Krippendorff's alpha (Krippendorff 2004, Hayes and Krippendorff 2007³⁵; α). The degree to which news readers agree on the sentiment, impact and implied duration of news-items was evaluated for the six (top) users³⁶. The three alpha values with 95% lower and upper limit confidence intervals (based on 10,000 bootstrap samples) were .6038 (.5340, .6699) for sentiment, .2383 (.1343, .3376) for impact and .0702 (-.0367, .1747) for duration ratings. The value for sentiment is large enough to indicate a moderately strong and consistent agreement among raters however the latter two alpha values indicate poor agreement for impact and virtually no agreement for duration. Moreover, table 7.3 reports agreements between individual users, where clearly there are users with much higher or lower agreement levels than the overall

33 Spearman's rho correlation between entity 1 and entity 2 sentiment is .480, at p (two-tailed) $< .001$ significance.

34 This is not entirely true, since user comments and news judgements are shared and visible to all. Hence some tendency for bias exists but it is expected individuals will act in line with their existing convictions most of the time; however, research into social news selection (Westwood and Messing 2011) showed some strength of social bias.

35 Hayes and Krippendorff (2007) provide a useful overview of inter-rater agreement measures. The custom SPSS kalpha macro written by Hayes and Krippendorff was used to compute the alpha and its confidence intervals – see <http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html>

36 This presented a sample of 35 unique news-items, since each item had to be rated by a minimum of five users.

agreements. For instance, user 6 tends to agree the least with all other users and at times has a tendency to consistently disagree (indicated by a negative alpha value, e.g. last row, right most column in table 7.3), however agreements between users 1 vs. 4, 1 vs. 3, 3 vs. 4 and 4 vs. 5 tend to be consistent and high (rows 2, 3, 10, 13).

For the generic purposes of collective intelligence it is desirable for the users' judgements to provide an estimate of the population average perception even though homogeneous subsets of users may produce a far more consistent perspective on the news data. For example Krippendorff's statistic for users 1, 2 and 5 shows a moderate agreement of .4420 (.2514, .6146) for duration judgements, as opposed to virtually no agreement (.0702) amongst all raters. Identifying homogenous subsets of users in terms of common news-interpretation is plagued with several issues and unfortunately was not further investigated in this study for reasons discussed in the limitations section 7.5.1.

Comparison	Sentiment - Kripp. α	Impact - Kripp. α	Duration - Kripp. A
User 1 vs. 2	.6729 (.4499, .8662)	.0722 (-.5041, .5342)	-.0878 (-.7236, .4489)
User 1 vs. 3	.8407 (.6960, .9421)	.3369 (-.0100, .6488)	.0880 (-.1386, .3267)
User 1 vs. 4	.8215 (.6568, .9451)	.4810 (.2312, .7101)	.7831 (.4925, .9880)
User 1 vs. 5	.8083 (.6314, .9410)	.2674 (-.1507, .6189)	.3540 (.0131, .6371)
User 1 vs. 6	.3593 (-.0349, .6797)	.0446 (-.3661, .3976)	-.0632 (-.4369, .2927)
User 2 vs. 3	.6360 (.4043, .8180)	.3962 (.0534, .6843)	.0940 (-.1057, .3036)
User 2 vs. 4	.6339 (.4143, .8243)	.3631 (-.0298, .7012)	.0373 (-.5389, .5049)
User 2 vs. 5	.5528 (.1592, .8390)	-.0079 (-.4826, .4253)	.1238 (-.3095, .5142)
User 2 vs. 6	.3337 (-.0146, .6214)	.1310 (-.3280, .5306)	-.1109 (-.6782, .3753)
User 3 vs. 4	.7900 (.6100, .9400)	.4876 (.2542, .7098)	.1233 (-.1293, .3829)
User 3 vs. 5	.7418 (.5159, .9193)	.3573 (.1839, .5408)	-.0850 (-.4894, .2502)
User 3 vs. 6	.2691 (.0001, .6205)	.0351 (-.3824, .4019)	.0124 (-.1242, .1606)
User 4 vs. 5	.8809 (.7469, .9702)	.4336 (.0891, .7252)	.1681 (-.2347, .5142)
User 4 vs. 6	.3304 (-.1786, .7054)	-.0272 (-.4036, .3304)	-.0038 (-.4519, .3906)
User 5 vs. 6	.3558 (.0187, .6704)	.3290 (-.0053, .6281)	-.2280 (-.6404, .1512)

Table 7.3 – Sentiment, Impact and Duration (Krippendorff's alpha, in brackets are the LL95%, UL95% confidence intervals based on 10,000 bootstrap samples) agreement for top six Newsmental users

Overall it would seem individuals have at least some common interpretation of news events, especially polarity judgments were shown to be highly reliable. Judgements regarding duration and impact of news events may be treated as indicative only. A better understanding of the

identities of various homogenous subsets of users, however, may provide more insight in this respect. Nevertheless, using a simple average of the judgement ratings throughout further analysis seems reasonable at this point also as relative to each other the ratings for duration and impact presented in table 7.3 seem to provide an intuitive interpretation of events.

7.4.2 Pre-processing

There are a number of pre-processing steps that can be applied to the news-item judgements, these are briefly discussed here. Given a judgement (or rating) J_i , where $J_i = \langle s, i, d, s_1, s_2 \rangle$, so that the record stands for the sentiment, impact, duration, sentiment for custom entity 1, sentiment for custom entity 2, respectively, with the following value ranges; $-2 \leq s \leq 2$, $0 \leq i \leq 100$, $0 \leq d \leq 4$, $-2 \leq s_1 \leq 2$ and $-2 \leq s_2 \leq 2$, we can compute a number of basic statistics on J_i where i ranges over time units, news-items or other units of analysis. For example J_i for all i of a given day or a given news can be aggregated, an average or standard deviation computed. The average essentially represents the consensus opinion for a feature, and standard deviation the average disagreement between users, for a day, news-item or some other aggregation unit for J_i . One could for example build a composite measure, taking into account the variability of judgements by normalising using the standard deviation, $\frac{\mu(s)}{1+\sigma(s)}$, where $\mu(s)$ and $\sigma(s)$ are the mean and standard deviation for $J_i \langle s \rangle$. Other more involved summaries of aggregate ratings are possible. However the values for s, i, d, s_1, s_2 are in a good range, intuitive, and easily interpretable, hence they shall not be pre-processed further except when J_i for certain i is aggregated – such as in section 7.4.3.1 by news story and section 7.4.3.2 by time unit.

7.4.3 Collective News Analysis

The time-period under analysis was rich in a number of noteworthy economic events. Given the features related to a rating J_i , events can be investigated for their collective interpretation. Entities extracted from news over the six week period (30th May – 11th July 2011) represent a useful semantic index of the topics covered by news stories. Entities were hence used as an index to filter stories on specific topics³⁷. User submitted tags may have also been used as an index for the ratings; however, due to their sporadic use such an index would be very incomplete.

³⁷ The top ten entity types during the covered period were (counts in brackets): Person (18,230), Company (15,373), Position (15,321), IndustryTerm (14,223), Country (11,706), Organisation (10,992), City (5,103), ProvinceOrState (1,701), MarketIndex (1,661), Continent (1,642).

7.4.3.1 Lagarde as head of IMF

During the month of June there was much speculation as to who would be the next head of the IMF (International Monetary Fund), with the main candidate being French economist Christine Lagarde. Table 7.4 presents chronologically ordered news items relating to Christine Lagarde, her candidacy and her subsequent election as head of IMF. The comments column provides some explanation of news ratings. In summary news supporting Lagarde in the run-up to her election was perceived as positive with her election being considered to be an important event (average impact rating – 79 out of 100).

News Headline	News Date	Sentiment avg.	Comments
Lagarde in Brazil to promote IMF candidacy	30-May-2011	1 (1 rater)	
Lagarde likely to be next IMF chief – report	05-Jun-2011	1 (1 rater)	
Inquiry threat may linger for IMF hopeful Lagarde	08-Jun-2011	0 (1 rater)	The overall news story received one neutral rating however entities Christine Lagarde and IMF were both rated at -1.
South Africa's Manuel opts out of IMF contest	10-Jun-2011	1 (3 raters)	
Lagarde in lead for IMF, South Africa's Manuel opts out	10-Jun-2011	2 (2 raters)	This and the former news were considered positive news. Since France was largely expected to provide an IMF chief, another, especially emerging economy country candidate would upset the existing balance of IMF and the world bank.
Lagarde still favourite as IMF nominations close	11-Jun-2011	1 (2 raters)	
Indonesia backs France's Lagarde for IMF job	12-Jun-2011	2 (1 rater)	Indonesia was the first major emerging market to back Lagarde.
Lagarde strengthens IMF bid with Indonesia backing	12-Jun-2011	2 (1 rater)	
Lagarde gets Indonesia backing for her IMF bid	13-Jun-2011	1 (3 raters)	
IMF could pick Lagarde as chief as soon as	28-Jun-2011	1 (2 raters)	

Tuesday			
Lagarde wins IMF top job, presses Greece on crisis	28-Jun-2011	1 (2 raters)	Average impact of the news was 79, which is high compared to other news-items.
Lagarde's selection marks a break with IMF's past	29-Jun-2011	1 (2 raters)	None of the four raters have rated Lagarde's election with a negative score.
Exclusive - Lagarde to give China bigger IMF job	06-Jul-2011	0.33 (3 raters)	One of the three raters perceived this as a negative news, even though overall the news was evaluated to be somewhat positive.
Lagarde wants more credible IMF	06-Jul-2011	1.33 (3 raters)	

Table 7.4 – Newsmental news-items relating to Christine Lagarde (as based on entity1 and entity2 index)

News-items in table 7.4 are covered by only 28 ratings; however, it is sufficient to indicate how the news was perceived. Five user-comments were also submitted within the 28 ratings, these are shown in table 7.5. Essentially these comments express some basic observations or further clarification of opinion relating to the rated news (comments are public and shared, see section 7.2.3). User-comments may be useful in providing further insight into the rating.

Lagarde would be a good head of IMF, as a woman it would also mean a lot symbolically
Good news for Lagarde, her chances have increased with Trevor Manuel falling out of the race.
good for the eu
It is good for the IMF and world of big business and finance to have more women, I think as a society we will benefit.
This was expected but is important!

Table 7.5 – comments left behind by several raters on news relating to Christine Lagarde

Unfortunately only 199 of the total 2,138 ratings contained a user-comment. A rudimentary sentiment scoring using a lexicon of positive and negative terms (compiled by Bing Liu, Liu 2010³⁸) to assign scores based on the sum of user-comment with lexicon term matches produced a .468 (sig at 0.001), .378 (sig at 0.001) and .298 (sig at 0.001) spearman correlation with the computed sentiment score and the rating's sentiment, entity 1 and entity 2 ratings respectively. As expected, this confirmed that user-comments are compatible with emotions

38 The sentiment lexicon (6800 terms) is available at <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>, and the code for sentiment scoring is accessible on http://www.newsmmental.com/thesis/newsmmental_project.html. Similar simple sentiment scoring method was used by others in the past, e.g. Gillam et al. 2002.

expressed in sentiment ratings to some extent; however, it seems that user-comments do not necessarily contain opinion words and / or these cancel themselves out with the comment possibly contrasting, comparing or simply providing additional facts / observations on news story.

7.4.3.2 Greece

Greece has been going through a lot of difficulty in repaying its debt and maintaining various commitments, to the extent that the EU (European Union) had to provide emergency funds to avoid a Greek bankruptcy (summer 2011), which could have major impact on the Eurozone and EU as a whole. Newsmental amassed valuable collective opinion, throughout the projects’ lifetime, concerning events of the Greek economic troubles. The median overall sentiment was found to be -1 for the entire time-frame, which points to an overall negative sentiment for news relating to Greece. The period can be broken down into six groups of ratings, a week of ratings for each of the six weeks of news data. A Kruskal-Wallis test³⁹ for overall news sentiment, news impact and news effect duration found that except for impact, the null hypothesis (i.e., that the distributions of ratings across all six weeks are the same) was rejected at $p < 0.001$ significance level. Figure 7.13 illustrates the distributions of news sentiment for the six weeks, from which it can be appreciated how sentiment ratings for different weeks have varied considerably.

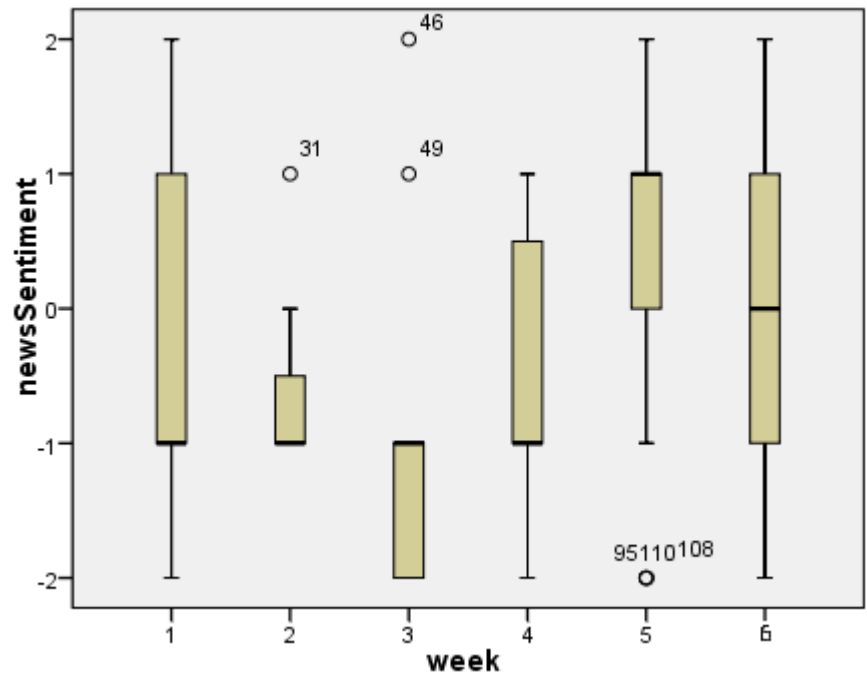


Figure 7.13 – Overall news sentiment, each box-plot represents a week of sentiment ratings relating to Greece

39 This is the appropriate non-parametric test to be used with more than 2 groups, to test whether samples originate from the same distribution, i.e. the samples originate from populations which have the same median. This test is an extension to the Man-Whitney U test.

During the first week (N=27) headlines such as, “Greece may run out of funds within weeks”, “Protesters take over finance ministry in Athens” were received very negatively. “Greece likely to get aid tranche” had five reviews, where three of them were negative and two positive, pointing to the mixed sentiments to seemingly good news. User-comments included “I hate Greece, they already received so much support and still they are unthankful..” or “i hate to imagine how much of our money goes to greece!!!”, these comments can provide some insight into the reasoning behind reader judgements.

In the second week (N=7) only 1 news-item was rated as positive, with Germany suggesting that private investors should share the burden of any further financial help for Greece.

On the 13th, Monday of the third week (N=17) Greece was downgraded by S&P rating agency to the country with the lowest credit rating in the world. This caused high market volatility the next day⁴⁰ and was perceived as highly negative and high impact news with long news-effect duration (avg. -1.12, 67 and 2.24 respectively). The next few days were followed by nationwide strikes, rioting, and the Greek PM offering to quit after mass protests. Interestingly one Newsmental user rated the PM’s offer to quit as positive, and explained his view in their user-comment; “This is good news, since it shows the political will in Greece to work towards a solution!!!“. Overall however fourteen out of sixteen ratings during the third week were negative.

The 20th June, beginning of the fourth week (N=40) was marked with euro-zone finance ministers delaying a decision to extend emergency loans to Greece, and offering an austerity measures ultimatum to Greece instead. This dragged world markets down on the day⁴¹ as it stirred wide-spread worry, and a negative sentiment was confirmed with 21 out of 24 Newsmental ratings for the 20th June being negative (avg. -1). The only non-negative sentiment for the day were three ratings (two of them neutral and one -1) relating to news on offering an ultimatum to Greece, which was by very few individuals perceived as a strong step from the EU and a good thing⁴². However a key late-night vote was well received and caused the markets to recover somewhat.

The fifth week (N=50) began on a mildly positive note the 27th June. Only 3 out of 13 news judgements for the day were negative (avg. sentiment 0.62), as the Greek prime minister indicated that rebel lawmakers in his parliament will likely back a key vote on austerity measures later in the week, and a plan by French banks for a Greek debt rollover was announced. The French plan was positively received the next day (3 positive ratings, avg. 1)

40 As measured on the representative Greek stock-index, Athens Composite Share Price Index (GD.AT) which accumulates the 42 largest listed Greek companies by turnover.

41 As measured by major market indices, S&P 500 and FTSE 100.

42 Expressed poignantly by the comment of one user who thought the news to be positive; “an ultimatum is necessary, thats why it is a good news, but what is happening with greece is terrible also for the eu”.

however the initiation of a 48 hour strike that turned violent over-shadowed any positive sentiment, and the remaining ratings for the day were all negative (avg. -1.4), a user-comment captured this sentiment, “These protests are very bad for the entire process of the Greek rescue.”. For the next four days news became dominated with stories of a successful Greek vote of key austerity measures and the final Eurozone approval of a related rescue package, 26 out of 30 ratings being positive (avg. 0.83).

In the sixth week (N=65) there was a wide mix of both, positively and negatively perceived news stories (avg. sentiment for the week, 0.05). To give some idea of the events; among negative news was a hearing of an appeal against euro bailouts at the highest German court (7 out of 7 negative ratings, avg. -1.17), or the lack of progress among European officials on securing a private sector contribution to the second Greek bailout. The worry over privatisation of state assets of Greece (and the related limitations imposed on Greek sovereignty) was however perceived positively (2 out of 3 ratings were positive, avg. 0.33) one user-comment pointed out the opportunity implied by such news – “Gives foreign investors a lot of opportunity, already I heard people began buying up cheap real estate in Greece!”. Further to this, the IMF approved several billions in rescue contributions which was perceived by all readers as positive news (6 positive ratings, avg. 1.17).

7.5 Discussion

The previous sections (7.4.3.1 and 7.4.3.2) qualitatively illustrated the utility of collective news analysis. The explanatory value of collective knowledge in the form of news analysis from a system such as Newsmental is evident. The judgements in aggregate provide a consensus news assessment, augmented with clarification or insight from user-comments. There have been significant efforts – especially in financial computation research – to automate and achieve reliable news analysis (e.g. Wuthrich et al. 1998, Fung et al. 2005, Mittermayer and Knolmayer 2006); however, plagued with a number of problems and with varying success. Arguably several human analysts will often disagree on news interpretation (Koppel and Shtrimberg 2006), how severe a news effect will be and what it means for affected actors or entities. Replicating human judgement, experience based induction and analytic abilities, as well as parsing an unstructured piece of text is exceedingly complex. A web 2.0 system such as Newsmental does not need to replicate human judgement as a function within some statistical / AI model rather it taps into the social participation of web 2.0 users at a relatively low cost. Despite the benefits provided by a system such as the one presented in this chapter, the complexity of building a web 2.0 application and maintaining it with high user participation is

plagued by a number of issues. In the next sub-section problems and limitations of the system are discussed, followed by a sub-section on future work and recommended improvements to a future collective intelligence style web 2.0 application.

7.5.1 Limitations

Despite collecting a considerable amount of news judgement data, a major limitation of this study was a relatively low count of judgements, which limited the analysis of collective intelligence. Altogether 1,070 news articles were judged during the six week study⁴³, only 536 received two or more ratings; with 2,138 total ratings averaging 1.98 rating per news story (the distribution is presented in figure E.1 within the appendix E). This meant that in order to quantitatively evaluate the links between financial market prices and sentiment ratings with statistical significance, or to predict values in the future, quantitative models could not be applied in a reliable manner, and hence instead the evaluation was mostly limited to an interpretative and reconstructive qualitative analysis. A generic linkage between the markets and collective news judgements and the interpretative value of the latter were established using these methods. However, the amount of contributed data was lower than expected, and three causes for this can readily be identified; 1-Timing of the (viral) marketing (see 7.3.2), 2- Initial technical issues in week two and three (see 7.3.1), 3-Late introduction of some type of social competition element (see 7.3.2). Focus group and individual user based feedback indicated that users were overall happy with the news analysis interface and the presentation of news stories. Despite this it is believed the user interface could benefit from re-designs. For example, a design that would put more emphasis on the display of user-comments embedded along the news on main-page would likely have helped to increase the ratio of users who contributed user-comments as well. Unfortunately the scale of the study did not allow to experiment with various other interface designs after the project launch.

Many news-items evoke a certain type of affective response within most individuals. The response can be much stronger from a well read and opinionated individual, hence news-rating habits in aggregate for an individual will differ from another who tends to read less news and / or subscribes to different set of fundamental (political, economical, etc.) opinions. A limitation of this study was the inability to investigate such differences (i.e. the identification of homogenous subsets of users in terms of their news-interpretation tendencies) due to lack of data, as mentioned above. Even though an attempt was made at collecting basic demographic and news reading habits information on a voluntary basis from users, in future studies more

⁴³ In the last two weeks the number of average daily news judgements submitted, approached 100 whereas within the first four weeks there were only 30 ratings per day.

systematic attention must be given to the choice of information which differentiates individuals affect the most and such information collected, with the aim of linking it to news-interpretation tendencies.

Newsmental was primarily advertised within the student community, as detailed in section 7.3, e.g. eleven student finance societies across the UK were approached. Although the system was also advertised with several sizeable non-student financial market groups, overall the user base of Newsmental contains a bias towards individuals from higher education institutions in the United Kingdom. This bias potentially impacts the evaluation of news. For instance, a group of experienced financial market traders or investors will likely exhibit different news evaluation behaviours due to their experience, and financial background, while a bias towards the UK implies that news is more readily evaluated from a British perspective. The identification of homogenous subsets of users (in terms of their news-interpretation tendencies), as discussed in the limitations paragraph above, highlights how to deal with biases among various user groups. In fact both points are closely related.

Hayes and Krippendorff (2007) point out that to measure reliability of ratings the data-generating process must be informed by instructions that are common to all observers who identify, categorize, or describe the units of interest. Users of Newsmental were not explicitly instructed on how to judge news items, this was done on purpose and partly efforts were made to design the website so as to imply the judgement criteria implicitly. Even though judging the sentiment of for example; a news item on S&P downgrading USA's credit rating, may well be self-explanatory, the judging of news duration and similarly news impact, will be much less so. It may not be clear whether a news reader refers to the duration people will talk about (buzz of the news), the duration of the news' effect on the markets (which would have long term effects to the debt markets, but possibly short to mid-term on the general stock-markets - O'Neil 2002), or some other "personal interpretation" of the news-item duration rating. Hence the fact that judgement criteria were implied implicitly must be taken into consideration when interpreting results.

The agreement statistics (in section 7.4.1) were very low for duration judgements. The low agreement may be contributed to by an artefact of the user-interface, since the news presentation interface only showed a graphic for collaborative sentiment and impact judgements but not duration. Hence some limitation is imposed by specifics of the user-interface design, discussed further in the next section.

7.5.2 Future work

The introductory chapter 2.3.2 presented substantial background on elements of social media interaction on which this chapter has built-on to deliver a custom web 2.0 solution. Although best efforts were made and system design guidelines followed in its development, there are naturally potential improvements to future versions of this system, and for that matter a new system design should carefully take into account all below presented suggestions.

As was mentioned in the previous section, the news analysis interface and the presentation of news stories design could have been optimised further, possibly by employing A/B testing over several iterations of the user interface design (Subraya 2006).

In order to judge a news-article the Newsmental website has to be used directly to submit news judgments – this could be seen as a limitation by some, who prefer to browse their favourite news-portals directly. A quicker and more preferable option for some users may hence be to employ a custom bookmarklet that allows to judge articles from within any news website by bringing up a temporary interface and subsequently sending the data to Newsmental's database. Bookmarklets provide a relatively platform \ browser independent way of adding programmatic functionality to a simple bookmark, and are in common use (see section 2.3.2.1, also). Assuming that the clustering algorithm would take care of organising the many stories from heterogeneous news sources, this seems to be feasible.

Pang and Lee (2008) suggested that within some web 2.0 applications it would be useful to provide a finer-grained object of contributions (e.g. identify and allow to rate particular aspects of a product in a review system), and further pointed out that, still, few systems use this capability. Within Newsmental the analysis of unstructured text was automated and the user was presented with the possibility to pass judgements on entities from within the news stories themselves. Unfortunately this feature fails on occasion to identify the most sensible objects / entities of interest in the article. A possible extension could rely on the users collectively picking an entity they deem important from a choice of system identified ones, and when enough users picked the same entity, it will be suggested by default.

Forte and Bruckman (2005) discuss the role of motivation and incentive and its implications for designing online communities. One of their suggestions is that regulars should be rewarded and leaders should be empowered. The empowering can be done by giving those users more rights, for example, to organise or otherwise manage content in an editorial fashion. Users of Newsmental, who are more involved than others may be given extra rights to organise, rank or otherwise edit certain sections of the news display. Although visitors to Newsmental were rewarded in form of a potential prize for regularly reading news, different ways of motivating

participation ought to be explored.

Social endorsements were shown by Westwood and Messing (2011) to be a much stronger predictor of news stories selection than did source cues. The mere presence of social endorsements reduced partisan selectivity to levels indistinguishable from chance (study was conducted on facebook, on a sample of USA users). In future evolutions of this project it might be desirable to increase transparency of user contributions (i.e. making user contributions more easily attributable and visible to other users), so that social endorsement and engagement can play a bigger role. To increase social exposure further, Newsmental accounts could be optionally linked with Twitter and Facebook platforms and allow sharing of judgements with the user's existing social networks.

Finally we conclude this section by reminding the reader that each of the entities identified in a news article has an RDF triplet id, and in fact a substantial subset of entities link to linked-data via this RDF id. Since entities are semantically identified it would not pose too much of a challenge to augment the current processing of news articles by tying in linked-data to perform basic semantic reasoning to produce more intelligent summaries of news stories from entities (Antoniou and Harmelen 2008).

7.6 Summary

A case study of designing a custom collective intelligence source in the form of a news-judgement platform was demonstrated in this chapter. The platform can be adapted to other kinds of information, and is currently licensed under the MIT open-source license. The system was designed following web 2.0 design conventions, and various considerations for such systems were discussed. In prior literature the author hasn't come across a full walkthrough of an actual web 2.0 system design, such as the one provided in this chapter. The system succeeded in engaging numerous users within a small community over a multiple-week period. Two well known events, relating to the financial crisis were selected and judgements of users for all news stories that were representative of the two events were discussed at length. This helped to illustrate the value of collective intelligence from Newsmental.

The chapter's main intention was to illustrate how to engineer a web 2.0 system with the main intention of it serving the purpose to amass CI, while various design choices would create value for the target web 2.0 community. In the previous chapter (chapter 6) several web 2.0 systems were presented as potential sources of CI, in this chapter, justifications were given for when a custom collective intelligence source may be more desirable.

In the next chapter (chapter 8), the framework for CI exploitation / usage will be presented. Newsmental is only one of many CI sources within this presented framework.

8 Collective Intelligence Overview and Characterisation



In this chapter the concept of collective intelligence is introduced and the assertion is presented that web 2.0 systems can essentially be regarded as voluntary, streamlined and efficient ways of polling for opinion and knowledge. This process bears similarities with implicit voting encapsulated in the trading of financial assets, which will be introduced using the Efficient Market Hypothesis from the field of Economics. The main contribution of this chapter is the introduction of a systematic step-by-step characterisation of how to approach the problem of utilising collective intelligence from web 2.0 applications. This description acts as a guideline on how to leverage and use collective intelligence from web 2.0 applications in real world problem domains. In other areas guiding frameworks exist as well, and it is believed that such guidelines may help in further research and application / systems development.

The chapter is structured as follows. First the need for a framework and a gap in the literature is identified, then some background on collective intelligence is provided, followed by the proposed model with its building blocks with a critical evaluation highlighting some issues. Finally a couple hypothetical applications of the framework in real world problem domains will be presented, and the chapter concluded. In terms of the entire thesis, the initial four chapters provided a detailed treatise on various aspects of web 2.0 to better the understanding of this relatively recent phenomenon. The final three chapters, including this one were concerned with issues relating to the financial markets and the potential collective intelligence represented

within web 2.0 applications.

8.1 Motivation

In section 2.4 substantial efforts to apply web 2.0 systems to improve communication and processes in various fields have been presented, it has been seen that RSS, Blogs, open APIs or social networks, for instance, can deliver much benefit within business, medicine, geography and other fields. Much has been published in academic literature within recent years on topics of implementation and effective social media use in a number of different areas (Murray et al. 2005, Alexander 2006, Anderson 2007, Boulos 2007, Dwyer 2007, McLean et al. 2007, DiMicco et al. 2008, Blumenthal et al. 2010). However, there has been far less research activity in the field of collective intelligence, and especially a discussion of its possible uses in different domains. The assertion is that instead of simply using web 2.0 tools, the value of UGC generated from such tools is recognised and leveraged in aggregate within Decision Support Systems (*DSS from now*) or systems that can automatically use the collective intelligence to infer decisions for practical purposes¹. This is different from looking at using the insights gained from analysis of UGC datasets to improve various web 2.0 systems themselves, even though many concepts and techniques are similar².

The data from web 2.0 applications is potentially very powerful. The reasons why user generated datasets from web 2.0 are important and valuable, and the concept of collective intelligence itself, will be discussed in the next (Background, 8.2) section. Earlier web applications allowed for subjects to easily share opinions and thoughts in natural text such as on forums³; however, with the uptake of web 2.0, focused, shared contributions at varying granularities became possible and more commonplace. In chapter 2 (section 2.3.2) a systematic taxonomy of web 2.0 applications was introduced, which assesses the elements of sharing (section 2.3.2.2) and the data-types of UGC being shared (section 2.3.2.1). Given the taxonomy and the understanding of the elementary collaborative activities (as discussed in 2.3.2), and some understanding of web 2.0 systems, a systematic set of guidelines to leverage collective

1 To some degree practical applications of Webometrics can be seen as a generalisation of this idea for the wider web.

2 Alag 2008, points out that so called social media “giants” use collective intelligence ubiquitously; “*Youtube recommends similar movies, Last.fm knows what one would like to listen to, Flickr which photos are ones favourites, and Amazon what other products someone might find interesting for purchase, based on what other users with a similar online profiles liked in the past*”. In the context of this thesis collective intelligence is understood at a higher-level, with a wider cross-application reach, as will become clear from the proposed framework. What Alag 2008 refers to is the field of Recommender Systems, covered by conferences such as the ACM Recommender Systems Conference series, see <http://www.recsys.acm.org/2011/index.shtml>

3 By some this is already considered to be social media, and in fact it is; however, in the earlier days of the Web, forums were quite basic, often in the form of mailing lists with a mirrored web presence.

intelligence for various application domains can be constructed. Indeed a systematic characterisation that can act as an initial guide to leveraging and using collective intelligence from social media (i.e. web 2.0) is needed. Such guiding quasi-frameworks exist in other domains. For example in Data Mining, many techniques and algorithms often require customisation in order to cope with different data types, data distributions, patterns / prediction problems. Yet an abstract model guiding the overall aspects of such analyses can still be useful and the CRISP-DM 1.0 was suggested by Chapman et al. (2000). The authors of CRISP-DM argue that at the time of publication the field of data-mining was still “*young and immature*”, and a need for a standard process was felt. Another example is from the field of time-series analysis and signal processing where one often has to deal with various time-series prediction problems. Many approaches would be undertaken by academics and there was effectively a need to standardise and bring a systematic process into place for time-series prediction problems, hence the ARMA / ARIMA model was introduced by Box and Jenkins (1970).

8.2 Background

8.2.1 Collective Intelligence

The idea of collective intelligence is not new, it emerged from writings by Douglas Hofstadter, Peter Russell, and Pierre Lévy (Hofstadter 1979, Russell 1983, Lévy 1994), but at an abstract level already H. G. Wells mentioned the idea of a collective “world brain” in his essay entitled, “*The Brain Organisation of the Modern World*” (Wells 1938). Recently with the emergence of the web, the idea has gained new momentum with numerous efforts to understand it, such as the *MIT Center for Collective Intelligence* initiative headed by Dr. Thomas W. Malone⁴. Collective intelligence in this sense is the application of basic data analysis, but also more sophisticated data-mining techniques to user generated datasets – datasets acquired within communities through web 2.0 based web applications⁵, for decision making. *Collective intelligence is essentially pattern based decision making based on collective knowledge, where collective knowledge can be collected via Web 2.0 systems.* This has become feasible within recent years and deserves much attention in the suggested context. More recently, in their position paper,

⁴ <http://cci.mit.edu/>

⁵ It is the user generated datasets that in aggregate hold latent value, and in fact user volunteered data has for a long time in society played quite an important role in science, for example the role of amateur astronomy in observations (Kanefsky et al. 2001). The innate powers of perception are unfortunately computationally irreplaceable; however, the world wide web has provided an interesting alternative for the collection of user generated data.

Zhang et al. (2010) present a road-way towards “*social and community intelligence*” SCI research. This is maybe the most closely related work to the *web 2.0* based Collective Intelligence characterisation presented in this chapter. Zhang et al. propose a research area which deals with similar issues; however, their work is clearly separate and different to ours. They suggest that the footprint left behind after mobile devices, GPS, pervasive sensing, and other mobile computing mediated activities (*i.e. related to field of sensor computing*) ought to be integrated into one collective intelligence data-store on top of which applications can be built, and suggest a general architecture for this. There are several parallels in ours and their work, such as the substantial importance of mostly participatory created datasets. In their work however the data-sources are mobile device sensors, GPS, video-monitoring, environment sensing, etc... whereas in our characterisation the data comes purely from web generated *i.e. web 2.0* mediated sources. The framework in this chapter presents specific details towards dealing with *web 2.0* data related issues, which Zhang et al. don’t consider, eventhought the potential use of static user generated datasets from the web as a complementary data-stream is mentioned briefly. Another differentiation is that in this chapter we discuss the concept of the Collective Brain and early related literature, which is a long standing and useful concept that often seems to be omitted in other work.

Ideas presented in this chapter are certainly not new, recently Bermingham and Smeaton (2010) have toyed around with the rough ideas of monitoring or tapping into the potential of user contributed datasets on a wider and applied scale, than has been done up to this point⁶. However, they don’t go as far to present a framework or any type of higher level guidelines to tap into this collective intelligence, instead, issues surrounding sentiment mining and how UGC has affected the text-mining field are discussed. Indeed sentiment mining is an extremely important element for processing free-text UGC data, and a number of researchers have looked into this area (Dave et al. 2003, Liu and Hu 2004, Liu et al. 2005, Pang and Lee 2005, Pang and Lee 2008, Thelwall et al. 2010). Sentiment mining is a specialisation of a generalisation of the problem known as intent mining, and represents one aspect of collective intelligence / UGC processing. Opinion is a kind of intent mining where the attitude is a positive or negative opinion; however, the intent can be preference [likes, dislikes] or agreement [assent, dissent], Attardi and Simi (2006).

The focus of discussion in this chapter is on how to bring the many aspects related to web 2.0, introduced so far together in order to apply these to domain-specific datasets that have been

⁶ An interesting point made by Bermingham and Smeaton relates to how mobile phone messaging (SMS) and instance messaging (IM) essentially represent instantaneous computer chatter, however intrinsically private. This they contrast with the public nature of the Web which allows to readily tap-into a global chatter, or “*collective intelligence*”.

acquired via web 2.0 systems use.

The usefulness and applicability of collective intelligence is highly dependent on the type of collaboratively collected datasets, since this is essentially the collective knowledge on which collective intelligence DSS Systems will be build. Each data-source has to be considered carefully, based on the principle of GIGO, garbage in, garbage out.

The following is a useful workable definition; “*Collective intelligence is a shared or group intelligence that emerges from the collaboration, competition, or sharing of many individuals in response to some challenge and is essentially pattern based decision making based on collective knowledge, where collective knowledge can be effectively collected via web 2.0 systems.*”. An implicit goal exists on practically all web 2.0 systems, in some form. The web 2.0 application can have an explicit goal (*e.g. Wikipedia – the goal is to amass knowledge into an Encyclopaedia*) or an undefined one (*e.g. Youtube – share Videos for entertainment or any reason really*). It is also important to understand what the actual data is that was accumulated via the web 2.0 system. It is useful to evaluate the basic atomic activities facilitated by the collaborative system to collect the dataset. Also essential and related to the atomic activity, is that only types of collaborative applications that leave a sizeable online collaborative footprint / trace are of significance to collective intelligence applications (*i.e. the more atomic activities, the better*). Either way, once there is a large enough dataset collected from contributions of individual users, a web 2.0 system can be considered to have evolved towards certain subjective fitness, i.e. *fitness* = $f(v)$ where *fitness* is a function of v , in which v is some vector of “**subjective**” function parameters (Sykora 2009). Consider as an example social picture sharing applications, such as the well known applications Flickr or Picassa. These allow users to submit images, tag, rate or leave comments on them. The ratings often refer to certain aspects of the visual appeal of images. Visual appeal represents the fitness and the vector v is the set of image features evaluated for – v can be difficult to quantify, as the aspects that make an image appealing differ from person to person, yet this might not matter since the picture will be “*appealing*” by consensus. This picture rating process used in conjunction with tagging⁷, for example, provides (*subjective measure of*) image fitness, in optimally user defined categories for the images (*based on the tags*)⁸. Sykora 2009 elaborates on collective intelligence from a Computational Intelligence point of view, and explains the possibilities for the related Computer Science area of Interactive Evolutionary Computation, for which these issues can be highly relevant.

⁷ Providing short keyword description of the image, see chapter 3, for example.

⁸ Further examples are provided in Sykora 2009, generally speaking this same concept applies to any kind of data (textual data, knowledge in a collaborative knowledge base, social networking applications, etc...)

8.2.2 Value of the Underlying Data

The collective intelligence represents the value locked-in the data contributed through web 2.0 applications. It represents a wide online opinion in the form of blog posts, ratings, tags and potentially many other types of web based social contributions. To simplify this discussion, one can argue that web 2.0 systems can quite simply be regarded as an efficient way of polling or surveying for public opinion. A classical survey would be composed of questions relating to the topic of the survey and usually some demographic questions to identify or group the respondents into relevant demographic groups. On a web 2.0 system, one would often contribute content (whether; videos, tags, ratings ...) which can be regarded as a vote. Usually there is also a profile associated with the user, which may contain valuable demographic information. Effectively a web 2.0 system can hence gather both, the topic of interest (social contributions) and the demographic information (user-profile). Akcora et al. (2010) for example point out that polling through Twitter has three advantages over traditional public opinion polls. *1–A classic opinion poll is not available over time continuum, unless re-pollled regularly (i.e. on twitter opinions can be tracked over time relatively easily), 2–Cost effective in reaching many individuals and 3–Capture opinions about the topics that are not asked in a questionnaire.* A simplistic view suggests that a web 2.0 system is a medium for streamlined, often voluntary and hence probably highly efficient way of polling knowledge from users, compared to more traditional surveying methods. In addition demographic information is sometimes provided by identifiable user-profiles; and even when profiles are not available, more recently, linguistic analysis techniques were developed, and became mature enough, to induce age, sex and other demographic background from anonymous textual contributions⁹ (Argamon et al. 2009).

Another way to look at UGC within web 2.0 systems is to consider a well known theory from economics explaining the price creation process for financial markets. Efficient Market Hypothesis, originally proposed in the 60s (Fama, 1965), essentially states that market participants have equal access to information, and as new information affecting a market comes out, this information is assimilated into the market price almost instantly. It was shown that with a relatively low number of rationally and well informed and well behaved participants, markets tend to become highly efficient. In other words the stock-market¹⁰ price creation

⁹ This type of analysis is still in its infancy. It is a separate but related problem to authorship identification in linguistics. For example, there is a new commercial system, Subtext³ which implements this task and is sold to social media marketing companies specifically, however the uses of such systems are much broader <http://blog.subtext3.com/2011/03/who-are-they-not-just-what-do-they-want/>.

¹⁰ An excellent treatment of EMH is provided in Fama (1965). Competing hypothesis have been suggested, for example the AMH (Lo 2004); however, given empirical evidence the EMH is still the most widely accepted hypothesis within Finance and Economics.

process through buy and sell orders is efficient in terms of optimal valuation of a stocks real price. Given all available public information, markets are extremely efficient in information transfer from a few knowledgeable participants, to incorporate this information into an optimal price. EMH applies to almost any financial asset, eventhought there are different levels of accepted efficiency, with more fluid (higher volume) markets generally being more efficient. This has been an important catalyst in the recent uptake in trend prediction markets, already mentioned in section 6.6.1. For the purpose of this work suffice it to say; given enough rational participants on a web 2.0 system, information will likely propagate highly efficiently, synonymous to what EMH would anticipate – see section 6.2.4 (on efficient information transfer).

8.3 The Framework

Models have been proposed to leverage collective intelligence, for example the *MIT Center for Collective Intelligence* has recently (Nagar and Malone 2011) proposed a model for combining human and machine based collective intelligence. This is a model based on prediction markets (see section 6.6.1) that combines predictions from groups of humans and artificial-intelligence agents to show that they are more robust than those from groups of humans or agents alone. Our model; however, is entirely concerned with web 2.0, and generic enough for heterogeneous web 2.0 systems to contribute, i.e. prediction markets would be one of the inputs into the framework. Cheong and Lee (2011) present their framework that similarly attempts to use collective intelligence; however, there is no attempt by the authors to generalise this to other web 2.0 applications, it is rather narrow, and only applies to a specific Micro-blogging website and field of application. An early draft of the here proposed model was presented in Sykora (2011), which has since evolved into the current version, presented in this chapter. As it will become evident the guidelines are generic in that virtually any application domain can be related to them. As a framework it is a guideline with an element / layer of abstraction, leaving some free choice to the designer, yet providing a workable blueprint to follow. Its main aim is to identify and engineer a complete collective intelligence dataset from a multitude of social media systems. The output from the guidelines can ultimately be used in decision support systems or prediction models. What follows is a formal description of the model.

First the domain of interest denoted D , for which collective intelligence is to be applied to, is assumed to be given. This can be the financial markets (as is the concern of this thesis), law enforcement, medicine or virtually any domain that relates to human individuals whose

collective intelligence matters in some way. The following three primitives, M , A , and U are defined.

Let M be the set of all existing web 2.0 applications. We write $M = \{ M_1, M_2, \dots, M_K \}$, where M_i is the i^{th} web 2.0 application, such as *Youtube*, *Twitter*, *Facebook*, *Flickr*, *PatientOpinion*, *Newsmental* (chapter 7), etc. If A_i denotes the set of all atomic activities supported by a specific web 2.0 application M_i , then $A = \{ A_1, A_2, \dots, A_K \}$ is the collection of all sets of atomic activities corresponding to M . Examples of atomic activities could be *submitting a micro-post*, *rating (score based rating)*, *rating (binary rating)*, *tagging*, *commenting*, etc. Similarly U_i denotes the set of all users on a social media application M_i , and $U = \{ U_1, U_2, \dots, U_K \}$ is hence the set of sets of users corresponding to M .

These three primitives are important because it is based on them that collective intelligence data extraction is driven on, hence **atomic activities**, **topical prevalence** and the **user base** are the three significant elements in the framework model. Where atomic activities determine what actual data is picked, the social media and users can topically limit the data, and the user base further limits the data selection based on user related criteria. First of all a set of social media which topically relates in some direct or indirect manner to the problem domain D has to be chosen, so the domain of interest is used to select the initial set $M_D \subset M$ of web 2.0 applications. Next, all atomic elements for M_D are joined into a set that represents all the atomic activity types $A_D = \bigcup A_i$, where $A_i = \{A_{i1}, A_{i2}, \dots, A_{ij}\}$, and since A represents atomic activity types it is possible for example for $A_{ip} = A_{iq}$ where $p \neq q$, but it is also likely that $|A_z| \neq |A_d|$ where $z \neq d$, as applications contain various different features. Similarly $U_i = \{U_{i1}, U_{i2}, \dots, U_{iL}\}$, where each U_{ij} defines a user on a web 2.0 application M_i . Every U_{ij} has specific user-information or user-profile associated with it, so we note that $\rho(U_{ij})$ is a record of attributes U_{ij} and contains these user specific details. In practice it is usually infeasible or unnecessary to consider all A_i , and all U_i and hence a filter ψ preserves only the atomic elements to be analysed, $A_i^* = \psi(A_i)$, similarly to produce the set U_i^* from U_i filter ϕ is applied. In every practical application of the collective intelligence framework some form of ψ (atomic elements filter) and ϕ (users filter) will usually have to exist, but note that these filters are higher level abstractions.

The ψ filter in an ideal scenario filters out the least elements (i.e. ideally $|A_i^*|$ is as large as possible), in order to have as complete a snapshot of user contributed content; however, usually one has to consider three reasons for including or excluding atomic activities, separately for each M_i . **I-Technical limitations** – for example in the case of *Youtube*, actual extraction of the Videos may be infeasible due to technical challenges in processing these and storage

limitations, **2-Relevance** – some atomic collaborative elements may not be relevant at all, for example a study solely on polarity of reviews would likely not require the full-text review posts to be retrieved if an explicit polarity rating is available, **3-Dependence** – on other atomic collaborative elements would require for all elements on which a target element depends on. For example a binary rating (useful or not) depends on the element that is being rated.

In any practical scenarios it would be infeasible to extract information (A_i^* contributions) from all the users of a web 2.0 system (see sections 8.3.1.1 and 8.3.1.2)¹¹, hence the need for φ filter. Most often users are randomly or systematically sampled, as was the case in chapter 6. Systematic sampling might be common where for example the importance of individual users is low but of the topic is high; items from search results, or topical categories relating to some topic(s) would act as φ and drive the selection of U_i^* . In cases where experts or specific groups of users of interest can be identified, the φ filter would provide the necessary rules to extract only such users. On Wikipedia for example users have profile pages where these are awarded badges and community rewards for their contributions, and it may hence make sense to retrieve information which only such users have contributed¹². Another example; Amazon classifies its users into numerous different types (section 6.4.4), hence building a φ filter system of rules for retrieving only certain types of users may be a common scenario. The φ can be either based on a-priori knowledge or established by gradual filtering of users through an iterative process. A highly important consideration for φ is to take into account the real-world (physical) person R_i , from set of all real people $R=\{R_1, R_2, \dots, R_n\}$, in order to establish whether a relationships between individual user U_{ij} and another user U_{uj} exists where $u \neq i$ across various M_j , such that R_i is the physical person behind user accounts U_{uj} and U_{ij} . In this thesis such users are referred to as *persistent users* across domains. Establishing such a relationship allows to identify and track the same physical individuals across various web 2.0 applications. Technically this is still a difficult task, if not only partially possible, as users are free to use pseudonyms, fake accounts, and even then name collisions are inevitable therefore a unique id for users has to be established across domains. Most web 2.0 systems have their own user-management systems, the best solution up till now seems to be the OpenId authentication protocol (Recordon 2010), that is if the web 2.0 platforms under analysis support these, unfortunately at this time its usage is still in its infancy¹³.

11 Consider a system such as Twitter. However, if a custom system is built, such as *Newsmental* in the previous chapter, all data is readily available (clearly an advantage of custom web 2.0 systems).

12 In fact a number of papers have provided a set of useful measures of article quality based on user-contributions, see Hu et al. (2007) or Lih (2004) for exact details. Contrary to wide-spread belief, Wikipedia contributors tend to have a strong sense of authorship and authors recognise one another and often claim ownership of articles, despite that contributions from users are not explicitly highlighted, see Forte and Bruckman (2005).

13 Some Web 2.0 systems allow the use of Facebook or Twitter account credentials which is a partial solution, but

For a given D the sets M_D , A_D^* and U_D^* essentially represent the valuable share of collective intelligence dataset from M_D . From here, one may go on to build decision support systems, or design forecasting models, based on the amassed collective intelligence. Visualisation and exploratory data analysis (Han and Kamber 2006) may be employed, although ultimately automated decision support with the use of classifiers / forecasting models may be sought. Some examples are provided in sub-section 8.3.2.

In summary then, A_D^* represents the spectrum of dataset, in other words the larger $|A_D^*|$ the more varied the data is, i.e. the more features / attributes there are in the *schema* (from a data-storage perspective). The set U_D^* defines the actual contributions towards A_D^* and can be regarded as the rows in the *schema*. Clearly the resulting final schema from this process will differ in the specifics. Number of tables will vary due to normalisation, customisation and system specifics, yet essentially it is useful to think about the collective intelligence data as A_D^* the attributes and U_D^* the data items in the final data-set.

$/M/$ is very large and changes often as new web 2.0 applications constantly emerge. A full list of these is naturally impractical and unnecessary¹⁴. The researcher should have good knowledge of individual social media applications, so that either specific domain based web 2.0 applications, or more general applications known to provide some potentially useful domain knowledge can be selected. Alternatively, it may well be that little satisfactory web 2.0 applications exist, in such case an implementation of a new custom web 2.0 application may be a feasible way forward. The set A does not change that much since individual elements of the set represent types of activities, and are defined by relatively fine / granular sharing activities. The general understanding of these tends to be therefore quite good. For example tagging has been studied and is relatively well understood (Golder and Huberman 2006), as is the analysis of micro-posts, i.e. extracting intent / sentiment from short texts (Thelwall 2009). As for activities A_i of a particular M_i application, this set may change over-time as new features for collaboration (*i.e. atomic activities*) are added to, or removed from a site. $/U/$ and $/U_i/$ would be highly dynamic as users join specific web 2.0 applications; however, there are some partial lists of high-profile users that are maintained¹⁵, in addition one may monitor for users with certain characteristics

is really more of a *quick fix* to the underlying problem. On the other hand, networking and submission patterns can very likely be exploited to identify real individuals, however this may be a relatively time consuming task.

14 Yet there are some partial lists, http://en.wikipedia.org/wiki/List_of_social_networking_websites, http://edutechwiki.unige.ch/en/List_of_web_2.0_applications, <http://www.web20searchengine.com/>, <http://www.go2web20.net/>, <http://www.seomoz.org/web2.0>. Professionals in many fields create, and maintain their own lists, such as, <http://larryferlazzo.edublogs.org/2011/07/14/the-best-web-2-0-applications-for-education-in-2011-so-far/>.

15 <http://www.realcelebrityprofiles.com>, <http://www.icims.com/blog/post/2009/10/20/21-HR-Leaders-in-Web-20-You-Must-Follow.aspx>, <http://www.tweeting-athletes.com/>, http://www.sportsin140.com/?page_id=13, Artists

$\rho(U_{ij})$, such as users with longest posts, most posts within a time period submitted, or other user characteristics.

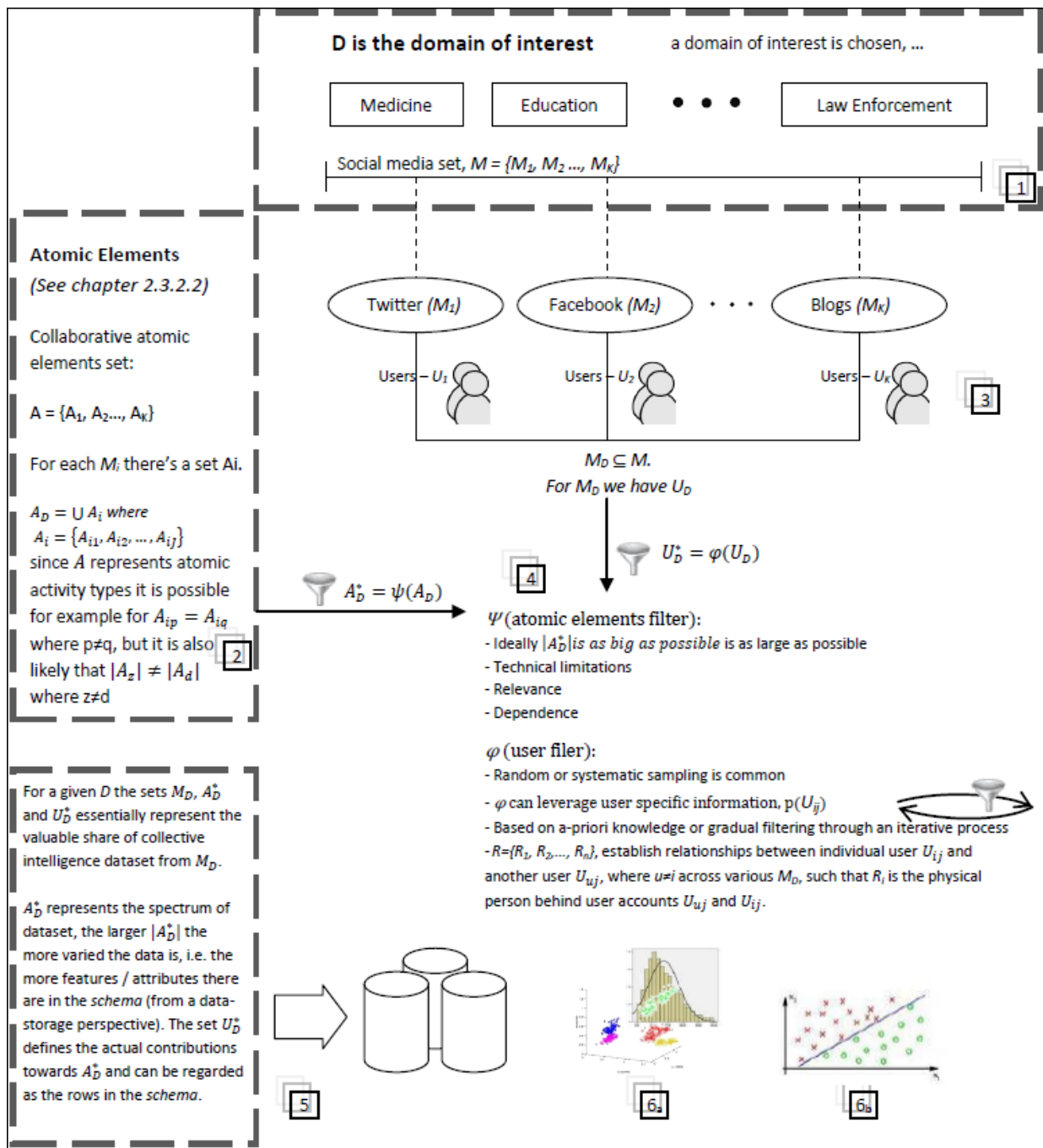


Figure 8.1 – Collective Intelligence framework / guidelines as characterised in section 8.3

The characterisation of collective intelligence use, discussed above, is summarised in figure 8.1, in order to assist in its interpretation. The diagram is read from the top, step 1 (see hint in the diagram; figure 8.1) and then from its left side, step 2, followed by step 3, under step 1 at top of the figure. Essentially in step 1 a decision is made about the set of social media relevant to a given domain of interest, which determines the atomic activity types (step 2) and users (step 3)

that are relevant to each of the selected social media applications, respectively. Step 4 is concerned with filtering atomic elements and users using the atomic elements (ψ) filter and users (φ) filter, as these were characterised earlier. Step 5 highlights the idea of persistence for the UGC data, as based on A_D^* and U_D^* . Finally step 6_a / 6_b simply highlights the subsequent use of the collective knowledge in DSS, where properties of the datasets can be analysed and used in classification / regression models, to effectively facilitate collective intelligence.

The characterisation in this section is concerned with data extraction; however, good working knowledge of web based social media and general domain of web 2.0 is highly relevant. This requirement means that web 2.0 relevant considerations (*that have been discussed throughout this thesis*) are important hence this is rightfully quite a separate task from mere data extraction. Some relevant technical issues related to data extraction are dealt within section 8.3.1, and several examples and hypothetical applications of the framework are illustrated in section 8.3.2.

8.3.1 Extracting Data (Technical Perspective)

The proposed guidelines suggest how to model extracted data in useful ways. However, the technical aspect of extracting the actual datasets from web 2.0 systems deserves further attention, since this is such a crucial step. Many research papers tend to mention a number of issues with the extraction of datasets from web 2.0 systems (e.g. Sykora 2009).

8.3.1.1 Request Limitations

The most restricting issue tends to be the limit imposed by web 2.0 systems on the number of allowed requests to a server¹⁶. Some websites limit the absolute number of requests allowed within an hour or day, while others limit the number by requiring a specific time-interval between consecutive requests. These limits are not always explicitly clarified (even in the terms and conditions), and tend to vary from one website to another. One common way around these limits is to reduce frequency of requests to satisfy the restrictions. This may be highly impractical with a lot of data to download (see Delicious in section 3.2.2). The requests are usually identified by either a unique developer-id or via an IP address. A download distributed request script which sends requests from multiple machines in a synchronised and non-abusive manner, or alternatively the use of a set of proxy servers sometimes allow circumvention of these limits, however depending on *terms and conditions* of the individual websites, this may

¹⁶ Reason often being the load on the servers caused by too many automated requests.

not be an option. In certain cases the website might be willing to provide exceptions to these limits for researchers, and applications developed within partnerships.

8.3.1.2 Result-set Limitations

Another significant limitation tends to be the number of results returned. This can substantially restrict feasibility of any analysis. For example in web-impact studies from the field of Webometrics a technique simply known as *query-splitting* can be used to return many more search results than the direct, imposed result-set limits generally allow (Thelwall 2009). Innovative ways to avoid these limits may be devised. For example in chapter 6.2, in order to increase the number of results returned on Youtube, related videos were also extracted with each video-item returned by a keyword query. Since the video similarity is based on a word-vector cosine measure (see section 7.2.1.3) of the video title's similarity, this technique is justified and increased results from around 6,000 direct results to around 90,000 results. In cases where there isn't possibility to retrieve more results than allowed, it must be kept in mind that any analysis performed on the results is only indicative, and essentially a convenience sample.

8.3.1.3 Private Content

On web 2.0 systems such as Twitter, Wikipedia or Yahoo pages are generally publicly accessible. However, for example, personal profiles on many social networks such as Facebook or Orkut are often private and require user-authentication¹⁷. This complicates access to datasets and in most cases it isn't possible to access private data unless users explicitly agree to sharing data with the requesting application or data collection script. Even then, there are ethical and privacy issues to consider.

8.3.1.4 Data Download

Thanks to the openness characteristic of web 2.0, APIs have become prevalent on the web (section 2.2.1 and 5.4). They allow to build software platforms on top of web 2.0 applications and to access data. An API is essentially a series of pre-defined HTTP requests that return responses in XML, JSON or SOAP formats. Extraction of datasets using APIs is substantially

¹⁷ A fraction of the profiles are public; however, which can be readily used in analyses.

simplified. Another technique often used when APIs do not support certain data-access or are not available, is to use a technique known as Web scraping. These are usually low level HTTP implementations which typically request HTML pages, and format these into structured data by extracting parts of the HTML content which contains the sought after content. Since a HTML page follows a DOM (Document Object Model¹⁸), the usual approach is to parse the HTML page into a DOM tree representation. Then using a-priori knowledge of the HTML specific parts of the DOM tree, it can be accessed for the required content on a particular page¹⁹. Simpler techniques avoid the DOM tree representation stage altogether and simply search for specific HTML tags to extract content in their proximity. This is often a reasonable approach if the webpage being scraped is clean, simple or when content is semantically annotated (i.e. semantic information is available within the markup or related files – see Antoniou and Harmelen 2008). Much useful information exists which covers in detail how to deal with varied HTML content and different HTTP requests, where for example matters are complicated with AJAX or JavaScript content or HTML forms, see (Heaton 2007) for a useful reference on how to deal with such issues. There are also a number of powerful tools described in Laender et al. (2002) which attempt to automate the code generation necessary for producing web scraping methods. A major disadvantage of Web scraping is the need for code to be maintained. Any larger changes to the underlying HTML page will likely render the web scraping code irrelevant, whereas this issue of maintenance doesn't exist with APIs.

8.3.1.5 Automatic Detection of Page Elements

In some situations when automated data retrieval from many heterogeneous webpages is required the techniques described above are often not suitable²⁰. It is a difficult problem to detect distinct semantic blocs of content from any given page (e.g. menu, adverts, footer, header, main content...). There are; however, a number of algorithms²¹ that can be used to help automate this task, with some accuracy. The VIPS (Vision-based Page Segmentation) method developed by Cai et al. (2003) uses visual representation of a web page as it automatically detects visual boundaries between (i.e. horizontal or vertical white/clear-space) content. This method has been developed by Microsoft researchers and seems to be one of the more popular

¹⁸ <http://www.w3c.org/DOM>

¹⁹ For example on a public video sharing website, the video title, description and all the comments with the authors and dates submitted can be extracted using this technique.

²⁰ For example, a generic and wide study of Blog articles over any blogging engine would require specific web scraping (or similar code) for every template or page layout, which is unfeasible in a practical scenario, unless the layouts are semantically annotated in the HTML, which is still rather rare.

²¹ Usually known under the name of webpage segmentation methods in the field of Web-mining

ones (Han and Kamber 2006)²². Feng et al. (2005) and Vineel (2009) propose alternative methods, and Vineel provides a useful overview of previous work in the area of webpage segmentation. The use of semantic content detection techniques were considered for <http://www.newsmental.com> in chapter 7²³; however, a more accurate method was feasible, similar to the one in section 8.3.1.4, above.

8.3.2 Example Applications

Several hypothetical, example applications of the framework within specific fields of interest will be presented in this section. The utility of this is to illustrate how one may employ the framework in a diverse set of fields and to stimulate ideas for future work. It is hoped that work by other researchers will help to elaborate on some of these ideas, since a full investigation of these would represent an additional separate body of research, outside of scope of this work. The examples discussed will be fairly generic and many details will be left out due to brevity, yet this section highlights future possible work, and possible interesting uses of collective intelligence.

8.3.2.1 Financial-markets

In the domain of financial markets, any model or decision support system tool that can aid in predicting assets, explaining asset behaviour given some events, or developing a less risky portfolio of assets, is generally of high interest to the practitioner (Schwager 1993). Chapter 6 discussed a number of web 2.0 systems in the context of finance. The discussion was concerned with an exploratory analysis of the collective intelligence data on these systems, although a more integrated exploitation of collective intelligence can be achieved. Given the domain of interest, a set of relevant M_D web 2.0 applications ought to be selected first. Since a strong relationship between markets and Youtube (section 6.2), Blogs, and Micro-Blogs (section 6.1) was shown to exist, one could select, Youtube, Twitter, and Blogger²⁴ based Blogs. In addition it seems that financial prediction markets and community websites discussed in section 6.6 are relevant, as well as Newsmental from chapter 7, and Amazon, where individual companies from the financial markets have a strong retail presence (i.e. numerous products are reviewed and

22 The software for VIPS (for research use) can be downloaded at <http://www.zjucadcg.cn/dengcai/VIPS/VIPS.html>

23 Semantic content detection techniques were considered for extracting data / articles / comments from news-sources in the Newsmental project.

24 Blogger is a Google based free Blog provider, the advantage of using a common provider for all Blogs, relates to the problems mentioned in section 8.3.1.5, at least when entire Blog posts are of interest to the analysis.

evaluated online for a given company).

Usually studies, and practical applications described in prior literature, would be constrained to a couple web 2.0 CI sources, at most (see section 6.1); instead the framework actively encourages a researcher to think of a web 2.0 application only as one of the building blocks in a CI system. The framework also encourages to think about a web 2.0 application, not in terms of what it provides for the user, but in terms of a source of CI to be exploited in a decision support system or similar models. In the case of Newsmental, in chapter 7, it was presented how an entire web 2.0 system can be engineered with the ultimate goal of exploiting the CI.

The filtered individual atomic elements and users (i.e. A_D^* , and U_D^*) need to be identified next. In the case of Youtube the choice of atomic elements may well be the same as in section 6.2, and in the case of Blogs and Twitter, the main element of interest is the post (in case of Blogs, comments and tags are also of interest). In case of the prediction markets, one would want to consider the individual votes, although usually the aggregated votes are only available. It is possible for the filter φ to emphasise certain types of users for U_D^* , although all users may of-course be considered. In making the above decisions, data can be extracted (see 8.3.1), and since all information is time-stamped, a temporal dimension for this data will be available²⁵.

In recent years within financial-markets, statistical classification models for prediction purposes became very popular. In Sykora et al. (2009) a currency trade forecasting system using kNN (k nearest neighbor), NN (neural network), and a combination of both algorithms was presented. The system was based on 140 technical analysis, price derived features at different time-intervals; however, it did not incorporate currency specific sentiments and related opinions. Additional features from Youtube, Blogs, and all other web 2.0 systems, could be added to the classification model. The frequency counts of submissions on specific topics may already improve predictions, although features derived from sentiment analysis within text based posts may prove to be further useful. Forecasting financial-markets using these models is of-course difficult, due to EMH (Fama 1965, Taylor et al. 2002). In such situations, it may be more interesting to help explain certain market behavior by exploring and visualizing datasets from web 2.0 collective intelligence sources instead. Another possible use of collective intelligence from web 2.0 would be to employ it in designing optimal portfolios. In modern portfolio theory, Markowitz's mean-variance portfolio optimisation uses correlations to compute risk-profit optimal portfolios (Elton et al. 2006). Correlations derived from web 2.0 could be used to provide alternative optimal portfolios. Tumasjan et al. (2010) showed that frequencies of several political parties mentioned together in Twitter posts, accurately represented real coalition and opposition links between German political parties. This is indicative that similarly

²⁵ Different time-zones must be dealt with properly as each application may provide dates in a different time-zone.

the mentions of two or more assets within single Twitter, Blog, Youtube posts, Amazon reviews, or Social Trading comments could be used to represent correlations between those entities, i.e. the more frequently they are mentioned together the more correlated they are. These are just a few suggestions from the field of finance.

8.3.2.2 *Medicine / Healthcare*

As in the section above, first one must identify applications of interest, M_D . For instance, let the social web applications Facebook, Twitter, and PatientOpinion, be considered. PatientOpinion was mentioned in section 2.4, Facebook contains several groups and group-pages dedicated to various disease / support groups, and there are many users with healthcare concerns on Twitter. As for atomic elements, in the case of Facebook one may consider any shared posts (i.e. status updates), shared links, uploaded files, binary votes (i.e. likes) associated with shared resources, on a wall, or profile pages; the larger the set of atomic elements considered, the better (section 8.3). The individual users who are important within the health domain should be selected; for instance on Twitter there may be influential politicians active in the healthcare industry, national health organisations, or well known researchers. On PatientOpinion, individual patient groups are of interest²⁶. Next the data items of all atomic elements and users are extracted and downloaded to some, usually, DB based storage. Finally more advanced data-mining²⁷, but also simpler exploratory analysis and visualisations can be applied to the datasets.

In a simple exploratory analysis of the dataset, for example one may plot correlation matrices of all atomic elements on PatientOpinion, from one interest group of patients against another group. A natural language toolkit to investigate the mood (negative or positive mood) reflected in Twitter and Facebook posts could be performed, and investigated for different diseases, and groups of users.

Data-mining analysis should be performed with the following abstraction in mind; the *dataset DB* is where the independent variables will originate from, and that the dependent (target) variable of focus is the specific domain of application. Let us assume that one wants to predict how much optimism (happiness / sadness) is connected with particular diseases. The independent variables could be the values from twitter posts, strength of interconnections between different diseases from the correlations, and other variables extracted from web 2.0

26 On PatientOpinion there are custom patient groups, which group together users who experience similar health issues, such as say diabetes patients, multiple sclerosis patients, stomach cancer patients, etc.

27 An exhaustive treatment of all relevant techniques is unfortunately out of scope for this thesis, and even (to some extent) unnecessary, as one can generalise, that any meaningful analysis should regard the acquired dataset to be the raw independent variables, and that the variable of focus (dependent variable / to be predicted) is the specific domain of application.

applications. The dependent variable could for instance be a mood index from 1 to 10, collected using some sort of stratified sampling; for each patient group, and enough to form train and validation and test data. Finally a statistical classification model can be trained, which would hopefully learn to judge the mood of each patient group, based on information collected from Facebook, Twitter, and PatientOpinion.

8.3.2.3 Policing / Law Enforcement

In a report by Her Majesty's Inspectorate of Constabulary (HIMC, Feb. 2011), it was pointed out that the British police forces were unprepared to deal with social media in times of public unrests. Later that year, early August 2011, widespread rioting spread across England. Collective intelligence from web 2.0 applications ought to be applied to policing and law enforcement. In order to stimulate ideas in this area, a brief description of some law enforcement related case-studies and suggestions of ways to extend these with web 2.0 related data, in order to motivate work in this and other areas of application is given. A useful project would naturally be one that is concerned with data solutions that reduce crime in certain areas (Chen et al. 2004). The data-features available within a typical crime record are illustrated in the list below.

The actors of a crime scene:

1. Suspect (known/unknown), once convicted he/she becomes a convict
2. Victim, usually reports the crime
3. Witnesses

Aspects for which there might be some information:

1. Location
2. Date/Time
3. Weapons used + any other details from the crime-scene
4. Suspect, Victim details
5. Modus operandi: statement / police officer summary

[Source: Nath 2006, Chen et al. 2004]

It is of interest to identify hotspots, or patterns unknown or unspotted a-priori by police officers. In Nath (2006) a statistical clustering algorithm was used to identify similar patterns of crime attribute-values, which would be descriptive of a common group of crimes. This helped to identify similar groups of patterns, crimes that weren't associated with each other initially, and was found to be very useful by the police forces taking part in the pilot study. Arguably, more insightful features should improve overall clustering performance. Since only basic information is available from standard police crime records, it is expected that clustering accuracy would

improve by incorporating collective intelligence derived features from web 2.0 sources. Below is a possible list of relevant web 2.0 systems.

- Information about local events (pubs events, etc.), from public Facebook events²⁸, Meetup.com²⁹ communities and similar online event planning websites.
- Flickr, Youtube, and other media sharing applications (for content that has been localised³⁰) can be searched when these are temporally compatible with the crime-incidents. Pictures taken by people at a specific times and locations could be highly indicative of a crime-scenes' context. Since comments and tags are commonly associated with such content, the screening process can be automated.
- Also there is the opportunity to locate people who might have been close to the crime scene, at the time of the crime. This can be done by searching public Blog, Twitter, or Facebook account posts that make references to locations and times.

More recently Cheong and Lee (2011) used Twitter to detect terrorism events, and instances of criminal activity. Incorporating a multitude of web 2.0 applications for such purposes seems to be a natural progression for these problems.

8.3.2.4 Sports Science Applications

Hambrick et al. (2010) and a similar study by Pegoraro (2010) used content analysis to evaluate the type of messages posted on Twitter by sports athletes; findings from both studies were compatible. It was found that athletes often use Twitter, to directly interact with their fans, and some athletes are heavy users of this Micro-blogging tool. These findings were confirmed by Kassing and Sanderson (2010), who investigated Tweets, over the duration of an entire sporting event (i.e. 3 week long, Giro d'Italia³¹). Their findings showed interesting insights from behind the scenes, which were largely ignored by official reports on the event, including changes to strategies of the cyclists during the race were revealed. Related to this, is the field of sports informatics, which is an emerging field where computer science is applied to sports science problems (Dabnichki and Baca 2008). Within this research field, Lames et al. (2008) introduces an area dedicated to the coaching process, usually in team-sports such as soccer, or handball. A possible application of the framework proposed in this chapter, would be the use of collective intelligence data generated by sports fans, and possibly sports analysts, regarding various

28 Significant amounts of web 2.0 user generated content, is private, and hence most applications will be restricted by the public availability of the data.

29 Meetup.com is a popular web 2.0 system for managing communities and wide range of events.

30 Latitude and Longitude information is usually attached in the forms of geo:tags (latitude / longitude coordinates).

31 Giro d'Italia is the Italian equivalent of the prestigious cycling event, Tour de France.

coaching parameters. The aim would be to provide decision support for the coach in positions of players and tactical changes a-priori, or during a game-play / sporting events.

8.3.2.5 English Language

In chapter 3 of this thesis, a new approach for neologism assessment (i.e. term emergence) was introduced. The method used collective Delicious tagging habits, and is in fact an example of the proposed framework in action. Links to specific web-resources, comments, and tags were retrieved, where the set of users was constrained by having bookmarked a link, where at least one other user used a certain tag, or combination of tags. Such data retrieval was sufficient, and a relatively simple analysis was used to assess neologism emergence. Visualisations of users' bookmarks over time, some basic text processing of tags and comments, and comparison of frequency counts were used to analyse the collective intelligence extracted from Delicious.

8.4 Discussion and Limitations

The proposed framework makes one main contribution. It provides a much needed (see section 8.1 Motivation) high-level guide for using web 2.0 generated datasets as collective intelligence, applied to specific domains. It is felt that more work is now needed with web 2.0 being applied in practical problem domains. Such work will show, to what extent and potential success collective intelligence from web-based collaborative or sharing systems can be used. Collective intelligence when applied to specific domains can help explain and possibly forecast in various fields of substantial interest.

Some work to this regard has been done over the last couple years. For example, Sakaki et al. (2010) used Twitter to detect earthquakes, and showed interesting results of collective knowledge dispersion³². A more applied example is (Gabrilovich 2006), who improved on the Bag-of-Words (BOW) approach to text classification (see section 7.2.1.3) by extending the feature-set of such problems with Wikipedia derived features. He reported impressive results compared to simpler BOW methods, and more recently (Hu et al. 2008) further improved on these results. Banerjee et al. (2007) did similar work, in clustering of short-texts, by enriching the text with features from Wikipedia article titles. Ginsberg et al. (2009) from Google have

³² Sakaki et al. (2010) developed a system that can detect an earthquake simply by monitoring Tweets. 96% of earthquakes registered by the Japan Meteorological Agency (JMA) with seismic intensity scale 3 or more were detected. The authors claim the system's subsequent notifications are delivered much faster than announcement broadcast by the JMA.

used relatively simple technique on search query data to anticipate flu outbreaks, Choi and Varian (2009) used an equivalent technique to anticipate unemployment rates consistently and more timely than other current methods allow. Collective intelligence approaches have also been applied to Climate Change (Malone et al. 2009)³³, and Tumasjan et al. (2010) built a system that used Twitter, to predict German federal elections of 2009, with an accuracy similar to traditional election polls. These interesting developments motivate the requirement for a guiding framework to help provide researchers new to this field with a starting point on how to approach the usage of collective intelligence from social media, and to others, a systematic guide in approaching such tasks.

The abstraction introduced by the guidelines can be useful in segmenting the cognitive thinking process of the researcher by the distinct tasks needed to accomplish a given applied collective intelligence task. Unfortunately this abstraction also has its disadvantages. The framework is too generic and potentially irrelevant for less common problems. It should be used with caution and is only a guideline which can help approach some problems. It may not cover every scenario. Social media and web 2.0 applications are quickly evolving and change with new features, which can be un-represented by the framework. It is hoped; however, this framework will be a starting point to open up more research questions, and improve over time.

8.5 Summary

It has been shown, and empirically established in previous chapters that web 2.0 is a real phenomenon with social and economic implications, and the importance and ongoing adoption within a number of fields was illustrated. However, the value of user generated contributions on such systems has been little understood, even though patterns in such data have been extensively studied in the literature. Potentially interesting and heterogeneous datasets are produced on web 2.0 systems by “average” users or employees of companies, and in aggregate this data can be treated as collective intelligence, for useful practical applications. In this chapter the assertion that web 2.0 systems can essentially be regarded as voluntary³⁴, streamlined and efficient ways of polling or opinion and knowledge sampling (i.e. polling, surveying), was presented. This process bears similarities with implicit voting encapsulated in the trading of financial assets, and the related Efficient Market Hypothesis. A systematic step-by-step characterisation of how to approach the problem of using CI from web 2.0 systems was

³³ See <http://mitsloanexperts.com/2011/02/06/solving-climate-change-with-crowdsourcing/> for a description of the initiative.

³⁴ Thelwall (2009); pp. 4, also refers to this as an advantage of Webometrics since “the data collection process is passive and relatively cheap”.

proposed. The characterisation acts as a guideline on how to leverage and use CI from web 2.0 applications in real world problems. In other areas guiding models exist and it is believed that such guidelines may help in further research and application development.

9 Conclusions and Future Work

In this chapter the conclusions of this work will be presented, and the main results will be summarised. Finally, the recommendations for future research related to this work will be discussed.

9.1 Conclusions

There has been much hype in vocational and academic circles surrounding the emergence of web 2.0; however, relatively little work was dedicated to substantiating the actual concept of web 2.0. This thesis has attempted to identify and critically evaluate the web 2.0 environment and what caused it to emerge; providing a rich literature review on the topic, a review of existing taxonomies, an evaluation of web 2.0 related “2.0” terms, a quantitative and qualitative evaluation of the concept itself via a survey study and via a historical analysis, an investigation of the collective intelligence potential that emerges from such application usage, development of a custom collective intelligence source, and a framework for harnessing the collective intelligence in a systematic manner was proposed.

Overall the thesis aimed at achieving two main goals (also see section 1.2.1). First, the task was to substantiate and evaluate web 2.0 and its wider context, and second to investigate the collective intelligence potential that emerges from web 2.0 applications use. These aims were achieved by satisfying the underlying research objectives, as presented in section 1.2.2. Each chapter in turn documented a research objective, which are summarised below.

Chapter 2 provided a synthesis of web 2.0 related research literature; in particular with a historical, social, economic and critical perspective. It was concluded from literature found that ideas behind web 2.0 were not recent but have been around for a while. Initially, however, technical issues presented a major barrier to a collaborative use of WWW, although at the turn of the millennia it was especially the ‘dot-com’ crash which marked a “turning point” for the web. In the years following the economic crisis an environment was created in which, trust into the web, viability of online business models, standardisation of technologies, emergence of productivity tools for developers, and advancements in architecture and integration of web and data systems slowly became a reality, and hence collaborative and social use of the WWW has emerged to be worthwhile. The social and economic implications of web 2.0 were also found to be noteworthy, such as the breakdown of traditional parasocial relationships, or the concept of peer / social production as a new organisational form with significant economic implications. Information saturation, quality of user generated content, “free-riding”, “slave labour”,

children's vulnerability, and privacy in general were all raised in literature as some valid concerns and criticisms of social media. We provided a synthesis of these criticisms and discussed possible approaches to mitigating and resolving some of the issues raised (see section 2.2.4). For instance motivations play an important role in understanding why and under what conditions "slave labour" likely takes place. The literature review further synthesised current use of web 2.0 systems in a variety of vocational fields, specifically in clinical practice, corporate use, politics, public service, education, journalism and geographic studies. In order to introduce a new web 2.0 taxonomy, as suggested in research objective 1.1 (section 1.2.2), several existing taxonomies were reviewed. The final contribution of this chapter was the two-step taxonomy, which is most importantly based on what atomic collaborative elements can be identified in a system and what dominant data objects are shared overall. This approach ensures that website elements which typically facilitate the sharing of content on web 2.0 will be used from the ground-up to categorise web 2.0 applications, rather than a direct category assignment. In order to provide further tangible evidence to support the concept of web 2.0 and social media, a new method of neologism emergence analysis (chapter 3), a survey (chapter 4), and an automated historical evolution analysis (chapter 5) were employed.

Chapter 3 evaluated a number of "2.0" terms and was a first study that looked at such a wider set of these terms. In order to investigate their emergence, a methodology for generic neologism emergence in the English language was proposed. The idea of the methodology was supported by some strong findings from prior literature – i.e. highly stable patterns of proportions for delicious based tags and their overall accurate representativeness of content. In addition to gauging the degree to which some "2.0" terms have taken hold in common use, the results were also interesting as they illustrated prevalence of web 2.0 technology in those fields.

In chapter 4, a survey was employed to assess the current use of web 2.0 applications in terms of several features of interest identified to be important for the uptake of web 2.0, in chapter 2. A revealing picture on the usage and prevalence of the actual terms "social media" and "web 2.0" was given, as well as the relationship of users to trust, motivation, commercial acumen and time spent on such applications. It was found that the role of trust and time spent online increases with web 2.0 competence of users and is very important to users, although only 57% of the respondents consider themselves able to judge trustworthy from non-trustworthy sites. Group buying which is a relatively new business model that evolved on the web is surprisingly popular among respondents, and users also tend to have more well defined motives the more they use web 2.0 applications. The chapter provides tables of the popularity of web 2.0 applications by their usage and by one of the factors in the study (*such as trust, time spent, etc.*) with the relevant tests for statistical significance. Similar breakdowns were also provided for

the activities that are commonly performed on web 2.0 systems, which lend support for the atomic activities identified for the taxonomy in chapter 2. The study quantifies and answers a number of important questions raised within chapter 2. In comparison to previous survey studies it has a considerably larger sample. Although there was a slight bias towards academic groups of respondents, the demographic details of the sample were still useful in detecting trends in groups of differently educated, aged, and technically skilled individuals. The study also provided survey design and reliability evaluation details, often left out in prior literature in this area.

Chapter 5 studied the historical evolution that has lead for web 2.0 to emerge, using a web based archives tool to investigate a selection of well known web applications throughout their development over time. A number of specific features conducive to what is perceived characteristic of web 2.0 applications were observed over a time-period reaching as far back as the late 90s. The outcomes of the study revealed support for a number of trends associated with the arrival of web 2.0, such as; increased AJAX / Javascript use, appearance of site Blogs, APIs, as well as decreased use of non-standard design elements, which was expected, but has not been shown in earlier work. Although web 2.0 implies a designated version and a discrete evolution, it was found that the changes associated with web 2.0 have emerged gradually and occurred over a prolonged time-period, much longer than initially expected. One valid criticism of only having used seven websites in this study was directly addressed by reviewing over 50 top ranking web 2.0 websites. These results were found to be largely compatible with the historical study. In addition to the results presented, the chapter's other important contribution was the evaluation and discussion of the Wayback Machine archives as a potential research tool for data driven historical studies.

The chapters 3, 4 and 5 provided much substance to various points raised in chapter 2; however they also provided context and background for the remaining chapters, 6, 7 and 8. For instance, a number of granular atomic activities that facilitate collaboration were described in literature, and these were identified within a proposed taxonomy in chapter 2. This taxonomy is of significance for the final collective intelligence framework proposed in chapter 8.

Chapter 6 introduces a definition for collective intelligence and prior research literature concerned with the analysis of user generated content which emerges from web 2.0 application use. Three web 2.0 applications were analysed and described in context of the financial crisis and the financial markets in general. An important and novel result from this chapter was the discovery of efficient information transfer to exist within web 2.0 applications (satisfying research objective 5.1, in section 1.2.2). The Stockmarket was used as a proxy of information efficiency, based on the efficient market hypothesis assumption. On Youtube videos are shared;

however, one must acknowledge the extra effort involved in preparing and submitting original video content, despite this overhead on effort, a strong correlation of video submissions with Stockmarket price volatility, as well as directional price movement was revealed. The analysis of Delicious confirmed this result to some extent, and an investigation into Amazon based UGC reported a number of interesting results, which were novel in the scope of existing literature. Another contribution of both, chapters 6 and 7, is that they emphasised the point of view that web 2.0 applications can and often should be regarded purely as sources of collective intelligence.

Chapter 7 demonstrated a web 2.0 application design, with the aim of ultimately providing UGC that other, existing web 2.0 applications do not provide, or at least not at the desired granularity. The system is a news analysis and opinion sharing system, with the aim of providing a custom source of collective intelligence. The potential of collective intelligence generated through its use was qualitatively evaluated and explored using two specific events of financial market significance – *the election for the head of IMF in 2011 and the financial crisis of Greece during the summer of 2011*. The web 2.0 system was found to be valuable and useful in providing the desired custom source of collective intelligence, although maintaining a community of users was deemed to be a challenge. Future work (section 7.5.2) on improving the system highlighted some ways of dealing with the relevant web 2.0 system issues. The overall conclusion of this chapter was that it can be feasible and indeed highly desirable to introduce a new web 2.0 platform where the aim is to satisfy a collective intelligence acquisition need for sharing specific type of knowledge, at a desired granularity.

Chapter 8 reviews CI work in the literature, and some prior work that is concerned with investigation of collective intelligence for a variety of higher-level purposes is identified, although little has been done in the effort of developing any set of guidelines or common processing steps in such type of work. Hence, a strong argument is made in favour of an abstract systematic step-by-step characterisation of how to approach the problem of utilising collective intelligence from web 2.0 applications. The chapter essentially presents a novel and abstract framework that is hoped will aid future researchers in harnessing collective intelligence. The main rationale of the framework is to outline several important elements to a social media application and systematically follow through a number of steps, which are outlined in the chapter. A crucial element of any tool that leverages collective intelligence is the actual extraction of web datasets, hence specific technical considerations are also presented and discussed. In order to illustrate the wide potential of such a framework and to stimulate further work, a variety of possible applications ranging from such disparate fields as Sports Sciences, and Law Enforcement were presented.

9.2 Future Work

This thesis presented a series of studies and methodologies that contribute to the existing body of research. However, there is much possible future work ahead. For example, the use of the proposed taxonomy and the framework in practice by other researchers will be interesting to observe. In particular section 8.3.2 outlines possible research applications of CI, using the framework. The potential of further work in these areas of application are significant, and some interesting work is already being done, as discussed in section 8.4. There are many issues connected to implementing CI based on the framework, some of these were discussed in detail, other relate to data-management, storage, or data-integration problems. With large datasets of UGC, new techniques of handling large datasets need to be assessed and investigated. As far as our understanding of web 2.0 is concerned, there is always scope for follow-up survey studies on larger or different demographic samples, and a historical content analysis that is performed by several human assessors would provide more accurate observations, and would likely be a worthwhile endeavour.

References

- Adamic L.A. and Glance N., 2005. *The political blogosphere and the 2004 U.S. election: divided they blog*, International Conference on Knowledge Discovery and Data Mining, Chicago (USA)
- Adewumi D., 2008. *Kenyan tech bloggers launch crisis-report site*, VentureBeat. 15th January 2008. Last Accessed: 1st August 2010
- Adler B., Chatterjee K., de Alfaro L., Faella M. and Pye I., 2008a. *Assigning trust to wikipedia content*. In WikiSym 2008: International Symposium on Wikis Proceedings, Porto (Portugal)
- Adler B., de Alfaro L., Pye I. and Raman V., 2008b. *Measuring author contributions to the wikipedia*. In WikiSym 2008 - International Symposium on Wikis Proceedings, Porto (Portugal)
- Akcora C. G., Bayir M. A., Demirbas M. and Ferhatosmanoglu H., 2010. *Identifying breakpoints in public opinion*, Proceedings of the First Workshop on Social Media Analytics, Washington DC (USA)
- Alag S., 2008. *Collective Intelligence in Action*, Manning Publications, USA
- Alexander B., 2006. *Web 2.0: A new wave of innovation for teaching and learning*, Learning Educause, **41** (2), pp. 32-44
- Ali-Hasan N. and Adamic L. A., 2007. *Expressing social relationships on the blog through links and comments*, Proceedings of the International Conference on Weblogs and Social Media (ICWSM)
- Anderson P., 2007. *What is web 2.0? Ideas, technologies and implications for education*, JISC (Joint Information Systems Committee) Technology and Standards Watch Report, London (United Kingdom)
- Andrew B. C., 2007. *Media-generated Shortcuts: Do Newspaper Headlines Present Another Roadblock for Low-information Rationality?*, The Harvard International Journal of Press/Politics, **12** (2), pp. 24-43
- Androutsopoulos I., Koutsias J., Chandrinou K. V., Paliouras G. and Spyropoulos C. D., 2000. *An evaluation of naive bayesian anti-spam filtering*, Proceedings of the workshop on Machine Learning in the New Information Age, Barcelona (Spain), pp. 9-17
- Antin J. and Cheshire C., 2010. *Readers are Not Free-Riders: Reading as a Form of Participation on Wikipedia*, Proceedings of the 2010 ACM conference on Computer supported cooperative work, Savannah, USA

- Antoniou G. and Harmelen van F., 2008. *A Semantic Web Primer*, MIT Press, USA
- Antweiler W. and Frank M. Z., 2004. *Is all that talk just noise? the information content of internet stock message boards*, Journal of Finance, **59** (1), pp. 1259-1269
- Argamon S., Koppel M., Pennebaker J. and Schler J., 2009. *Automatically Profiling the Author of an Anonymous Text*, Communications of the ACM, **52** (2), pp. 119-123
- Argawal R. and Srikant R., 1994. *Fast Algorithms for Mining Association Rules in Large Databases*, Proceedings of VLDB-94 Conference, Santiago de Chile, Chile
- Arms W. Y., Aya S., Dmitriev P., Kot B. J., Mitchell R., Walle L., 2006. *Building a research library for the history of the web*, Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries, Chapel Hill (USA)
- Askatas N. and Zimmermann K.F., 2009. *Google econometrics and unemployment forecasting*, Applied Economics Quarterly, **55** (2), pp. 107-120
- Attardi G. and Simi M., 2006. *Blog Mining through Opinionated Words*, Proceedings of the Fifteenth Text Retrieval Conference (TREC 2006), Gaithersburg (USA)
- Bao S., Wu X., Fei B., Xue G., Su Z. and Yu Y., 2007. *Optimising Web Search Using Social Annotations*, Proceedings of the 16th International conference on World Wide Web, Banff (Canada)
- Balog K., Mishne G. and Rijke M., 2006. *Why are they excited? Identifying and explaining spikes in blog mood levels*, Technical report, University of Amsterdam
- Banerjee S., Ramanathan K. and Gupta A., 2007. *Clustering Short Texts using Wikipedia*, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, Netherlands
- Barberis N., Shleifer A. and Vishny R., 1998. *A model of investor sentiment*, Journal of Financial Economics, **49** (307), pp. 307-343
- Barbry E., 2007. *Web 2.0: Nothing changes... but everything is different*, Journal of Communications and Strategies, **65** (1)
- Benkler Y., 2002. *Coase's Penguin, or, Linux and the Nature of the Firm*, Yale Law Journal **112**
- Benkler Y., 2006. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, Yale University Press, USA
- Berlanga A. J., Bitter-Rijkema M., Brouns F. and Sloep P. B., 2011. *Personal Profiles: Enhancing Social Interaction in Learning Networks*, International Journal of Web Based Communities, **7** (1), pp. 66-82

- Bermingham A. and Smeaton A. F., 2010. *Crowdsourced real-world sensing: sentiment analysis and the real-time web*, Sentiment Analysis Workshop at Artificial Intelligence and Cognitive Science Proceedings, Galway (Ireland)
- Berners-Lee T. and Fischetti M., 1999. *Weaving the Web: The Past, Present and Future of the World Wide Web by its Inventor*, Orion Business Books, USA
- Bird S., Klein E. and Loper E., 2009. *Natural Language Processing with Python*, O'Reilly Media, USA
- Bishr M., 2009. *WWW: The Darwinian Imperative*, Proceedings of the WebSci'09: Society On-Line, Athens, Greece
- Bishop S. R. and Baudains P., 2010. *Booklet: Global System Dynamics and Policies (GSD)*, available from <http://www.globalsystemdynamics.eu/>, Last accessed on 1st March 2011
- Blumenstock J. E., 2008. *Size matters: word count as a measure of quality on Wikipedia*, Proceedings of the 17th international conference on World Wide Web, Beijing (China), pp. 1095-1096
- Blumenthal J., Allee N. and Mayman G., 2010. *Public Health 2.0: Collaborative partnerships for integrating social technologies into the practice community*, Proceedings of the 10th International Congress on Medical Librarianship, Brisbane (Australia)
- Bohannon J., 2010. *Google Opens Books to New Cultural Studies*, Science Journal, **330**, pp. 1600
- Bollen J., Mao H. and Zeng X.-J., 2011. *Twitter mood predicts the stock market*, Journal of Computational Science, **2** (1), pp. 1-8
- Bond P., 2008. *Study: Young People Watch Less TV*, Hollywood Reporter, 17th December 2008, Last Accessed: 1st August 2010
- Boulos M. K., Maramba I. and Wheeler S., 2006. *Wikis, blogs and podcasts: a new generation of Web-based tools for virtual collaborative clinical practice and education*, BMC Medical Education, **6** (41)
- Box G. and Jenkins G., 1970. *Time series analysis: Forecasting and control*, San Francisco: Holden-Day (USA)
- Brabham D. C., 2008. *Crowdsourcing as a Model for Problem Solving: An Introduction and Cases*, International Journal of Research into New Media Technologies, **14** (1), pp. 75-90
- Brin S., Page L., 1998. *The anatomy of a large-scale hypertextual Web search engine*, Computer networks and ISDN systems **30** (7), pp. 107-117

- Brownell C., 2008. *Subprime Meltdown: From U.S. Liquidity Crisis To Global Recession*, CreateSpace Publishing, USA
- Budd A., Moll C. and Collison S., 2009. *CSS Mastery: Advanced Web Standards Solutions*, Friends of Ed, USA
- Burns R., 2011. *The Naked Trader: How Anyone Can Make Money Trading Shares*, Harriman House, UK
- Butuc M. G., 2009. *Semantically Enriching Content Using OpenCalais*, EDITIA Journal, **9** (1), pp. 77-88
- Cai D., Yu S., Wen J. R. and Ma W. Y., 2003. *VIPS: A vision based page segmentation algorithm*, MSR-TR 2003, Microsoft Research Asia
- Calais, 2010. *English Semantic Metadata: Entity/Fact/Event Definitions and Descriptions*, online only, <http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions>, last accessed: 1st May 2011
- Cameron J., Banko K. M. and Pierce W. D., 2001. *Pervasive Negative Effects of Rewards on Intrinsic Motivation: The Myth Continues*, Behaviour Analyst **24**, pp. 1-44
- Casey M. E. and Savastinuk L. C., 2007. *Library 2.0: A Guide to Participatory Library Service*, Information Today, USA
- Cha M., Kwak H., Rodriguez P., Ahn Y. and Moon S., 2007. *I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system*, Proceedings of ACM Internet Measurement Conference (IMC), San Diego (USA)
- Chan W.S., 2003. *Stock price reaction to news and no-news: Drift and reversal after headlines*, Journal of Financial Economics, **70** (2), pp. 223-260
- Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shearer C. and Wirth R., 2000. *CRISP-DM 1.0: Step-by-step data mining guide*, CRISP-DM Consortium, August 2000
- Chen Y. and Barnes S., 2007. *Initial trust and online buyer behaviour*, Industrial Management & Data Systems **107** (1), pp. 21-36
- Chen H., Chung W., Xu J. J. and Yi-Qin G. W., 2004. *Crime Data Mining: A General Framework and Some Examples*, IEEE Computer Society Journal, **37** (4), pp. 50-56
- Cheong M. and Lee V. C., 2011. *A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter*, Information Systems Frontiers, **13** (1), pp. 45-59
- Chesley P., Vincent B., Xu L. and Srihari R.K., 2006. *Using verbs and adjectives to*

automatically classify blog sentiment, Proceedings of AAAI-CAAW-06, the Spring Symposium on Computational Approaches to Analyzing Weblogs

Chevalier J. and Mayzlin D., 2006. *The Effect of Word of Mouth on Sales: Online Book Reviews*, Journal of Marketing Research, online: <http://www.nber.org/papers/w10148>

Chi E. H., 2008. *The Social Web: Research and Opportunities*, IEEE Computer, **41** (9), pp. 88-91

Choi H. and Varian H., 2009. *Predicting initial claims for unemployment benefits*, Google Inc. Technical Report

Choudhury M. D., Sundaram H., John A. and Seligmann D., 2008. *Can blog communication dynamics be correlated with stock market activity?* Proceedings of the nineteenth ACM conference on Hypertext and hypermedia, New York (USA)

Cialdini R. B., 2000. *Influence: Science and Practice*, (4th ed.) HarperCollins, USA

Clarkson P. M. and Joyce D. and Tutticci I., 2006. *Market reaction to takeover rumour in internet discussion sites*, Journal of Accounting and Finance, **46** (1), pp. 31-52

Coase R., 1937. *The Nature of the Firm*, Economica **4** (16), pp. 386-405

Cohen J., 1960. *A coefficient of agreement for nominal scales*, Educational and Psychological Measurement, **20** (1), pp. 37-46

Coleman J. S., Katz E. and Menzel H., 1966. *Medical Innovation: A Diffusion Study*, Bobbs-Merrill, USA

Conover W. J., 1998. *Practical Nonparametric Statistics*, Wiley Publishing, USA

Cormode G. and Krishnamurthy B., 2008. *Key differences between Web 1.0 and Web 2.0*, Internet Journal - First Monday, **13** (6)

Crockford D., 2008. *JavaScript: The Good Parts*, Yahoo Press, USA

Cronbach L. J., 1951. *Coefficient alpha and the internal structure of tests*, Psychometrika, **16** (3), pp. 297-334

Dabnichki P. and Baca A. (ed.), 2008. *Computers in Sport*, WIT Press, UK

Daft R. L. and Lengel R. H., 1986. *Organisational information requirements, media richness, and structural design*, Management Science, **32** (5), pp. 554-571

Danescu-Niculescu-Mizil C., Kossinets G., Kleinberg J. and Lee L., 2009. *How Opinions are Received by Online Communities: A Case Study on Amazon.com Helpfulness Votes*, Proceedings

of the 18th international conference on World Wide Web, Madrid (Spain), pp. 141-150

Das S., Martinez-Jerez A. and Tufano P., 2005. *Information: A clinical study of investor discussion and sentiment*, Journal of Financial Management, **34** (3), pp. 103-137

Daugherty T., Eastin M. S. and Bright L., 2008. *Exploring consumer motivations for creating user-generated content*, Journal of Interactive Advertising, **8** (2), pp. 1-24

Dave K., Lawrence S. and Pennock D., 2003. *Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews*, Budapest, Hungary

Dawson C. W., 2009. *Projects in Computing and Information Systems: a Student's Guide*, Addison Wesley, London, UK

Deci E. L., 1972. *Intrinsic Motivation, Extrinsic Reinforcement, and Inequity*, Journal of Personality and Social Psychology **22** (1), pp. 113-120

Deci E. L., Koestner R. and Ryan R., 1999. *A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation*, Psychological Bulletin **125** (6), pp. 627-668

DiMicco J., Millen D. R., Geyer W., Dugan C., Brownholtz B. and Muller M., 2008. *Motivations for social networking at work*, Proceedings of the 2008 ACM conference on Computer supported cooperative work, San Diego (USA), pp. 711-720

Dor D., 2003. *On newspaper headlines as relevance optimizers*, Journal of Pragmatics, **35** (1), pp. 695-721

Dowman M., Tablan V., Cunningham H. and Popov B., 2005. *Web-assisted annotation, semantic indexing and search of television and radio news*, Proceedings of the 14th international conference on World Wide Web, Chiba (Japan)

Dwyer P., 2007. *Building Trust with Corporate Blogs*, In Proceedings of International Conference on Weblogs and Social Media (ICWSM'07), Boulder (USA)

Ederington L. H. and Lee J. H., 1993. *How markets process information: News releases and volatility*, Journal of Finance, **48** (4), pp. 1161-1191

Efimova L. and Grudin J., 2007. *Crossing boundaries: A case study of employee blogging*, Proceedings of the 40th Annual Hawaii International Conference on System Sciences, Waikoloa (Hawaii)

Ehmann K., Large A. and Beheshti J., 2008. *Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia*, Internet Journal - First Monday, **13** (10)

Elton E. J., Gruber M. J., Brown S. J. and Goetzmann W. N., 2006. *Modern Portfolio Theory*

and *Investment Analysis*, Wiley and Sons, USA

Ennals R., Byler D., Agosta J. M. and Rosario B., 2010, *What is disputed on the web?*, Proceedings of the 4th workshop on Information credibility, Raleigh (USA), pp. 67-74

Eysenbach, G., 2001. *What is e-health?*, Journal of Medical Internet Research **3** (2), online: e20

Fallows J., 2006. *Homo Conexus: A veteran technology commentator attempts to live entirely on Web 2.0 for two weeks*, TechnologyReview.com, 1st July 2006, Last Accessed: 31st July 2010

Fama E., 1965. *Random walks in stock market prices*. Financial Analysts Journal, **51** (5), pp. 55-59

Farrell H. and Drezner D.W., 2008. *The power and politics of blogs*. Public Choice, **134** (1), pp 15-30

Feldman R., Aumann Y., Liberzon Y., Ankori K., Schler J., and Rosenfeld B., 2001. *A domain independent environment for creating information extraction modules*, Proceedings of the tenth international conference on Information and knowledge management, Atlanta, USA

Feng J., Haffner P. and Gilbert M., 2005. *A learning approach to discovering Web page semantic structures*, Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, Seoul, South Korea

Flanagin A. J. and Metzger M. J., 2008. *The credibility of volunteered geographic information*, GeoJournal **72** (3), pp. 137-148

Fleiss J. L., 1971. *Measuring nominal scale agreement among many raters*, Psychological Bulletin, **76** (1), pp. 378-382

Forsyth E., Lin J. and Martell C., 2008. *The NPS Chat Corpus*, Last accessed on 1st July 2011, online: <http://faculty.nps.edu/cmartell/NPSChat.htm>

Forte A. and Bruckman A., 2005. *Why do people write for Wikipedia? Incentives to contribute to open-content publishing*, Proceedings of GROUP 2005 Workshop: Sustaining Community, the role and design of incentive mechanisms in online systems, Orlando, USA

Foster A. and Rosenzweig M., 1995. *Learning by doing and learning from others: Human capital and technical change in agriculture*, Journal of Political Economy, **103** (6), pp. 1176–1210

Fowler J. H and Christakis N. A., 2010. *Cooperative behavior cascades in human social networks*, Proceedings of the National Academy of Sciences (PNAS), **107** (12), pp. 5334-5338

Francis J. C., 1988. *Management of Investments*, McGraw-Hill, USA

Freeman E., Freeman E., Bates B. and Sierra K., 2004. *Head First Design Patterns*, O'Reilly Media, USA

Freeman B. and Chapman S., 2007. *Is youtube telling or selling you something? Tobacco content on the youtube video-sharing website*, Technical report, Tobaccocontrol

Friedman J., 2010. *What Caused the Financial Crisis*, University of Pennsylvania Press, USA

Fukuda K., Saito S., Takagi H. and Asakawa C., 2005. *Proposing new metrics to evaluate web usability for the blind*, Proceedings of the Human factors in computing systems conference, New York, USA

Fulgoni G. M., Mörn M. P. and Shaw M. 2010. *How Online Advertising Works: Whither the Click in Europe? A U.K. & European Perspective on the Latent Impact of Display Advertising*, ComScore Inc., Technical Report, USA

Fung G., Yu J. and Lu H., 2005. *The Predicting Power of Textual Information on Financial Markets*, IEEE Intelligent Informatics Bulletin, 5 (1), pp. 1-10

Gabrilovich E., 2006. *Feature Generation for Textual Information Retrieval Using World Knowledge*, PhD Thesis, Haifa – Israel Institute of Technology, Israel

Gelman L., 2004. *Internet Archive's webpage snapshots held admissible as evidence*, Packets 2 (3), <http://cyberlaw.stanford.edu/packets002728.shtml>, Last Accessed 1st July 2011

Gilbert M., 2006. *Rationality in Collective Action*, Philosophy of the Social Sciences Journal 36 (1), pp. 3-17

Giles J., 2005. *Internet encyclopaedias go head to head*. Nature, 438:900–901

Gillam L. and Ahmad K. and Ahmad S. and Casey M. and Cheng D. and Taskaya T. and Oliveira P. and Manomaisupat P., 2002. *Economic News and Stock Market Correlation: A Study of the UK Market*, Proceedings of Terminology and Knowledge Engineering Conference, Nancy, France

Ginsberg J., Mohebbi M. H., Patel R. S., Brammer L. and Smolinski M. S., 2009. *Detecting influenza epidemics using search engine query data*, Nature, 457 (7232), pp. 1024-1014

Gloor P. and Krauss J. S. and Nann S. and Fischbach K. and Schoder D., 2008. *Web science 2.0: Identifying trends through semantic social network analysis*, Technical report, Social Science Research Network

Goffman E., 1959. *The presentation of self in everyday life*, Doubleday Anchor Books, USA

Golder S. A. and Huberman B. A., 2006. *Usage patterns of collaborative tagging systems*,

Journal of Information Science **32** (2), pp.198-208

Goodchild M. F., 2007. *Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0*, International Journal of Spatial Data Infrastructure Research **2** (1), pp. 24-32

Goodman L. A., 1961. *Snowball sampling*, Annals of Mathematical Statistics **32** (1), pp. 148-170

Gramlich W., 1995. *Public Annotations*, In Proceedings of the 1st Workshop on WWW and Collaboration, <http://www.w3.org/Collaboration/Workshop/Proceedings/P10.html>, Last accessed 31st July 2010

Grishman R. and Sundheim B., 1996. *Message Understanding Conference-6: a brief history*, Proceedings of the 16th conference on Computational linguistics, Stroudsburg, USA

Gross R. and Acquisti A., 2005. *Information revelation and privacy in online social networks*, Proceedings of the 2005 ACM Workshop on Privacy in the Electronic Society, New York (USA), pp. 71–80

Gruhl D., Guha R., Libe-Nowell D. and Tomkins A., 2004. *Information diffusion through blogspace*, Proceedings of the 13th international conference on World Wide Web, New York (USA), pp. 491-501

Gruhl D., Meredith D. N., Pieper J. H., Cozzi A. and Dill S., 2006. *The web beyond popularity: a really simple system for web scale RSS*, Proceedings of the 15th international conference on World wide web, Edinburgh (Scotland), pp. 183-192

Gunning R., 1952. *The technique of clear writing*, McGraw-Hill International Book Co, USA

Hackett S. and Parmanto B., 2005. *A longitudinal evaluation of accessibility: higher education web sites*, Internet Research **15** (3), pp.281-294

Halvey M. J. and Keane. M. T., 2007a. *Analysis of online video search and sharing*, In Proceedings of the eighteenth conference on Hypertext and hypermedia, Manchester, UK

Halvey M. J. and Keane. M. T., 2007b. *Exploring social dynamics in online media sharing*, Proceedings of the 16th international conference on World Wide Web, Banff, Canada

Hambrick M. E., Simmons J. M., Greenhalgh G. P. and Greenwell T. C., 2010. *Understanding Professional Athletes' Use of Twitter: A Content Analysis of Athlete Tweets*, International Journal of Sport Communication, **3** (1), pp. 454-471

Han J. and Kamber M., 2006. *Data Mining: Concepts and Techniques*, Morgan Kaufman, USA

Hayes A. F. and Krippendorff K., 2007. *Answering the Call for a Standard Reliability Measure for Coding Data*, Communication Methods and Measures, **1** (1), pp. 77-89

Hearst M. A., 2009. *Social Technology: Technology used by groups of people. How society is being changed by technology-mediated interactions*, Presentation at the NSF CISE, March 18th 2009

Heaton J., 2007. *HTTP Programming Recipes for C# Bots*, Heaton Research, USA

Hill A., 2009. *Analysing Survey Data*, Presentation to UNLV (Angelina Hill, Associate Director of Academic Assessment) at the Academic Assessment Workshop, May 14th-15th 2009

HMIC, Feb. 2011. *Policing Public Order: an overview and review of progress against the recommendations of Adapting to Protest and Nurturing the British Model of Policing*, HMIC Government Report, UK

Hoffman D.L., Novak T.P. and Peralta M., 1999. *Building consumer trust online*, Communications of the ACM **42** (4), pp. 80-85

Hofstadter D., 1979. *Gödel, Escher, Bach: an eternal golden braid*, Penguin Press, UK

Howe J., 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, Crown Business Publishers, USA

Howell B. A., 2006. *Proving Web History: How to use the Internet Archive*, Journal of Internet Law **9** (8), pp. 3–9

Hu M., Lim E. P., Sun A., Lauw H. W. and Vuong B. Q., 2007. *Measuring article quality in Wikipedia: models and evaluation*, Proceedings of CIKM – Information and Knowledge Management Conference, Lisboa, Portugal

Hu J., Fang L., Cao Y., Zeng H. J., Li H., Yang Q. and Chen Z., 2008. *Enhancing text clustering by leveraging Wikipedia semantics*, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore

Hu N., Zhang J. and Pavlou P. A., 2009. *Overcoming the J-shaped distribution of product reviews*, Communications of the ACM, **52** (10), pp. 144-147

Hughes, B. and Joshi, I. and Wareham, J., 2008. *Health 2.0 and Medicine 2.0: Tensions and Controversies in the Field*, Journal of Medical Internet Research **10** (3), online: e23

Huijboom N., Frissen V., Broek T. and Punie Y., 2010. *The Impact of Social Computing on Public Services: a Rationale for Government 2.0*, European Journal of ePractice **9** (3), pp. 5-19

IBM SPSS inc., 2004. CHAID and Exhaustive CHAID Algorithms, <http://support.spss.com/productsext/spss/documentation/statistics/algorithms/14.0/TREE-CHAID.pdf>, Last Accessed 31st July 2010

- Jackson N. A. and Lilleker D. G., 2009. *Building an Architecture of Participation? Political Parties and Web 2.0 in Britain*, Journal of Information Technology & Politics, **6** (3), pp. 232-250
- Jaeger R., 2002. *All About Hedge Funds : The Easy Way to Get Started*, McGraw-Hill, USA
- Johnson T.J., Kaye B.K., Bichard S.L., and Wong W.J., 2007. *Every blog has its day: Politically-interested internet users' perceptions of blog credibility*, Journal of Computer-Mediated Communication, **13** (1), pp. 100-122
- Jones A. L., 2006. *Have internet message boards changed market behaviour?*, Info, **8** (1), pp. 67-76
- Jordan R., 2003. *Crossroads of Twilight (Wheel of Time, Book 10)*, Tor Books, USA
- Kaiser H. F., 1960. *The application of electronic computers to factor analysis*, Educational and Psychological Measurement, **20**, pp. 141-151
- Kambil A. and Van Heck E., 2002. *Making Markets: How Firms Can Design and Profit from Online Auctions and Exchanges*, Harvard Business Press, USA
- Kambil A., 2003. *You Can Bet on Idea Markets*, Last Accessed on 1st April 2011, <http://hbswk.hbs.edu/archive/3808.html>
- Kanefsky B., Barlow N. G. and Gulick V. C., 2001. *Can distributed volunteers accomplish massive data analysis tasks?*, Technical Report, <http://bit.ly/9mtG1i>, Last Accessed: 1st August 2010
- Kaplan A. M. and Haenlein M., 2010. *Users of the world, unite! The challenges and opportunities of Social Media*, Journal of Business Horizons **53** (2), pp. 59-68
- Kassing J. W. and Sanderson J., 2010. *Fan-Athlete Interaction and Twitter Tweeting Through the Giro: A Case Study*, International Journal of Sport Communication **3** (1), pp. 113-128
- Keen A., 2007. *The Cult of the Amateur: How the Democratization of the Digital World is Assaulting Our Economy, Our Culture, and Our Values*. Doubleday Currency Publishing, USA
- Kennedy G., Dalgarno B., Gray K., Judd T., Waycott J., Bennett S., Maton K., Krause K. L., Bishop A. and Chang R., 2007. *The net generation are not big users of Web 2.0 technologies: Preliminary findings*, Proceedings of Conference on ICT: Providing choices for learners and learning, Singapore, Republic of Singapore
- Kittur A., Suh B., Pendleton B.A. and Chi E.H. 2007. *He Says, She says: Conflict and coordination in Wikipedia*, Proceedings of the SIGCHI conference on Human factors in computing systems, San Jose, USA

Klein M., 2007. *Achieving Collective Intelligence via Large-Scale On-Line Argumentation*, MIT CCI Working Paper No. 4647-07, available online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1040881

Kneepkens E. W. E. M. and Zwaan R. A., 1994. *Emotions and literary text comprehension*, *Poetics*, **23** (1), pp. 125-138

Koltay T., 2011. *New media and literacies: Amateurs vs. Professionals*, *Internet Journal - First Monday* **16** (1)

Koppel M. and Shtrimberg I., 2006. *Computing attitude and affect in text: theory and applications - Good News or Bad News? Let the Market Decide*, eds. Shanahan J., Qu Y. and Wiebe J., Springer, **20** (1), pp. 297-301

Krippendorff K., 2004. *Content analysis: An introduction to its methodology*, Thousand Oaks, CA: Sage, USA

Kuznetsov S., 2006. *Motivation of Contributors to Wikipedia*, *ACM SIGCAS Journal of Computers and Society*, **36** (2), pp. 1-7

Kuzma J. 2010. *Asian Government Usage of Web 2.0 Social Media*, *European Journal of ePractice* **9** (3), pp. 71-83

Laender A. H. F., Ribeiro-Neto B. A., De Silva A. S. and Teixeira J.S., 2002. *A brief survey of web data extraction tools*, *ACM Sigmod Record*, **31** (2), pp. 84-93

Lames M., Dabnichki P., and Baca A., 2008. *Computers in Sport: Coaching and computer science*, WIT Press, UK, pp. 99-119

Lanier J., 2010. *You Are Not A Gadget: A Manifesto*, Allen Lane Publishers, USA

Laningham S., 2006, *Tim Berners-Lee Originator of the Web and director of the World Wide Web Consortium talks about where we've come, and about the challenges and opportunities ahead*, developerWorks Interviews, 28th July 2006, <http://www.ibm.com/developerworks/podcast/dwi/cm-int082206txt.html>, Last accessed 31st July 2010

Leadbeater C., 2009. *We-Think: Mass innovation, not mass production*, Profile Books Press, USA

Lerman K., 2007. *Social Information Processing in News Aggregation*, *IEEE Internet Computing Journal*, **11** (6), pp. 16-28

Lévy P., 1994. *L'intelligence collective. Pour une anthropologie du cyberspace*, La Découverte, France

- Li C. and Bernoff J., 2008. *Groundswell: Winning in a World Transformed by Social Technologies*. Harvard Business School Press, USA
- Li C. and Bernoff J., 2008. *Harnessing the power of the Oh-So-Social Web*, MITSloan Management Review Journal **49** (3), pp. 36-42
- Lih A., 2004. *Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as news resource*, Proceedings of the 5th International Symposium on Online Journalism, Hong Kong, China
- Likert R., 1932. *A Technique for the Measurement of Attitudes*, Archives of Psychology, **140** (1), pp. 1-55
- Lindmark S., 2009. *Web 2.0: Where does Europe stand?*, European Comission (JRC) Joint Research Centre - (IPTS) Institute for Prospective Technological Studies Report, Seville (Spain)
- Liu B., 2010 (2nd ed). *Sentiment Analysis and Subjectivity*, Handbook of Natural Language Processing, CRC Press - Taylor and Francis Group, USA
- Liu B. and Hu M., 2004. *Mining and summarizing customer reviews*, Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, USA
- Liu B., Hu M. and Cheng J., 2005. *Opinion Observer: Analyzing and Comparing Opinions on the Web*, Proceedings of the 14th international conference on World Wide Web, Chiba, Japan
- Lo A., 2004. *The adaptive market hypothesis: Market efficiency from an evolutionary perspective*, Journal of Portfolio Management, **30** (1), pp. 15–29
- Malinen S., 2009. *Heuristics for supporting social interaction in online communities*, IADIS WWW/Internet Conference Proceedings, Rome, Italy
- Malone T. W., Laubacher R., Introne J., Klein M., Abelson H., Sterman J. and Gary Olson, 2009. *The Climate Collaboratorium: Project Overview*, CCI Working Paper No. 2009-003, online: <http://cci.mit.edu/publications/CCIwp2009-03.pdf>
- Malouf R. and Mullen T., 2006. *A preliminary investigation into sentiment analysis of informal political discourse*, Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, California, USA
- Malouf R. and Mullen T., 2007. *Graph-based user classification for informal online political discourse*, Proceedings of the 1st Workshop on Information Credibility on the Web 2007, Miyazaki, Japan
- Mares M. and Woodard E. H., 2006. *In Search of the Older Audience: Adult Age Differences in Television Viewing*, Journal of Broadcasting and Electronic Media **50** (4), pp. 595-614

- Martin E., 2006. *Survey Questionnaire Construction*, U.S. Census Bureau – Research Report Series on Survey Methodology, December 21st 2006, pp. 1-13
- Marwick A. E., 2008. *To catch a predator? The MySpace moral panic*, Internet Journal - First Monday **13** (6)
- Matsuo Y. and Yamamoto H., 2009. *Community Gravity: Measuring Bidirectional Effects by Trust and Rating on Online Social Networks*, Proceedings of the 18th international conference on World wide web, Madrid (Spain)
- Mayer-Schonberger V., 2009. *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press, USA
- McCown F., Nelson M. L., 2007. *Search Engines and Their Public Interfaces: Which APIs are the Most Synchronized?*, Proceedings of the 16th international conference on World wide web, Banff (Canada)
- McLean R., Richards B. H. and Wardman J., 2007. *The effect of Web 2.0 on the future of medical practice and education: Darwinkinian evolution or folksonomic revolution?*, The Medical Journal of Australia **187** (3), pp. 174-177
- Meier P. and Brodock K., 2008. *Crisis Mapping Kenya's Election Violence: Comparing Mainstream News, Citizen Journalism and Ushahidi*, Harvard Humanitarian Initiative, HHI, Harvard University, 23rd October 2008, Last Accessed: 1st August 2010
- Michel J.B., Shen Y.K., Aiden A.P., Veres A., Gray M.K., Pickett J.P., Hoiberg D., Clancy D., Norvig P. and Orwant J., 2010. *(Supporting Online Material for) Quantitative analysis of culture using millions of digitized books*, Science Journal, Published online 16 December 2010 [DOI:10.1126/science.1199644], online: <http://www.sciencemag.org/cgi/content/full/science.1199644/DC1>
- Millard D. E. and Ross M., 2006. *Web 2.0: Hypertext by any other name?*, Proceedings of the 17th conference on Hypertext and Hypermedia, Odense, Denmark
- Mishne G. and Rijke M., 2006a. *Capturing global mood levels using blog posts*. Technical report, University of Amsterdam
- Mishne G. and Rijke M., 2006b. *Moodviews: Tools for blog mood analysis*, Technical report, University of Amsterdam
- Mishne G., Balog K., Rijke M., and Ernsting B., 2007. *Moodviews: Tracking and searching mood-annotated blog posts*, Proceedings of International Conference on Weblogs and Social Media, Tokyo, Japan
- Mittermayer M. A. and Knolmayer G., 2006. *Text mining systems for market response to news:*

A survey, Working paper at Bern University, available at <http://www2.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-184.pdf>

Moore D. S. and McCabe G. P., 2001. *Introduction to the Practice of Statistics*, Freeman and Company, USA

Murphy J. J., 1999. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*, Prentice Hall, USA

Murphy C. and Persson N., 2008. *HTML and CSS Web Standards Solutions: A Web Standardistas' Approach*, Friends of Ed, USA

Murphy J., Hashim N. and O'Connor P., 2008. *Take me back: validating the Wayback Machine*, Journal of Computer-Mediated Communication **13** (1), pp. 60–75

Murray E., Burns J., See Tai S., Lai R. and Nazareth I., 2005. *Interactive Health Communication Applications for people with chronic disease*, Cochraine Report, John Wiley and Sons, USA

Nagar Y. and Malone T., 2011. *Combining Human and Machine Intelligence for Making Predictions*, CCI Working Paper No. 2011-02, online: <http://cci.mit.edu/publications/CCIwp2011-02.pdf>

Nanno T. and Okumura M., 2006. *HTML2RSS: automatic generation of RSS feed based on structure analysis of HTML document*, Proceedings of the 15th international conference on World wide web, Edinburgh (Scotland), pp. 1061-1062

Nath S. V., 2006. *Crime pattern detection using data mining*, Proceedings of 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 41-44

Nielsen, July 2009. *Nielsen news release*, http://www.nielsen-online.com/pr/pr_090713.pdf, Last accessed 1st October 2010

Nielsen, February 2010. *Facebook Users Average 7 hrs a Month in January as Digital Universe Expands*, Nielsenwire blog – http://blog.nielsen.com/nielsenwire/online_mobile/facebook-users-average-7-hrs-a-month-in-january-as-digital-universe-expands, Last accessed 1st October 2010

Nielsen, September 2011. *August 2011 – Top US Web Brands*, http://blog.nielsen.com/nielsenwire/online_mobile/august-2011-top-us-web-brands, Last accessed 1st October 2011

Ochoa X. and Duval E., 2008. *Quantitative Analysis of User-Generated Content on the Web*, Proceedings of the First International Workshop on Understanding Web Evolution, Beijing (China)

- Orchard L. M., 2006. *Hacking Delicious (eXtreme Tech)*, Wiley Publishing, USA
- Osimo D., 2008. *Web 2.0 in Government: Why and How?*, Institute for Prospective Technological Studies (IPTS), JRC, European Commission - Technical Report
- O'Neil W. J., 2002. *How to Make Money in Stock - A Winning System in Good Times or Bad*, McGraw-Hill, USA
- O'Reilly T., 2005. *What Is Web 2.0*. O'Reilly Network, 30th October 2005. Last Accessed 31st July 2010
- O'Reilly T., 2006. *Levels of the Game: The Hierarchy of Web 2.0 Applications*, O'Reilly Network, 17th July 2006,. Last Accessed: 31st July 2010
- O'Sullivan D., 2009. *Wikipedia: A New Community of Practice?*, Ashgate Publishers, UK
- Pang B. and Lee L., 2005. *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*, Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Michigan, USA
- Pang B. and Lee L., 2008. *Opinion Mining and Sentiment Analysis*, Foundations and Trends in Information Retrieval, **2** (1-2), pp. 1-135
- Paroutis S. and Saleh A., 2009. *Determinants of knowledge sharing using Web 2.0 technologies*, Journal of Knowledge Management, **13** (4), pp. 52-63
- Parycek P. and Sachs M., 2010. *Open Government – Information Flow in Web 2.0*, European Journal of ePractice **9** (3), pp. 59-70
- Pascu C., 2008. *An Empirical Analysis of the Creation, Use and Adoption of Social Computing Applications: IPTS Exploratory Research on the Socio-economic Impact of Social Computing*, European Commission (JRC) Joint Research Centre - (IPTS) Institute for Prospective Technological Studies Report, Seville (Spain)
- Passmore C., Dobbie A. E., Parchman M. and Tysinger J., 2002. *Guidelines for Constructing a Survey*, Family Medicine, **34** (4), pp. 281-286
- Pegoraro A., 2010. *Look Who's Talking - Athletes on Twitter: A Case Study*, International Journal of Sport Communication **3** (1), pp. 501-514
- Pennock D. M., Lawrence S. C., Giles L. and Nielsen F. A., 2001. *The real power of artificial markets*, Science **291** (5506), pp. 987-988
- Petersen S. M., 2008. *Loser Generated Content: From Participation to Exploitation*, Internet Journal - First Monday **13** (3)

Priedhorsky R., Chen J., Lam S.T.K., Panciera K., Terveen L. and Riedl J., 2007. *Creating, Destroying and Restoring Value in Wikipedia*, Proceedings of the 2007 International ACM Conference on Supporting Group Work, Sanibel Island, USA

Radwanick S., 2010. *The 2009 U.S. Digital Year in Review: A Recap of the Year in Digital Marketing*, ComScore Inc., Technical Report, USA

Recordon D., 2010. *OpenID: The Definitive Guide*, O'Reilly Media, USA

Reese S. D., Rutigliano L., Hyun K. and Jeong J., 2007. *Mapping the blogosphere: Professional and citizen-based media in the global news arena*, Journalism **8** (3), pp. 235-261

Resnick P. and Zeckhauser R., 2002. *Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System*, The Economics of the Internet and E-Commerce – Advances in Applied Microeconomics, **11** (1)

Russell P., 1983. *The Global Brain Awakens*, Global Brain Publishing, USA

Sabherwal S. and Sarkar S. K. and Zhang Y., 2008. *Online talk: does it matter?*, Journal of Managerial Finance, **34** (2), pp. 423-436

Sakaki T., Okazaki M., and Matsuo Y., 2010. *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web, Raleigh, USA

Salganik M. J. and Heckathorn D. D., 2004. *Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling*, Sociological Methodology, **34** (1), pp. 193–239

Scharl A. and Tochtermann K., 2007. *The Geospatial Web: How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society*, Springer, UK

Schau H. J. and Gilly M. C., 2003. *We are what we post? Self-presentation in personal web space*, Journal of Consumer Research, **30** (3), pp. 385-404

Schneider J., Passant A. and Breslin J., 2010. *A Qualitative and Quantitative Analysis of How Wikipedia Talk Pages Are Used*, Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, Raleigh, USA

Scholz T., 2008. Market Ideology and the Myths of Web 2.0, Internet Journal - First Monday **13** (3)

Schroth C. and Janner, T, 2007. *Web 2.0 and SOA: Converging Concepts Enabling the Internet of Services*, IT Professional **9** (3), pp. 36-41

Schwager J. D., 1993. *Market Wizards: Interviews with Top Traders*, Collins Publishers, USA

- Schwarz N. and Sudman S., 1996. *Answering Questions: Methodology for determining cognitive and communicative processes in survey research*, Jossey Bass, USA
- Segaran T., 2007. *Programming Collective Intelligence: Building Smart Web 2.0 Applications*, O'Reilly Media, USA
- Servan-Schreiber E., Wolfers J., Pennock D. M. and Galebach B., 2004. *Prediction market, does money matter?*, *Electronic Markets*, **4** (3), pp. 243-251
- Shirky C., 2009. *Here Comes Everybody: The Power of Organising Without Organisations*, Penguin Press, reprint edition
- Shirky C., 2010. *Cognitive Surplus: Creativity and Generosity in a Connected Age*, Allen Lane Publishers, USA
- Short J., Williams E. and Christie B., 1976. *The social psychology of telecommunications*, John Wiley & Sons Publishers, USA
- Shrout P.E. and Fleiss J.L., 1979. *Intraclass correlations: uses in assessing rater reliability*, *Psychological bulletin*, **86** (2), pp. 420-428
- Siegel J. J., 2002. *Stock for the long run - Guide to Financial Market Returns and Long-Term Investment Strategies*, McGraw-Hill, USA
- Silver D., 2008. *History, Hype, and Hope: An Afterward*, *Internet Journal - First Monday* **13** (3)
- Slot M. and Frissen V., 2007. *Users In The "Golder" Age of The Information Society*, *Observatorio Journal* **1** (3), pp. 201-224
- Solove D. J., 2008. *The Future of Reputation: Gossip, Rumour, and Privacy on the Internet*, Yale University Press, USA
- Soros G, 1995. *Soros on Soros: Staying Ahead of the Curve*, Wiley Publishing, USA
- Spector R., 2002. *Amazon.com: Get Big Fast*, Harper Paperbacks, USA
- Starbird K. and Palen L., 2010. *Pass It On?: Retweeting in Mass Emergency*, Proceedings of the 7th International ISCRAM Conference on Information Systems for Crisis Response and Management, Seattle, USA
- Stekauer P. and Lieber R., 2006. *Handbook of word-formation*, Springer, Germany
- Stolcke A., Ries K., Coccaro N., Shriberg E., Bates R., Jurafsky D. P., Martin T. R., Van Ess-Dykema C. and Meteer M., 2000. *Dialogue act modeling for automatic tagging and recognition of conversational speech*, *Computational Linguistics*, **26** (3), pp. 339-373

Subraya B. M., 2006. *Integrated Approach to Web Performance Testing: A Practitioner's Guide*, IRM Press Publishing, USA

Surowiecky J., 2005. *The Wisdom of Crowds*, Anchor Publishing, USA

Sykora M., 2009. *Power of Web 2.0 Mass Collaboration in Computational Intelligence and its' uses, an example from Finance*, Proceedings of the UKCI 9th Annual Workshop on Computational Intelligence, Nottingham (England)

Sykora M. and Panek M., 2009. *Financial news content publishing on Youtube.com*. Proceedings of the 3rd International Workshop on Soft Computing Applications, 29 July–1 August, Szeged (Hungary) – Arad (Romania), pp. 99-104.

Sykora M., Wang X., Archer R., Parish D. and Bez H. E., 2009. *Case Based Reasoning Approach for Transaction Outcomes Prediction on Currency Markets*, Proceedings of the 3rd IEEE Conference on Soft Computing and Applications - SOFA2009, Szeged / Arad, Hungary / Romania

Tapscott D. and Williams A. D., 2008. *Wikinomics – How mass collaboration changes everything*, Atlantic Books, USA

Taylor J. G., Shadbolt J., Adcock C., Attew D., Burgess N., Hazarika N., Larsson S. and Towers N., 2002. *Neural Networks and the Financial Markets: Predicting, Combining and Portfolio Optimisation*, Springer, USA

Tetlock P. C., 2007. *Giving content to investor sentiment: The role of media in the stock market*, Journal of Finance, **62** (3), pp. 1139–1168

Thelwall M., Buckley K., Paltoglou G. Cai D. and Kappas A., 2010. *Sentiment strength detection in short informal text*, Journal of the American Society for Information Science and Technology, **61** (12), pp. 2544–2558

Thelwall M., 2009. *Introduction to Webometrics: Quantitative Web Research for the Social Sciences (Synthesis Lectures on Information Concepts)*, Retrieval, and Services, Morgan & Claypool, UK

Thomas J.D. and Sycara K., 2000. *Integrating genetic algorithms and text learning for financial prediction*, Proceedings of the Genetic and Evolutionary Computing 2000 Conference Workshop on Data Mining with Evolutionary Algorithms, pp. 72-75, Las Vegas, Nevada, USA

Thurman N., 2008. *Forums for citizen journalists? Adoption of user generated content initiatives by online news media*, **10** (1), pp. 139-157

Tomasello M., Rakoczy H. and Wameken F., 2008. *The Sources of Normativity: Young Children's Awareness of the Normative Structure of Games*, Developmental Psychology **44** (3), pp. 875-881

Tumarkin R. and Whitelaw R.F., 2001. *News or noise? Internet postings and stock prices*, Financial Analysts Journal, **57** (1) pp. 41-51

Turner A., 2006. *Introduction to Neogeography*, O'Reilly Media, USA

Turner T., 2007. *A Beginner's Guide to Day Trading Online*, Adams Media Corporation, USA

Tumasjan A., Sprenger T. O., Sandner P. G. and Welpe I. M., 2010. *Predicting elections with twitter: What 140 characters reveal about political sentiment*, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Menlo Park, California, USA

Ure J., 1971. *Lexical density and register differentiation*, Application of linguistics, Cambridge: Cambridge University Press, pp. 443-452

Ushahidi, 2010. *About: Introduction and History*, www.usahidi.com/about, Last accessed 1st August 2010

Van De Belt, 2010. *Definition of Health 2.0 and Medicine 2.0: A Systematic Review*, Journal of Medical Internet Research **12** (2), online: e18

Veronin M. A., 2002. *Where Are They Now? A Case Study of Health-related Web Site Attrition*, Journal of Medical Internet Research, **4** (2), available online: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1761933>

Vineel G., 2009. *Web page DOM node characterisation and its Application to Page Segmentation*, 3rd IEEE International Conference on Internet Multimedia Services Architecture and Applications (IMSAA), Bangalore, India

Vosen G. and Hagemann S., 2007. *Unleashing Web 2.0: From Concepts to Creativity*, Morgan Kaufmann, USA

Wagner C. and Prasarnphanich P., 2007. *Innovating Collaborative Content Creation: The Role of Altruism in Wiki Technology*, Proceedings of the 40th Annual Hawaii International Conference on System Sciences HICSS07, Waikoloa, Hawaii

Wang C. Y., 2009. *An Empirical Study of Continuing Usage in Group-Buying Websites*, Master Thesis, National Central University (Taiwan)

Weick K., Sutcliffe K. and Obstfeld D., 2005. *Organizing and the process of sensemaking*, Journal of Organizational Science **16** (4), pp. 409-421

Wells H. G., 1994. *World Brain: the Idea of a Permanent World Encyclopedia*, Alan Mayne Publishers

Westwood S. J. and Messing S., 2011, *Selective Exposure in the Age of Social Media*:

Endorsements Trump Partisan Source Affiliation when Selecting Online News Media, Working paper at Stanford University, available at <http://stanford.edu/~seanjw/papers/socialNews.pdf>

Wilde E., 2010. *RESTful web services: principles, patterns, emerging technologies*, Proceedings of the 19th international conference on World wide web, Tutorial Session, Raleigh (USA)

World Wide Web Consortium (W3C), 1995. *Proceedings of the 1st Workshop on WWW and Collaboration*, 11th-12th September 1995, <http://www.w3.org/Collaboration/Workshop/>, Last accessed 31st July 2010

Wright A., Bates D., Middleton B., Hongsermeier T., Kashyap V., Thomas S. M. and Sittig D. F., 2009. *Creating and sharing clinical decision support content with Web 2.0: Issues and examples*, Journal of Biomedical Informatics **42** (2), pp. 334-346

Wunsch-Vincent S. and Vickery G., 2007. *Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking*, OECD (Organisation for Economic Co-operation and Development) Research Report, Paris (France)

Wuthrich B., Cho V., Leung S., Permuntilleke D., Sankaran K. and Zhang J., 1998. *Daily stock market forecast from textual web data*, Systems, IEEE International Conference on Man, and Cybernetics, San Diego, USA

Wysocki P., 1999. *Cheap talk on the web: the determinants of postings on stock message boards*, Working Paper, online: <http://mit.edu/wysockip/www/cheaptalk.pdf>

Ye N., 2003. *The Handbook of Data Mining*, Lawrence Erlbaum Associates Publishers, USA

Zeldman J., 2007. *Designing with web standards*, New Riders Publishers in association with AIGA, USA

Zemke S., 2003. *Data Mining for Prediction: Financial Series Case*, PhD thesis, The Royal Institute of Technology, Department of Computer and Systems Sciences, Sweden

Zhang D., Guo B., Li B. and Yu Z., 2010. *Extracting Social and Community Intelligence from Digital Footprints: An Emerging Research Area*, Ubiquitous Intelligence and Computing (Springer), 6406, pp. 4-18

Zimmer M., 2008. *Critical Perspectives on Web 2.0*, Internet Journal - First Monday **13** (3)

Zwaan R. A., Britton B. K. and Graesser A. C., 1993. *Toward a model of literary comprehension*, Models of understanding text, Hillsdale NJ Erlbaum, USA

APPENDIX A: *Defining Web 2.0*

Table A.1 – Top 10 words (text normalised as per methodology) based on all available data for the year 2005

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Library	1.Health	1.Medical	1Health.	1.Medical
2.Blog	2.Blog	2.Blog	2.Blog	2.Medical	2.Health	2.Search	2.Search
3.Librarian	3.Information	3.Librarian	3.Information	3.Search	3.Search	3.Medical	3.Health
4.Technology	4.Web	4.Technology	4.Web	4.Medicine	4.Engine	4.Medicine	4.Engine
5.Reference	5.Librarian	5.Reference	5.Librarian	5.Blog	5.Technology	5.Reference	5.Pubmed
6.Internet	6.Search	6.Internet	6.Search	6.Gadget	6.Journal	6.Blog	6.Interface
7.Information	7.Resource	7.Information	7.Resource	7.Reference	7.Internet	7.Gadget	7.Technology
8.Search	8.Technology	8.Search	8.Technology	8.Science	8.Emerge	8.Pubmed	8.Alternative
9.Bookmark	9.Email	9.Bookmark	9.Email	9.Technology	9.Information	9.Science	9.Journal
10.News	10.Professional	10.News	10.Professional	10.Searchengine	10.Site	10.Web	10.Site

Table A.2 – Top 10 words (text normalised as per methodology) based on all available data for the year 2006

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Library	1.Health	1.Medical	1.Health	1.Health
2.Blog	2.Blog	2.Blog	2.Web	2.Medicine	2.Health	2.Search	2.Search
3.Librarian	3.Web	3.Technology	3.Technology	3.Medical	3.Search	3.Medical	3.Medical
4.Technology	4.Information	4.Education	4.Bookmark	4.Blog	4.Technology	4.Medicine	4.Engine
5.Search	5.Librarian	5.Metadata	5.Information	5.Search	5.Site	5.Blog	5.Site
6.Reference	6.Search	6.Tag	6.Metadata	6.Reference	6.Journal	6.Research	6.Computer
7.Research	7.Resource	7.Book	7.Blog	7.Gadget	7.Physician	7.Reference	7.People
8.Tool	8.Technology	8.Web	8.System	8.Healthcare	8.Information	8.Science	8.Web
9.Firefox	9.Provide	9.Bookmark	9.Use	9.Science	9.Blog	9.Healthcare	9.Information
10.Book	10.Open	10.Folksonomy	10.Education	10.Technology	10.Internet	10.News	10.Pubmed

Table A.3 – Top 10 words (text normalised as per methodology) based on all available data for the year 2007

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Library	1.Health	1.Medical	1.Health	1.Health
2.Blog	2.Blog	2.Javascript	2.Web	2.Medical	2.Health	2.Medical	2.Medical
3.Technology	3.Web	3.Ajax	3.Use	3.Medicine	3.Information	3.Medicine	3.Search
4.Librarian	4.Librarian	4.Blog	4.Dialog	4.Healthcare	4.Doctor	4.Reference	4.Information
5.Book	5.Resource	5.Book	5.Book	5.Reference	5.Search	5.Science	5.Site
6.Reference	6.Social	6.Technology	6.Blog	6.Blog	6.Share	6.Search	6.Science
7.Tool	7.Online	7.Education	7.Javascript	7.Community	7.Online	7.Healthcare	7.Disease
8.Opensource	8.Information	8.Tool	8.Technology	8.Search	8.Community	8.Blog	8.Web
9.Catalogue	9.Technology	9.Prototype	9.Online	9.Doctor	9.Physician	9.Research	9.Engine
10.OPAC	10.Book	10.Tag	10.Information	10.Social	10.Social	10.Community	10.Medicine

Table A.4 – Top 10 words (text normalised as per methodology) based on all available data for the year 2008

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Library	1.Health	1.Health	1.Health	1.Health
2.Blog	2.Web	2.Javascript	2.Web	2.Medicine	2.Medical	2.Science	2.Medical
3.Technology	3.Blog	3.Book	3.Book	3.Medical	3.Information	3.Medicine	3.Search
4.Tool	4.Information	4.Technology	4.Use	4.Healthcare	4.Web	4.Medical	4.Information
5.Opensource	5.Online	5.Blog	5.Online	5.Science	5.Doctor	5.Healthcare	5.Web
6.Education	6.Librarian	6.Education	6.Information	6.Reference	6.Share	6.Reference	6.Site
7.Reference	7.Use	7.Web	7.Javascript	7.Community	7.Patient	7.Research	7.Online
8.Librarian	8.Resource	8.Tool	8.Free	8.Research	8.Body	8.Google	8.Google
9.Wiki	9.Site	9.Ajax	9.Create	9.Blog	9.Site	9.Search	9.Human
10.Web	10.Book	10.Wiki	10.Technology	10.Socialnetwork	10.Medicine	10.Blog	10.Use

Table A.5 – Top 10 words (text normalised as per methodology) based on all available data for the year 2009

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Library	1.Health	1.Health	1.Health	1.Health
2.Blog	2.Web	2.Education	2.Book	2.Medicine	2.Medical	2.Medicine	2.Medical
3.Technology	3.Use	3.Tool	3.Web	3.Healthcare	3.Information	3.Medical	3.Information
4.Book	4.Book	4.Book	4.Use	4.Medical	4.Online	4.Science	4.Search
5.Education	5.Information	5.Javascript	5.Free	5.Health2.0	5.Site	5.Reference	5.Site
6.Tool	6.Librarian	6.Resource	6.Online	6.Reference	6.Use	6.Healthcare	6.Web
7.Reference	7.Blog	7.Reference	7.Create	7.Research	7.Free	7.Research	7.Online
8.Resource	8.Resource	8.Technology	8.Site	8.Science	8.Community	8.Search	8.Use
9.Librarian	9.Site	9.Framework	9.Make	9.Blog	9.Doctor	9.Education	9.Free
10.Research	10.Online	10.Web	10.Framework	10.Socialmedia	10.Patient	10.Health2.0	10.Share

Table A.6 – Top 10 words (text normalised as per methodology) based on all available data for the year 2010

Library 2.0		Library & Web 2.0		Medicine 2.0		Medicine & Web 2.0	
Tags	Comments	Tags	Comments	Tags	Comments	Tags	Comments
1.Library	1.Library	1.Library	1.Book	1.Health	1.Health	1.Health	1.Health
2.Book	2.Book	2.Book	2.Library	2.Healthcare	2.Medical	2.Medicine	2.Medical
3.Blog	3.Web	3.Tool	3.Web	3.Medicine	3.Http	3.Science	3.Information
4.Education	4.Use	4.Education	4.Free	4.Medical	4.Information	4.Medical	4.Site
5.Technology	5.Site	5.Reference	5.Site	5.Packrati.us	5.Twitter	5.Healthcare	5.Http
6.Tool	6.Information	6.Resource	6.Use	6.Reference	6.Free	6.Reference	6.Search
7.Resource	7.Online	7.Web	7.Read	7.Research	7.Patient	7.Research	7.Free
8.Packrati.us	8.Librarian	8.Technology	8.Online	8.Community	8.Social	8.Blog	8.Web
9.Research	9.Resource	9.Javascript	9.Tool	9.Health2.0	9.Online	9.Community	9.Patient
10.Librarian	10.Blog	10.Webdesign	10.Search	10.Twitter	10.Site	10.Video	10.Online

As per methodology, described in sub-section 3.3.3, all textual data was normalised. The web 2.0 terms in Figure A.1 and A.2 were selected by a human expert from the top 30 term occurrences, and are based on the technologies, applications and functionalities that web 2.0 applications facilitate. An apparent upward trend in such term usage can be appreciated from the figures.

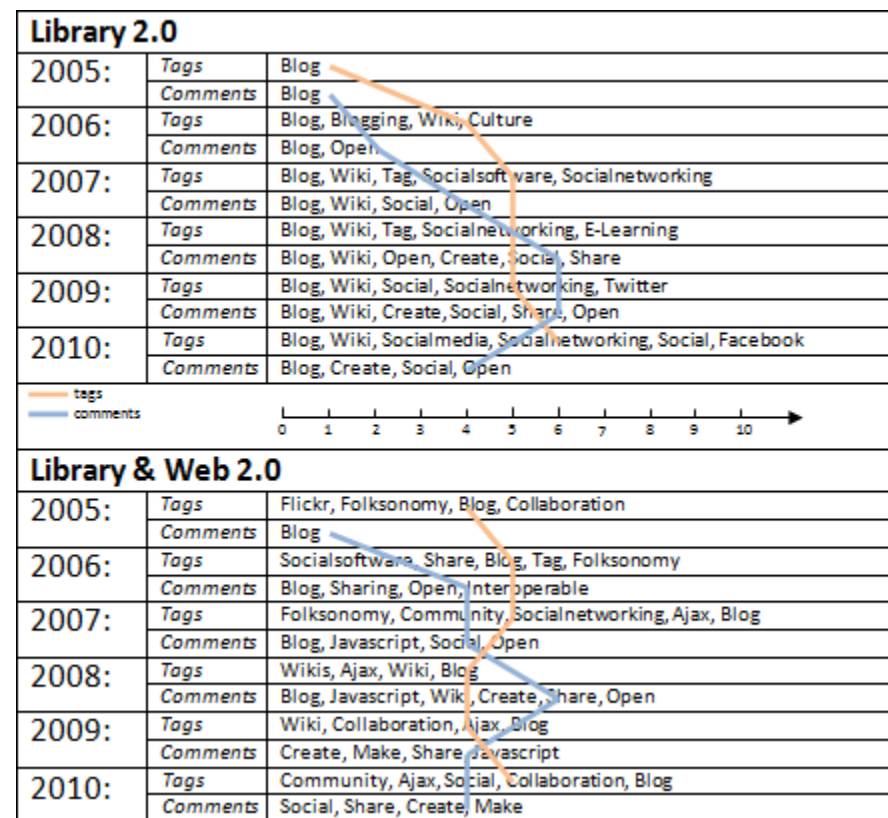


Figure A.1 – Social Web / Web 2.0 terms taken from top 30 most occurring words, for all available data

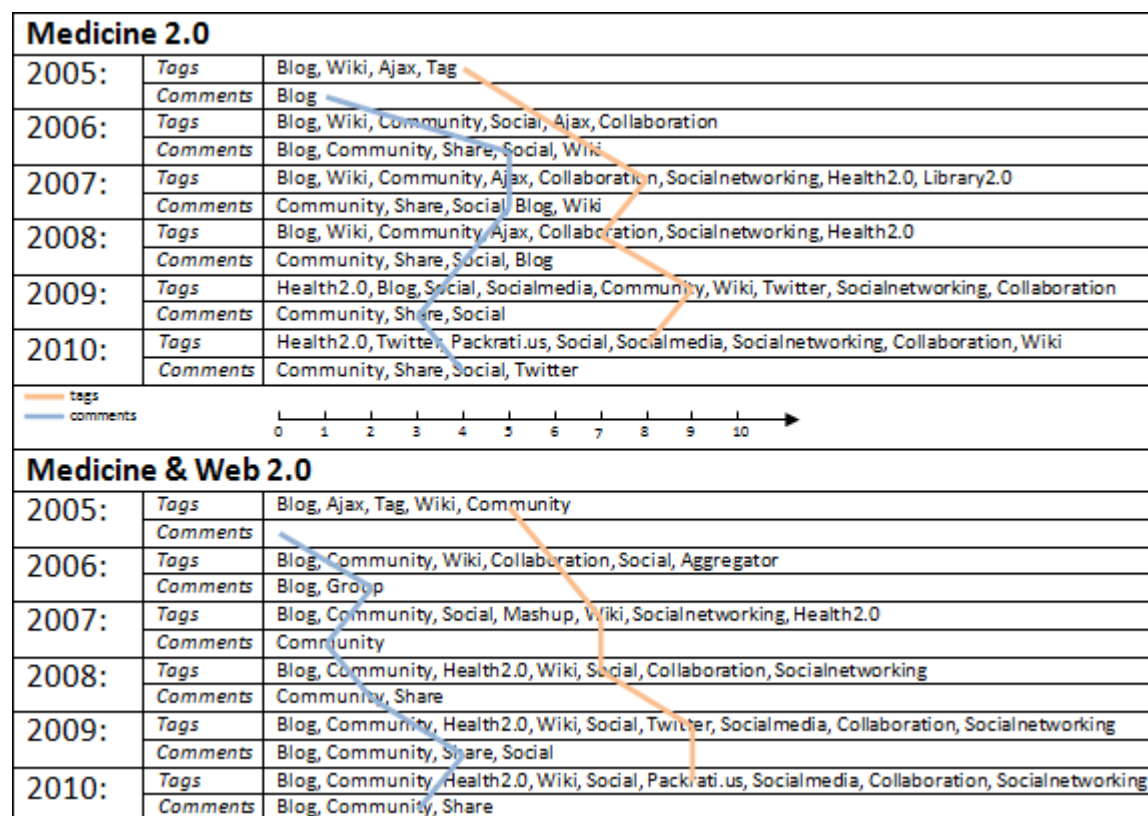


Figure A.2 – Social Web / Web 2.0 terms taken from top 30 most occurring words, for all available data

APPENDIX B: *Survey Analysis and Results*

In this section, further tables detailing full output from the Principal Component Analysis performed in chapter 4 / section 4.4.3 are provided.

Table B.1 – Communalities

	Initial	Extraction
Q5 Trust (likert)	1.000	.626
Q7 Time (relative)	1.000	.793
Q8 Time (likert)	1.000	.843
Q10 Wikipedia useful (likert)	1.000	.854
Web2.0 competence	1.000	.967
Q4 score (business)	1.000	.363
Q6 score	1.000	.772
Trust	1.000	.967
Time	1.000	.989
Varied motives	1.000	.410
Q1 score	1.000	.545
Q2 score	1.000	.684
Q3 score	1.000	.781

The communalities table B.1 reports percentage of variance within each variable that is explained by the resulting factors (i.e. the resulting factors explain 62.60% variance in the variable Q5-Trust [likert]).

Table B.2 - Total Variance Explained

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.809	36.990	36.990
2	1.957	15.053	52.043
3	1.819	13.994	66.037
4	1.011	7.775	73.813
5	.837	6.437	80.250
6	.707	5.437	85.687
7	.622	4.782	90.469
8	.576	4.432	94.901
9	.343	2.640	97.541
10	.320	2.459	100.000
11	4.673E-15	3.594E-14	100.000
12	5.939E-16	4.568E-15	100.000
13	-9.472E-18	-7.286E-17	100.000

Each row in Table B.2 presents the factors that emerged with their total Eigenvalues (the first

four were selected based on their Eigenvalues being larger than 1.0 cutoff).

Table B.3 - Component Matrix

	Component			
	1	2	3	4
Q5 Trust (likert)	.366	.196	.656	.152
Q7 Time (relative)	.679	.450	-.357	.048
Q8 Time (likert)	.589	.639	-.296	-.002
Q10 Wikipedia useful (likert)	.190	-.064	.150	.890
Web2.0 competence	.827	-.491	-.201	-.044
Q4 score (business)	.453	-.207	.238	-.241
Q6 score	.630	.131	.570	-.180
Trust	.642	.183	.718	-.075
Time	.697	.612	-.358	.023
Varied motives	.565	-.175	-.092	.228
Q1 score	.440	-.568	-.105	.134
Q2 score	.714	-.371	-.128	-.144
Q3 score	.790	-.323	-.222	-.064

This is the unrotated solution which identifies individual variables and the resulting factors on which they load. It provides a breakdown of the variance within each variable among the resulting four factors, e.g. $(.365)^2 + (.196)^2 + (.656)^2 + (.152)^2 = .133 + .038 + .430 + .023 = .62$, which equals to the variance explained in the variable Q5-Trust (likert), see first row in the communalities table earlier in this appendix. In order to prevent a single variable from loading too highly on more than one factor, orthogonal rotation was finally performed, which is not included in appendix since it is discussed in table 4.5, chapter 4.

Correlation Analysis

Kendall's tau b correlation coefficient matrix for the numerical scores, representing all the answers to the survey questions are shown in the table below.

Table B.4 – Correlations (Kendall's tau b)

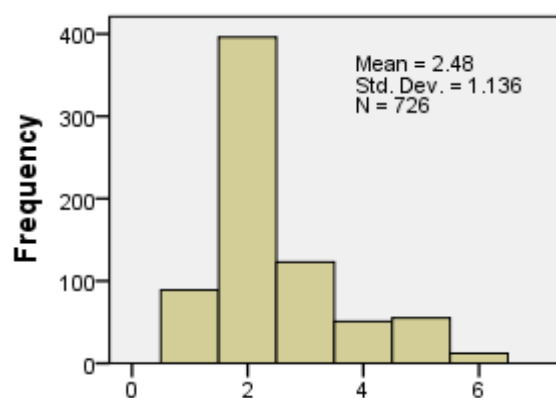
	Age group	Qualification	Expertise area	web2.0-competence	Q4 score business	Trust	Time	Varied motives	Q10 Wikipedia (likert)
Age group	1.000	.284** .000	.075* .016	-.129** .000	-.039 .230	-.240** .000	-.337** .000	-.140** .000	-.057 .073
Qualification level	.284** .000	1.000	.038 .227	.029 .325	-.034 .296	-.155** .000	-.142** .000	.003 .924	.018 .585
Expertise area	.075* .016	.038 .227	1.000	.207* .000	.053 .090	.020 .500	-.069 .018	.081 .008	.068 .030
web2.0 competence	-.129** .000	.029 .325	.207* .000	1.000	.266** .000	.229** .000	.236** .000	.365** .000	.088 .003
Q4 score (business)	-.039 .230	-.034 .296	.053 .090	.266** .000	1.000	.237* .000	.095* .002	.122 .000	.081 .012
Trust	-.240** .000	-.155** .000	.020 .500	.229** .000	.237* .000	1.000	.225** .000	.213 .000	.094** .002
Time	-.337** .000	-.142** .000	-.069 .018	.236** .000	.095* .002	.225** .000	1.000	.229** .000	.050 .097
Varied motives	-.140** .000	.003 .924	.081* .008	.365** .000	.122 .000	.213* .000	.229** .000	1.000	.135* .000
Wikipedia	-.057 .073	.018 .585	.068 .030	.088* .003	.081 .012	.094 .002	.050 .097	.135* .000	1.000

** . Correlation is significant at the 0.0005 level (2-tailed).

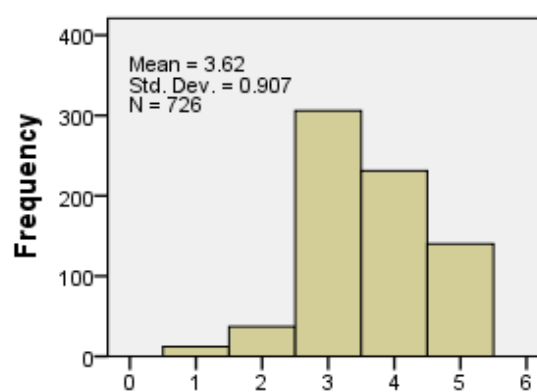
* . Correlation is significant at the 0.05 level (2-tailed).

Each item has two rows, first row represents Kendall's tau b, second row represents signif. test value.

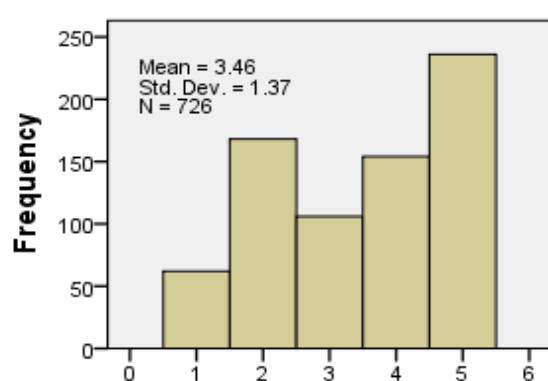
Distribution histograms for all the survey responses



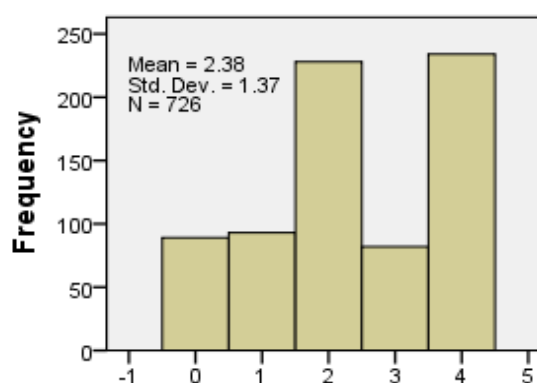
Age group



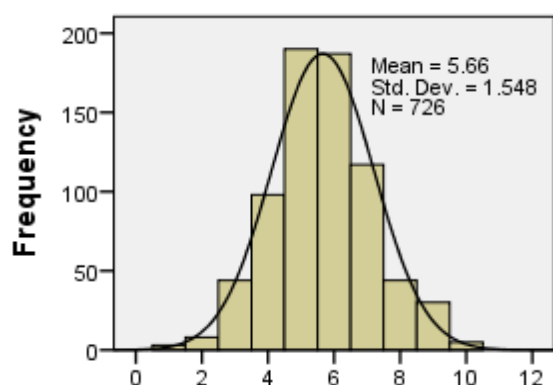
Qualification level



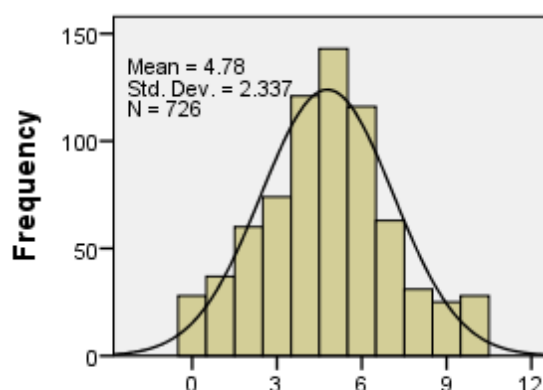
Expertise area



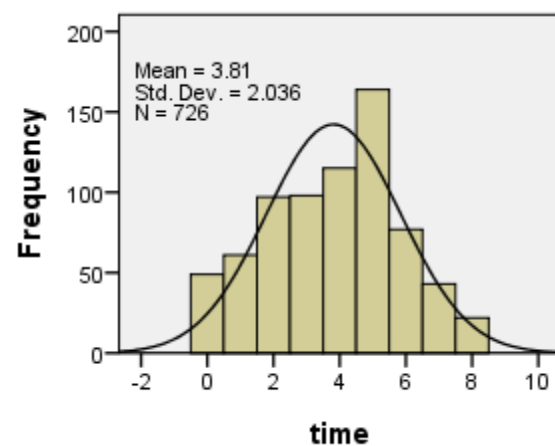
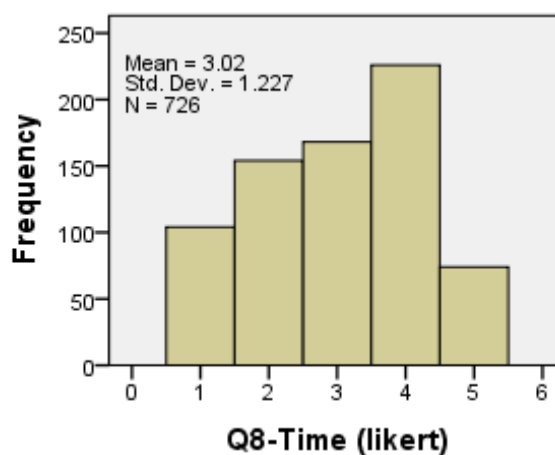
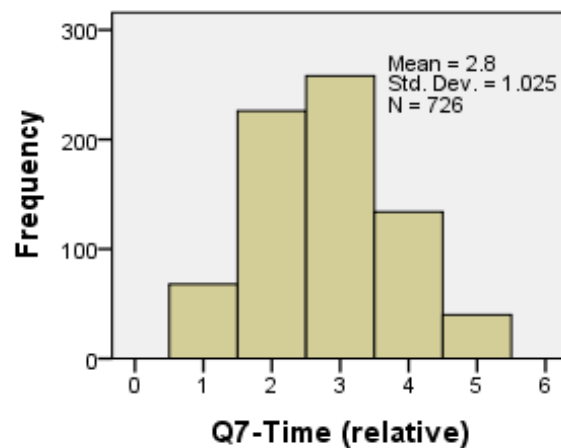
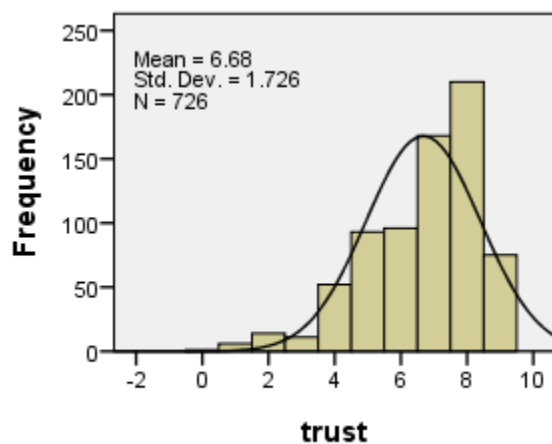
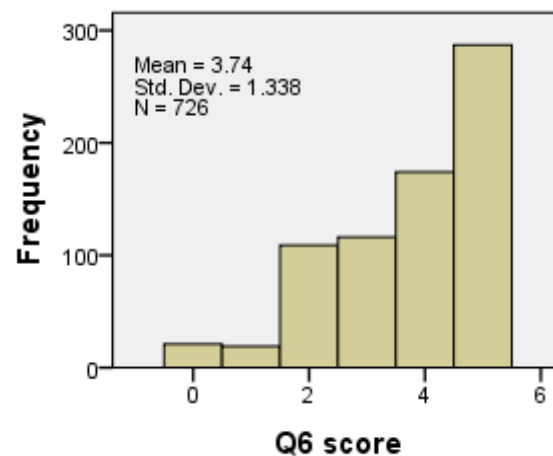
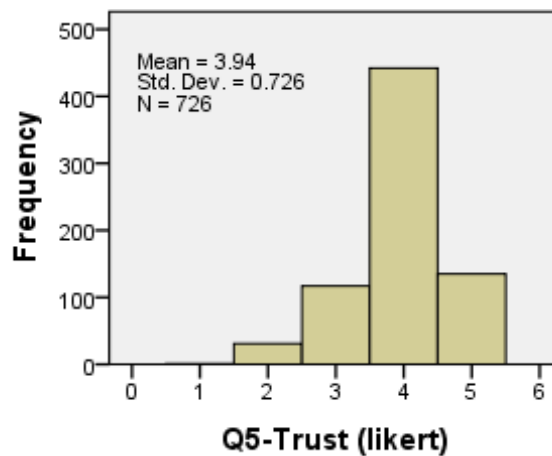
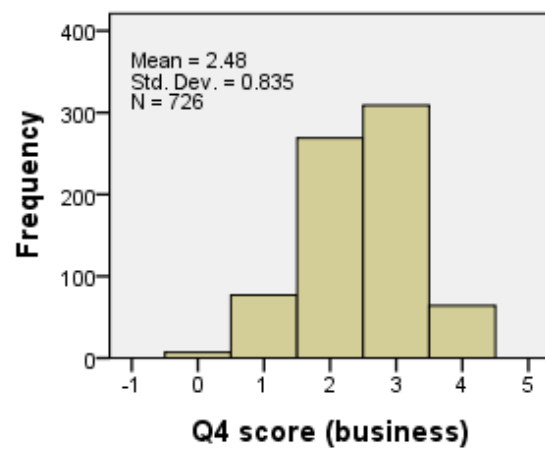
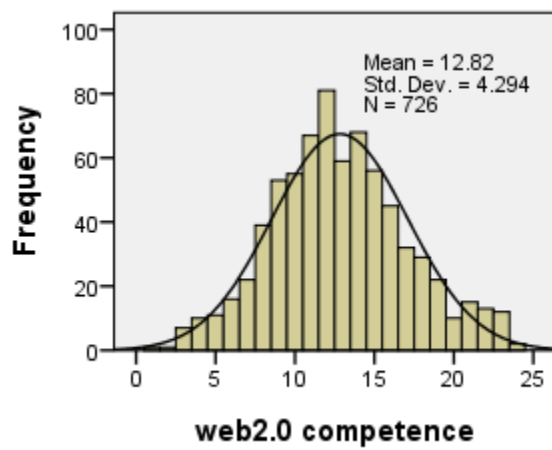
Q1 score



Q2 score



Q3 score



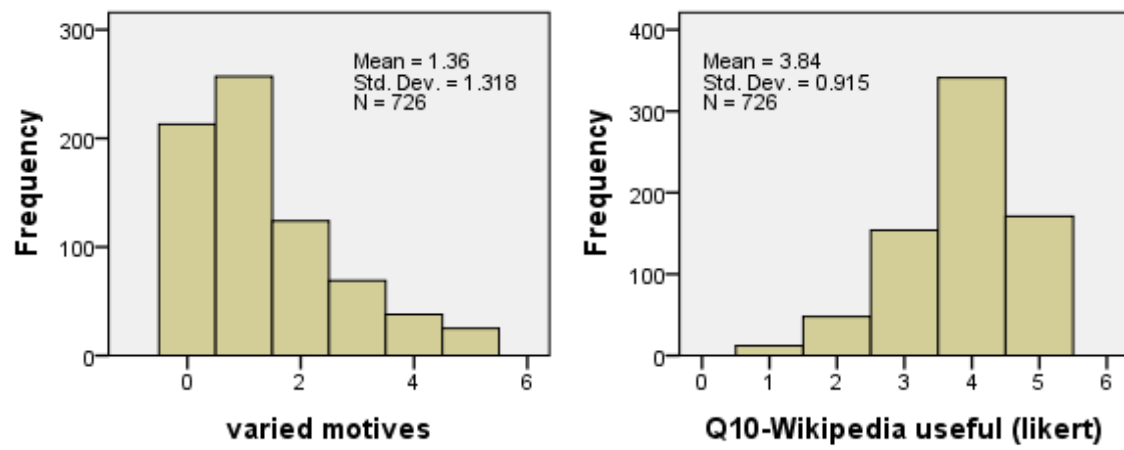


Figure B.1 – Distributions for all variables in the survey sample, $N = 726$

Correlation Matrix: Application Popularity Rankings per Age Group

Table B.5 – Correlation Matrix (Spearman's rho) of web 2.0 app rankings

		19≤_years	20-30 years	31-40 years	41-50 years	51-60 years	61 ≥ years
19≤_years	Correlation Coef.	1.000	.988**	.879**	.879**	.867**	.855**
	Sig. (2-tailed)	.	.000	.001	.001	.001	.002
	N	10	10	10	10	10	10
20-30 years	Correlation Coef.	.988**	1.000	.903**	.867**	.855**	.879**
	Sig. (2-tailed)	.000	.	.000	.001	.002	.001
	N	10	10	10	10	10	10
31-40 years	Correlation Coef.	.879**	.903**	1.000	.939**	.952**	.927**
	Sig. (2-tailed)	.001	.000	.	.000	.000	.000
	N	10	10	10	10	10	10
41-50 years	Correlation Coef.	.879**	.867**	.939**	1.000	.952**	.855**
	Sig. (2-tailed)	.001	.001	.000	.	.000	.002
	N	10	10	10	10	10	10
51-60 years	Correlation Coef.	.867**	.855**	.952**	.952**	1.000	.891**
	Sig. (2-tailed)	.001	.002	.000	.000	.	.001
	N	10	10	10	10	10	10
61 ≥ years	Correlation Coef.	.855**	.879**	.927**	.855**	.891**	1.000
	Sig. (2-tailed)	.002	.001	.000	.002	.001	.
	N	10	10	10	10	10	10

** . Correlation is significant at the 0.01 level (2-tailed).

Essentially the matrix shows that as rankings in one age group change they also tend to change in the same direction in another age group – i.e. all correlations are higher than $r = .855$, and all are also significant at p (two-tailed) $< .01$. Consider, for example, the first age group 19 years or less (row 1, table 1.7), and compare this with all the other age groups, we get correlation coefficients $r = .988, .879, .879, .867$ and $.855$ respectively. Even though all coefficients are very high, they do decrease with age, suggesting a slight tendency for application preferences to change at age extremes. Consider the 41-50 age-group (row 4, table 1.7), here the highest correlations are with 31-40 (.939) and 51-60 (.952) year olds, the rankings in these groups are similar in that they change at similar rate; however, moving away from this middle age group of users application preferences become somewhat different than within that group.

Popularity of web 2.0 applications, actions, and other

Table B.6 – Web 2.0 applications across different age groups

		Age group						Total
		19_less _years	20_30 _years	31_40 _years	41_50 _years	51_60 _years	61_more_ _years	
Q2-1twitter	Count	36	190	51	18	11	3	309
	% within Age grp.	40.4%	48.0%	41.5%	35.3%	20.0%	25.0%	
Q2-2youtube	Count	87	390	116	47	47	11	698
	% within Age grp.	97.8%	98.5%	94.3%	92.2%	85.5%	91.7%	
Q2-3facebook_my space	Count	86	380	105	31	39	4	645
	% within Age grp.	96.6%	96.0%	85.4%	60.8%	70.9%	33.3%	
Q2-4delicious	Count	6	28	11	1	4	0	50
	% within Age grp.	6.7%	7.1%	8.9%	2.0%	7.3%	.0%	
Q2-5flickr_picassa	Count	33	165	60	22	21	2	303
	% within Age grp.	37.1%	41.7%	48.8%	43.1%	38.2%	16.7%	
Q2-6wikipedia	Count	87	384	115	48	50	9	693
	% within Age grp.	97.8%	97.0%	93.5%	94.1%	90.9%	75.0%	
Q2-7digg_reddit	Count	11	61	12	3	1	0	88
	% within Age grp.	12.4%	15.4%	9.8%	5.9%	1.8%	.0%	
Q2-8craigslist	Count	3	43	17	1	3	3	70
	% within Age grp.	3.4%	10.9%	13.8%	2.0%	5.5%	25.0%	
Q2-9ebay	Count	73	319	100	46	37	8	583
	% within Age grp.	82.0%	80.6%	81.3%	90.2%	67.3%	66.7%	
Q2-10amazon	Count	81	358	117	49	51	11	667
	% within Age grp.	91.0%	90.4%	95.1%	96.1%	92.7%	91.7%	
Total	Count	89	396	123	51	55	12	726

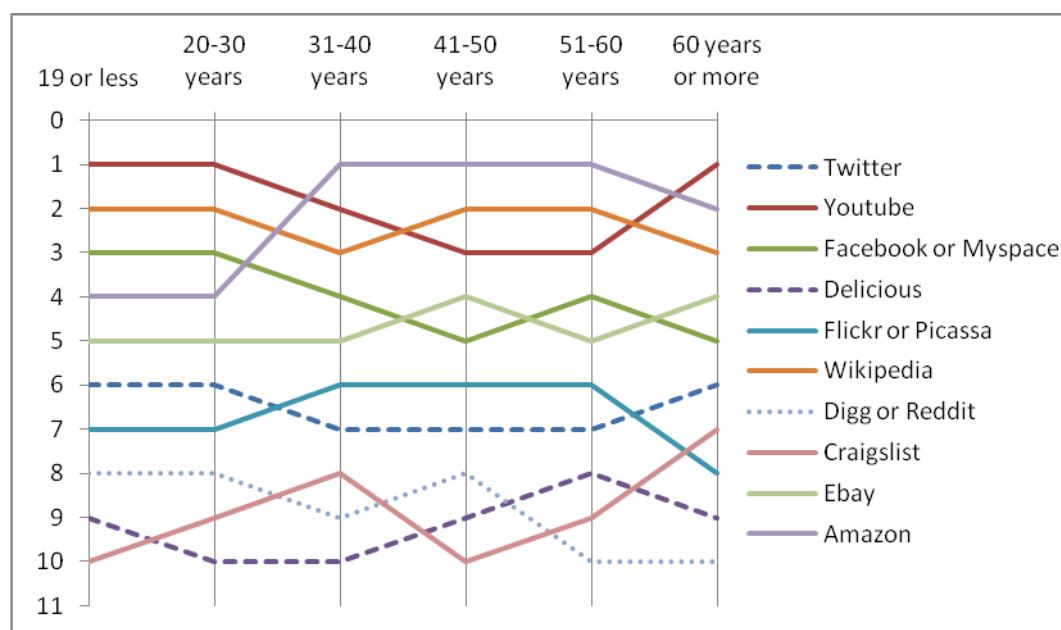


Figure B.2 – Rankings of web 2.0 applications over increasing age-group (1 [highest] to 10 [lowest])

Table B.7 – Web 2.0 applications across different qualification levels

		Qualification level					Total
		still_in_school	finished_school	undergraduate	Postgraduate	phd_dr	
Q2-1twitter	Count	2	8	144	102	53	309
	% within Qual. I.	16.7%	21.6%	47.1%	44.2%	37.9%	
Q2-2youtube	Count	11	30	300	224	133	698
	% within Qual. I.	91.7%	81.1%	98.0%	97.0%	95.0%	
Q2-3facebook_my space	Count	11	24	289	207	114	645
	% within Qual. I.	91.7%	64.9%	94.4%	89.6%	81.4%	
Q2-4delicious	Count	0	0	22	18	10	50
	% within Qual. I.	.0%	.0%	7.2%	7.8%	7.1%	
Q2-5flickr_picassa	Count	3	8	113	101	78	303
	% within Qual. I.	25.0%	21.6%	36.9%	43.7%	55.7%	
Q2-6wikipedia	Count	11	30	300	218	134	693
	% within Qual. I.	91.7%	81.1%	98.0%	94.4%	95.7%	
Q2-7digg_reddit	Count	2	1	47	29	9	88
	% within Qual. I.	16.7%	2.7%	15.4%	12.6%	6.4%	
Q2-8craigslist	Count	2	1	25	31	11	70
	% within Qual. I.	16.7%	2.7%	8.2%	13.4%	7.9%	
Q2-9ebay	Count	8	27	260	176	112	583
	% within Qual. I.	66.7%	73.0%	85.0%	76.2%	80.0%	
Q2-10amazon	Count	11	35	288	200	133	667
	% within Qual. I.	91.7%	94.6%	94.1%	86.6%	95.0%	
Total	Count	12	37	306	231	140	726

Table B.8 – Web 2.0 applications across different expertise areas

		Expertise area					Total
		theoretical	arts_humaniities	business_economics	engineering	computers	
Q2-1twitter	Count	20	73	36	49	131	309
	% within Epxr. ar.	32.3%	43.5%	34.0%	31.8%	55.5%	
Q2-2youtube	Count	59	161	102	145	231	698
	% within Epxr. ar.	95.2%	95.8%	96.2%	94.2%	97.9%	
Q2-3facebook_my space	Count	56	150	100	131	208	645
	% within Epxr. ar.	90.3%	89.3%	94.3%	85.1%	88.1%	
Q2-4delicious	Count	1	9	4	6	30	50
	% within Epxr. ar.	1.6%	5.4%	3.8%	3.9%	12.7%	
Q2-5flickr_picassa	Count	25	68	25	69	116	303
	% within Epxr. ar.	40.3%	40.5%	23.6%	44.8%	49.2%	
Q2-6wikipedia	Count	60	158	102	150	223	693
	% within Epxr. ar.	96.8%	94.0%	96.2%	97.4%	94.5%	
Q2-7digg_reddit	Count	0	7	7	22	52	88
	% within Epxr. ar.	.0%	4.2%	6.6%	14.3%	22.0%	
Q2-8craigslist	Count	3	17	15	10	25	70
	% within Epxr. ar.	4.8%	10.1%	14.2%	6.5%	10.6%	
Q2-9ebay	Count	54	128	77	127	197	583
	% within Epxr. ar.	87.1%	76.2%	72.6%	82.5%	83.5%	
Q2-10amazon	Count	61	151	93	144	218	667
	% within Epxr. ar.	98.4%	89.9%	87.7%	93.5%	92.4%	
Total	Count	62	168	106	154	236	726

Table B.9 – Web 2.0 activities across different age groups

		Age group						Total
		19_less _years	20_30 _years	31_40 _years	41_50 _years	51_60 _years	61_more_ years	
Q3-1bog_post	Count	30	151	42	13	10	1	247
	% within Age grp.	34.5%	38.2%	35.9%	31.0%	20.4%	12.5%	
Q3-2comment_on	Count	72	331	91	24	25	0	543
	% within Age grp.	82.8%	83.8%	77.8%	57.1%	51.0%	.0%	
Q3-3rated	Count	44	226	62	19	18	1	370
	% within Age grp.	50.6%	57.2%	53.0%	45.2%	36.7%	12.5%	
Q3-4uploaded_file	Count	83	371	102	36	36	6	634
	% within Age grp.	95.4%	93.9%	87.2%	85.7%	73.5%	75.0%	
Q3-5tagged_content	Count	81	345	75	17	19	1	538
	% within Age grp.	93.1%	87.3%	64.1%	40.5%	38.8%	12.5%	
Q3-6edited_shared_resource	Count	22	101	28	11	8	1	171
	% within Age grp.	25.3%	25.6%	23.9%	26.2%	16.3%	12.5%	
Q3-7joined_community	Count	75	342	89	31	38	5	580
	% within Age grp.	86.2%	86.6%	76.1%	73.8%	77.6%	62.5%	
Q3-8rss	Count	24	124	51	13	13	1	226
	% within Age grp.	27.6%	31.4%	43.6%	31.0%	26.5%	12.5%	
Q3-9openId_disqus	Count	8	43	10	1	4	0	66
	% within Age grp.	9.2%	10.9%	8.5%	2.4%	8.2%	.0%	
Q3-10api_mashup	Count	13	55	17	2	4	2	93
	% within Age grp.	14.9%	13.9%	14.5%	4.8%	8.2%	25.0%	
Total	Count	87	395	117	42	49	8	698

Table B.10 – Web 2.0 activities across different qualification levels

		Qualification level					Total
		still_in_sc hool	finished_s chool	undergrad uate	Postgra duate	phd_dr	
Q3-1bog_post	Count	4	9	109	77	48	247
	% within Qual. l.	36.4%	31.0%	36.3%	34.4%	35.8%	
Q3-2comment_on	Count	8	15	249	179	92	543
	% within Qual. l.	72.7%	51.7%	83.0%	79.9%	68.7%	
Q3-3rated	Count	4	13	161	132	60	370
	% within Qual. l.	36.4%	44.8%	53.7%	58.9%	44.8%	
Q3-4uploaded_file	Count	10	24	287	199	114	634
	% within Qual. l.	90.9%	82.8%	95.7%	88.8%	85.1%	
Q3-5tagged_content	Count	6	17	274	158	83	538
	% within Qual. l.	54.5%	58.6%	91.3%	70.5%	61.9%	
Q3-6edited_shared_resource	Count	1	4	77	61	28	171
	% within Qual. l.	9.1%	13.8%	25.7%	27.2%	20.9%	
Q3-7joined_community	Count	7	22	264	179	108	580
	% within Qual. l.	63.6%	75.9%	88.0%	79.9%	80.6%	
Q3-8rss	Count	4	8	84	78	52	226
	% within Qual. l.	36.4%	27.6%	28.0%	34.8%	38.8%	
Q3-9openId_disqus	Count	0	3	33	23	7	66
	% within Qual. l.	.0%	10.3%	11.0%	10.3%	5.2%	
Q3-10api_mashup	Count	0	3	43	27	20	93
	% within Qual. l.	.0%	10.3%	14.3%	12.1%	14.9%	
Total	Count	11	29	300	224	134	698

Table B.11 – Web 2.0 activities across different expertise areas

		Expertise area					Total
		theoretical	arts_hum anities	business_e conomics	engineer ing	compute rs	
Q3-1bog_post	Count	16	43	30	47	111	247
	% within Epxr. ar.	26.7%	26.4%	29.4%	32.2%	48.9%	
Q3-2comment_on	Count	47	129	79	106	182	543
	% within Epxr. ar.	78.3%	79.1%	77.5%	72.6%	80.2%	
Q3-3rated	Count	32	69	52	73	144	370
	% within Epxr. ar.	53.3%	42.3%	51.0%	50.0%	63.4%	
Q3-4uploaded_file	Count	52	149	91	132	210	634
	% within Epxr. ar.	86.7%	91.4%	89.2%	90.4%	92.5%	
Q3-5tagged_cont ent	Count	51	131	82	102	172	538
	% within Epxr. ar.	85.0%	80.4%	80.4%	69.9%	75.8%	
Q3-6edited_share d_resource	Count	11	31	20	34	75	171
	% within Epxr. ar.	18.3%	19.0%	19.6%	23.3%	33.0%	
Q3-7joined_comm unity	Count	49	136	84	111	200	580
	% within Epxr. ar.	81.7%	83.4%	82.4%	76.0%	88.1%	
Q3-8rss	Count	14	28	21	41	122	226
	% within Epxr. ar.	23.3%	17.2%	20.6%	28.1%	53.7%	
Q3-9openId_disqu s	Count	4	6	2	12	42	66
	% within Epxr. ar.	6.7%	3.7%	2.0%	8.2%	18.5%	
Q3-10api_mashup	Count	5	4	6	12	66	93
	% within Epxr. ar.	8.3%	2.5%	5.9%	8.2%	29.1%	
Total	Count	60	163	102	146	227	698

Table B.12 – Business related activities across different age groups

		Age group						Total
		19_less _years	20_30 _years	31_40 _years	41_50 _years	51_60 _years	61_more _years	
Q4-1online_advert	Count	51	225	74	25	28	3	406
	% within Age grp.	58.0%	57.4%	60.2%	50.0%	51.9%	25.0%	
Q4-2bough_online	Count	86	376	121	48	53	12	696
	% within Age grp.	97.7%	95.9%	98.4%	96.0%	98.1%	100.0%	
Q4-3paypal_webmoney	Count	74	326	101	43	39	9	592
	% within Age grp.	84.1%	83.2%	82.1%	86.0%	72.2%	75.0%	
Q4-4group_buying	Count	8	64	24	4	4	0	104
	% within Age grp.	9.1%	16.3%	19.5%	8.0%	7.4%	.0%	
Total	Count	88	392	123	50	54	12	719

Percentages and totals are based on respondents.

Table B.13 – Business related activities across different qualification levels

		Qualification level					Total
		still_in_school	finished_school	undergraduate	postgraduate	phd_dr	
Q4-1online_advert	Count	8	18	172	140	68	406
	% within Qual. I.	66.7%	48.6%	56.4%	61.7%	49.3%	
Q4-2bough_online	Count	11	37	297	218	133	696
	% within Qual. I.	91.7%	100.0%	97.4%	96.0%	96.4%	
Q4-3paypal_webmoney	Count	10	30	266	172	114	592
	% within Qual. I.	83.3%	81.1%	87.2%	75.8%	82.6%	
Q4-4group_buying	Count	1	6	38	41	18	104
	% within Qual. I.	8.3%	16.2%	12.5%	18.1%	13.0%	
Total	Count	12	37	305	227	138	719

Percentages and totals are based on respondents.

Table B.14 – Business related activities across different expertise areas

		Expertise area					Total
		theoretical	arts_humanities	business_economics	engineering	computers	
Q4-1online_advert	Count	37	81	60	83	145	406
	% within Epxr. ar.	59.7%	48.8%	56.6%	54.2%	62.5%	
Q4-2bough_online	Count	61	161	104	147	223	696
	% within Epxr. ar.	98.4%	97.0%	98.1%	96.1%	96.1%	
Q4-3paypal_webmoney	Count	54	135	78	131	194	592
	% within Epxr. ar.	87.1%	81.3%	73.6%	85.6%	83.6%	
Q4-4group_buying	Count	11	19	13	17	44	104
	% within Epxr. ar.	17.7%	11.4%	12.3%	11.1%	19.0%	
Total	Count	62	166	106	153	232	719

Percentages and totals are based on respondents.

Table B.15 – Web 2.0 applications across increasing trust awareness levels

		Trust score based trust-level			Total
		Low trust	Normal trust	High trust	
Q2-1twitter	Count	32	130	147	309
	% within trust. l.	38.1%	36.4%	51.6%	
Q2-2youtube	Count	80	339	279	698
	% within trust. l.	95.2%	95.0%	97.9%	
Q2-3facebook_my space	Count	65	309	271	645
	% within trust. l.	77.4%	86.6%	95.1%	
Q2-4delicious	Count	7	17	26	50
	% within trust. l.	8.3%	4.8%	9.1%	
Q2-5flickr_picassa	Count	30	139	134	303
	% within trust. l.	35.7%	38.9%	47.0%	
Q2-6wikipedia	Count	75	340	278	693
	% within trust. l.	89.3%	95.2%	97.5%	
Q2-7digg_reddit	Count	8	24	56	88
	% within trust. l.	9.5%	6.7%	19.6%	
Q2-8craigslist	Count	4	32	34	70
	% within trust. l.	4.8%	9.0%	11.9%	
Q2-9ebay	Count	50	286	247	583
	% within trust. l.	59.5%	80.1%	86.7%	
Q2-10amazon	Count	66	325	276	667
	% within trust. l.	78.6%	91.0%	96.8%	
Total	Count	84	357	285	726

Table B.16 – Web 2.0 activities across increasing trust awareness levels

		Trust score based trust-level			Total
		Low trust	Normal trust	High trust	
Q3-1bog_post	Count	21	110	116	247
	% within trust. l.	28%	32.3%	41.1%	
Q3-2comment_on	Count	50	254	239	543
	% within trust. l.	66.7%	74.5%	84.8%	
Q3-3rated	Count	34	164	172	370
	% within trust. l.	45.3%	48.1%	61.0%	
Q3-4uploaded_file	Count	64	303	267	634
	% within trust. l.	85.3%	88.9%	94.7%	
Q3-5tagged_content	Count	46	248	244	538
	% within trust. l.	61.3%	72.7%	86.5%	
Q3-6edited_share d_resource	Count	19	69	83	171
	% within trust. l.	25.3%	20.2%	29.4%	
Q3-7joined_community	Count	51	266	263	580
	% within trust. l.	68.0%	78.0%	93.3%	
Q3-8rss	Count	22	95	109	226
	% within trust. l.	29.3%	27.9%	38.7%	
Q3-9openId_disqus	Count	8	17	41	66
	% within trust. l.	10.7%	5.0%	14.5%	
Q3-10api_mashup	Count	6	40	47	93
	% within trust. l.	8.0%	11.7%	16.7%	
Total	Count	75	341	282	698

Table B.17 – Web 2.0 applications across increasing time score levels

		Time score based time-spent			Total
		Little time	Normal time	Much time	
Q2-1twitter	Count	55	160	94	309
	% within time. s.	26.6%	42.4%	66.2%	
Q2-2youtube	Count	190	366	142	698
	% within time. s.	91.8%	97.1%	100.0%	
Q2-3facebook_my space	Count	150	356	139	645
	% within time. s.	72.5%	94.4%	97.9%	
Q2-4delicious	Count	12	22	16	50
	% within time. s.	5.8%	5.8%	11.3%	
Q2-5flickr_picassa	Count	75	158	70	303
	% within time. s.	36.2%	41.9%	49.3%	
Q2-6wikipedia	Count	191	366	136	693
	% within time. s.	92.3%	97.1%	95.8%	
Q2-7digg_reddit	Count	15	40	33	88
	% within time. s.	7.2%	10.6%	23.2%	
Q2-8craigslist	Count	15	35	20	70
	% within time. s.	7.2%	9.3%	14.1%	
Q2-9ebay	Count	160	302	121	583
	% within time. s.	77.3%	80.1%	85.2%	
Q2-10amazon	Count	189	347	131	667
	% within time. s.	91.3%	92.0%	92.3%	
Total	Count	207	377	142	726

Table B.18 – Web 2.0 activities across increasing time score levels

		Time score based time-spent			Total
		Little time	Normal time	Much time	
Q3-1	blog_post Count	42	142	63	247
	% within time. s.	23.0%	38.1%	44.4%	
Q3-2	comment_on Count	104	313	126	543
	% within time. s.	56.8%	83.9%	88.7%	
Q3-3	rated Count	80	199	91	370
	% within time. s.	43.7%	53.4%	64.1%	
Q3-4	uploaded_file Count	150	347	137	634
	% within time. s.	82.0%	93.0%	96.5%	
Q3-5	tagged_content Count	90	313	135	538
	% within time. s.	49.2%	83.9%	95.1%	
Q3-6	edited_shared_resource Count	35	93	43	171
	% within time. s.	19.1%	24.9%	30.3%	
Q3-7	joined_community Count	132	320	128	580
	% within time. s.	72.1%	85.8%	90.1%	
Q3-8	rss Count	55	124	47	226
	% within time. s.	30.1%	33.2%	33.1%	
Q3-9	openid_disqus Count	11	31	24	66
	% within time. s.	6.0%	8.3%	16.9%	
Q3-10	api_mashup Count	18	54	21	93
	% within time. s.	9.8%	14.5%	14.8%	
Total	Count	183	373	142	698

Chi-Square test tables used in Activity vs. Trust levels

Tables B.19 – test tables, web 2.0 activities vs. trust levels

Blog post - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-1	BlogPost No	63	247	169	479
	Yes	21	110	116	247
Total		84	367	285	726

X-squared = 10.3495, df = 2, p-value = 0.005658

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	1.0179932	0.7466223	-1.3882925
Yes	-1.4176341	-1.0397292	1.9333044

Comment on - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-2	Comment on No	34	103	46	183
	Yes	50	254	239	543
Total		84	367	285	726

X-squared = 25.3301, df = 2, p-value = 3.16e-06

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	2.7874653	1.3717215	-3.0485505
Yes	-1.6182108	-0.7963273	1.7697790

Rated - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-3Rated	No	50	193	113	356
	Yes	34	164	172	370
Total		84	367	285	726

X-squared = 17.3539, df = 2, p-value = 0.0001705

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	1.372700	1.356075	-2.262967
Yes	-1.346479	-1.330172	2.219741

Uploaded file- Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-4Uploaded	No	20	54	18	92
File	Yes	64	303	267	634
Total		84	367	285	726

X-squared = 21.7634, df = 2, p-value = 1.88e-05

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	2.8674497	1.3024492	-3.0144435
Yes	-1.0923080	-0.4961467	1.1483028

Tagged content - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-5Tagged	No	38	109	41	188
content	Yes	46	248	244	538
Total		84	367	285	726

X-squared = 40.051, df = 2, p-value = 2.009e-09

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	3.483757	1.721674	-3.818235
Yes	-2.059376	-1.017744	2.257098

Edited shared resource - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-6Edited	No	65	288	202	555
shared resource	Yes	19	69	83	171
Total		84	367	285	726

X-squared = 8.4907, df = 2, p-value = 0.01433

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.09797616	0.91323843	-1.07529674
Yes	-0.17650981	-1.64525271	1.93721029

Joined community - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-7Joined community	No	33	91	22	146
	Yes	51	266	263	580
Total		84	367	285	726

X-squared = 52.8926, df = 2, p-value = 3.27e-12

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	3.919031	2.266776	-4.664628
Yes	-1.966261	-1.137289	2.340342

RSS - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-8RSS	No	62	262	176	500
	Yes	22	95	109	226
Total		84	367	285	726

X-squared = 11.0878, df = 2, p-value = 0.003911

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.5454589	1.0288301	-1.4476050
Yes	-0.8113211	-1.5302924	2.1531825

OpenID / DISCUSS - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-9OpenID / Discuss	No	76	340	244	660
	Yes	8	17	41	66
Total		84	367	285	726

X-squared = 17.7831, df = 2, p-value = 0.0001375

Fisher's exact test is also significant at p-value = 9.859e-05 (two-tailed)

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	-0.04161252	0.85786405	-0.93753894
Yes	0.13159034	-2.71280433	2.96475845

API / Mashup - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q3-10API / Mashup	No	78	317	238	633
	Yes	6	40	47	93
Total		84	367	285	726

X-squared = 6.6973, df = 2, p-value = 0.03513

Fisher's exact test is also significant at p-value = 0.03693 (two-tailed)

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.5562422	0.3248579	-0.6655664
Yes	-1.4511901	-0.8475275	1.7364079

Chi-Square test tables used in Application vs. Trust levels

Tables B.20 – test tables, web 2.0 applications vs. trust levels

Twitter - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-1Twitter	No	52	227	138	417
	Yes	32	130	147	309
Total		84	367	285	726

X-squared = 15.6828, df = 2, p-value = 0.0003931

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.5401708	1.5325940	-2.0085518
Yes	-0.6275089	-1.7803931	2.3333067

Youtube - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-2Youtube	No	4	18	6	28
	Yes	80	339	279	698
Total		84	367	285	726

X-squared = 3.896, df = 2, p-value = 0.1426

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.42242743	1.14035423	-1.50563060
Yes	-0.08460644	-0.22839736	0.30155722

Chi-squared approximation may be incorrect since the counts are too small (i.e. top-left corner cell)
Fisher's exact test confirms the insignificance, at p-value = 0.1239 (two-sided)

Facebook / MySpace – Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-3Facebook	No	19	48	14	81
	Yes	65	309	271	645
Total		84	367	285	726

X-squared = 24.232, df = 2, p-value = 5.471e-06

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	3.1450439	1.2944432	-3.1561881
Yes	-1.1145238	-0.4587179	1.1184731

Delicious - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-4Delicious	No	77	340	259	676
	Yes	7	17	26	50
Total		84	367	285	726

X-squared = 5.0097, df = 2, p-value = 0.08169

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	-0.1373685	0.4161196	-0.3911484
Yes	0.5050980	-1.5300531	1.4382351

Flickr / Picassa - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-5Flickr	No	54	218	151	423
Picassa	Yes	30	139	134	303
Total		84	367	285	726

X-squared = 5.6733, df = 2, p-value = 0.05862

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.7229771	0.6930818	-1.1682062
Yes	-0.8542272	-0.8189048	1.3802837

Wikipedia - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-6Wikipedia	No	9	17	7	33
	Yes	75	340	278	693
Total		84	367	285	726

X-squared = 10.2732, df = 2, p-value = 0.005878

Fisher exact test is also significant at p-value = 0.00961 (two-tailed)

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	2.65188000	0.19182423	-1.65438859
Yes	-0.57868766	-0.04185948	0.36101719

Digg / Reddit - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-7Digg /	No	76	333	229	638
Reddit	Yes	8	24	56	88
Total		84	367	285	726

X-squared = 25.4619, df = 2, p-value = 2.958e-06

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.2539434	1.0880956	-1.3556727
Yes	-0.6837635	-2.9297870	3.6502605

Craigslist - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-8Craigslist	No	80	325	251	656
	Yes	4	32	34	70
Total		84	367	285	726

X-squared = 4.197, df = 2, p-value = 0.1226

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	0.4705145	0.1348232	-0.4063363
Yes	-1.4403763	-0.4127313	1.2439087

Ebay - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-9Ebay	No	34	71	38	143
	Yes	50	286	247	583
Total		84	367	285	726

X-squared = 30.235, df = 2, p-value = 2.72e-07

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	4.29110550	0.08130828	-2.42062813
Yes	-2.12521530	-0.04026878	1.19884164

Amazon - Trust

		Q6-Trust			Total
		Low trust	Normal trust	High trust	
Q2-10Amazon	No	18	32	9	59
	Yes	66	325	276	667
Total		84	367	285	726

X-squared = 29.6658, df = 2, p-value = 3.615e-07

Fisher's exact test is also significant at p-value = 1.2e-06 (two-tailed)

Residuals (i.e. con):

	Low trust	Normal trust	High trust
No	4.2765542	0.5546655	-2.9425145
Yes	-1.2719120	-0.1649659	0.8751484

Chi-Square test tables used in Activity vs. Time levels

Tables B.21 – test tables, web 2.0 activities vs. time levels

Blog post - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-1BlogPost	No	165	235	79	479
	Yes	42	142	63	247
Total		183	373	142	726

X-squared = 26.3884, df = 2, p-value = 1.861e-06

Residuals (i.e. con):

	Little time	Normal time	Much time
No	2.4323458	-0.8710018	-1.5175391
Yes	-3.3872294	1.2129373	2.1132904

Comment on - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-2Comment on	No	103	64	16	183
	Yes	104	313	126	543
Total		183	373	142	726

X-squared = 94.3657, df = 2, p-value = 2.2e-16

Residuals (i.e. con):

	Little time	Normal time	Much time
No	7.035776	-3.183016	-3.308406
Yes	-4.084488	1.847840	1.920633

Rated - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-3Rated	No	127	178	51	356
	Yes	80	199	91	370
Total		183	373	142	726

X-squared = 22.8474, df = 2, p-value = 1.093e-05

Residuals (i.e. con):

	Little time	Normal time	Much time
No	2.5306258	-0.5049101	-2.2327079
Yes	-2.4822875	0.4952656	2.1900601

Uploaded file- Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-4Uploaded	No	57	30	5	92
File	Yes	150	347	137	634
Total		183	373	142	726

X-squared = 59.6455, df = 2, p-value = 1.117e-13

Fisher's exact test is also significant at p-value = 1.022e-12 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	6.0075430	-2.5715292	-3.0632996
Yes	-2.2884750	0.9795819	1.1669138

Tagged content - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-5Tagged	No	117	64	7	188
content	Yes	90	313	135	538
Total		183	373	142	726

X-squared = 149.3359, df = 2, p-value = 2.2e-16

Fisher's exact test is also significant at p-value = 2.2e-16 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	8.659062	-3.403184	-4.909573
Yes	-5.118687	2.011746	2.902228

Edited shared resource - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-6Edited	No	172	284	99	555
shared resource	Yes	35	93	43	171
Total		183	373	142	726

X-squared = 8.907, df = 2, p-value = 0.01164

Residuals (i.e. con):

	Little time	Normal time	Much time
No	1.0935410	-0.2475465	-0.9169594
Yes	-1.9700784	0.4459695	1.6519563

Joined community - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-7Joined community	No	75	57	14	146
	Yes	132	320	128	580
Total		183	373	142	726

X-squared = 48.6204, df = 2, p-value = 2.768e-11

Residuals (i.e. con):

	Little time	Normal time	Much time
No	5.172347	-2.160903	-2.723980
Yes	-2.595076	1.084171	1.366678

RSS - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-8RSS	No	152	253	95	500
	Yes	55	124	47	226
Total		183	373	142	726

X-squared = 2.8098, df = 2, p-value = 0.2454

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.7904581	-0.4121955	-0.2827474
Yes	-1.1757356	0.6131038	0.4205613

OpenID / DISCUSS - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-9OpenID / Discuss	No	196	346	118	660
	Yes	11	31	24	66
Total		183	373	142	726

X-squared = 14.3984, df = 2, p-value = 0.0007472

Fisher's exact test is also significant at p-value = 0.001407 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.5699240	0.1767810	-0.9761562
Yes	-1.8022579	-0.5590305	3.0868769

API / Mashup - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q3-10API / Mashup	No	189	323	121	633
	Yes	18	54	21	93
Total		183	373	142	726

X-squared = 4.4084, df = 2, p-value = 0.1103

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.6339338	-0.3147561	-0.2525317
Yes	-1.6538811	0.8211726	0.6588343

Chi-Square test tables used in Application vs. Time levels

Tables B.22 – test tables, web 2.0 applications vs. time levels

Twitter - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-1Twitter	No	152	217	48	417
	Yes	55	160	94	309
Total		207	377	142	726

X-squared = 54.1048, df = 2, p-value = 1.784e-12

Residuals (i.e. con):

	Little time	Normal time	Much time
No	3.03589311	0.03117003	-3.71623980
Yes	-3.52675478	-0.03620979	4.31710407

Youtube - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-2Youtube	No	17	11	0	28
	Yes	190	366	142	698
Total		207	377	142	726

X-squared = 17.1845, df = 2, p-value = 0.0001855

Fisher's exact test is also significant at p-value = 0.0001172 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	3.1911227	-0.9283572	-2.3402103
Yes	-0.6391382	0.1859373	0.4687121

Facebook / MySpace – Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-3Facebook	No	57	21	3	81
	Yes	150	356	139	645
Total		207	377	142	726

X-squared = 24.232, df = 2, p-value = 5.471e-06

Fisher's exact test is also significant at p-value = 2.2e-16 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	7.055111	-3.247539	-3.226616
Yes	-2.500152	1.150846	1.143431

Delicious - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-4Delicious	No	195	355	126	676
	Yes	12	22	16	50
Total		207	377	142	726

X-squared = 5.2827, df = 2, p-value = 0.07127

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.1625126	0.2115819	-0.5409637
Yes	-0.5975516	-0.7779771	1.9890992

Flickr / Picassa - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-5Flickr	No	132	219	72	423
Picassa	Yes	75	158	70	303
Total		207	377	142	726

X-squared = 5.9209, df = 2, p-value = 0.0518

Residuals (i.e. con):

	Little time	Normal time	Much time
No	1.03737160	-0.04433117	-1.18026013
Yes	-1.22569728	0.05237911	1.39452596

Wikipedia - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-6Wikipedia	No	16	11	6	33
	Yes	191	366	136	693
Total		207	377	142	726

X-squared = 7.1722, df = 2, p-value = 0.02771

Fisher's exact test is also significant at p-value = 0.0338 (two-tailed)

Residuals (i.e. con):

	Little time	Normal time	Much time
No	2.14867875	-1.48235340	-0.17891407
Yes	-0.46888014	0.32347603	0.03904225

Digg / Reddit - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-7Digg /	No	192	337	109	638
Reddit	Yes	15	40	33	88
Total		207	377	142	726

X-squared = 21.9052, df = 2, p-value = 1.751e-05

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.7481748	0.3129905	-1.4133105
Yes	-2.0145224	-0.8427528	3.8054550

Craigslist - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-8Craigslist	No	192	342	122	656
	Yes	15	35	20	70
Total		207	377	142	726

X-squared = 4.6358, df = 2, p-value = 0.09848

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.36257437	0.07313667	-0.55693059
Yes	-1.10994126	-0.22389174	1.70491988

Ebay - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-9Ebay	No	47	75	21	143
	Yes	160	302	121	583
Total		207	377	142	726

X-squared = 3.3564, df = 2, p-value = 0.1867

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.97524337	0.08615525	-1.31786224
Yes	-0.48299957	-0.04266930	0.65268519

Amazon - Trust

		Q7-Time			Total
		Little time	Normal time	Much time	
Q2-10Amazon	No	18	30	11	59
	Yes	189	347	131	667
Total		207	377	142	726

X-squared = 0.1317, df = 2, p-value = 0.9363

Residuals (i.e. con):

	Little time	Normal time	Much time
No	0.28713533	-0.11521685	-0.15894526
Yes	-0.08539840	0.03426724	0.04727273

2-Way tables used in Activity vs. Motivation Chi-Square tests

Tables B.23 – 2-way tables, web 2.0 activities vs. motivation exists

		Motivation exists		Total
		No	Yes	
Q3-1bog post	No	153	298	451
	Yes	36	211	247
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-2comment on	No	71	84	155
	Yes	118	425	543
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-3rated	No	122	206	328
	Yes	67	303	370
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-4uploaded file	No	27	37	64
	Yes	162	472	634
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-5tagged content	No	61	99	160
	Yes	128	410	538
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-6edited shared resource	No	163	364	527
	Yes	26	145	171
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-7joined community	No	51	67	118
	Yes	138	442	580
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-8rss	No	147	325	472
	Yes	42	184	226
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-9openId / disqus	No	185	447	632
	Yes	4	62	66
Total		189	509	698

		Motivation exists		Total
		No	Yes	
Q3-10api / mashup	No	180	425	605
	Yes	9	84	93
Total		189	509	698

2-Way tables used in Application vs. Motivation Chi-Square tests

Tables B.24 – 2-way tables, web 2.0 applications vs. motivation exists

		Motivation exists		Total
		No	Yes	
Q2-1twitter	No	156	261	417
	Yes	57	252	309
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-2youtube	No	12	16	28
	Yes	201	497	698
Total		213	513	726

			Motivation exists		Total
			No	Yes	
Q2-3facebook / myspace	No		46	35	81
	Yes		167	478	645
Total			213	513	726

		Motivation exists		Total
		No	Yes	
Q2-4delicious	No	203	473	676
	Yes	10	40	50
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-5flickr / picassa	No	151	272	423
	Yes	62	241	303
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-6wikipedia	No	15	18	33
	Yes	198	495	693
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-7digg / reddit	No	203	435	638
	Yes	10	78	88
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-8craigslist	No	198	458	656
	Yes	15	55	70
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-9ebay	No	48	95	143
	Yes	165	418	583
Total		213	513	726

		Motivation exists		Total
		No	Yes	
Q2-10amazon	No	23	36	59
	Yes	190	477	667
Total		213	513	726

2-Way table used in self-presentation motive vs. Twitter Chi-Square test

Table B.25 – a 2-way table for self-presentation motive vs. twitter

		Q9-Motivation self presentation		Total
		No	Yes	
Q2-1twitter	No	197	64	261
	Yes	136	116	252
Total		333	180	513

Survey - Print Copy

Web 2.0 Research (PhD) Survey

Thank you for taking out time to complete this short (1-2 minute) survey and helping me in the process with my PhD research.
(Read the numbered questions carefully, and answer as indicated!)

Background Information

1. Check each statement that you agree with:

- ☐ I have heard of the term "Web 2.0". [if you are not sure, leave the checkbox empty]
- ☐ I have heard of the term "Social Media".
- ☐ I think I have a rough idea of what the term "Web 2.0" means.
- ☐ I think I have a rough idea of what the term "Social Media" means.

2. Check the applications that you use or have used in the past:

- ☐ Twitter.com
- ☐ Youtube.com
- ☐ Facebook.com or Myspace.com
- ☐ Delicious.com
- ☐ Flickr.com or Picassa.com
- ☐ Wikipedia.com
- ☐ Digg.com or Reddit.com
- ☐ Craigslist.com
- ☐ Ebay.com
- ☐ Amazon.com

3. Check all the activities that you have done on the web at some point:

- ☐ Submitted a blog post (i.e. I have/had a blog)
- ☐ Commented on a book, blog post, picture, status...
- ☐ Rated a book, blog, picture, etc...
- ☐ Uploaded a picture, video or audio file...
- ☐ Tagged some content (link, picture, video, liked a status...)
- ☐ Edited a shared resource, e.g. a Wikipedia article, OpenStreetMap,...
- ☐ Joined a specific online community, e.g. a facebook group, groups.yahoo.com,...
- ☐ Used RSS or an RSS reader / aggregator
- ☐ Used a cross-website system such as OpenID or DISQUS
- ☐ Knowingly used an API (Application Programming Interface) or built a 'Mashup'

4. Have you ever:

- ☐ Clicked / followed an online advert on purpose
- ☐ Bought something online (Trip/Hotel-booking, Book, Movie, Music, Clothes, Membership,...)
- ☐ Used PayPal, Webmoney or some other web-based payment system
- ☐ Used a group buying website (such as Groupon.com, BuyWithMe.com, Twango.com,...)

Trust

5. I trust most websites that I use on a regular basis.

Strongly Agree ☐, Agree ☐, Neutral ☐, Disagree ☐, Strongly Disagree ☐

6. Check all the opinions that you agree with:

- ☐ I feel comfortable sharing my personal details (email, pictures, opinions,...) on web-pages.
- ☐ I feel comfortable sharing my personal details on web-pages that use all the appropriate security precautions and measures.
- ☐ I feel comfortable purchasing products online.
- ☐ I feel comfortable purchasing from online stores that I know.
- ☐ Looking at a web-page I can usually judge whether it is a trustworthy page or not.

Time

7. Compared to your friends, in your free time how much time per day do you spent on sites such as Facebook, Digg, Youtube, LinkedIn, Amazon, Ebay, Craigslist, Twitter, Myspace, Reddit, Delicious, MySpace, Flickr or Picassa:

- ☐ Probably **much** more than **most** of my friends
- ☐ Probably more than **some** of my friends
- ☐ Same as majority of my friends
- ☐ Less than most of my friends
- ☐ No time at all, or nearly no time

8. I spent too much time on Twitter, Facebook, Youtube, Wikipedia,...

Strongly Agree ☐, Agree ☐, Neutral ☐, Disagree ☐, Strongly Disagree ☐

Motivation

9. Would you contribute to any of the websites mentioned because (tick all that apply best to you):

- ☐ I want to contribute content for the greater common good
- ☐ I want to contribute content for greater good but I expect similar action in return
- ☐ I want to contribute to my community and to help raise awareness within it from my actions
- ☐ I want to build my online profile (i.e. web reputation)
- ☐ I want to show my experience and autonomy / knowledge in a certain topic

Final Question

10. I consider wikipedia.com to be a useful body of encyclopaedic reference:

Strongly Agree ☐, Agree ☐, Neutral ☐, Disagree ☐, Strongly Disagree ☐

Optional (Personal Details)

Please select what best describes you:

Age: 19 years or less | Qualification Level: Still in school | Expertise: Computer & Information Science

Finished

Thank you for your time. Your completed survey will make a contribution to my research project.

Thank you for taking my research-survey! *[Please do not take it again, as only one response from an individual is valid for this study.]*

You can leave me your email if you are interested into the outcomes of this survey study. *[These may not be available for several months.]*

Email:

[Home](#) | [Contact](#)

Thanks for your time on completing my Survey | [Valid XHTML](#) | [Valid CSS](#)

Figure B.3 – Survey - Print Copy; all survey questions

APPENDIX C: *Historical Website Study*

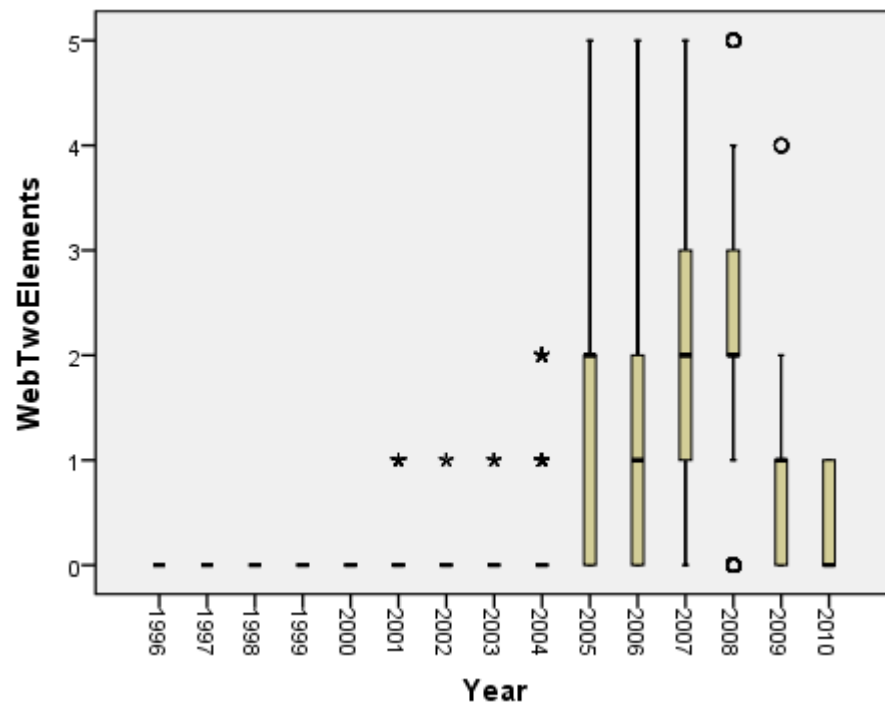


Figure C.1 – Box-plots of web 2.0 elements (measured by keyword count) over the sampling period for all websites

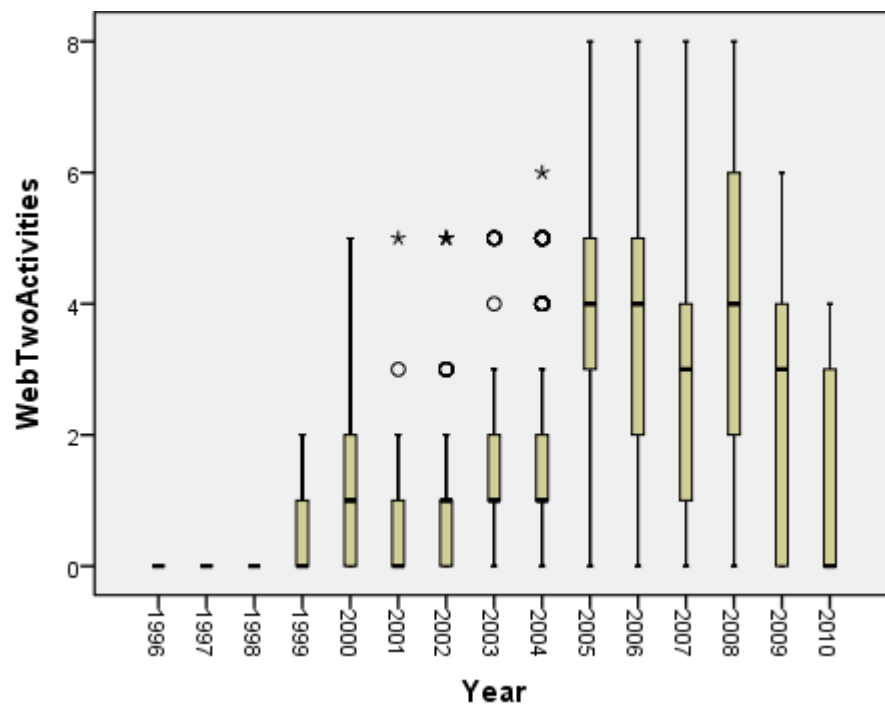


Figure C.2 – Box-plots of web 2.0 activities (measured by keyword count) over the sampling period for all websites

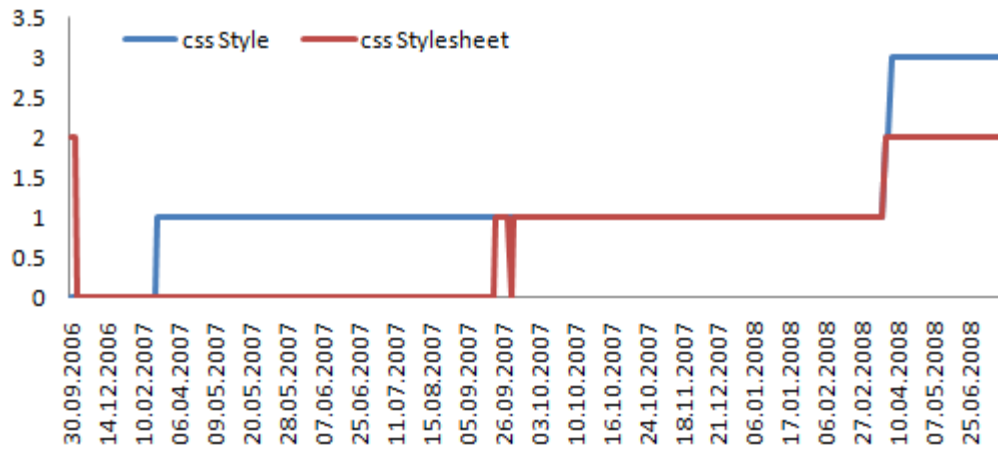


Figure C.3 – Page level CSS and CSS Stylesheet elements, over time, for Twitter

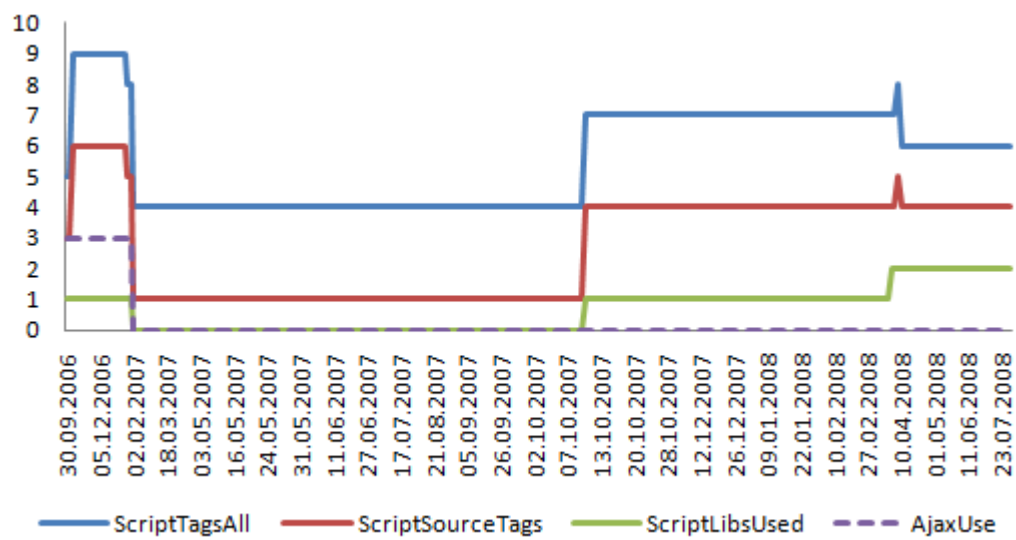


Figure C.4 – Script related elements, over time, for Twitter

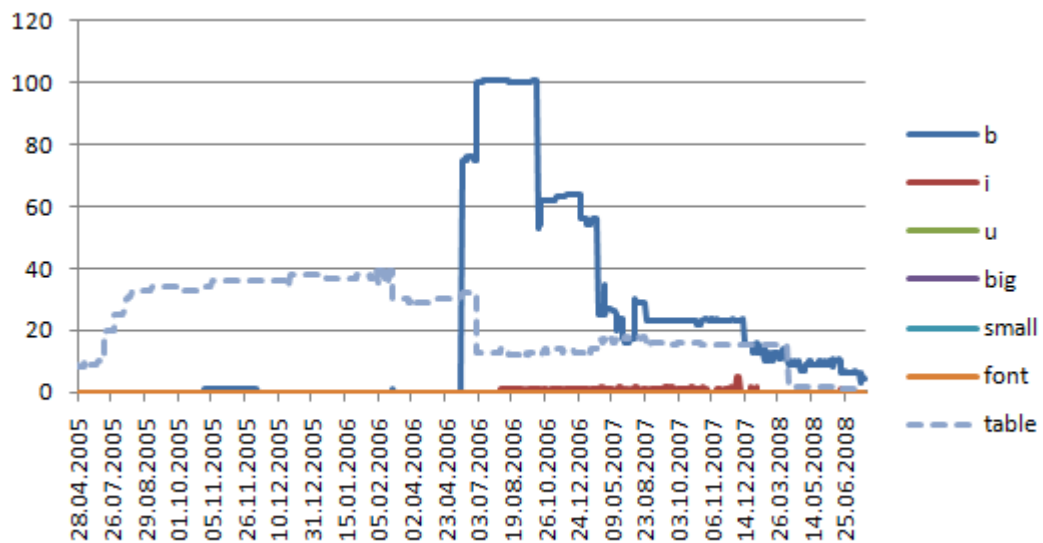


Figure C.5 – old / depreciated elements, over time, Youtube

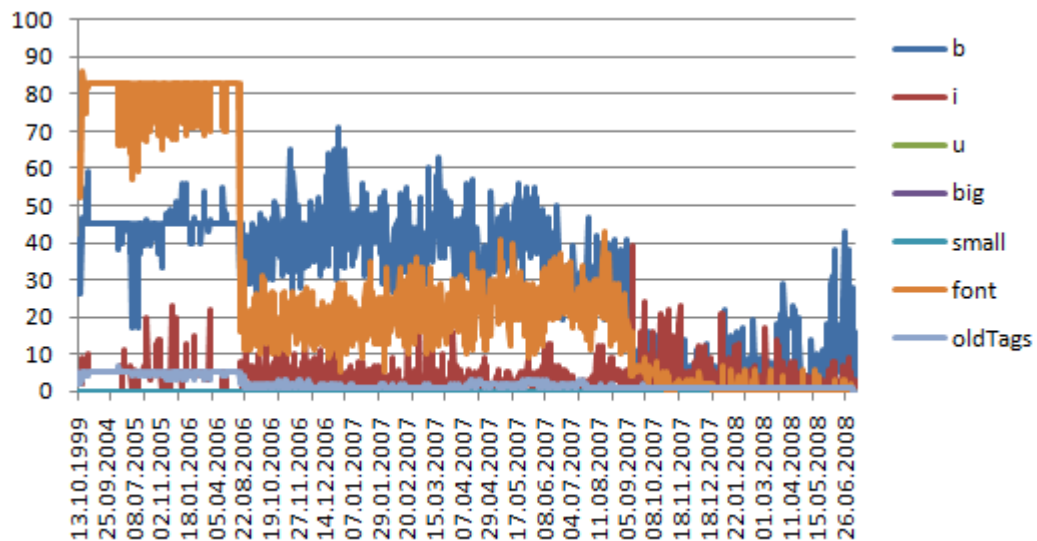


Figure C.6 – old / depreciated elements, over time, Amazon, from as early as 1999

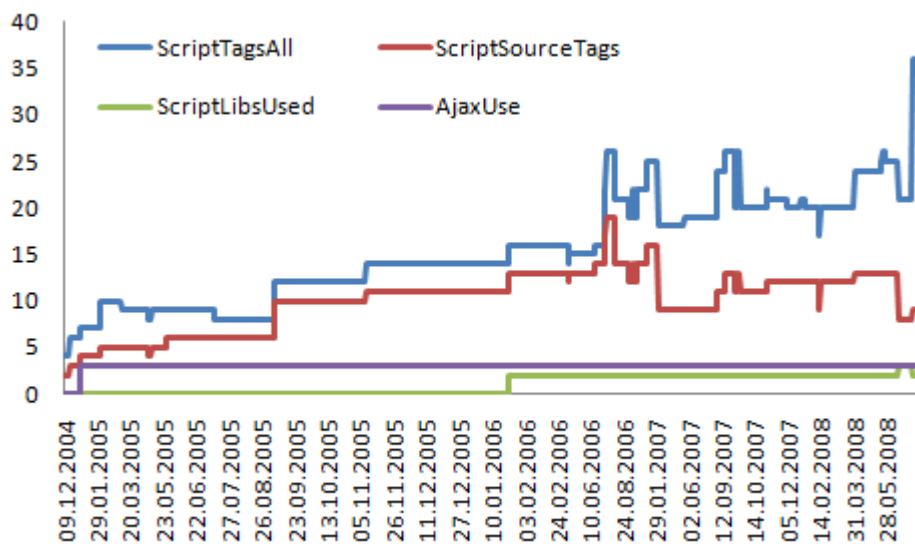


Figure C.7 – script elements and AJAX use, over time, Digg

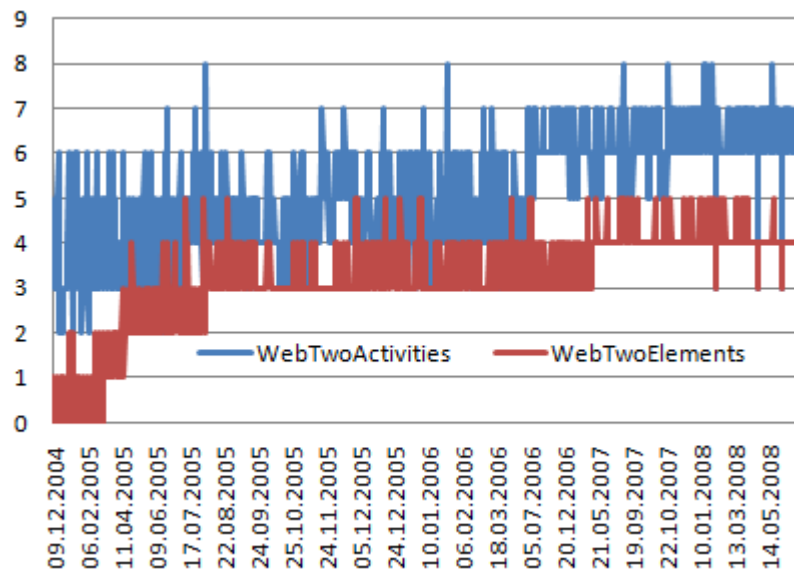


Figure C.8 – Web 2.0 elements and activities, over time, Digg

APPENDIX D: *Collective Intelligence Sources*

The image shows a screenshot of a YouTube video page with various annotations highlighting specific data points:

- Video title:** "Dow Closes Below 10,000, a Four-year Low"
- Video date:** "06 October 2008"
- Author:** "Associated Press"
- Rating and number of ratters:** "★★★★★ 12 ratings"
- View count:** "2,069 views"
- Video responses:** "Video Responses (0)"
- Comment, date and author:** "AcaiBerryMONAVIE (6 months ago) listen to HEAVEN 17 'WE LIVE SO FAST' 1983 toward your Nostalgic Future"
- Video description, tags and the category:** "Wall Street suffered through another extraordinary and traumatic session Monday, with the Dow Jones Industrial Average plunging as much as 800 points, their largest one day point drop, before recovering..."
- Related videos:** "FLO Rida feat. T.Pain Music from the Movie Step...", "Dow slumps below 10,000", "Wall 'Street Beating' Dow Falls As Much As 800 ...", "Teen Decapitated by Six Huge Roller-coaster", "Dow Drops Below \$600 in Massive Sell-off"

Figure D.1 – An example Youtube video, with some of the related Meta and other UGC data

Table D.1 – The three Youtube negative / positive term dictionaries (*note that the terms are word-stems*)

Dictionary	Good words	Bad words
Comments	good, great, interest, invest, happi, hope, save, nice, rich, profit, wealth, bailout, increas, buy, posit, secur, grow, enjoy, demand, rise, rais, easi, awesom, amaz, love, glad, support, luck, pass, benefit, success, incom, strong, opportun, chanc, drive, perfect, surpris, recommend, promis, intellig, cheap, health, logic, super, stimulus, prosper, expand	fuck, conspiraci, shit, pay, debt, collaps, kill, sell, problem, fear, stupid, wrong, idiot, evil, fall, lose, hell, hate, bullshit, blame, destroy, cut, sad, lost, fail, war, bad, credit, crash, poor, wors, drop, corrupt, crisi, leav, depress, death, asshol, damn, dead, worri, lower, wast, brainwash, low, fake, afraid, murder, fuckin, hurt, loss, bastard, worthless, warn, fraud, fucker, illeg, bubbl, failur, suffer, damag, worst, bankrupt, deflat, recess, threat, negat, risk, danger, ruin, victim, useless, horribl, paranoia, destruct, unemploy, hyperinfl
Video descriptions	profit, rebound, success, reward, achiev, great, support, expand, good, uptrend, bailout, top, opportun, interest, demand, benefit, rise, save, earn, win, increas, return, nice, enjoy, hope, easi, improv, rose, boost, posit	risk, loss, sell, stop, lose, collaps, crisi, recess, down, drop, bottom, downtrend, low, crash, lower, meltdown, declin, fall, fail, debt, conspiraci, loser, fell, fear, worst, bubbl, unemploy, attack, wors, bankrupt, lowest, bad, troubl, suffer, decreas, failur, problem
Video titles	bailout, rise, opportun, good, jump, up, high, skyrocket, soar, lift, higher	crash, crisi, fall, drop, recess, meltdown, lose, fail, worst, down

Table D.2 – Review of all statistics (80,214 Amazon reviews)

		Rating					
		1	2	3	4	5	Total
Usefulness	Mean	.4887	.5519	.5796	.6944	.7191	.6692
	Median	.5000	.6346	.6667	.6667	.6667	.6667
	Mode	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	Standard Deviation	.3191	.3227	.3299	.3090	.2900	.3127
Review helpful	Mean	12.67	7.54	7.61	7.07	6.67	8
	Median	4	3	2	2	2	2
	Mode	2	2	2	2	2	2
	Standard Deviation	232.13	26.27	64.72	87.52	42.13	94
Out of (helpful)	Mean	24.44	13.82	11.77	9.28	9.03	11
	Median	10	6	4	3	3	3
	Mode	3	3	3	3	3	3
	Standard Deviation	246.09	36.38	67.74	90.25	45.98	100
Review title length	Mean	29.32	29.85	30.23	27.05	23.95	26
	Median	26	27	27	24	21	23
	Mode	18	25	20	15	14	14
	Standard Deviation	16.09	15.3	16.21	14.96	13.64	15
Review length	Mean	787.54	998.42	1,029.84	971.82	749.62	834
	Median	547	698	681	588	430	504
	Mode	263	413	153	109	99	99
	Standard Deviation	838.65	1,076.91	1,195.93	1,174.06	1,005.05	1051
Row N %		11.0%	5.3%	7.6%	20.4%	55.6%	100.0%
Total N		8,836	4,270	6,107	16,391	44,610	80,214

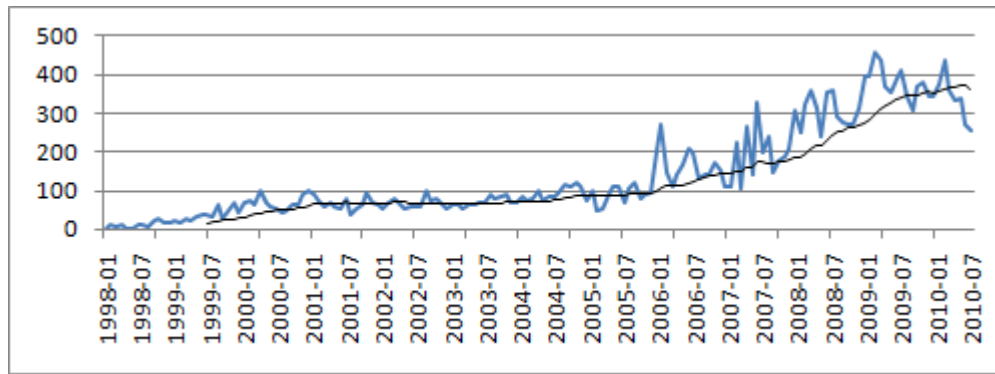


Figure D.2 – Currency / FOREX related book reviews (monthly), correlates with monthly books published on the topic (Pearson $r=.606$, p (two-tailed) $< .01$)

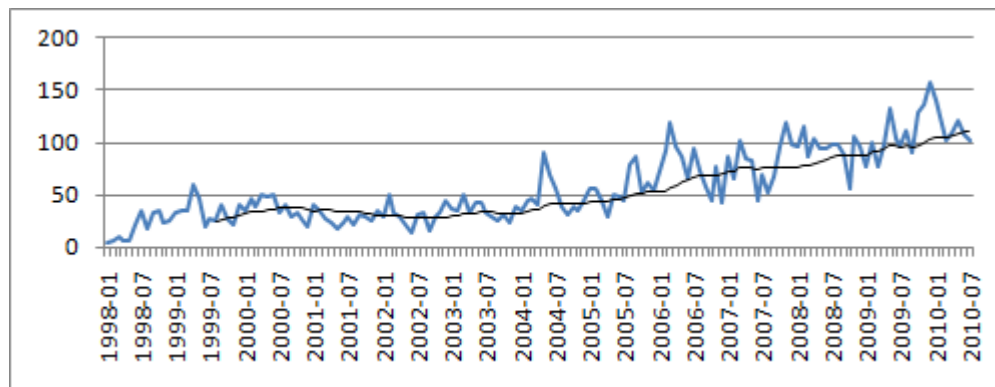


Figure D.3 – Commodities related book reviews (monthly), correlates with monthly books published on the topic (Pearson $r=.473$, p (two-tailed) $< .01$)

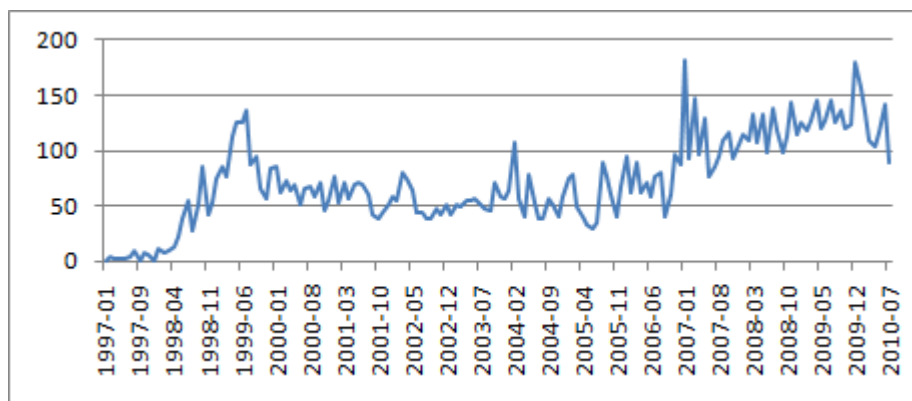


Figure D.4 – Trading book related reviews (monthly)

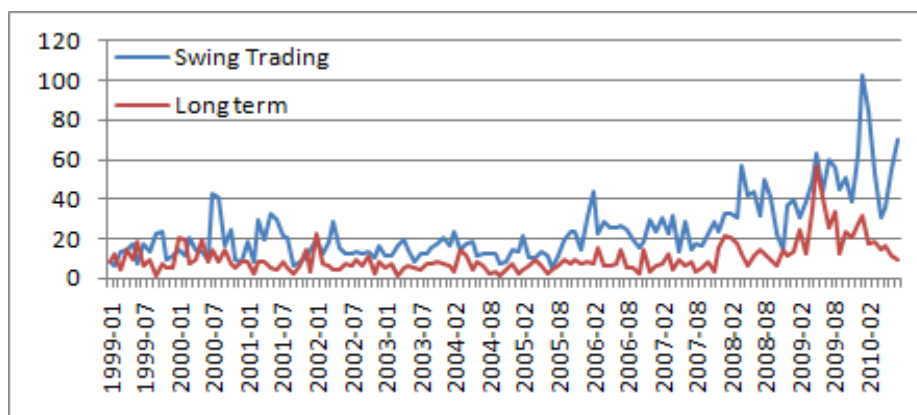


Figure D.5 – Long term and Swing term (shorter term) trading books related reviews (monthly)

10 of 14 people found the following review helpful:

☆☆☆☆☆ **Missing a critical part of the problem**, December 29, 2008

By **M. J Lane** (USA) - [See all my reviews](#)

VINE™ VOICE REAL NAME

This review is from: **Financial Shock: A 360° Look at the Subprime Mortgage Implosion, and How to Avoid the Next Financial Crisis (Hardcover)**

Customer review from the Amazon Vine™ Program ([What's this?](#))

The author of this book is employed by Moody's (a rating agency) and excuses himself from discussing the role that the rating agencies played. In fact you cannot fully explain how the financial meltdown occurred without fully disclosing the role the ratings agencies played and their culpability. It is not possible to avoid future meltdowns without adding reform of the debt rating system to the plan.

I was frankly amazed that a Moody's employee would even have the nerve to write a book like this. I was further disgusted to see that he excuses himself from discussing the role the ratings agencies played (supposedly to avoid "a conflict of interest"). I noticed that the book had a first printing in July 2008 so it was probably done sometime in 2007. That may have meant the book was mostly done before it was fully clear how bad the securitized products, that Moody's and their ilk had unleashed upon the public, were. These financial products, that are currently causing a slow-motion destruction of the financial world, were, for the most part, stamped with "US Government quality" triple A ratings.

By and large, people are not NEARLY angry enough at Moody's, S&P etc. and that is probably because they don't understand just how culpable these ratings agencies were. This book, and this author, had a unique opportunity to really explain that and did not. However, the truth did come out in congressional hearings aired on C-SPAN.

[edited on 2/3/09 for clarity]

Help other customers find the most helpful reviews

Was this review helpful to you?

[Report abuse](#) | [Permalink](#)


 [Comments \(4\)](#)

Figure D.6 – A contextual review for a book on the subprime mortgage, where the author of the review focuses his review only partly on the quality of the book, but rather discusses it in the context of the financial crisis and the role of the ratings agencies

APPENDIX E: *Custom Source of Collective Intelligence*

Table E.1 – Gender (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	34	70.8	72.3	72.3
	Female	13	27.1	27.7	100.0
	Total	47	97.9		
Missing	System	1	2.1		
Total		48	100.0		

Table E.2 – Age groups (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	13-19	2	4.2	4.3	4.3
	20-29	29	60.4	61.7	66.0
	30-39	9	18.8	19.1	85.1
	40-49	4	8.3	8.5	93.6
	50-59	3	6.3	6.4	100.0
	Total	47	97.9	100.0	
Missing	System	1	2.1		
Total		48	100.0		

Table E.3 – Reading frequency (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Only sometime	1	2.1	2.2	2.2
	Once every few days	12	25.0	26.1	28.3
	Every day	27	56.3	58.7	87.0
	Every few hours	6	12.5	13.0	100.0
	Total	46	95.8	100.0	
Missing	System	2	4.2		
Total		48	100.0		

Table E.4 – Financial experience (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	None at all	11	22.9	23.9	23.9
	Interested	17	35.4	37.0	60.9
	Knowledgeable	13	27.1	28.3	89.1
	Financial Guru	5	10.4	10.9	100.0
	Total	46	95.8	100.0	
Missing	System	2	4.2		
Total		48	100.0		

Table E.5 – Location (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	United Kingdom	32	66.7	69.6	69.6
	United States and Canada	4	8.3	8.7	78.3
	European Union	6	12.5	13.0	91.3
	Middle East	1	2.1	2.2	93.5
	Africa	1	2.1	2.2	95.7
	Australia & New Zealand	1	2.1	2.2	97.8
	South America & Mexico	1	2.1	2.2	100.0
	Total	46	95.8	100.0	
Missing	System	2	4.2		
Total		48	100.0		

Table E.6 – News interests (Newsmental users)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Politics	5	10.4	10.9	10.9
	Finance	15	31.3	32.6	43.5
	Technology	16	33.3	34.8	78.3
	World events	10	20.8	21.7	100.0
	Total	46	95.8	100.0	
Missing	System	2	4.2		
Total		48	100.0		

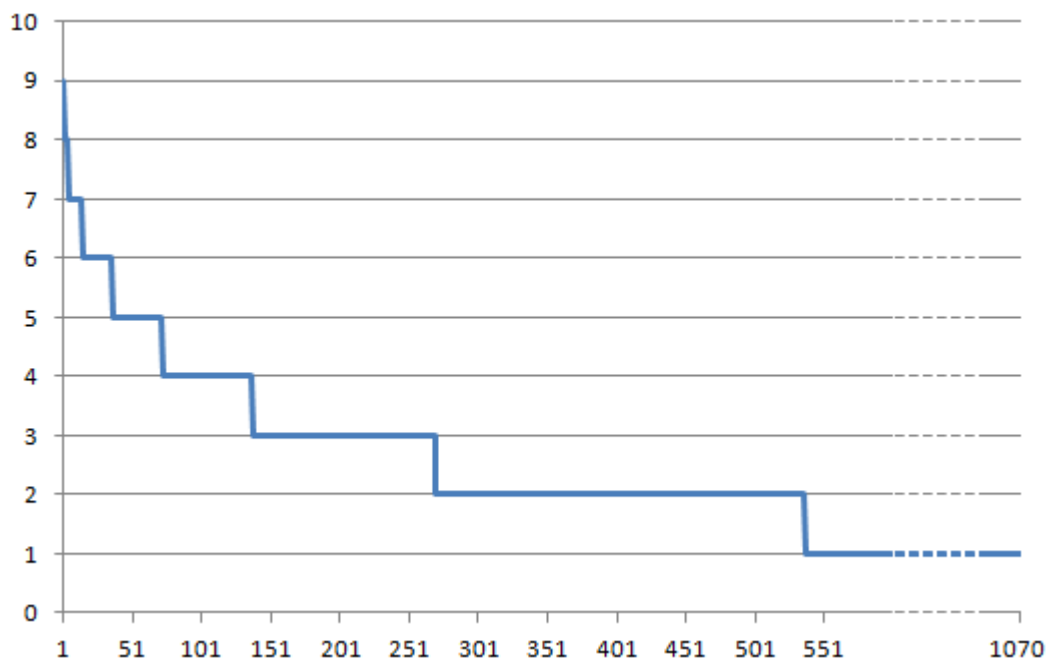


Figure E.1 – Number of ratings for each of the 1,070 rated news stories (y-axis: no. of ratings, x-axis: news story item – ordered by ratings count)

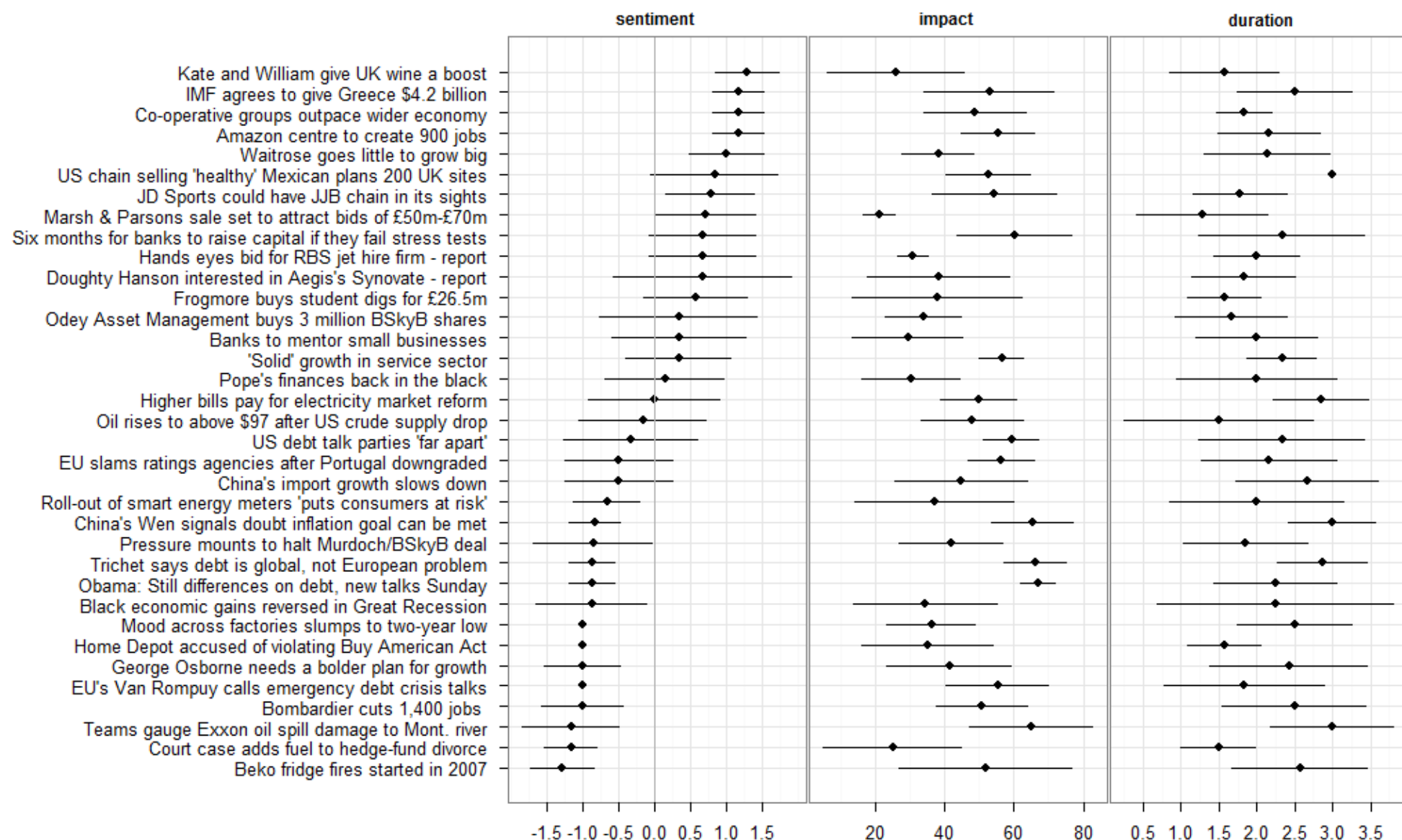


Figure E.2 – Sentiment, impact and duration dot-plots (from left to right) of news stories with six or more ratings (35 headlines), ordered by decreasing average sentiment (standard deviations shown)

APPENDIX F: *Publications*

Sykora M., 2011. *Web 2.0: Common uses and potential applications*, Journal of Interdisciplinary Social Sciences, **5** (10), 2011

Sykora M. and Panek M., 2009. *Media Sharing Websites and the US Financial Market*, Proceedings of the IADIS Internet / WWW 2009 Conference, Rome, Italy

Sykora M., 2009. *Power of Web 2.0 Mass Collaboration in Computational Intelligence and its' uses, an example from Finance*, Proceedings of the 9th Annual Workshop on Computational Intelligence (UKCI), 2009, Nottingham, United Kingdom

Sykora M. and Panek M., 2009. *Financial News Content Publishing on Youtube.com*, Proceedings of the 3rd IEEE Conference on Soft Computing and Applications - SOFA2009, Szeged / Arad, Hungary / Romania

Sykora M., Wang X., Archer R., Parish D. and Bez H. E., 2009. *Case Based Reasoning Approach for Transaction Outcomes Prediction on Currency Markets*, Proceedings of the 3rd IEEE Conference on Soft Computing and Applications - SOFA2009, Szeged / Arad, Hungary / Romania

Sykora M. and Singh S., 2007. *Developing Trading Strategies based on Risk-analysis of Stocks*, Progress in Pattern Recognition - WAPR2007, Southampton, United Kingdom