Augmenting Social Media Items with Metadata using Related Web Content

Sheila Kinsella

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

Supervisor:
Dr. John Breslin

Internal Examiner:
Prof. Dr. Stefan Decker

External Examiner:
Dr. Fabien Gandon

Digital Enterprise Research Institute (DERI),
National University of Ireland, Galway (NUIG)
January 2012

Abstract

The Web has shifted from a read-only medium where most users were solely consumers of information, to an interactive medium where collaborative technologies allow anyone to publish or edit content. In this environment, social media such as social network sites, blogs, wikis, and content-sharing websites have flourished and now masses of users are contributing to the pool of human knowledge that is the Web. This large-scale user participation means that the content-creation capacity of the Web has exploded and there is now wide coverage of news, niche interests and hyperlocal content, all available in real-time. In short, Web 2.0 services have successfully harnessed collective intelligence and a huge and diverse information source has emerged.

The downside of social media as an information source is that often the individual items are very short, informal and lacking in metadata. Despite the wealth of information available in online communities, locating objects of interest can still be challenging. The search and navigation of social media could be greatly improved by augmenting the content of social media items with annotations to provide additional context or descriptors.

This thesis investigates the potential of using related data from the Web to enrich social media items with metadata and thus make it easier to find or browse information in social media. We provide three methods by which social media items can be augmented with novel metadata, specifically tags, locations and categories. Our approaches make use of existing Web data retrieved from HTML documents, APIs and Linked Data. We describe how Semantic Web technologies can be used to represent social media posts and their metadata in a uniform way and thus allow enhanced search and browsing over online community data integrated from heterogeneous sources.

Declaration

I declare that this thesis is composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. The research presented in this thesis was supported by Science Foundation Ireland under Grant No. SFI/02/CE1/I131 (Lion) and Grant No. SFI/08/CE/I1380 (Lion-2), and by the European Commission under contract 215032 (OKKAM) and contract FP7-248984 (GLOCAL).

Sheila Kinsella
January 18, 2012

Acknowledgements

I would like to thank my advisor John for his guidance and encouragement over the last few years, as well as my examiners Stefan and Fabien for their valuable feedback and discussions.

I am grateful to all my friends and colleagues at DERI for their help and inspiration. Thanks especially to my collaborators Alexandre, Andreas, Conor, Mengjiao and Uldis. Special thanks also to all the current and past members of the Social Software Unit, who provided lots of much appreciated support and a great working environment. I am also grateful to my colleagues at EPFL and Yahoo! Research for the valuable internships - especially Adriana, Gleb, Sebastian and Karl in Lausanne, and Vanessa and Neil in Barcelona.

Thanks to Richard for always being there for me. Thanks to my friends, especially Fiona, for lending an ear and taking my mind off things when needed.

Most importantly, thanks to my family for the many years of support and encouragement for my study, through school and my undergraduate years as well as through this PhD.

Core Publications

Papers in Conference Proceedings

- Sheila Kinsella, Mengjiao Wang, John Breslin, Conor Hayes: *Improving Categorisation in Social Media using Hyperlinks to Structured Data Sources.* The 8th Extended Semantic Web Conference (ESWC 2011), Springer, 2011
- Sheila Kinsella, Alexandre Passant, John Breslin: *Topic Classification in Social Media using Metadata from Hyperlinked Objects.* The 33rd European Conference on Information Retrieval (ECIR 2011), Springer, 2011
- Sheila Kinsella, Alexandre Passant, John G. Breslin: *Using Hyperlinks to Enrich Message Board Content with Linked Data.* The 6th International Conference on Semantic Systems (I-SEMANTICS 2010), ACM, 2010
- Sheila Kinsella, Uldis Bojars, Andreas Harth, John G. Breslin, Stefan Decker: *An Interactive Map of Semantic Web Ontology Usage.* The 12th International Conference on Information Visualisation (IV08), IEEE Computer Society, 2008

Papers in Workshop Proceedings

- Sheila Kinsella, Vanessa Murdock, Neil O'Hare: *"I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets.* The 3rd International Workshop on Search and Mining User-generated Contents (SMUC 2011) at the 20th International Conference on Information and Knowledge Management (CIKM 2011), ACM, 2011
- Sheila Kinsella, Adriana Budura, Gleb Skobeltsyn, Sebastian Michel, John G. Breslin, Karl Aberer: *From Web 1.0 to Web 2.0 and Back – How did your Grandma use to Tag?* The 10th International Workshop on Web Information and Data Management (WIDM 2008) at the 17th International Conference on Information and Knowledge Management (CIKM 2008), ACM, 2008

Book Chapters

- Sheila Kinsella, Alexandre Passant, John G. Breslin, Stefan Decker, Ajit Jaokar: *The Future of Social Websites: Sharing Data and Trusted Applications with Semantics.* In Marvin Zelkowitz (ed.), Advances in Computers, 76. Elsevier, 2009
- Sheila Kinsella, Andreas Harth, Alexander Troussov, Mikhail Sogrin, John Judge, Conor Hayes, John G. Breslin: *Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects.* In Thomas Friemel (ed.), Why Context Matters: Applications of Social Network Analysis. VS Verlag, 2008
- Sheila Kinsella, John G. Breslin, Alexandre Passant, Stefan Decker: *Applications of Semantic Web Methodologies and Techniques to Social Networks and Social Websites.* In Cristina Baroglio, Piero A. Bonatti, Jan Maluszynski, Massimo Marchiori, Axel Polleres, Sebastian Schaffert (ed.), Reasoning Web, Fourth International Summer School 2008. Springer, 2008

Papers in Poster Proceedings

- Sheila Kinsella, Alexandre Passant, John G. Breslin: *Ten Years of Hyperlinks in Online Conversations.* The Web Science Conference: Extending the Frontiers of Society On-Line (WebSci10), 2010

Additional Publications

Journal Articles

- Aidan Hogan, Andreas Harth, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, Stefan Decker: *Searching and Browsing Linked Data with SWSE: The Semantic Web Search Engine.* Journal of Web Semantics, Volume 9, Issue 4, Elsevier, 2011.

Papers in Conference Proceedings

- Peyman Nasirifard, Sheila Kinsella, Krystian Samp, Stefan Decker: *Social People-Tagging vs. Social Bookmark-Tagging.* The 17th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2010), Springer, 2010.
- Andreas Harth, Sheila Kinsella, Stefan Decker: *Using Naming Authority to Rank Data and Ontologies for Web Search.* The 8th International Semantic Web Conference (ISWC 2009), Springer, 2010.

Papers in Workshop Proceedings

- Benjamin Heitmann, Sheila Kinsella, Conor Hayes, Stefan Decker: *Implementing Semantic Web Applications: Reference Architecture and Challenges.* The 5th International Workshop on Semantic Web Enabled Software Engineering (SWESE 2009) at the 8th International Semantic Web Conference (ISWC 2009), CEUR-WS.org, 2009.

Book Chapters

- Andreas Harth, Aidan Hogan, Jürgen Umbrich, Sheila Kinsella, Axel Polleres, Stefan Decker: *Searching and Browsing Linked Data with SWSE.* In Roberto De

Virgilio, Francesco Guerra, Yannis Velegrakis (ed.), Semantic Search over the Web, Springer, 2012.

- Alexandre Passant, Sheila Kinsella, Uldis Bojars, John G. Breslin, Stefan Decker: *Understanding Online Communities by Using Semantic Web Technologies.* In Ben Kei Daniel (ed.), Handbook of Research on Methods and Techniques for Studying Virtual Communities: Paradigms and Phenomena, IGI Global, 2010.

Papers in Poster Proceedings

- Uldis Bojars, Andreas Harth, Sheila Kinsella, Srinivas Raghavendra: *An Empirical Investigation of Social Networks in the Blogosphere.* The 1st International Conference on Weblogs and Social Media (ICWSM 2007), 2007.

Contents

List of Figures

List of Tables

List of Listings

Part I

Prelude

Chapter 1

Introduction

1.1 Motivation and Problem Statement

In recent years, the Web has seen an immense growth of applications that enable social interaction, often called social media. Web 2.0 technologies have lowered the barriers to online contribution, and now the average Web user can be an active publisher of user-generated content rather than a passive consumer. Much of the content generated is simple chatter, but there is also a wealth of useful information available on the Social Web. Normal people who witness important events can become citizen journalists providing real-time news and photos. Experts from various domains share their specialist knowledge and experiences. Thus social media has become an important everyday information resource, in particular for niche and real-time information needs.

While social media is a unique and useful data source of information, it also presents challenges for search and navigation. Social media items are typically much shorter and more informal than typical documents (Gruhl et al., 2009). Much of the useful snippets of information occur in conversations where users have a shared context and common knowledge which allows them to communicate without explicitly stating all relevant information. This means that a search for posts on a certain topic may well overlook relevant items which do not use the same vocabulary as that of the searcher. The ability to find social media items of local interest is another feature which is not yet adequately provided on the Web.

In order to allow users to more effectively explore the social media space, metadata or "data about data" is required to provide search interfaces with more information about the content and context of items. Some social media sites allow users to annotate items with metadata themselves. However users are not always willing to put in the effort necessary to provide detailed metadata. There are also some existing techniques which analyse content in order to automatically generate annotations. These approaches are useful but often do not take full advantage of the rich network of content which exists on the Social Web.

A key trend materialising on the Web, which can be useful for discovering metadata, is the move towards structured and interlinked data. In the early days of the Web, information was provided primarily as human-readable text. With the emergence of Web 2.0, APIs became a popular way of providing structured data to enable content reuse and mashups. The Semantic Web is an initiative to add machine-readable and semantically-rich descriptions of data to enable automatic processing and reasoning over Web data. The Linked Data project is a more recent effort to give practical guidelines for exposing and connecting data on the Web. The result of these activities is that increasing amounts of structured, connected data from various domains are becoming available and ready to use. We can also represent social media metadata using standard structured formats, enabling posts from diverse sources to be easily integrated and explored in a uniform way.

In this thesis, we investigate techniques for augmenting social media items with metadata, using related data from the Web. We focus in particular on the potential of exploiting structured Web data to enhance items with metadata. We present approaches to automatically generate novel metadata for items in social media, which can be represented using standard formats for the purpose of enhancing search and navigation across the Social Web. This thesis contributes to enabling networked knowledge - *i.e.*, interlinking and semantically enriching existing information, in order to create valuable new knowledge. Our methods make use of existing links and structure on the Web and turn implicit background knowledge (such as anchortext or hyperlinks) into explicit metadata (such as tags or topic categories). The approaches proposed in this thesis can empower social media sites to better organise their content and provide more advanced functionality via queries over structured metadata.

1.2 Approach

In this thesis, we will present methods which make use of interlinked content and objects on the Web to infer metadata for social media posts. We use the terms "post" and "item" interchangeably to describe a single message or content item posted by a user, such as a blog post, photo or social bookmark. The new metadata assigned to posts can help improve the accessibility of information created by online communities and make social media a more useful information resource. In particular, we will examine the problems of automatic tag generation, geolocation and topic classification. For the task of tag generation, we exploit existing community annotations on the Web in the form of anchortext. For geolocation, we make use of geotagged social media items, building language models from their content in order to estimate the most probable location of a new, non-geotagged item. Finally, for classifying the topic of the post, we extract metadata from hyperlinks, which often represent objects or concepts that are the focus of the post. Figure 1.1 illustrates the enrichment process graphically. A flowchart detailing how the metadata generation process can be applied to a post will be presented in Chapter 8.1.

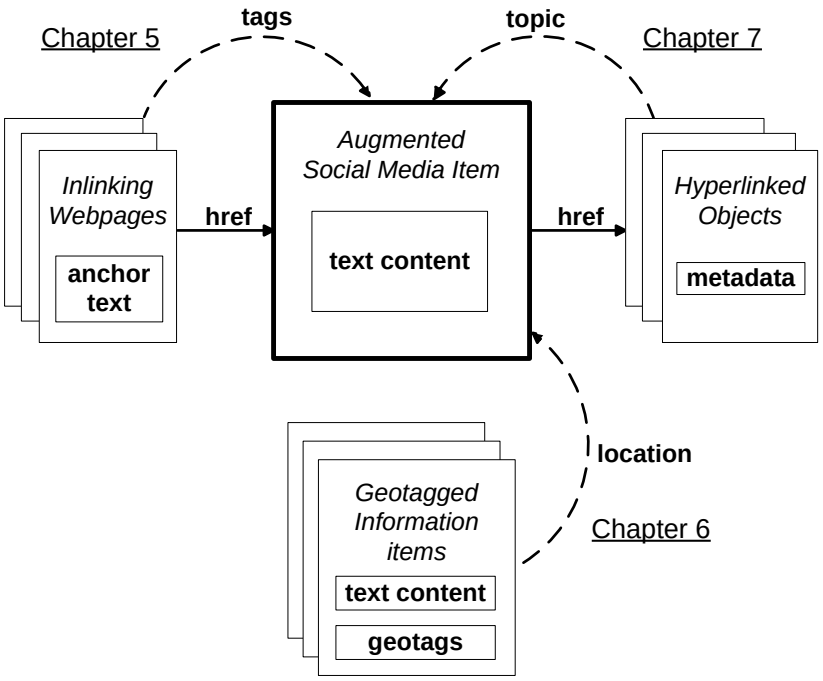

Figure 1.1: Enriching a social media post with metadata using related Web content

1.3 Contributions

This thesis explores the potential of using related Web data to add novel metadata to social media items for enhanced retrieval. The main contributions of this thesis are three distinct and complementary methods for augmenting information items in social media with relevant data, each of which makes use of related information on the Web:

- *Tag Prediction.* Social bookmarking sites make use of user-provided tags to organise content. However, many bookmarked items have already been annotated on the Web, via anchortext. The collection of anchortexts describing a particular item can be used as a source of keywords from which we can automatically predict tags for a bookmarked item. A vector space model approach for inferring tags for Web resources from anchortext is described, and results of an automatic evaluation on a large dataset and of a human evaluation are presented.

- *Geolocation.* Many social media items are geotagged with the location of the creator. By aggregating the content of items by location, language models corresponding to particular places can be generated. These models can then be used to predict the location of a new, non-geotagged item. We present an approach for location prediction which involves resolving geotags of tweets to semantically-meaningful locations, building language models, and using these to infer the location of an item or a user. Evaluation on a large dataset has been performed and achieved promising results, especially for certain levels of geographic granularity.

- *Topic Classification.* Users of social media often link to external objects in order to provide context to a conversation. These hyperlinks can be vital for understanding the topic of an information item. Data from these hyperlinks, especially structured data, provides a useful source of additional textual information for classifying the topic of a social media post. We propose an approach for using metadata extracted from structured data sources to provide additional textual input to a text classifier, and present the results of an evaluation which confirms the improvements gained using this method. We also demonstrate how the usefulness of different metadata

types varies and how the best sources of text features can be experimentally identified.

In the process of investigating the datasets used for evaluating each of the approaches, the thesis also presents the following contributions:

- A study of the characteristics of anchortext and social bookmarks on the Web, and a comparison of the two behaviours.
- A summary of the typical properties of tweets from a microblogging site.
- An analysis of the changes in hyperlinking behaviour observed over a ten year span of a message board site.
- A description of the properties of objects and structured object metadata retrievable via external hyperlinks in a message board dataset.

1.4 Example Scenario

We now present a scenario where a post is augmented with tag, location and topic information from related Web content, and this metadata is represented using common structured formats. Figure 1.2 shows a hypothetical example of a post about a rugby match which the poster attended.

To a human reader who is from Galway and familiar with rugby, it is easy to understand the context of the post. For example, they can see immediately that it is a review of a match. The team concerned is Connacht, which they know is also a province in Ireland, and the match was held in the Sportsground which they know is located in Galway City. They also know that the topic of the post is rugby. All of these pieces of information would be useful items of metadata in order to allow this post to be found by another user from Galway City who is interested in rugby.

Automatically deriving these metadata from the post content is not trivial. The poster does not mention a geographical location for the post – they instead mention the province of the team, which gives only a broad indication of the location, and the stadium name, which is ambiguous. The poster also does not

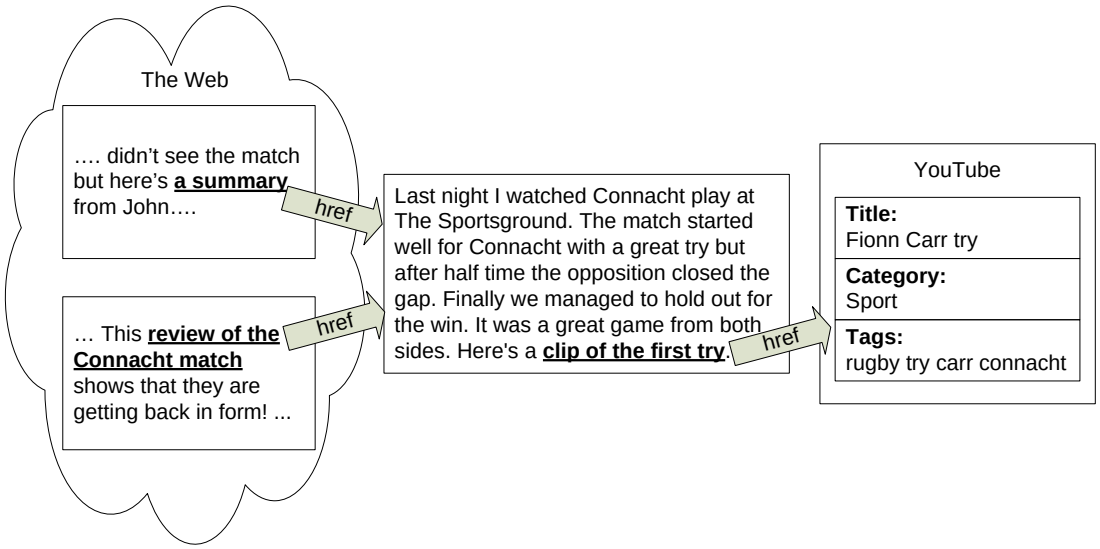

Figure 1.2: Example of a social media post (middle) and related Web data

Tags	connacht, match, review, summary
Location	Galway, Ireland
Category	Rugby

Table 1.1: Examples of metadata that could be extracted for the post of Figure 1.2 by exploiting interlinked Web data

mention the sport involved, nor the fact that the post provides a review of a match.

However, by taking into account relevant information on the Web, it is possible to infer these pieces of information. The anchortext of incoming hyperlinks to the post provide community annotations which can be reused as tags, such as *summary* and *review*. By considering the text of the post and of other geotagged posts, it is possible to detect implicit location clues such as local venues or slang. The fact that many items mentioning the words *Connacht* and *Sportsground* originate from Galway City can be used to infer that this post is also likely to relate to Galway City. Finally, the hyperlink in the post gives additional information regarding the topic of the post – for example, the tag *rugby*. Table 1.1 summarises the metadata which could be inferred for the post of Figure 1.2, by exploiting data from interlinked or related Web content.

We can represent the post of Figure 1.2 and the inferred metadata from Table 1.1 using the Resource Description Framework (RDF), in order to make it

easily queryable. We propose expressing this information using SKOS, DC and SIOC, which are well-known ontologies – schemas which formally describe the concepts in a domain and the relationships between them. RDF, ontologies and related technologies are described in more detail in Section 3.4. Listing 1.1 shows an N-Triples representation of the example post and its inferred metadata.

```
@prefix ex: <http://example.org/> .
@prefix content: <http://purl.org/rss/1.0/modules/content/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix sioc: <http://rdfs.org/sioc/ns#> .

ex:post111 rdf:type sioc:Post .
ex:post111 content:encoded ''Last night I watched Connacht play at The
      Sportsground. The match started well for Connacht with a great try but
      after half time the opposition closed the gap. Finally we managed to
      hold out for the win. It was a great game from both sides. Here's a
      [url=''http://www.youtube.com/watch?v=[...]'']clip of the first try.[/url]'' .
ex:post111 sioc:links_to <http://www.youtube.com/watch?v=[...]> .
ex:post111 dc:subject ''connacht'' .
ex:post111 dc:subject ''match'' .
ex:post111 dc:subject ''review'' .
ex:post111 dc:subject ''summary'' .
ex:post111 dc:spatial <http://sws.geonames.org/2964180/> .
ex:post111 sioc:topic <http://www.dmoz.org/Sports/Football/Rugby_Union/> .
```

Listing 1.1: N-Triples representation of example post and inferred metadata

We suggest mapping the location and topic information to URIs (Uniform Resource Identifiers), since these are clearly understood semantic concepts. A known location can be automatically mapped to a URI from a geographic database such as Geonames by specifying the information necessary to identify it such as its name, place type and coordinates. A set of predefined topics can be represented by creating a suitable SKOS (Simple Knowledge Organization System) concept hierarchy, or could be mapped to URIs from some appropriate existing categorisation system (such as the ODP - Open Directory Project).[1] We represent the tags as literals since these are plain text values which do not inherently correspond to a particular concept. The tags that we infer are expressed using the DC

[1] `http://www.dmoz.org/`, accessed July 2011

(Dublin Core) subject property for indicating keywords, however, it would also be possible to disambiguate these terms to semantic concepts represented by URIs. The predicted location is expressed using a URI from Geoname's geographical database and DC's spatial property for indicating the spatial characteristics of a resource. Finally the topic assigned to the post is represented using a category from the ODP category hierarchy, and is connected to the post via the topic property from the SIOC (Semantically-Interlinked Online Communities) vocabulary. Since all of the metadata are predictions, it would also be possible to add confidence values to the inferred metadata, for example using the approach of (Dividino et al., 2009).

1.5 Use-cases

This section presents two use-cases which show how such augmented items can be used for enhanced search and browsing, enabled by SPARQL, a query language for RDF. The use of RDF and SPARQL means that the enrichment and querying approach is platform independent and can be carried out over heterogeneous data integrated from multiple sources. Thus posts from microblogs, message boards, mailing lists, content sharing sites and other social media could all be explored via one interface.

1.5.1 Local search

In 2009 the Volvo Ocean Race, a round-the-world yacht race, had a summer stopover in Galway, Ireland. It was a major event for the city and attracted many spectators, both locals and visitors. In this search scenario, a blogger wishes to write a blog post reporting about his experience at the Volvo Ocean Race stopover. He wishes to enrich the post with media, such as photos or videos, and quotes from other spectators. In order to find such items, he needs to perform a search over social media from heterogeneous sources. One way would be to perform individual searches on YouTube (for videos), Flickr (for photos), Twitter (for quotes), and perhaps additional services. A preferable way would be to perform one single search, ideally being able to specify the time and location of interest.

Listing 1.2 provides an example of a query that could be used to perform such a search, assuming a dataset which has tag and location data available, and is represented using the RDF, SIOC, and DC vocabularies. The query specifies that the returned resources should be any type of post, that it be related to the appropriate location (Galway) and from the relevant time period (May 23 – June 6, 2009) The post should also be assigned at least one of a set of related tags (*volvooceanrace*, *vor*, *oceanrace*, *yacht*). We use the proposed SPARQL 1.1 keyword EXISTS to ensure that one of these tags is assigned to the post.

The results returned by this query could include any type of social media items which have been represented in RDF, including blog posts, microblog messages and Flickr or YouTube uploads. In this scenario, the user information need is quite specific in terms of both topic and temporal scope, and tags are more likely to be of use for locating the required results than a taxonomy of fixed categories.

1.5.2 Local browsing

In our second scenario, an Irish sports fan from the province of Munster is interested following social media conversations relating to sport in her region. Listing 1.3 presents a query to perform such a request over an RDF social media dataset. The query will return posts that are related to any location which is Munster. The posts must be on the topic of sport, which we once again define using the ODP category hierarchy, although other URIs which identify this concept could also be specified. We use the proposed SPARQL 1.1 feature of retrieving path matches of arbitrary length to include any topics which are below Sports in a concept hierarchy (such as Sports/Tennis).

In this scenario, the user's information need is quite broad, and corresponds well to a topic which is likely to belong in a taxonomy of topic categories. Hence we propose attempting to fulfil this request using a category hierarchy rather than tags, since posts relating to specific sports may not have such a general tag as *sport*, and to supply a list of tags that would cover such a broad area would require excessive time and effort (we discuss an alternative approach to this problem in Section 8.3). However tags could be useful if the user decided to narrow down the area of interest – for example, she could specify *u21* to find youth events or *womens* to find events specifically for females.

```
PREFIX rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .
PREFIX sioc: <http://rdfs.org/sioc/ns#> .
PREFIX dc: <http://purl.org/dc/terms/> .

SELECT ?post WHERE {
  ?post rdf:type sioc:Post .
  ?post dc:spatial <http://sws.geonames.org/2964180/> .
  ?post dc:created ?date .
  FILTER (str(?date) > ''2009−05−23T00:00:00'') .
  FILTER (str(?date) < ''2009−06−06T23:59:59'') .
  FILTER EXISTS {
    { ?post dc:subject ''volvooceanrace'' } UNION
    { ?post dc:subject ''vor'' } UNION
    { ?post dc:subject ''oceanrace'' } UNION
    { ?post dc:subject ''yacht'' }
  }
}
```

Listing 1.2: SPARQL query for a local search scenario

```
PREFIX rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .
PREFIX sioc: <http://rdfs.org/sioc/ns#> .
PREFIX gn: <http://www.geonames.org/ontology#> .
PREFIX skos: <http://www.w3.org/2004/02/skos/core#> .

SELECT ?post WHERE {
  ?post rdf:type sioc:Post .
  ?post dc:spatial ?location .
  ?location gn:parentADM1 <http://sws.geonames.org/2963597/> .
  ?post sioc:topic ?topic .
  ?topic skos:broaderTransitive+ <http://www.dmoz.org/Sports/> .
}
```

Listing 1.3: SPARQL query for a local browsing scenario

1.6 Thesis Outline

The remainder of this thesis is divided into three parts, which are structured as follows:

Part II – Background

This part of the thesis provides necessary background material for investigating the problem of using Web data to automatically generate metadata for social media items. We introduce relevant concepts from the following areas:

- **Chapter 2** gives a history of online communities and describes the most common forms of social media on the Web. It also introduces the theory of object-centred sociality.
- **Chapter 3** reviews different sources of structured data available on the Web, including Web 2.0 APIs, microformat-enhanced data, and Linked Data. It also introduces the fundamentals of the World Wide Web and of the burgeoning Semantic Web. The chapter ends with an overview of existing work on applying Semantic Web technologies to the Social Web.
- **Chapter 4** introduces metadata and the various types of metadata that are most relevant for social media items. It also reviews existing methods for generating metadata in online communities – from content-based approaches to more advanced techniques which incorporate additional information from related Web content.

Part III – Core

The Core part of the thesis is composed of the following three chapters which each investigate a different ways of augmenting items in social media with metadata generated from related information on the Web.

- **Chapter 5** investigates automatic tag prediction for bookmarked documents, based on anchortext from the Web. The approach is evaluated on a dataset combined from a social bookmarking service and a Web crawl.

- **Chapter 6** investigates location prediction for social media items, based on language models generated from geotagged social media. The approach is evaluated on a large microblogging dataset.
- **Chapter 7** investigates topic classification of social media items, based on hyperlinks to external sources of structured data. The approach is evaluated on datasets from two different forms of social media: a message board and a microblogging service.

Part IV – Conclusion

- **Chapter 8** concludes the thesis by summarising of the work presented and explaining how the complementary approaches of Chapters 5, 6 and 7 can be used in combination to improve information-finding in social media. Potential areas for future research are discussed.

Part II

Background

Chapter 2

Online Communities and Social Media*

The main contributions of this thesis are methods for enriching social media items with structured metadata in the form of tags, locations and topics. To provide context for the approaches proposed in the core part of this thesis, we begin by introducing relevant background information. The current chapter will provide a brief history of online communities and an outline of the various existing types of social media. Next, Chapter 3 will summarise approaches for expressing Web data, including Social Web data, in a structured, machine-processable way. Finally, in Chapter 4 we will outline existing approaches to add structure to social media by enriching posts with metadata, covering both manual annotations and automatically generated annotations.

Social media is a term for the wide range of technologies which enable the interaction of groups and individuals on the Web. For many people, they are a fundamental part of daily life, and are often the primary method used to communicate with family, share media items, research product reviews, and organise protests. The Internet has been used as a means of communication since its inception, but it is only in the past decade that it has rapidly become ingrained in our everyday social lives. We begin this chapter with a brief account of the history of online communities, and then describe some of the most prominent forms of social media currently existing on the Web. The chapter finishes with

*This chapter is partially based on (Kinsella, Breslin, et al., 2008), (Kinsella, Harth, et al., 2008) and (Kinsella et al., 2009)

an introduction to object-centred sociality, the theory of objects as mediators of social ties between individuals.

2.1 History of Online Communities

The Internet has long been used as a medium to facilitate social interactions and the formation of virtual communities. Since the 1980's, Usenet mailing lists allowed people to connect with each other and enabled communities to develop, often around topics of interest. Later, throughout the 1990's, technologies such as IRC (Internet Relay Chat), instant messaging, Internet forums and blogging continued the trend of using the Internet to build communities. The social networks formed via these technologies were often not explicitly stated, but were implicitly defined by the interactions of the people involved.

Early social media focused mainly on enabling communication, often between strangers, but with the emergence of social network sites (SNSs) there was a new focus on enabling users to maintain a public profile and a list of connections that can be browsed by others. In 1997, the first website with these features, SixDegrees, was launched. Through the 2000's, more successful SNSs began to appear such as Friendster (an early SNS previously popular in the US, now widely-used in Asia), LinkedIn (an SNS for professional relationships) and Myspace (a music-oriented service). These services, where explicitly-stated networks of friendship form a core part of the website, have become part of the daily lives of millions of users, and generated huge amounts of investment. In the past decade, the popularity of these sites has grown hugely and continues to do so. boyd and Ellison (2008) summarised the history of social network sites, and suggested that in the early days of SNSs, when only the SixDegrees service existed, there simply were not enough users to generate a critical mass: "While people were already flocking to the Internet, most did not have extended networks of friends who were online". One report from Miniwatts Marketing Group (2009) shows the growth in the number of Internet users over time. Between 2000 (when SixDegrees shut down) and 2003 (when Friendster became the first successful SNS), the number of Internet users had doubled.

Other forms of social websites have also emerged and the term Social Web has been coined to describe them collectively. Content sharing sites with social

network functionality such as YouTube (a video-sharing site), Flickr (for sharing images) and Last.fm (a radio and music community site) have achieved immense popularity. Wikis have provided an accessible way of collaborative knowledge management and one example, Wikipedia, has become the world's largest online encyclopaedia. The Social Web is also strongly associated with Web 2.0, the phase of the Web where websites have become interactive and enable collaboration. Web 2.0 technologies such as RSS (enabling content syndication) and Ajax (enabling interactive webpages) have played an integral role in the development of various Social Web services.

Nowadays, engaging in social media is one of the most common uses of time spent online. A report from comScore, Inc. (2011) states that in 2010, 15.6 percent of time spent online globally among those aged 15 and older is accounted for by social network sites. They also report that 90% of U.S. Internet users visited a social network site each month. According to Alexa's[1] Web traffic statistics, Facebook was the second most popular website in the world as of July 2011, with the social media services YouTube, Wikipedia, Blogger and Twitter also featuring in the top ten. One of biggest current trends in social media is the growing popularity of mobile applications, with smartphones enabling users to remain constantly connected to their social networks. An interesting consequence of mobile usage of social media is the emergence of location-based services. These geosocial services allow users to connect and coordinate with local like-minded people, and to keep abreast of news, events and restaurant reviews in their immediate surroundings.

2.2 Prominent Forms of Social Media

In this section, we describe some of the most common forms of social media, and where applicable, we introduce the most popular websites or platforms of each type.

2.2.1 Social network services

Social network services enable users to maintain a profile and define connections to friends. They allow people to maintain an explicit representation of their

[1] `http://www.alexa.com/`, accessed July 2011

social network which is visible online. Often users may communicate via private messages or public comments on profile pages. Some services also allow users to create groups sharing some common interest or affiliation. Social network services often include other features such as photo-sharing or blogging. The most widely used social network service is Facebook[2] which was launched in 2004 and as of January 2011 had accumulated 600 million users (Business Insider Inc., 2011).

Compared to some of the other types of social media, social network sites generally have less public content because of the more sensitive nature of the data. Most social network services, including Facebook, offer the option of private profiles and some enable users to have fine-grained control over what personal content is made publically available. Due to the personal nature of the communications and the privacy features of these websites, social network sites are not very useful as a general information source compared to other social media such as blogs and wikis.

2.2.2 Message boards

A message board, also known as an Internet forum, is a website which enables conversations. Message boards are related to earlier technologies like Usenet, a network of servers that enables users to post articles and reply within threads, and Bulletin Board Systems, software that allowed users to connect by terminal and exchanges messages in public boards.

Message boards are organised into a hierarchical structure of forums, which may themselves contain subforums. Within a forum, a user can create a thread, which is a container for a single conversation, and other users can reply with follow-up posts. Generally, each forum corresponds to a particular topic. In Chapter 7, we make use of categorised posts from a message board in order to investigate how topic classification of social media items can be improved.

2.2.3 Blogs

A blog, or weblog, is a user-created website consisting of journal style entries displayed in reverse-chronological order. Entries may contain text, links to other websites, and images or other media. Blog posts may have tags or categories

[2] `http://www.facebook.com/`, accessed July 2011

assigned for the purpose of organisation. Often there is a facility for readers to leave comments on individual entries, which make blogs an interactive medium. Blogs may be written by individuals, or by groups of contributors. A blog may function as a personal journal, or it may provide news or opinions on a particular subject. As of July 2011, Blogpulse reported indexing over 164 million blogs.[3]

2.2.4 Wikis

A wiki is a website which allows users to edit content through the same interface they use to browse it, usually a Web browser, while some desktop-based wikis also exist. This facilitates collaborative authoring in a community, especially since editing a wiki does not require advanced technical skills. A wiki consists of a set of webpages which can be connected together by links. Normally, anyone can edit an existing wiki article, and if an article does not exist for a particular topic, anyone can create it. If someone vandalises an article, there is a revision history so that the contents can be reverted or fixed by the community. One of the most well-known wikis is the Wikipedia free online encyclopaedia.[4] As of July 2011, the Wikipedia project consists of over 280 different wikis, corresponding to a variety of languages.

Wikis often allow articles to be organised in a hierarchical category structure. Similar to the content of articles, the category structure is also community-generated. Another feature of wikis which promotes structure is templates. Infobox templates encourage users to provide standardised information across related articles. For example, a tennis player infobox could include fields such as name, country, birth date and career prize money. The DBpedia project (Auer et al., 2007) takes advantage of these structured templates in order to create a structured representation of Wikipedia. Our experiments in Chapter 7 make use of DBpedia data in order to improve the categorisation of social media posts which link to Wikipedia articles.

[3] `http://www.blogpulse.com/`, accessed July 2011
[4] `http://www.wikipedia.org/`, accessed July 2011

2.2.5 Social bookmarking services

A social bookmarking system is a way of storing, organizing and sharing Internet bookmarks. Users create a collection of bookmarks, which they can access through a browser interface. These lists can be publicly available, enabling other members of the community with similar interests to view the links. Bookmarks can be organized by assigning keywords or tags to each one. These freely chosen terms generate a "folksonomy" or informal system of social classification.

One of the most popular social bookmarking services is Delicious[5] which was founded in 2003. As of December 2008, Delicious had over 5.3 million users and 180 million unique bookmarked URLs (Delicious Blog, 2008). In Chapter 5, we describe experiments carried out on a dataset from Delicious where we aim to automatically predict tags for bookmarks.

2.2.6 Content sharing services

Content sharing services enable users to share some type or types of media such as photos, videos and music. Usually uploaded content can be commented on by other members of the community. Users can subscribe to feeds from other users or can join groups corresponding to topics of interest.

The most well-known photo sharing service is Flickr[6], which was launched in 2005. In September 2010, Flickr's five billionth photo was uploaded (Flickr Blog, 2010). The most popular video-sharing service is YouTube[7], which was founded a year later. In November 2011, YouTube reported that over 35 hours of video were being uploaded to YouTube every minute (YouTube Blog, 2010). Flickr and YouTube both allow content to be annotated with titles, tags and descriptions, while YouTube additionally requires each video to be assigned to one of a small set of categories. Our experiments in Chapter 7 make use of this metadata in order to classify the topic of social media items which link to Flickr photos and YouTube videos.

[5] `http://www.delicious.com/`, accessed July 2011
[6] `http://www.flickr.com/`, accessed July 2011
[7] `http://www.youtube.com/`, accessed July 2011

2.2.7 Microblogs

Microblogging is a form of blogging where posts are limited to a much shorter length than traditional blogs. A typical post could consist of one sentence or a single image. The most famous micro-blogging service is Twitter[8] which was founded in 2006 and by June 2011 was generating 200 million posts per day (Twitter Blog, 2011). Twitter allows users to share 140 character messages, also known as status updates and tweets. Users are automatically shown the tweets of other users who they "follow". They can also keep track of conversations by searching for topics or usernames of interest. Status updates can be either publically available or restricted to a user's connections. Users can create status updates on the Twitter website, or using one of many applications which interface with Twitter.

There are Twitter-specific syntaxes which have evolved to allow users add some conversational notation to their tweets. Tweets can contain mentions of usernames, specified by prefixing the username with an @ symbol as in @exampleuser. A reply to another person usually contains a mention of their username at the beginning of the tweet. Twitter users often "retweet" other user's status updates to spread a message to their own followers. The original, unofficial convention for retweeting was to copy and paste the original, and prefix it with "RT @username:", optionally adding the retweeter's own comment beforehand. Twitter has since introduced an official method of retweeting, where users simply press a button to perform a re-tweet. These official retweets appear with a special icon beside them and are also annotated as retweets in the API output.

As a publically available source of large amounts of online conversations, Twitter has become a popular source of data for researchers performing analysis of online communities. We make use of data from Twitter for experiments reported later in this thesis, on the topics of location prediction (Chapter 6) and topic classification (Chapter 7).

[8] `http://twitter.com/`, accessed July 2011

2.3 Object-Centred Sociality

Object-centred sociality suggests that people interact and form communities based around common objects of interest. The theory that objects form a central part of social life and mediate ties between people was put forward by Knorr Cetina (1997). Rather than viewing a social network a graph of links between people, we should recognise them as a network of people connected by shared objects around which they socialise. The idea of object-centred sociality can be applied to online communities as well as offline social networks. On the Web, as in real life, social connections are formed through the interactions of people around objects – via the content they create together, comment on, link to, or for which they use similar annotations.

Jyri Engeström, co-founder of the micro-blogging site Jaiku, has proposed that the success of social network sites is partially dependant on whether these networks involve object-centred sociality, that is, whether people are connecting via items of interest such as their jobs, locations or favourite hobbies (Engeström, 2005). Engeström suggests that websites which base their functionality mainly on the creation of links between people, without objects of interest to hold communities together, are unlikely to be sustainable. For example, websites such as Friendster focus on defining the actual social links between people rather than the objects that give context to these connections, such as shared workplaces or schools. This can lead to the problem of users' social networks becoming out-of-date or irrelevant, since a person may change jobs or lose interest in old friends. Figure 2.1 shows an example of the type of explicit relationships that are found in social networking services which do not consider object-centred sociality and therefore lack contextual information about connections.

In contrast, Engeström believes that services which are built around objects are more suited to handling the dynamics of user interests and social activities and can therefore remain relevant and useful to their users. Some of the most popular types of social websites are those whose purpose revolves around objects such as photos (Flickr), videos (YouTube), and bookmarks (Delicious). These objects of interest in these services provide motivation for users to form connections and to continue visiting the sites for updates. Figure 2.2 shows an example of how an implicit relationship can be formed through one person, Bob, commenting

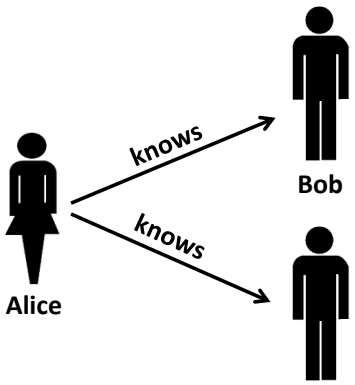

Figure 2.1: A social network without information about shared objects

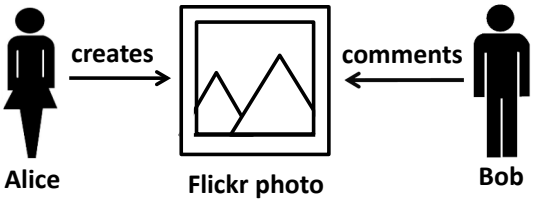

Figure 2.2: Object-centred sociality forms implicit social ties via shared objects

on a photo created by another user, Alice. Rather than being connected simply through online social network relationships (*i.e.*, by explicitly-defined friendships or contacts), people are bound together through "social objects" of common interest.

Flickr, YouTube and Delicious are examples of services which enable centralised object-centred socality – they base their functionality around internal objects that exist within their own website. There also exist websites which encourage object-centred sociality around external objects. These websites allow users to link to objects on the Web and post comments about them, thereby reusing existing objects but providing a medium for people to meet and interact with other users who share an interest in those objects. Facebook is a prime example of a service which makes use of distributed object-centred sociality. Other websites place a "Like" button on their pages, which a Facebook user can click to import an object into their Facebook profile and attach a message to it. Thus Facebook can take advantage of existing content on the Web to provide the impetus for conversations within its own website.

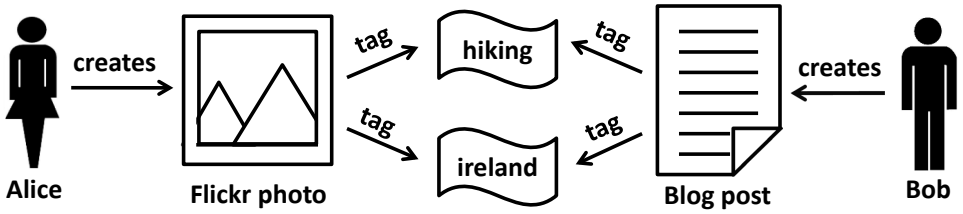

Figure 2.3: Object-centred sociality forms implicit social ties via shared annotations

Adding annotations to items in social networks (*e.g.*, using topic tags, geographical pinpointing, or categories) is an especially useful aid for browsing and locating both interesting items and related people with similar interests. On Flickr, people can look for photos categorized using an interesting tag, or connect to photographers in a specific community of interest. On the Upcoming[9] website, events are also tagged by interest, and people can connect to friends or other like-minded individuals who are attending social or professional events in their locality. Figure 2.3 shows an example of how two people can be indirectly tied through creating posts tagged with similar annotations, *hiking* and *ireland*. These shared annotations indicate a latent connection via common interests, which could be used to help Alice or Bob find each other if they are interested in starting a real-world group around their activity.

The example of Figure 2.3 shows how people can be implicitly connected through manually annotated content, but much of the Social Web is made up of content which does not have manually assigned metadata. For social media posts which do not have user-generated annotations, metadata can be automatically inferred using the methods that we present in later chapters. This metadata can aid users in locating objects of interest and through these objects encourage the formation of new ties. Thus the approaches proposed by this thesis can support object-centred sociality, make content and people easier to find, and therefore increase the value of online communities.

[9] `http://upcoming.yahoo.com/`, accessed July 2011

2.4 Conclusions

In this chapter, we have provided an overview of how social media has evolved on the Web, and discussed the relevance of object-centred sociality to online communities. Early social media technologies focused on enabling conversation, while later developments include enabling users to create public profiles, articulate their social networks, share multimedia content, and collaboratively edit documents. We have also described some of the most prominent types of social media sites that exist on the Web today, and given an overview of their general functionality. The techniques that we propose later in this thesis are applicable to many of these social media services, and the experiments that we have conducted to evaluate our approaches make use of data acquired from services mentioned in this chapter. The next chapter will introduce methods to represent structured information on the Web, including social media data. Representing data from online communites in a machine-processable way is vital for extracting meaning from the data and making the Social Web more searchable and navigable.

Chapter 3

Structured data on the Web*

The previous chapter has provided an overview of social media on the Web. In this chapter, we will outline approaches for representing Web data, including social data, in structured formats. Methods for expressing social media data in a structured, machine-readable way are relevant for two main reasons; firstly, so that the metadata we infer for posts can be easily interpreted and used to augment user search and navigation tasks; and secondly, as we show in Chapters 6 and 7, so that social media items and their metadata can be processed in order to automatically infer new metadata.

The Web began as a source of mostly human-readable information, but gradually more and more structured data has become available, enabling new applications that automatically process Web content. Web APIs are now a common way of exchanging structured information, and the amount of semantically-rich RDF data is growing. We begin this chapter with an introduction to the fundamentals of the World Wide Web architecture, on which data publishing methods depend. We then describe different approaches to publishing structured data that have emerged on the Web, from Web APIs to microformats to Semantic Web technologies. The chapter concludes with a discussion of the new applications and analyses that are enabled by expressing Social Web data using Semantic Web techniques.

*This chapter is partially based on (Kinsella, Bojars, et al., 2008), (Kinsella, Breslin, et al., 2008) and (Kinsella et al., 2009)

3.1 Basic Architecture of the World Wide Web

The World Wide Web is a distributed information management system of interlinked hypertext documents, *i.e.*, documents that contain references which can be followed to other documents, or parts of documents. The Web operates over the Internet, the global system of interconnected computer networks which originated in DARPA (formerly ARPA – the Advanced Research Projects Agency) in the 1960s. Tim Berners-Lee proposed the "Web" in 1989 while working at the particle physics laboratory CERN (Berners-Lee, 1989). The key components of the Web are:

- **Uniform Resource Identifiers (URIs).** A universal schema for identifying and addressing resources (Berners-Lee et al., 2005)
- **HyperText Transfer Protocol (HTTP).** A data transfer protocol for client-server communication (Fielding et al., 1999)
- **HyperText Markup Language (HTML).** A simple markup language for representing hypertext content in webpages. URIs are used to identify resources which are linked to within documents (Connolly & Masinter, 2000)

This decentralised architecture enables the Web to be accessed by content producers and consumers who may use different clients and servers but who use the same technologies to identify resources, transfer data and represent content. In 1990, Berners-Lee used this framework in developing the first Web server for hosting Web content, and the first Web browser for displaying (and in this case also editing) Web documents. The first webpage also went live in 1990, providing a description of the World Wide Web project.[1]

3.2 Web APIs

A Web API (application programming interface) is an interface to a service that enables HTTP requests to be made to the service, and returns response messages, enabling data exchange between the two parties. The APIs provided by

[1] `http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html`, accessed July 2011

Web 2.0 sites typically conform to the REST (Representational State Transfer) style, which means that they are simple services where resources are identified with URIs and all data transfer is made solely using HTTP methods. Web APIs typically return data in a structured format such as JSON (JavaScript Object Notation) or XML (Extensible Markup Language), but do not use common vocabularies, so each API has its own individual format for requests and responses.

Web APIs are a simple and popular way of providing programmatic access to structured data on the Web. ProgrammableWeb reported in January 2011 that their database of APIs covered 2,647 services and that this number had doubled during 2010 (ProgrammableWeb Blog, 2011). These APIs are typically lightweight, easy to learn, and enable a simple way of accessing and integrating Web data. However, since every API has its own specialised methods for data access, programmers must create custom code for each API that they need to access and so the integration of data from multiple sources requires considerable effort.

The lack of standardisation between APIs can be seen by comparing request and response formats of popular social media services. Listing 3.1 shows the response returned by a query to the Twitter API for a user's contact list, obtained by issuing the request `http://twitter.com/friends/ids/sheilak.xml`. Listing 3.2 shows the response of a similar query to the Flickr API, retrieved by issuing the request `http://api.flickr.com/services/rest/?method=flickr.contacts.getPublicList&api_key=[...]&user_id=25634193@N07`. The two queries require different parameters, and the response formats are unrelated, despite the fact that they return similar information. For example, a user ID in Twitter's model is denoted using the tag `id`, but in the Flickr model the same data is denoted as an attribute, `nsid`. In addition, despite the fact that both queries involve accounts belonging to the same user, the user IDs are completely different and cannot easily be resolved to the same person. This lack of common semantics in heterogeneous Web APIs adds complexity to the integration of data from multiple sources, and is a problem that Semantic Web technologies aim to solve.

In the experiments reported in this thesis, we made use of several public Web APIs, in cases where there was no alternative source of more semantically rich data. These included APIs from social media sites (Flickr, Twitter and

```
<?xml version=''1.0'' encoding=''UTF-8'' ?>
<ids>
 <id>1153581</id>
 <id>822659</id>
 <id>8139122</id>
 [...]
</ids>
```

Listing 3.1: Twitter API response to a request for a user's contact list

```
<rsp stat=''ok''>
 <contacts page=''1'' pages=''1'' per_page=''1000'' perpage=''1000'' [...] />
  <contact nsid=''33669349@N00'' username=''Alexandre Passant'' [...] />
  <contact nsid=''62377636@N00'' username=''Cloudie'' [...] />
  <contact nsid=''43184127@N00'' username=''cygri'' [...] />
  [...]
 </contacts>
</rsp>
```

Listing 3.2: Flickr API response to a request for a user's contact list

YouTube), APIs providing geographic data from Yahoo! Geo, and Amazon's product information API. In cases where the data was also represented using Semantic Web technologies, we choose this more semantically-rich representation in order to more easily identify and extract the relevant information.

3.3 Microformats

Microformats are an approach to add semantics to HTML (or XHTML – eXtensible HyperText Markup Language) by reusing existing HTML tags. The use of microformats to embed metadata within the content of a webpage enables the automatic extraction of marked-up data items. Existing microformats allow the representation of geographic data (the geo microformat), contact information (hCard), social relationships (XFN), events (hCalendar), reviews (hReview) and more.[2] Listing 3.3 gives an example of a portion of a webpage where the geo microformat has been used to markup geographic coordinates. A human reader

[2] `http://microformats.org/`, accessed July 2011

will simply see a sentence listing the coordinates, while a machine processing the page will be able to extract the explicitly identified latitude and longitude.

```
DERI is located at
<span class="geo">
  <span class="latitude">53.290</span>,
  <span class="longitude">−9.074</span>
</span>
```

Listing 3.3: Portion of a webpage marked up with geographic coordinates

Microformats have become widely recognised as a method for publishing structured data on the Web. Google integrates data from microformat-enhanced webpages in their search engine, displaying facts extracted from documents as "rich snippets" in results lists. Figure 3.1 shows an example of a search result for a camera review which has been marked-up with the hReview microformat. Without even clicking through to the review a user can immediately see the price range of the camera and the fact that it has achieved an average rating of four stars. Microformats were also supported by the now-defunct Yahoo! SearchMonkey service which enabled developers to enhance search results using structured data. Major websites publishing data marked up with microformats include LinkedIn, Facebook, Yahoo! (*e.g.*, their Upcoming events website[3]) and the Associated Press, who proposed the hNews format which has been adopted by the microformats community.

Nikon Coolpix S8100 (gold) Review | ZDNet
★★★★☆ 7 reviews - Price range: $269.95 - $299.95
12 Jan 2011 **...** Gadget gift ideas and picks by the ZDNet Reviews experts you **...**
www.**zdnet**.com/**reviews**/product/digital-**camera**s/**nikon**.../34182245 - Cached

Figure 3.1: Google rich snippet for a microformat-enhanced product review

The popularity of microformats is due to the fact that they provide a simple and non-proprietary way of facilitating automated processing of Web content. Microformats are maintained by an open community with no official standards body. As the example in Listing 3.3 shows, microformats do not require much

[3] `http://upcoming.yahoo.com/`, accessed July 2011

expertise beyond knowledge of HTML. The structured data provided by microformats is embedded within the human-readable text of the webpage so additional files or services are unnecessary. Individual microformat specifications are typically short, focused, and simple. The geo microformat, for example, contains only two properties: latitude and longitude. They are easy to use both for data publishers and for the developers of applications which consume the microformat-enhanced data. Since microformats define common standards, once a application can parse a format such as hReview, it can automatically extract data from any website which publishes data marked up with the microformat.

Microformats are an easy way of adding simple semantics to Web documents, but their simplicity also results in limitations. There are only a limited number of microformats, and the creation of custom formats or extension of existing formats outside of the microformats community is discouraged, since only known vocabularies can be parsed. Therefore a publisher is limited to the existing predefined microformats. To enable the representation of any arbitrary data, an alternative approach is necessary.

3.4 The Semantic Web

The Semantic Web is an effort to extend the human-readable documents of the World Wide Web into a network of interlinked and semantically-rich machine-readable knowledge. A Scientific American article from Berners-Lee et al. (2001) describes the Semantic Web as "an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation". The standard format for publishing data on the Semantic Web is the RDF language. Data is represented within ontologies, also known as vocabularies or schemas, which formally describe a set of concepts and the relationships between those concepts.

3.4.1 Resource Description Framework (RDF)

The RDF language (Manola & Miller, 2008) enables information to be represented on the Web in the form of statements, or triples, which are each composed of a subject, a predicate, and an object. The subject of a triple denotes a resource,

identified either by a URI, or unidentified – in which case it is termed a blank node. The predicate represents a relationship, identified by a URI. The object denotes the value of the triple, and it may be a literal value, an identified resource with a URI, or an unidentified resource. Since a resource may occur as both a subject and an object (and in the case of properties, also as a predicate), RDF data forms a directed labelled graph. Figure 3.2 shows an example of a simple RDF graph. The graph contains two triples; one linking two resources together with a particular property, and another linking a resource to a value with another named property. Since URIs are globally unique identifiers, their use in RDF data enables multiple data sources to be easily interlinked. Two different publishers can reuse the same identifier to provide information about a particular resource.

There are two important vocabularies which extend RDF to define the semantics of ontologies – RDF Schema and the Web Ontology Language (OWL). RDF Schema (Brickley & Guha, 2004) enables the description of classes, properties and the relationships between them. OWL (McGuinness & Harmelen, 2004) allows additional richer descriptions to be added: for example, expressing the disjointness of classes and the cardinality of properties. The RDF Schema is sufficient for simple ontologies, while OWL allows more complex data to be modelled. Reasoning engines such as SAOR (Bonatti et al., 2011) can be used to make inferences over Semantic Web data, *i.e.*, use logic theory in order to infer new facts.

Data can be expressed in RDF using various syntaxes, or serialisations. One of the most commonly used syntaxes is RDF/XML. Listing 3.4 shows a possible serialisation of the graph of Figure 3.2 in RDF/XML. The document begins with an XML declaration. In the second line, we declare the top-level rdf:RDF element and the namespaces that are used in the document. The prefixes rdf and foaf are used later in the document to abbreviate resource URIs.

There are also text-based serialisations of RDF, which enable data to be more easy read by humans. Widely used text-based syntaxes include N-Triples (Grant & Beckett, 2004) and Turtle (Beckett & Berners-Lee, 2011). Listing 3.5 shows an N-Triples serialisation of the graph of Figure 3.2. Again, namespace prefixes have been declared to enable abbreviation of URIs.

RDF data can be queried using the SPARQL – the SPARQL Protocol and RDF Query Language (Prud'hommeaux & Seaborne, 2008). The SPARQL syn-

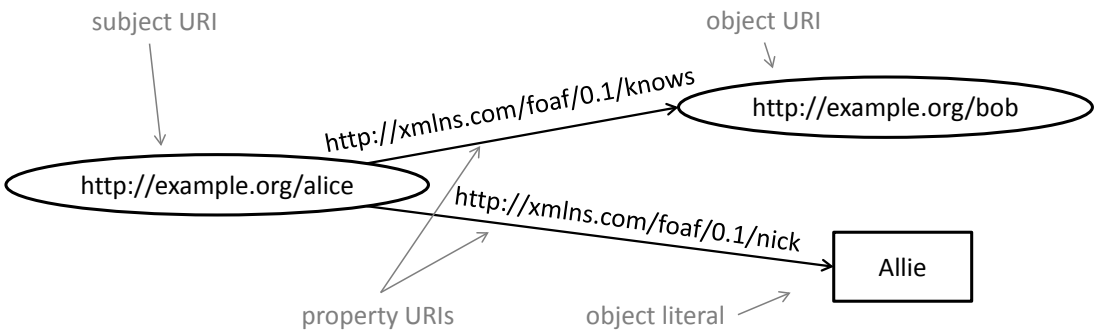

Figure 3.2: A simple RDF graph

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdf:Description rdf:about="http://example.org/alice">
    <foaf:nick>Allie</foaf:nick>
    <foaf:knows rdf:resource="http://example.org/bob" />
  </rdf:Description>
</rdf:RDF>
```

Listing 3.4: RDF/XML representation of Figure 3.2

```
@prefix ex: <http://example.org/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .

ex:alice foaf:nick "Allie" .
ex:alice foaf:knows ex:bob .
```

Listing 3.5: N-Triples representation of Figure 3.2

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .

SELECT ?nick WHERE {
   ?person foaf:nick ?nickname .
}
```

Listing 3.6: Example SPARQL query

tax is similar to the SQL language for querying relational databases. Listing 3.6 gives an example of a simple SPARQL query which returns the nicknames of all persons in an RDF graph. SPARQL can be used to query RDF data from multiple diverse sources and can return results sets, as for the query of Listing 3.6, boolean values, or RDF graphs (*e.g.,* as for the query of Listing 7.2). A specification for a new version of the language, SPARQL 1.1 (Harris & Seaborne, 2011), is currently in progress and proposes important new features including aggregation and property paths.

3.4.2 Linked Data

In 2006, Tim Berners-Lee outlined an ideal approach for creating interlinked data on the Web (Berners-Lee, 2006). He put forward the following four principles for publishing Linked Data:

- Use URIs to identify things.
- Use HTTP URIs so that they can be looked up (dereferenced).
- When a URI is looked up, the publisher should provide useful information about the thing which the URI identifies, using standards such as RDF or SPARQL.
- Include links to other URIs, so that further related entities can be discovered.

Since then, the Linking Open Data community project (Bizer et al., 2009) has worked to make more open data sources on the Web available as Linked Data. Prominent data sources which have been published as Linked Data include the Wikipedia online encyclopaedia, the DBLP computer science bibliography, and the Geonames geographical database. In Chapter 7, we make use of three datasets published as Linked Data as sources of structured metadata in order to improve topic categorisation of social media posts.

3.4.3 RDFa

An alternative approach to publishing RDF data is RDFa (Adida & Birbeck, 2008), which embeds RDF triples directly within XHTML documents. List-

ing 3.7 shows the graph of Figure 3.2 expressed using RDFa. RDFa is related to microformats since both approaches involve the insertion of structured data into human-readable webpages. RDFa is extensible, allowing publishers to create their own vocabularies as needed, but it is not quite as simple as microformats and requires the use of XHTML, unlike microformats which can be embedded within HTML. Therefore either approach may be useful, depending on the situation. It is also possible to extract RDF data from microformats, using the GRDDL mechanism – Gleaning Resource Descriptions from Dialects of Languages (Connolly, 2007).

```
<div about="http://example.org/alice">
   Yesterday I met <span property="foaf:nick">Allie</span> and her friend
   <a rel="foaf:knows" href="http://example.org/bob">Bob</a>.
</div>
```

Listing 3.7: RDFa representation of Figure 3.2

The amount of RDFa data on the Web has greatly increased recently, in particular due to some key initiatives. Facebook's Open Graph Protocol[4] encourages developers to embed RDFa descriptions of objects in order to enable users to click a "Like" button, connecting the object to the user's Facebook profile. As a result many websites including those of Microsoft, IMDb, and the NHL have adopted the Open Graph Protocol. The Drupal content management system has added RDFa output in Drupal 7, which means that major websites including that of the White House are now publishing RDFa. Another large source of RDFa data is from vendors who have adopted the Good Relations vocabulary (Hepp, 2008) for product, price, and company data. Companies publishing Good Relations include Best Buy and O'Reilly, and Good Relations data is supported by Google's Rich Snippets.

3.4.4 Ontologies

An ontology is a formal specification of the concepts of a domain and the relationships that exist between them (Gruber, 1993). The term *ontology* originates from the field of philosophy, where it means a "systematic account of Existence".

[4] `http://ogp.me/`, accessed July 2011

The explicit definitions of concepts and relationships within an ontology allows a community to represent data using common models.

Ontologies usually contain a set of classes, which describe the concepts within a domain. These may be organised into a hierarchy with sub-class and super-class relationships. Classes may have instances, which are specific things that are members of the class. Classes and instances can be described using attributes, which are data values, or using relationships to connect them to other classes or instances. These connections between classes, instances and data values are the means to describe the semantics of a domain.

The open and distributed nature of the Web means that publishers are free to create or extend ontologies as they wish. This has the advantage of allowing arbitrary data from any domain to be easily expressed using RDF data; however, it also means that data is not curated and may be erroneous. We found in a study of ontology usage that there are often inconsistencies in how people use RDF ontologies (Kinsella, Bojars, et al., 2008). Some of the most popular classes observed were not defined in official specifications. Another study by Hogan et al. (2010) details commonly occurring errors in Linked Data, and provides recommendations for data publishers on how to avoid making these errors, and for data consumers on how to handle such noisy data. The authors point out that while there is now a rich vein of community-created structured data available on the Web, the open nature of the creation process means that data quality is an issue which will need to be addressed in order to attract more wide-spread adoption of Semantic Web technologies.

There are many ontologies which have become widespread on the Web and are considered to be the standard way for representing data for their respective domains. We now describe some of the most well-known Semantic Web vocabularies, in particular those which are most relevant for publishing social media data.

Dublin Core (DC)

The Dublin Core vocabulary (DCMI Usage Board, 2010) provides a set of metadata elements for describing resources. It includes terms for expressing information such as a resource's creator, title, format, license, language and provenance. Publishers of RDF data frequently use the DC schema to describe the standard

metadata of resources, often in conjunction with other vocabularies for representing more domain-specific information.

Simple Knowledge Organization System (SKOS)

The SKOS languages (Miles & Bechhofer, 2009) enable the representation of taxonomies and other controlled structured vocabularies. Concepts can be organised in collections, and may be related by hierarchical links (skos:broader, skos:narrower) and associative links (skos:related). Additional properties are provided to define mappings between different vocabularies (*e.g.*, skos:closeMatch and skos:narrowMatch). SKOS allows schemes of related concepts to be expressed in a simpler way than using OWL.

Friend-of-a-Friend (FOAF)

The FOAF project (Brickley & Miller, 2010) was started in 2000 and defines a widely-used vocabulary for describing people and the relationships between them, as well as the things that they create and do. Anyone can create their own FOAF file describing themselves and their social network, and the information from multiple FOAF files can easily be combined to obtain a higher-level view of the network across various sources. This means that a group of people can articulate their social network without the need for a single centralised database, following the distributed principles used in the architecture of the Web. Therefore FOAF can be useful for enabling social network portability – the ability to reuse ones own profile across various social network sites and applications.

Prominent social network services that expose FOAF data include LiveJournal (a social network and blogging community site) and identi.ca (an open source micro-blogging service). People can also create their own FOAF document and link to it from their homepage, and exporters are available for some major social websites as Flickr, Twitter, and Facebook. Such FOAF documents usually contain personal information, links to friends, and other related resources.

Semantically-Interlinked Online Communities (SIOC)

The SIOC initiative (Breslin et al., 2005) was started in 2004 and aims to interlink community content on the Web. Discussions from various platforms including

blogs, message boards and mailing lists can be expressed using a combination of FOAF, DC and SIOC. SKOS schemas may be used to represent the topics of posts. Using the SKOS taxonomies, related conversations from different platforms can be interlinked via their topic. The motivation behind SIOC is to allow improved linkages and interoperability between the many diverse and heterogeneous forms of online communication that exist on the Web.

There are many exporters available for exposing SIOC data for online communities. Platform specific exporters have been developed for the Drupal content management site, for vBulletin discussion forums, and for the WordPress blogging software, among others. There are also exporters available for mailing lists and IRC logs. The American news magazine Newsweek exposes SIOC data for articles on their website.

3.4.5 The Semantic Web and social media

Semantic Web technologies enable more structure to be embedded in social media items, and more connections to be expressed between items. Through the use of common formats like FOAF and SIOC, increased interoperability and data portability are possible between social media sites. Semantically-rich social media data also enables new applications and analysis involving online communities.

Semantically-enhanced social applications

Semantic Web technologies are useful for generating recommendations from user-generated content, since they allow complex networks of resources and relationships from multiple sources to be integrated and reasoned over. Passant and Raimond (2008) detail how the Semantic Web could be used to generate cross-platform music recommendations based on tags and social networks, among other features. NoTube (Schopman et al., 2010) is a project which aims to develop a framework based on Semantic Web technologies for personalised television recommendation and content delivery. The project proposes to integrate profile data from the Web, attention data from various social media sites, and other relevant data, and to generate recommendations based on user interests and social networks. Stankovic et al. (2010) propose using Linked Open Data as a source of information for recommending experts, based on the premise that integrating

data from multiple structured sources on the Web can provide a more comprehensive view of expertise than other systems have.

Another application area where Semantic Web techniques can be of use is in enhancing visualisation and exploration of user-generated content. The Twitris system (Jadhav et al., 2010) uses spatial, temporal and topical information from tweets to extract event descriptions for enhanced exploration of event related data. The approach behind Twitris involves obtaining tweets related to an event, extracting key descriptors from the text, and slicing the data into spatio-temporal slices. This enables event data to be displayed to the user in a way that allows them to see what is going on, and where and when it is taking place. Twitris uses DBpedia to enrich their semantic models, and exposes the resulting data as Linked Data. A related application presented by Rowe and Mazumdar (2010) extracts social data from multiple heterogeneous sources, generates a consistent metadata descriptions in RDF for each item, and uses SPARQL queries to query the integrated dataset. Similar to Twitris, the user interface enables user-generated content to be explored along the facets of location, time and subject.

Semantics applied to social network analysis

Social network analysis (SNA) is the sub-field of sociology which uses network theory to analyse networks of people based on the ties between them. San Martín and Gutierrez (2009) have studied the requirements of a data model for social networks and concluded that RDF and SPARQL are suitable technologies for representing and transforming social networks. Analysts may use Semantic Web technologies to simply fuse data from many heterogeneous sources, or new methods may be developed which take advantage of the rich representations of concepts and relationships on the Semantic Web – "semantic social network analysis".

Existing work has been performed which shows how the Semantic Web can be used to perform SNA on distributed social networks from many sources. For example, Finin et al. (2005) conducted an analysis of a social network extracted from personal profile data expressed using the FOAF ontology. Flink (Mika, 2005a) is a related application which extracts and aggregates social network data from FOAF documents, mailing lists, HTML pages and bibliographic databases, and allows the resulting networks to be browsed, along with visualisations and

SNA metrics. Both of these works involved multiple references to the same person to be fused, using OWL inverse-functional properties – properties for which only one subject can be linked to any given object, meaning that two subjects which link to the same object must represent the same entity.

The rich graphs of typed nodes and links that can be described using Semantic Web technologies is useful for detecting implicit links between people via shared objects. Ontocopi (Alani et al., 2003) is an application for community detection that exploits implicit semantic relations in ontologies, such as the fact that two people attend the same conferences. We performed a related community detection study (Kinsella, Harth, et al., 2008) on a semantic social network extracted from FOAF documents and composed of nodes with classes including persons, workplaces and projects. Both of these approaches made use of spreading activation algorithms to identify related concepts and therefore communities. Semantic Web technologies have also been used for the task of conflict of interest detection. Aleman-Meza et al. (2006) presented an application which integrates data from social networks and bibliographic data in order to detect potential conflicts of interest arising due to explicit or implicit relationships between reviewers and authors.

Recent research has made use of Semantic Web technologies for various other SNA applications. Erétéo et al. (2009) proposed an ontology for representing concepts for social network analysis, and SPARQL extensions for querying social network patterns. They applied these techniques to perform an analysis of a social network dataset, which showed that networks derived from different types of relationships display significantly different properties, indicating that it is important to considering the semantics of relationships in social networks. S. Wang and Groth (2010) presented a framework for measuring the dynamics of influence between content networks and social networks. They extracted networks using on Semantic Web technologies and applied longitudinal social network analysis to identify how the properties from each network affect each other time. Finally, the Live Social Semantics application (Barrat et al., 2010) is a platform that allows the dynamics of online and offline networks to be studied and compared. Data is gathered from sensors, social network services, and publications and is integrated in an RDF store. The data can then be queried in order to investigate the correlations between online activity, face-to-face activity and academic seniority.

3.5 Conclusions

This chapter has outlined various techniques for expressing Web data in a structured way, from Web APIs, which are generally simple but non-standard, to microformats, which are standard but non-extensible, to RDF, which enables shared, extensible vocabularies, often evolving through community consensus. We have also described how Semantic Web technologies can be used for novel applications and analyses using Social Web data. These examples provide motivation for using Semantic Web vocabularies to express the metadata generated by the approaches that we propose in the core part of this thesis. The next chapter will focus on how structured information can be added to unstructured social media items by attaching metadata, either manually assigned by users or automatically assigned via analysis of the post content or related items.

Chapter 4

Enriching Social Media Items with Metadata

The first chapter of the background part of this thesis has introduced social media and online communities, while the second has outlined approaches for representing structured data on the Web. In this chapter, we explain the motivation for adding metadata to social media items and present existing approaches for doing so. We begin by defining the term metadata, introducing the field of information retrieval, and describing how information retrieval in social media can be improved by exploiting metadata. Next, we summarise common types of user-assigned annotations. Finally, we discuss approaches for automatically generating metadata. We differentiate between those approaches which analyse posts as isolated objects, and those which exploit the fact that social media posts are not isolated items but are in fact part of a network of related and interlinked objects.

4.1 Metadata and Information Retrieval

Metadata is "data about data", which may provide additional information about the content, format or context of a particular data item. For example, the metadata of a digital photograph may include a timestamp, the image resolution, and the latitude and longitude at which the photo was taken. The metadata for a blog post may mention the author, the title and the creation date. Certain metadata for Web resources are also often described as annotations, since effectively they

are notes that are attached to resources.

Information retrieval (IR) is the process of searching for documents or data within document collections, databases, or on the Web (Baeza-Yates & Ribeiro-Neto, 1999). User information seeking behaviour on the Web can be divided into two categories, browsing and searching, where browsing involves an unfocused information need and searching is focused on a particular information need (Bodoff, 2006). The information retrieval process begins when a user formulates a query based on an information need, and enters the query into the IR system – for example, a Web search engine. On the Web, queries are generally keyword based. The system matches the user query against object representations which are stored in an index and returns a results list, often ranked by relevance. These results are displayed to the user in their browser. The user may choose to refine their query and repeat the process, or if the system allows it, they may choose to apply a filter to the results list. Examples of the search options offered by Google include the ability to restrict results to videos, to content which has been created in the last 24 hours, or to content related to a specified city.

Metadata is an important resource for information retrieval, especially for multimedia content (Baeza-Yates & Ribeiro-Neto, 1999). Search within social media can likewise be greatly enhanced by making use of metadata. For images and videos, textual annotations such as tags allow them to be located by the keyword searches that Web users are accustomed to performing. For textual social media content, metadata may still provide useful information which is not extractable from the content of the post. For example, a geographic annotation can provide location information which was not explicitly specified, or "votes" assigned by other users may indicate the quality of a post. Sometimes users provide such annotations themselves, however the added work involved in doing so is a disincentive which means that automatic means of inferring metadata are also required and are often employed.

In this thesis, we focus on the generation of tags, geotags and topic categorisations. These three approaches are complementary ways of organising information. Location annotations may only be relevant for certain geographically-focused items, while tagging and topic classification provide more generic annotations. The topic classification method that we propose assumes a set of predefined categories *i.e.*, a taxonomy. This is a top-down approach, which has the advantage

of ensuring that all participants are using a homogenous scheme. Tagging, or the creation of folksonomies, is a bottom-up free-form approach which has the advantage of enabling arbitrary descriptors at any level of detail. Additionally, tags may be related to any aspect of the item whereas the category suggestions that we propose correspond specifically to the topic of the content. Depending on the user information need, either or both of these forms of metadata may be useful. Tagging has been described as a "social indexing process" (Lu et al., 2009), with each item being described by any number of tags, describing any aspect of the item at any level of detail. Categories tend to be broader and limited to a fixed vocabulary, providing a starting point to narrow down the search space by topic. Therefore tags are particularly for trying to find a specific item or exploring a quite precise topic, while categories are especially useful for general browsing, or for filtering search results to a particular theme.

4.2 User-assigned Social Annotations

The simplest way of acquiring metadata for social media items is to enable user-generated annotations. The ability to annotate an object may be restricted to the item creator, or it may be open to any member of the community. For quality ratings, annotation is generally only allowed by users who are not the original author of the content. Some types of user-assigned metadata are optional, and some are compulsory. The quality of user-assigned metadata is unreliable, since users are often careless, lack expertise in assigning annotations, and in the case of spammers, may intentionally assign fraudulent information. However studies (Bao et al., 2007; Zhou et al., 2008) have found that nevertheless, user-generated annotations are a useful source of information for improving information retrieval on the Web.

The selection of metadata types available varies across social media platforms, with each website having its own unique features. Practically all social media posts are automatically annotated with their creation time, enabling posts to be ordered or filtered by time. Most posts are also automatically labelled with the username of their author, although some sites allow anonymity. Posts on social network services and microblogs usually have no additional metadata beyond author and temporal information. We now describe in detail some forms of user-

generated metadata or social annotations that are of particular relevance to social media.

4.2.1 Titles

A title is the name assigned to a social media item. Usually a title is a short phrase or sentence. Titles are generally required for wiki articles, blog posts, content sharing services, social bookmarking services, and at least for the initial posts of threads in message boards.

4.2.2 Descriptions

Descriptions are usually consist of a sentence or multiple sentences. Creating a description enables the user to provide more detailed information about a resource. Descriptions are most common in content sharing services and are usually optional. This means that they are not as reliable a source of textual features as titles – a study by Figueiredo et al. (2009) found that approximately half of the description fields were empty on the social music recommendation websites Last.fm and the scientific reference sharing website CiteULike.

4.2.3 Tags

Tags are free-text keywords that can be attached to items and enable an informal way of organising resources and making them easily searchable. The vocabulary of tags that emerges within a service is called a folksonomy, a portmanteau of folks and taxonomy, since it is a system of classification derived from collaborative free-text tagging by a community.

The practice of tagging has become increasingly widespread on the Web since the emergence of social bookmarking tools in 2004 (Hammond et al., 2005). Tags are a key feature of many social media sites, including Delicious, Flickr, YouTube and often blogs. Although microblogging websites such as Twitter do not provide a formal method of annotations, a convention has emerged in these communities for adding informal tags. Tweets are tagged with a topic, event or other annotation, by prefixing a tag with a hash to make a hashtag *e.g.*, *#twitter*. Twitter

have also announced a feature, Twitter Annotations[1], that would enable developers to implement structured metadata within tweets. Twitter suggests the use of labelled annotations such as rating for reviews and postal_code for places, however as of July 2011 this functionality has not yet been released.

The motivations of users to assign tags are varied. Golder and Huberman (2006) identified several functions of tags describing different aspects of an item: topic (*e.g.*, *microsoft*), type (*book*), owner, category refinement (*e.g.*, numbered items), characteristics, self reference (*mystuff*) and task organisation (*toread*). Bischoff et al. (2008) suggested two more functions: time and location. Sen et al. (2006) collapse the categories of Golder and Huberman into three broad classes: factual tags, subjective tags and personal tags. In their study, the distribution of tags across classes was 63% factual, 29% subjective, 3% personal and 5% other. Ames and Naaman (2007) conducted an interview-based investigation into the motivations of taggers, and developed a taxonomy of motivations based on two dimensions: Sociality (intended for self or others) and Function (intended for organisation or communication). They describe a wide range of motivations reported by participants that go beyond previously suggested taxonomies: for example, using certain tags to gain more views, or tagging a photo with the names of people appearing in the photo, so that those users can later find pictures of themselves.

Since tagging has become widespread on the Web, many researchers have also investigated the effectiveness of tags for search. Yanbe et al. (2007) proposed a search system which incorporates tag popularity and tag sentiment of bookmarked webpages, and conducted a preliminary evaluation of the proposed search enhancement. Bao et al. (2007) introduced two algorithms exploiting tags to improve the quality of query-dependent and query-independent rankings of webpages. Zhou et al. (2008) developed a framework for modelling document content and annotations, detecting topical information in tags, and integrating this topic-level information to improve the performance of traditional information techniques. Heymann, Koutrika, and Garcia-Molina (2008) investigated the potential of social annotations to improve Web search, and summarised their findings in terms of positive indicators (*e.g.*, Delicious is a good source of recent, unindexed documents) and negative indicators (*e.g.*, Delicious covers only a tiny

[1] `http://dev.twitter.com/pages/annotations_overview`, accessed July 2011

fraction of the Web).

Many other studies of various aspects of tagging behaviour have been conducted. Mika (2005b) proposed a tri-partite model of tagging composed of relationships involving resources, users and tags, and showed how this model can be used to automatically detect communities and observe their emergent semantics. Golder and Huberman (2006) performed a study of the changes in user activity in the Delicious community over time. They observed wide variation in the behaviour of users, and the popularity of tags. Marlow et al. (2006) developed a taxonomy of different types of tagging systems, and performed a case study of the Flickr website. They compared their results to Golder and Huberman's study of Delicious data, and concluded that tagging patterns vary across different types of platforms. Halpin et al. (2007) conducted a study focusing on tag distribution, tag dynamics and tag-tag correlations. They observed that tagging distributions tend to stabilise into power laws, where a small fraction of tags are very popular, but most tags appear rarely, in what is known as the long tail of a power law distribution.

There have been a number of studies investigating the relationship of tagging to traditional Web metadata and anchortext. Heymann, Koutrika, and Garcia-Molina (2008) analysed the occurrence of tags in the pages they annotate, and the text of the pages linked to or from the annotated page. They found that at least 80% of the time a tag was present in at least one of these places, indicating that a substantial amount of tags are redundant for search as the terms would be discovered in the page content regardless. They did not investigate the frequency of tag occurrence in anchortext, and they did not attempt to recover tags from anchortext. Lipczak and Milios (2010) found that tags are strongly influenced by document titles, even where other synonyms exist. They also found that users reuse their own previous tags where possible, indicating that more value is placed on maintaining a consistent collection of personal tags than on collaboration. Noll and Meinel (2007) compared the metadata supplied by authors of documents (title and keywords of a HTML page) to the metadata created by the readers of documents (Delicious tags). They found that less than 60% of tags are found in the document content or metadata. They did not however investigate anchortext. In follow-up work (Noll & Meinel, 2008), the authors found that anchortext provides less novel information than tags, but that the similarity

between tags and anchortext was quite low, so they provide different kinds of information. They concluded that if one is interested in identifying new data, both metadata types should be considered. Bischoff et al. (2008) also examined the overlap between tags, anchortext and webpage content, as part of a study on the usefulness of different types of tags for search. They found that for both tags and anchortext, approximately 45% of terms also occurred in the document content. They observed that 43% of tags also occur in anchortext. Finally, Liu et al. (2008) performed a comparative analysis of tags and anchortext and found that they shared some similar characteristics but that the overlap between them is not high, and tags tend to be more diverse.

4.2.4 Geotags and location information

A geotag is an item of geographical metadata. Usually a geotag consists of a latitude and longitude coordinate, but on the Web place names are often used instead in order to preserve the privacy of users. Frequently geotags are not manually entered by the user, but instead their smart-phone or GPS-enabled device will attach the user's geographic coordinates at the time of item creation. Users may also manually specify locations, either one location associated with their profile or individual locations assigned to each content item that they create. Geotags can enable users to easily find precise location-specific information. For example, Flickr's map interface allows users to zoom to a given place and browse photos taken in that area.

With the increase of smartphone usage, geotagging has become an important feature of many social media sites. The photo-sharing website Flickr allows photos to be tagged with geographic coordinates (geotagged), either manually by the user or automatically by the device taking the photo. YouTube's video-sharing service allows users to input coordinates when uploading a video to the website. Twitter allows contributions to be geotagged by the user's phone, or manually via their website. It is also possible to allow Twitter to access your browser location information to geotag your tweets. Twitter application developers have two options when representing geotags of tweets: they can simply include the latitude and longitude of the tweet, or they can make use of Twitter's reverse geocoding function to include a description of a place, *e.g.*, at a neighbourhood

level. It is also common to include both representations. Twitter profiles also have a Location field where users can enter their geographic information, but this self-reported data is often unreliable.

4.2.5 Category information

Categories provide a way of dividing items into different classes of some sort, often by topic. Categories allow users to browse items relating to a particular area of interest. Some websites such as Wikipedia use a hierarchical structure, where categories are organised into a taxonomy and may contain more specific sub-categories. Others such as YouTube use flat categories, where categories are unconnected and have no sub-categories. Another way in which categorical information differs across sites is that some websites offer a fixed set of categories, whereas others (such as blog platforms) allow users to freely assign categories. The way in which categories are implemented is not consistent, even between social media technologies which provide similar functions. For example, the video-sharing website YouTube requires videos to be assigned to one of ten broad categories, while the photo-sharing website Flickr does not have a formal category structure, although photos may be assigned to sets which can be used as an informal way of attaching categorical metadata.

The user-generated category structure of Wikipedia has proven to be a very useful resource for information retrieval and data mining on the Web, since it provides a large hierarchical classification of concepts with a very broad scope. Wikipedia's articles and category structure have been used in identifying the topic of documents (Schonhofen, 2006), computing semantic relatedness (Strube & Ponzetto, 2006) and disambiguating named entities (Cucerzan, 2007), among other applications.

4.2.6 Quality ratings

Many social media sites allow users to rate items created by other users, resulting in collaboratively-generated quality ratings. This feature is especially common in social review sites. Examples of this type of functionality can also be seen in the Slashdot[2] technology news website, where comments can be voted up

[2] `http://slashdot.org/`, accessed July 2011

or down, or in the Yahoo! Answers[3] community question-and-answer site, where answers can be similarly voted up or down. In the user interfaces of these services, items with the highest ratings are generally displayed most prominently. The Flickr Favourites feature and Facebook's Like button also allow users to indicate favoured items.

Several approaches have been proposed which make use of collaborative ratings in social media. Ratings may be used in aggregate as an indicator of quality, for example as in the study of Bian et al. (2008) which showed that incorporating community feedback features such as number of votes was useful in improving the ranking of answers on a question-and-answer website. They may also be used on the level of individual users, as an indicator of preferences, as in the study by Matsuo and Yamamoto (2009) who found that product-ratings on a community review site can be used to estimate trust between users as well as improving product recommendation.

4.2.7 Comments

On blogs, content sharing sites, and social network sites, users are often able to respond to their friend's content by adding comments to an item such as a blog post or a photo. Comments can be viewed as additional content, since they allow long discussions to evolve, whose scope may extend far beyond the original post. However they may also be viewed as a type of metadata, since they provide a way for other users to add annotations to the original post.

Comments can be a useful resource for search in social media, as shown by Mishne and Glance (2006) in their study of a blog collection. Yee et al. (2009) showed that user comments also improved the quality of search results in YouTube, especially for popular videos. They point out a problem with using comments for search, which is that unlike author-assigned metadata, it takes time for comments to accumulate to the extent that they yield enough terms to be useful. Comments have also been used to assess the trustworthiness of items (Nakamura et al., 2008) and to detect polarising items (Siersdorfer et al., 2010). The frequency of interactions between users through comments may give insight into the social network of the users involved (Gilbert & Karahalios, 2009).

[3] `http://answers.yahoo.com/`, accessed July 2011

4.2.8 Attention measures

Attention measures are properties of an item that are not assigned by one single user, but are generated by aggregating the actions of all of the users who interact with the item. While quality ratings enable users to explicitly assess the quality of items, attention measures are an implicit indicator of the popularity of items. The number of comments received by items is one possible attention measure. On Twitter, the number of retweets a post receives is also an example of an attention measure. Another common measure is number of views, which is used by Flickr and YouTube as an indicator of item popularity. The number of times that an item has been shared is a possible measure of attention, and is used by services such as TweetMeme[4] to assess the popularity of stories on Twitter. Social media sites usually incorporate these attention measures into their user interfaces, so that popular items are ranked highest or are highlighted visually (*e.g.*, marked as "Top Tweets" on Twitter). Retweet trees have been shown to be a good feature for predicting information credibility on Twitter (Castillo et al., 2011).

4.3 Automatic Metadata Generation: From Isolated Items to Interlinked Items

In this section we give an overview of existing approaches for automatically inferring metadata for social media items. The types of metadata which we focus on for automatic generation are tags, locations, and categories. We do not attempt to predict titles and descriptions, since they provide a summarisation of the author's individual view of his or her own content. It is also not appropriate to generate quality ratings, comments or attention measures, since they indicate the reactions of the community to a content item. The three metadata types of tags, locations, and categories are also the types which most previous works on metadata generation in social media have focused on. We will describe the existing work related to each of these tasks in this section. Later, after introducing each of our three metadata prediction approaches in Chapters 5, 6 and 7, we compare our proposed approach to the most relevant of the existing methods described below.

[4] `http://tweetmeme.com/`, accessed July 2011

The most straightforward way to add metadata to a social media item is to consider just the content of the post, whether textual (*e.g.*, a tweet) or otherwise (*e.g.*, a photo or video). Metadata generation approaches which are purely content-based do not take advantage of the interlinked nature of the Web and do not exploit related content, however for some applications they provide acceptable results. More advanced approaches for metadata generation take advantage of the networked nature of Web content to help annotate each item, rather than annotating individual items in isolation. This can involve learning from already annotated items, propagating metadata between related items, or gleaning clues from hyperlinks. We break down automatic metadata generation techniques into the following general categories, though some approaches use a combination of these techniques.

- **Content analysis of isolated items.** Methods that extract metadata based on the item content. They may make use of collection statistics or additional resources such as ontologies but do not consider related items or their annotations.

- **Learning from specific related objects.** Approaches that identify related objects and generate metadata based on their metadata or content. These approaches make use of implicit links between resources (*e.g.*, text similarity).

- **Learning from models of previously annotated items.** Approaches that make use of supervised learning techniques to construct models from a training set of annotated items, and use these models to make predictions for a test set of unannotated items.

- **Incorporating information from hyperlinks.** Techniques that use information extracted from incoming hyperlinks, outgoing hyperlinks, or both. These approaches make use of explicit links between resources.

4.3.1 Tag prediction

Content analysis of isolated items

There is a body of work on keyphrase extraction which could be applied to the task of tag prediction, as proposed by Medelyan et al. (2009). However this type of approach has the disadvantage of only generating tags which already occur as terms in the document, thus limiting the potential for improving retrieval in a collection. Some other approaches combine extracted keyphrases with information gleaned from other resources, enabling novel terms to be generated. PIRATES (Pudota et al., 2010) is a framework that processes documents to extract keyphrases and then applies knowledge from domain ontologies to generate new potential tags. P-TAG (Chirita et al., 2007) is a system that suggests personalized tags for webpages based on both the content of the document and the data on the user's desktop.

Learning from specific related objects

One approach to learning from related items is to identify the tagged items that are most closely related to a target document, and generate tags based on the tags or content of the related documents. The simplest way to determine the related documents is based on text similarity. The approach of Lu et al. (2009) applies cosine similarity between Web page content to propagate tags between closely-related URLs. Oliveira et al. (2008) also use cosine similarity to find related pages, but they suggest tags based on the content of the target document and the set of related documents, rather than re-using existing tags. AutoTag (Mishne, 2006) and TagAssist (Sood et al., 2007) are tools for automatically tagging blog posts based on the tags of the most similar existing posts.

Another commonly used approach to identifying related items is tag co-occurrence, which is useful for augmenting the existing tag set of a resource. Sigurbjörnsson and Zwol (2008) proposed a method that uses tag co-occurrence to predict additional tags for tagged images. Budura et al. (2009) introduced an approach which makes use of metrics including document similarity and tag co-occurrence to propagate tags on a graph of linked documents. They evaluate their approach on a citation network of scientific publications and on the Web graph.

Alternative approaches identify relevant items by making use of additional, platform-specific information such as geographic or social network data. For example, SpiritTagger (Moxley et al., 2008) is a tag suggestion tool that suggests tags for a geotagged photo by identifying a set of similar photos from the surrounding geographic region and determining which of these tags are most relevant to the set as a whole. Rae et al. (2010) proposed a method that makes use of social networks for tag prediction, and showed that tag recommendation can be improved by considering tag co-occurrence within a user's own communities as well as globally.

Learning from models of previously annotated items

An alternative approach to making use of previously tagged items for tag prediction is to view the set of documents tagged with a certain term as a class. Then a classifier is trained using these documents, in order to determine whether a given test document also belongs to the class, and therefore should be annotated with the relevant tag. This approach has been investigated by J. Wang and Davison (2008) in their study of webpages, and by Hassanali and Hatzivassiloglou (2010) in their study of political blogs. Classifiers have also been used to predict tags using non-textual features, such as audio features in music (Eck et al., 2007). The disadvantage of these approaches is that the set of potential tags is limited to a predefined set that occurred in the training set in sufficient numbers to build a useful classifier.

Incorporating information from hyperlinks

A method of using anchortext as a source of information for tag prediction has been investigated in a paper by Heymann, Ramage, and Garcia-Molina (2008). The authors presented a method for automatically tagging webpages in Delicious based on the anchortext of incoming links, Web page content, surrounding hosts, and previously assigned tags. In their experiments, tags were predicted from the set of 100 most frequent tags by training a classifier which runs a binary classification task for each tag. The classifier was trained on a set of bookmarks that have a large number of attached tags. Hence, this approach takes advantage of both previously annotated items, and external hyperlinks.

4.3.2 Location prediction

Content analysis of isolated items

The simplest approach to determining the geographical focus of Web content is the use of gazetteers, which are databases of geographic names. The Web-a-where system of Amitay et al. (2004) detects mentions of places names in webpages and disambiguates these to determine where each mention refers to. The system then determines the overall geographic focus of the page as a whole. Their evaluation showed that Web-a-where could correctly determine the correct city or state in 65% of cases. Fink et al. (2009) adapted the algorithm of Amitay et al. (2004) and used the place names mentioned in blogs to determine the location of the author of each blog. They achieved an accuracy of 63% where a prediction was considered correct if it was within 100 miles of the correct one, or within the same state or province. Both of these systems only make use of explicit geographic information to assign a location to textual content.

Related techniques have also been applied on search query logs to investigate the geographic intent of queries. Jones et al. (2008) investigated the geographic intent of Web searchers by comparing the locations mentioned in a query to the IP address that issued the query. The locations were identified based on a database of official and colloquial place names. They found that locations in search queries are often relatively close to source of the query, but that the distance varies greatly for different topics. For example, people tend to specify nearby locations when they search for "pizza", but often mention distant places when they search for "maps".

Learning from specific related objects

Hays and Efros (2008) used visual features of Flickr photos to predict their locations with a nearest-neighbour classification approach. They report correctly placing approximately 13% of photos within 25 kilometres (city scale) and 25% of photos within 750 kilometres (country scale), based solely on the visual properties of the photos.

Learning from models of previously annotated items

Some recent papers have investigated techniques for geolocating Twitter users based on models built using tweets originating from known locations. Cheng et al. (2010) proposed a probabilistic framework for estimating a Twitter user's city-level location based on tweet content. Their approach does not use geotags from individual tweets, but instead uses place information reported in a user's Location field. First, words that have a local focus are identified, and their geographic distribution is modelled. Next, geographic-smoothing is applied to these models. Then for each user the city where they are most probably located can be determined, based on their tweet content. For their experiments, city models were built from over 4 million tweets from approximately 131,000 users who have reported United States cities in their Location field. The test set was limited to around 5,000 users who have reported coordinates in their Location field and have 1,000 or more tweets. The authors also reported that 100 tweets was typically enough to give acceptable results. Their method correctly placed 51% of Twitter users within 100 miles of their correct location. They did not report results on a more local scale and their evaluation only considered cities in the United States. They also did not tackle the problem of placing individual tweets, rather than users.

Eisenstein et al. (2010) introduced a technique for identifying geographically-salient terms from geotagged data and thus detecting coherent linguistic communities. Their system also allows them to predict the location of an author of textual content. The approach involves building a cascading topic model for each topic, and then generating regional variants. This enables the authors to generate models of regions and thus predict the location of tweet authors. Their experiments were limited to authors from the contiguous United States who created 20 or more status updates. The dataset used contains approximately 9,500 users and 380,000 tweets. For modelling purposes, they defined each user's ground truth location as the location of their first tweet in the sample. The evaluation shows that the system could correctly place users within a mean of 900 kilometres from their correct location, and could identify the correct state of the user in 24% of cases. They were also to able to identify terms with strong regional biases, often relating to sports teams, athletes, place names, and slang.

Hecht et al. (2011) conducted an investigation into usage of the Location field in Twitter, and report on experiments that attempt to predict the home country and state of Twitter users. Their location prediction approach involves using a Multinomial Naive Bayes model to classify tweets. They incorporated an algorithm to bias the models towards terms with a regional focus. Their experiments used a limited dataset of four countries, and the state-level experiments were restricted to the United States. A high-precision, low-recall geocoder was used to obtain location ground truth data from the users' Location fields. After filtering out users with less than 10 tweets, a dataset of almost 100,000 users remained. Their approach correctly placed users to their home states with an accuracy of up to 30%. The evaluation found that their models could correctly place the users at a much better accuracy than random, indicating that users implicitly reveal location information in their tweets.

Much of the previous work on placing social media items has focused on Flickr photos. Flickr was one of the first sources of geotagged content on the Web, and as a photo-sharing site, users are often motivated to provide geographic data for their photos, in order to share more information about their travels. An analysis by Sigurbjörnsson and Zwol (2008) found that over 13% of Flickr tags could be classified as locations using Wordnet.

One photo-placing study by Serdyukov et al. (2009) used a language modelling approach to predict the locations of Flickr photos. The models were built from 120,000 photos and the locations were defined as cells from grids of varying size placed on the world map. In the experiments reported by this paper, their approach was able to correctly place 7% of photos within 1 kilometre, and 19% within 10 kilometres.

Another photo-placing study by Crandall et al. (2009) investigated the use of visual, textual and temporal features to classify photos within specific cities. They identified the top ten landmarks in 100 cities, then for each city they conducted a 10-way classification task on geotagged photos in order to predict the landmark relevant to each photo. An average accuracy of 54% was achieved for this task.

There has also been related work carried out using query logs from search engines. Backstrom et al. (2008) modelled the spatial variation in search queries, identifying those queries that have a natural geographic focus, and accurately

detecting the location of the geographic focus. Rather than using external geographic information, they observed the patterns that emerged from the aggregate query logs. Yi et al. (2009) built language models for US cities using explicit geographic queries and use these to identify implicit geographic intent in queries where no placename is mentioned. They did not report results on a more fine-grained granularity such as postal codes.

4.3.3 Topic classification

There is a wide range of work on classification of social media, for purposes including automatic spam detection, sentiment analysis, and quality assessment. In this section we focus on the work that involves topic classification, *i.e.*, categorising posts in terms of predefined topics.

Learning from specific related objects

Genc et al. (2011) present an approach for improving the classification of tweets into predefined categories from Wikipedia. They first map tweets to their most similar Wikipedia article, based on occurrences of words from the tweet within Wikipedia articles and their titles. They then identify the predefined category that is closest to the category of the article, using the Wikipedia category structure. They do not make use of hyperlinks within posts.

Yin et al. (2009) propose improving object classification within a website by bridging heterogeneous objects across websites so that category information can be propagated from one domain to another. They use tags as bridges to connect the unlabelled objects to labelled ones, based on the intuition that users are likely to use similar tags to describe objects with the same topic, even if the objects are of a different type. They improve the automatic classification of Amazon products by learning from the tags of resources contained within the Open Directory Project categories. This work exploits external metadata by improving classification of items in one domain using text and category information from another domain.

Learning from models of previously annotated items

One approach to classifying social media items based on a training set is to simply consider the content of a post, especially in mediums without rich metadata, such as Twitter. Garcia Esparza et al. (2010) investigated short message categorisation based purely on their content. They carried out experiments on a dataset from Twitter and a dataset from a product review website. Sharifi (2010) performed a similar study on tweet content categorisation. The labelled dataset used in those experiments was obtained by hand-selecting Twitter accounts corresponding to certain predefined categories and retrieving the tweets generated by each account. Rodrigues et al. (2008) classified questions on a question and answering site using the text of the initial question. Pal and Saha (2010) performed multi-label categorisation of blog entries in order to identify posts relevant to certain product groups. They then performed sentiment analysis in order to assess the tone of communication about those products in the blogosphere. All of these studies operated only on the plain text of the posts.

Other studies have improved classification of social media by making use of the metadata of the post. For example, Berendt and Hanser (2007) investigated automatic domain classification of blog posts with different combinations of body, tags and title. They identified tags and body nouns as the most useful features for classification. Sun et al. (2007) investigated the classification of entire blogs, as opposed to individual blog posts. Their experiments compared blog tags against blog title and descriptions as sources of text for a classifier. They concluded that blog tags were a better descriptor of blog topic than titles and descriptions but that classification was improved by including all three. These papers both studied the categorisation of social media posts using the items' own content and metadata.

A relevant study has been performed by Figueiredo et al. (2009), who assess the quality of various textual features in Web 2.0 sites for classifying objects within that site. For Last.fm artists, and videos from Yahoo! Video and YouTube, they found that tags were always the best single feature for classification, although a combination of features generally performs better. They found that a bag-of-words approach for combining feature vectors of different text sources tended to slightly outperform a concatenation approach. They did not investigate the use

of any external data.

Huang et al. (2010) conducted a study which attempted to identify extremist videos on YouTube. They tested various classifiers based on many lexical, syntactic and content-based features, from user-generated text including object titles, descriptions and comments. They included certain tags and categories as binary features in their classifier, but did not compare the improvements gained from different text sources.

A study of event-driven classification of images carried out by Firan et al. (2010) who compared the timestamp and textual features from Flickr for classifying photos into events. The events were extracted from the YAGO ontology (Suchanek et al., 2007) and the Upcoming events website. A combination of all features provided the best results. The authors also noted that classification based on all text features (title, tags and description) performed only marginally better than classification based on tags alone.

There is also existing work on classifying Web documents based on structural parts of their content. Riboni (2002) tested different text sources for webpage representation – the body, title and meta tag content – and found that classification based on a combination of the meta tag and title gave best results. Othman et al. (2010) performed a similar study but found that adding the document body to the meta tag and title resulted in the highest accuracy. Golub and Ardö (2005) compared page title, headings, meta tag and body text with the aim of determining how they affect automated classification. They found that title was the best single indicator of topic, but for best results, all of these elements should be included. These results show that for Web documents, the structure of their text content is relevant for topic categorisation.

Incorporating information from hyperlinks

Some model-based approaches to determining the topic of social media items also integrate information from hyperlinks. Antonelli and Sapino (2005) proposed a rule-based approach to classifying message board topics by type (*e.g.*, announcement, question, answer). Their rules use metrics including the similarity of the post to the title and meta tag of hyperlinked webpages. Irani et al. (2010) studied the problem of identifying posters who aim to dishonestly gain visibility by misleadingly tagging posts with popular topics. They built models that correspond

to topics in order to identify messages that are tagged with a topic but are in fact spam. They took advantage of hyperlinks by augmenting their models with text from webpages linked to within posts.

Similarly, some studies on Web document classification have compared the effects of incorporating various parts of neighbouring websites on classification accuracy. Many of these focus on citing or inlinking pages, those which contain hyperlinks to the target webpage. An early study by Attardi et al. (1999) proposed Web document classification using the titles, incoming anchortext, and text surrounding hyperlinks from citing pages, and reported that they achieved encouraging experimental results. Fürnkranz (1999) compared the results of webpage classification based on content, to the results of classification based on parts of citing webpages. They considered anchortext, the paragraph containing the hyperlink, and the headings that structurally precede the hyperlink. The highest-scoring classifier was that based solely on the anchortext and headings of citing pages. Ghani et al. (2001) investigated website classification using a method which exploits patterns in hyperlink structure. They noted that results were improved by the inclusion of metadata, titles and words from hyperlinked pages. Glover et al. (2002) performed a study which compared document text, incoming anchortext, and the text surrounding incoming anchortext as sources for a text classifier. Their results showed that classification based on anchortext gave comparable results to classification based on the text of the target document, and a combination of text from both the original document and the citing document gave best results. Another study by Sun et al. (2002) compared document text, title, and incoming anchortext and found that the combination of title and anchortext resulted in the highest accuracy. Lim et al. (2005) investigated the task of genre classification on the Web (*e.g.*, personal homepages, FAQs, image collections). They represented each document using combinations of features from the URL, anchortext, titles, meta tag and body. Qi and Davison (2008) performed experiments which compared different weightings of titles, anchortext, extended anchortext and full text, from the target page and its neighbours. The term neighbours includes citing pages, cited pages, co-citing pages and co-cited pages. They found that using fielded information as opposed to the full text of webpages improved classification accuracy.

4.4 Conclusions

This chapter has completed the background section of the thesis by describing how structured information can be attached to social media items by way of manually assigned or automatically generated metadata. In the following core chapters, we will focus on the tasks of predicting tag, location and topic metadata for social media posts. In Chapter 8.1, we will show how the three approaches can be combined to augment a social media item with each of these three complementary metadata types.

Part III

Core

Chapter 5

Tag Prediction using Anchortext*

5.1 Introduction

On the HTML Web, publishers can provide references to external resources of interest by creating hyperlinks. Web authors may optionally label these resources using anchortext, the text which is visible and clickable in a hyperlink. Often anchortext provides a description of the topic of the page which is the target of the hyperlink. Anchortext has been extensively used by search engines as a rich source of Web page annotations to improve search quality, for example by Brin and Page (1998), who claim that anchortext can often provide a better description of a Web page than the page itself. A study by Craswell et al. (2001) found that in their experiments, ranking based on link anchortext was twice as effective as ranking based on document content. Eiron and McCurley (2003) argue that the reason why anchortext is so effective for Web search is because it provides a short and precise summarization of the target pages. Thus anchortext provides a way of implicitly annotating resources, although it can also be used for simple navigational purposes, for example *click here*, or *read more*.

The creation of anchortext is limited to those Web users who author content on the Web, which substantially restricts the user community. However it is not essential to be able to write HTML or host Web pages in order to create anchortext. WYSIWYG (What You See Is What You Get) interfaces and simple syntaxes such as BBcode for message boards and Wiki markup for wikis make

*This chapter is partially based on (Kinsella, Budura, et al., 2008)

it easier for non-technical users of Web 2.0 services to contribute anchortext annotations for Web resources.

Social bookmarking services, on the other hand, enable explicit annotation of Web resources. Users may annotate resources with keywords or terms that somehow relate to the document. These tags can be used to search or navigate a collection of bookmarks, and provide easy-to-read summaries for the bookmarked documents. Tag clouds (such as those shown in Figures 5.1 and 5.2) are often used as an interface to allow users to browse articles by their annotations. Social annotations have also been proven effective for Web search (Bao et al., 2007; Zhou et al., 2008). Like anchortext, social bookmarks are often used to describe the topic of the document, however users also often use bookmarks for personal annotations, such as *toRead*. Compared to anchortext, social bookmarking services provide an easier to use and more accessible way of annotating Web resources and thus open up the annotation process to a larger community.

Despite the advantages of tagging over anchortext as a source of annotations, anchortext is still a potentially useful source of information about online resources. It is reasonable to expect that given the similarities between anchortext and tagging, many of the tags manually assigned on social bookmarking sites could in fact be inferred from anchortext already available on the Web. It is also possible that anchortext can provide novel information about Web resources in combination with tags.

Figures 5.1 and 5.2 show examples of tag clouds for the websites of Virgin Radio and the British Broadcasting Corporation (BBC), which were generated from the datasets presented later in this chapter. The tag clouds on the left are generated from the most popular occurring keywords from anchortext for these websites. The tag clouds on the right are generated from the most popular occurring keywords from Delicious for these websites. It can be seen that each set of tags reflects the usage pattern of their source. For example, the tag cloud from anchortext for the Virgin Radio homepage contains the term *listen* which is often used when linking to online radio stations, while the tag cloud from Delicious contains organisational terms like *entertainment*. The names of organisations occur in both sets of tag clouds, but are more frequent in anchortext. Despite these differences, there is a substantial overlap of terms derived from anchortext with those assigned in Delicious.

(a) (b)

Figure 5.1: Tag clouds for the Virgin Radio website (`http://www.virginradio.co.uk/`) generated from (a) anchortext; (b) social bookmarks

(a) (b)

Figure 5.2: Tag clouds for the BBC website (`http://www.bbc.co.uk/`) generated from (a) anchortext; (b) social bookmarks

We propose taking advantage of the anchortext already existing on the Web in order to automatically generate annotations for untagged Web resources. These annotations could be presented to a user who bookmarks a page as tag suggestions, reducing the effort involved in bookmarking. One study by Suchanek et al. (2008) found that approximately one third of tag applications seemed to be a consequence of tag suggestions, indicating that tag suggestions are a useful feature for social bookmarking service users. Predicted tags could also be used to bootstrap a new system or augment the tag set in an existing system. Thus Web documents can be annotated with useful tags without duplicating the effort that has already been made to annotate these resources via anchortext.

5.2 Approach

The first step of our approach to automatically tag a social media item is the retrieval of HTML documents which link to the item. From these documents, we extract the anchortext of the hyperlink(s) which describe the item to be tagged. If the item to be tagged is a HTML document, we also include text from the META element. We then perform preprocessing on these Web annotations (for example, stopword removal) and we rank the remaining terms and select the top k as tags. Figure 5.3 shows a flowchart of the processes involved in tag prediction for a social media item, which we now describe in more detail.

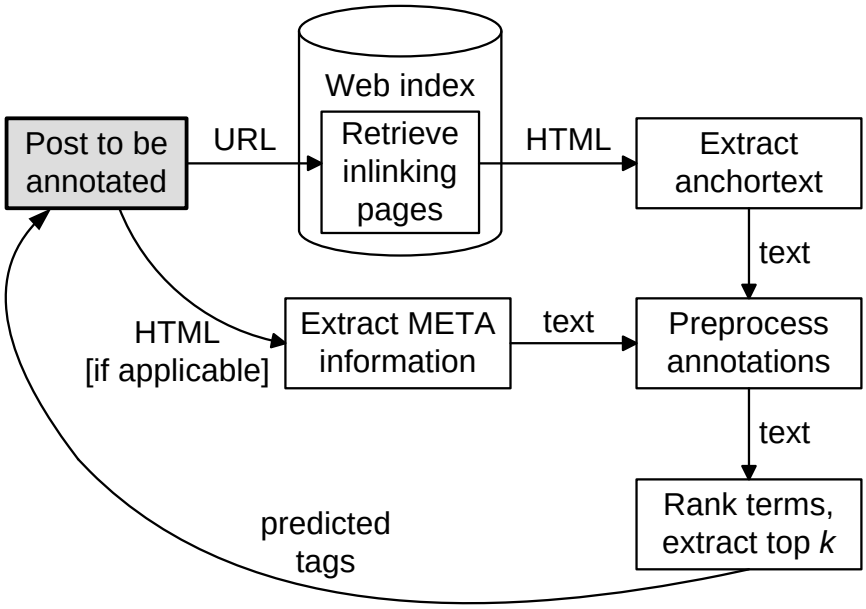

Figure 5.3: Flowchart for the tag prediction approach

5.2.1 Data collection and pre-processing

In order to assemble a collection of textual annotations describing a Web page, we identify incoming hyperlinks to the document and retrieve all available anchortext. This data could be extracted from a Web crawl, if available. Alternatively a search engine could be used to retrieve all incoming hyperlinks to a document, which would then be downloaded to obtain their content. Although this study focuses on social annotations, we also make use of HTML metadata within the

original document, if present. We process the page header and include any text from the TITLE element and text from the keywords and description properties within the META element. The aim is to make use of any already available annotations on the Web. We use the phrase "Web annotations" to refer collectively to the incoming anchortext and HTML metadata of a webpage.

The following preprocessing steps are applied to the extracted annotations. The text is converted to lower-case and punctuation is removed. All terms are omitted except those which are at least two characters long, contain one or more letters, and consist only of alphanumeric characters. We remove stopwords according to a widely-used stopword list of frequently occurring words in the English language[1] and additionally omit some Web-related terms that frequently occur in HTML documents (*e.g.*, *www*, *click*). Finally, a stemmer is applied to ensure that words with the same root (such as *environments* and *environmental*) are considered to be the same.

5.2.2 Tag indexing and ranking

The document annotations are indexed using the vector space model. Document vectors are typically built from document content, but for our approach we generate the vectors from the preprocessed anchortext and metadata. Each document is represented as $d = (t_1, t_2 ... t_{n_d})$ where the entries in the vector are terms in the vocabulary, the value t_i denotes the weight of term i in document d, and n_d is the number of terms in d.

We experimented with two different weighting methods for determining the value for $t_{i,j}$ – term frequency (tf) and term frequency-inverse document frequency (tf-idf). The term frequency $\text{tf}_{t,d}$ for a term t and a document d is defined as the number of times that t occurs in the annotations of d. The default method of ranking tags for display in the Delicious website is tf. The tf measure is a very simple approach to determining term weights, and does not consider the fact that certain terms tend to be less meaningful than others. For example, if a term occurs for almost every document in a collection, it is unlikely to be highly relevant for many of them.

In order to attenuate the effect of very common terms, we can take into

[1] `ftp://ftp.cs.cornell.edu/pub/smart/english.stop`, accessed July 2011

account the document frequency (df) of a term, *i.e.*, the number of documents in the collection whose annotations contain that particular term. The inverse document frequency idf_t for a term t is calculated by dividing the total number of documents in the collection N by the natural logarithm of df_t, and adding one to the denominator to avoid division by zero if a term does not occur in the corpus:

$$\text{idf}_t = \log \frac{N}{1 + \text{df}_t} \tag{5.1}$$

Thus idf_t is a measure of the rarity of t.

Term frequency-inverse document frequency (tf-idf) is the product of the previous two measures, where the idf factor is used to decrease the influence of very popular terms:

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \tag{5.2}$$

For tag ranking, the tf and tf-idf measures can both be applied either to terms extracted from Web annotations or to terms obtained from social bookmarking services.

After generating a vector of term weights for each document, using either tf or tf-idf, we consider the top ranked k terms in each document vector as the candidates for tag suggestion, where k is the desired number of predicted tags. In the remainder of this chapter, "predicted tags" denotes those tags that we have extracted from Web annotations, while "Delicious tags" denotes those that users have assigned to URLs in Delicious.

5.2.3 Comparison with previous approaches

In Section 4.3.1 we reviewed existing methods of automatic tag generation. Most of those are either based on document content (Medelyan et al., 2009; Pudota et al., 2010), or else learning from previously tagged resources (Mishne, 2006; J. Wang & Davison, 2008; Budura et al., 2009; Lu et al., 2009), but they generally do not use the already existing social annotations which are available for many resources as anchortext.

The tag prediction approach which is most related to ours is that of Heymann, Ramage, and Garcia-Molina (2008). Their method used a classifier to predict tags for Web resources from the 100 most popular tags on Delicious, based on

the anchortext of incoming links, document content, inlinking and outlinking hosts and domains, and tags which were previously assigned to the resource. The classifier was trained using documents which had a large number of annotations. Our method in contrast considers only anchortext and metadata extracted from the Web page header and we use an easy-to-apply method which does not rely on classification, and therefore is not limited in terms of the number of distinct tags we can consider. We do not have any initial conditions on the number of times a resource has been bookmarked, our only requirement being that its URL is present in both the of datasets that we use to run our experiments: a crawl of Delicious and a crawl of the Web. In addition, we do not restrict ourselves to a predefined set of tags. By doing so we could expect a gain in precision; however, this would restrict our general approach to only a small number of commonly used tags, and would exclude niche, domain-related and highly specific tags.

Our approach could be combined with others which make use of alternative features such as document content (Pudota et al., 2010), tag co-occurrence (Sigurbjörnsson & Zwol, 2008), document similarity (Budura et al., 2009) or social networks (Rae et al., 2010). In particular, anchortext could be of use in cases where there is little or no textual content (*e.g.*, photos or short documents), and no preexisting tags.

5.3 Data Corpus

5.3.1 Dataset description

For the purpose of evaluating the approach, a large collection of interlinked Web documents with corresponding tags was required. Two datasets were combined to generate the corpus:

Web collection. The WEBSPAM-UK2007 dataset (Yahoo! Research, 2007) is a crawl of the `.uk` domain consisting of approximately 106 million pages. A summary version containing a subset of 12 million pages is also provided. In the following experiments, the full dataset was used for calculating the indegree distribution, while for all other experiments the summary version was used.

Delicious dataset. The experiments in this chapter make use of a crawl of Delicious that was carried out in 2007. The dataset contains tags for approxi-

mately 4.5 million URLs.

For our analysis, we require only those URLs that (a) exist in the Delicious dataset and are tagged; and (b) have incoming hyperlinks in the Web collection which incorporate anchortext. Over 184,000 URLs fulfil these requirements. For each of these URLs, the Delicious dataset contains the complete set of Delicious tags assigned to that resource at the time of the crawl. However the `.uk` domain represents only a small fraction of the Web. Therefore for most URLs, the Web collection contains only a small fraction of the corresponding anchortext available on the Web. In addition, many of the hyperlinked URLs are from outside of the `.uk` domain and are not themselves present in the Web collection, so HTML document metadata for these is missing. Therefore the quality of the tags that can be inferred is probably lower than could be achieved with a complete crawl, however we aim to show that even with this partial Web dataset we can still report interesting results.

In total, we were able to extract approximately 50MB of textual annotations from anchortext, and approximately 4MB from HTML document metadata. Hence anchortext has a much larger influence on the tags predicted in this study.

5.3.2 Data characteristics

We now report and compare some characteristics of the tags extracted from Delicious, and the tags predicted from Web annotations (anchortext and HTML metadata).

Table 5.1 shows the top 20 terms from Web annotations and from Delicious tags for the documents which occur in both the Web collection and Delicious dataset. Therefore the table compares the most popular terms that were applied to the same set of documents as tags and as Web annotations. The text has been lower-cased and punctuation removed, but stopwords have not been omitted. There is some similarity between the two lists (*uk* appears on both, *lib* and *dems* occur for Web annotations and *politics* occurs for Delicious) but in general they have different characteristics. Most of the top terms from Web annotations are related to Web technologies (*e.g.*, *phpbb*, *xhtml*) or navigation (*e.g.*, *click*, *online*), while most of the top Delicious terms appear to be used for broad categorisation (*e.g.*, *reference*, *fun*). The Web annotations list also consists of many stopwords

Web annotations			Delicious Tags	
	Term	Count	Term	Count
1.	the	168,332	1. reference	22,239
2.	of	94,513	2. software	18,311
3.	phpbb	90,952	3. design	16,164
4.	xhtml	70,563	4. tools	14,843
5.	and	67,366	5. music	12,935
6.	national	65,925	6. art	12,526
7.	for	57,302	7. blog	12,183
8.	to	53,673	8. news	11,953
9.	by	51,975	9. research	10,982
10.	lib	49,027	10. technology	10,947
11.	web	47,953	11. programming	10,802
12.	dems	47,824	12. science	10,366
13.	css	45,572	13. fun	9,488
14.	valid	44,420	14. politics	9,065
15.	online	40,669	15. uk	8,996
16.	here	40,041	16. blogs	8,581
17.	website	35,350	17. webdesign	8,535
18.	adobe	34,916	18. education	8,441
19.	uk	32,160	19. imported	8,437
20.	reader	30,701	20. cool	8,357

Table 5.1: Top 20 Web annotation terms and Delicious tags ordered by number of occurrences

(*e.g.*, *the*, *and*), since unlike tags, anchortext is often composed of full sentences. With regard to the most popular terms overall, the two data sources appear quite different.

Figure 5.4 shows the cumulative distribution of inlinks in the Web collection and Figure 5.5 shows the cumulative distribution of the number of times URLs are tagged in the Delicious dataset. In both cases it appears that part of the curve follows a power law distribution. A line has been fitted to each curve proportional to the power law function. Since the figures show cumulative distributions, the function fitted is $k^{\alpha-1}$ rather than the usual power law function, $k^{-\alpha}$. The slope of the line fitted in Figure 5.4 in the range of $y \in [0.01, 10^{-7}]$ is 0.99, indicating a power law exponent α of 1.99. This is comparable to previous exponent of 2.1 calculated for the whole World Wide Web (Broder et al., 2000). The slope of the line fitted in Figure 5.5 in the range of $y \in [0.1, 10^{-6}]$ is 1.16, indicating a power

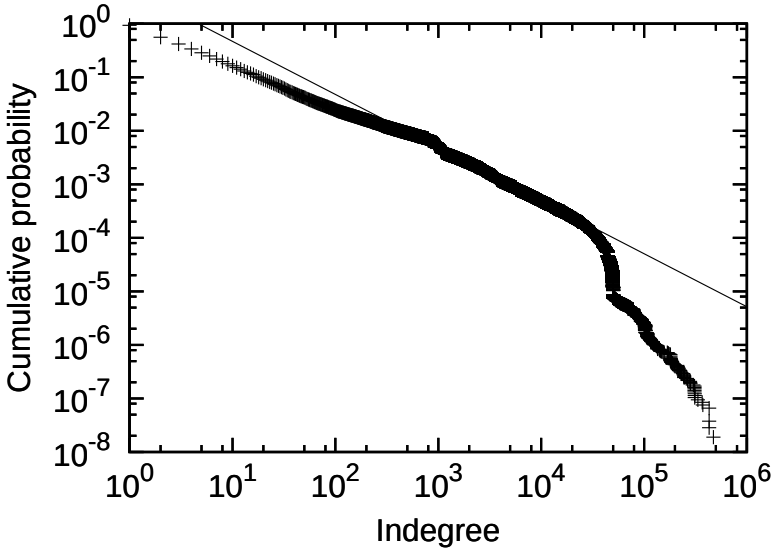

Figure 5.4: Cumulative distribution of inlinks of documents in Web collection

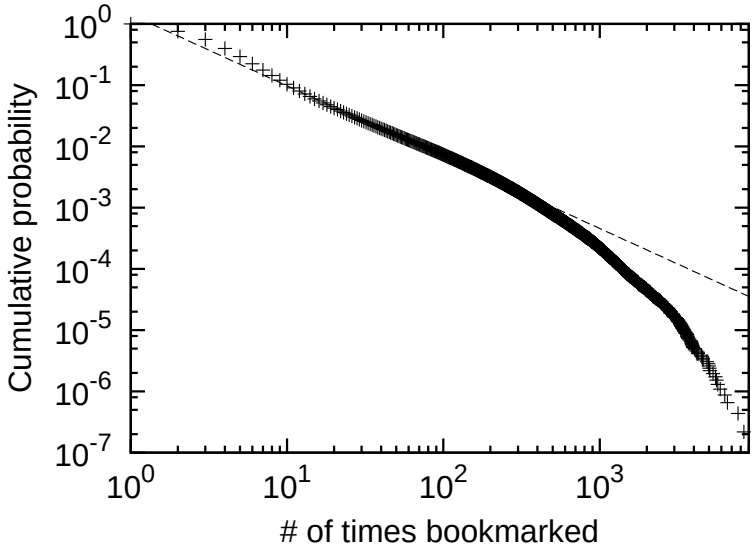

Figure 5.5: Cumulative distribution of document bookmarks in Delicious dataset

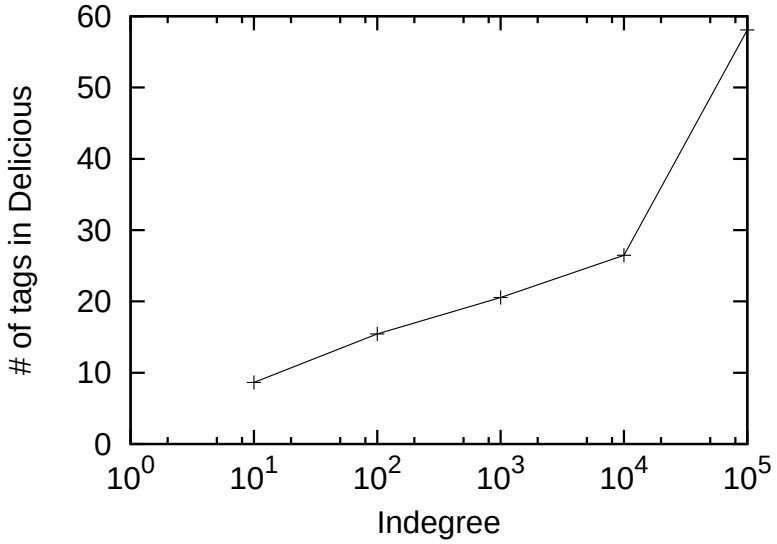

Figure 5.6: Number of Delicious tags per document, averaged by indegree

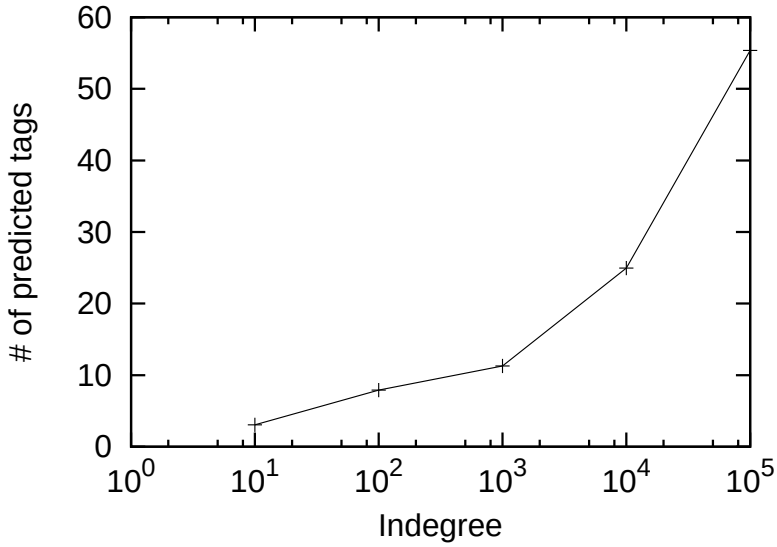

Figure 5.7: Number of predicted tags per document, averaged by indegree

law exponent α of 2.16. From Figures 5.4 and 5.5 it can be seen that the two datasets share a quite similar distribution, indicating that user patterns in tagging follow user patterns in creating hyperlinks to pages. Many Web documents are linked to or tagged very few times, but a small number of documents attract many links or tags.

Figure 5.6 shows the relationship between the indegree of Web documents and the number of tags that they receive in Delicious. The tag counts are averaged across documents in order to smooth the distribution. The figure shows that documents that have an indegree between 1 and 10 receive an average of 9 tags, while documents that have an indegree between 10,000 and 100,000 receive an average of 58 tags. The average number of tags assigned to documents increases with indegree, showing that pages which receive many inlinks also tend to receive many tags.

Figure 5.7 shows the relationship between the indegree of Web documents and the number of tags that we can predict using the approach described in Section 5.2. The predicted tag counts are again averaged across documents to smooth the distribution. Since a higher number of inlinks generally results in a greater amount of anchortext, the more popular a document is, the greater the number of tags that can be inferred from anchortext. Comparing Figures 5.6 and 5.7 it can be seen that on average, documents of all indegrees tend to have a higher number of tags in Delicious than can be inferred from Web annotations, at least for the dataset that we analysed. Note that this dataset only covered documents from the `.uk` domain; with a more complete Web crawl, a greater number of terms from anchortext and HTML metadata would be available.

5.4 Evaluation

In order to assess the quality of the predicted tags, two sets of experiments were carried out.

Firstly, an automatic evaluation was performed, where tags predicted from Web annotations were compared to the tags which were actually assigned to each document on Delicious. This allows us to measure the extent to which our approach generates the same tags that users provide. This evaluation assumes that tags that have been assigned by Delicious users are indeed relevant annotations

for their corresponding URLs.

Secondly, predicted tags and actual tags were presented to human evaluators who were requested to judge their relevance. This manual evaluation was motivated by the fact that even if a tag is not present in Delicious for a certain document, it may still be a valid annotation. The human evaluation also enables us to assess the extent to which Delicious annotations are considered relevant by users.

5.4.1 Automatic evaluation

For the automatic evaluation of predicted tags, the ground truth for each document was considered to be the set of tags retrieved from Delicious. Precision@k and recall@k were calculated relative to this ground truth. These metrics are not a measure of absolute relevance, but a measure of how the predicted tags compare to the tags users assigned in Delicious. For the purpose of calculating relative recall@k, we retain the tag ranks as defined by default on the Delicious website, *i.e.* according to their frequency (tf). Relative precision and relative recall are defined as follows:

Relative precision@k: The average proportion of the top-k predicted tags that are also assigned to the corresponding document in Delicious, averaged over all documents.

Relative recall@k: The average proportion of the top-k tags in Delicious that are also predicted from a document's Web annotations, averaged over all documents.

We report the results of relative precision@k and relative recall@k for all values of k from 1 to 5. Therefore, we include in our experiments only those documents that have at least 5 predicted tags and 5 Delicious tags. After excluding documents which do not met these criteria there were 14,058 documents remaining in the dataset.

Table 5.2 and Table 5.3 show the results of relative precision@k and relative recall@k for the tf and tf-idf weightings of predicted terms. The relative recall@k results are the same for both ranking schemes, since for recall the ranks of the

predicted tags are irrelevant. The relative precision@k results shows that tf provides a better ranking than tf-idf for the purpose of tag prediction from Web annotations. Using the tf method, 48% of the top 1 predicted tags are among those which have been assigned by users in Delicious. Regarding the other 52% of top 1 predicted tags, it is unclear from the automatic evaluation how many of these would in fact serve as useful annotations. It is likely that some of these would be appropriate tags, but have not yet been assigned in Delicious, particularly in those cases where a document has very few tags. The relative recall results shows that on average, 41% of the top 1 tags and 28% of the top 5 tags assigned to a document in Delicious can be inferred from anchortext or metadata. This indicates that a significant portion of the tags in Delicious are somewhat redundant for search purposes, as search engines would already have located these terms on the Web. There is also a large portion of Delicious tags which are novel compared to the Web annotations and would thus be potentially valuable for search purposes.

	k=1	k=2	k=3	k=4	k=5
tf	**0.48**	**0.45**	**0.42**	**0.39**	**0.37**
tf-idf	0.36	0.36	0.35	0.35	0.34

Table 5.2: Relative precision@k for tf and tf-idf

k=1	k=2	k=3	k=4	k=5
0.41	0.35	0.32	0.29	0.28

Table 5.3: Relative recall@k for tf and tf-idf

5.4.2 Human evaluation

For the human evaluation, a test set consisting of 80 documents was sampled from the corpus. Since for some URLs there are very many predicted tags, and for others there are much less, we wanted to ensure that the evaluation included an even spread of URLs across this range. Therefore, 20 Web pages were chosen from each of the following ranges: (5, 6-18, 19-60, 61+).

A group of 25 judges were asked to assess the relevance of the top ranked tags for these pages. The top 5 tags from Delicious and the top 5 predicted tags were

manually rated, for both the tf and tf-idf weighting methods. Each document and tag pair was evaluated by three judges. The judges were asked to assign each tagging instance a score of either 0, 1 or 2, where 0 indicates that a tag is "not relevant", 1 indicates that a tag is "quite relevant", and 2 indicates that a tag is "very relevant".

Figure 5.8 shows the distribution of 0, 1 and 2 scores for each method of tag extraction and ranking. The predicted tags have a greater number of non-relevant ratings than the Delicious tags (for tf, 27% versus 18%). However the gap in "very relevant" ratings assigned is lower, with 42% of predicted tags and 47% of Delicious tags judged as relevant, for tf. Delicious tags also receive slightly more "quite relevant" ratings (for tf, 36% versus 31%). The two ranking methods tf and tf-idf yield fairly similar results from the user's point of view, with tf resulting in slightly less non-relevant ratings.

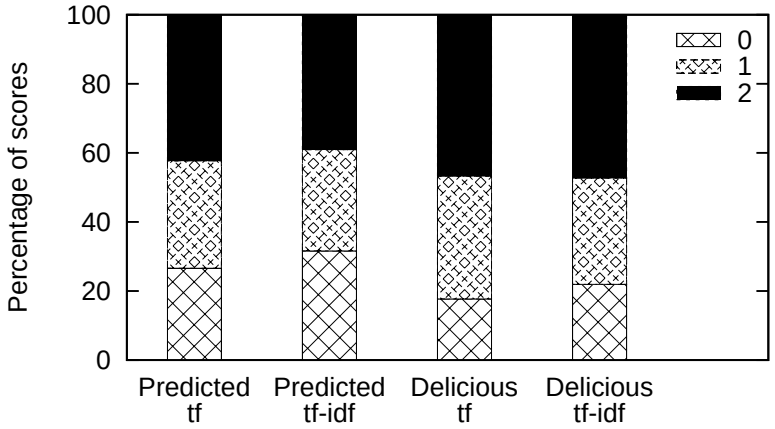

Figure 5.8: Distribution of user 0, 1 and 2 scores for each method of tag extraction and ranking

The relevance of both the Delicious tags and the predicted tags were measured using precision. Recall is not reported since the total number of relevant tags for a document is potentially unlimited. It is possible to ask evaluators to confirm a tag as relevant (as is required for measuring precision), but it is not feasible for them to draw up an exhaustive list of possible relevant tags for each document (as is required for measuring recall). Since the relevance of each tag was manually evaluated, we consider the resulting measure as absolute precision, as opposed to the relative precision that was earlier reported with respect to the Delicious tags. Absolute precision@k is defined as follows:

Precision@k: The average proportion of the top-k tags that are judged as relevant by evaluators.

Table 5.4 shows the results where tags that scored on average 1.5 or above are considered to be relevant. This is a high relevance threshold and the results are therefore an indicator of accuracy when only very relevant tags are desired. For Delicious tags, the best precision@5 score is 0.43, indicating that on average, only 43% of the top 5 tags assigned to documents are evaluated as highly relevant. This indicates that many of the tags that users assign in Delicious are not considered very useful by other people, although they may be useful to the original creator, or they may be quite general descriptions of the resource's characteristics. For the predicted tags, the best precision@5 is 0.36, which is quite low but is still adequate in comparison to the precision@5 of 0.43 achieved by the user-assigned Delicious tags. The differences between predicted tags (tf) and Delicious tags (tf) were statistically significant for precision@3 and precision@4, but not for other values of k. Significance was determined using a paired t-test with $p < 0.05$.

Table 5.5 shows the results where tags that scored on average 1.0 or above are considered to be relevant. This is a lower relevance threshold and the results are therefore an indicator of accuracy where somewhat relevant tags are acceptable. The more relaxed relevance threshold results in higher scores for both Delicious tags and predicted tags. The best precision@5 score for Delicious tags is 0.78, and the best precision@5 score for predicted tags is 0.66. As in Table 5.4, the Delicious tags are considered more relevant than the predicted tags, with statisically significant differences for precision@2-5, however once again the gap is not extremely large.

Table 5.6 shows the precision obtained when we consider only those URLs for which at least 19 tags can be predicted, with a relevance threshold of 1.0. In these cases precision@1 for tags predicted by tf rises from 0.78 to 0.85. Our method of tag prediction even outperforms Delicious tags for precision@1, however for higher values of k the Delicious tags are judged as more relevant. The differences between predicted tags and Delicious tags are not significant for precision@1-4, however for precision@5 Delicious tags are rated significantly better. These results indicate that for documents where there is a large amount of anchortext available, the accuracy of predicted tags improves considerably. With a large

Tags	Ranking	k=1	k=2	k=3	k=4	k=5
Predicted	tf	0.48	**0.46**	**0.39**	**0.39**	**0.36**
Predicted	tf-idf	**0.54**	0.41	0.38	0.35	0.33
Delicious	tf	**0.61**	**0.52**	**0.50***	**0.45***	0.42
Delicious	tf-idf	0.55	0.49	0.47	0.44	**0.43**

Table 5.4: Precision@k when tags with average rating 1.5 or above are considered to be relevant. * indicates statistical significance w.r.t. predicted tags (tf) with paired t-test $p<0.05$

Tags	Ranking	k=1	k=2	k=3	k=4	k=5
Predicted	tf	0.78	**0.76**	**0.69**	**0.67**	**0.66**
Predicted	tf-idf	**0.80**	0.70	0.66	0.61	0.60
Delicious	tf	**0.86**	**0.84***	**0.82***	**0.80***	**0.78***
Delicious	tf-idf	0.75	0.75	0.75	0.75	0.74

Table 5.5: Precision@k when tags with average rating 1.0 or above are considered to be relevant. * indicates statistical significance w.r.t. predicted tags (tf) with paired t-test $p<0.05$

Tags	Ranking	k=1	k=2	k=3	k=4	k=5
Predicted	tf	**0.85**	**0.80**	**0.73**	**0.69**	**0.67**
Predicted	tf-idf	0.80	0.66	0.60	0.53	0.52
Delicious	tf	**0.83**	**0.83**	**0.82**	**0.78**	**0.76***
Delicious	tf-idf	0.68	0.73	0.73	0.73	0.73

Table 5.6: Precision@k for URLs for which we can predict at least 19 tags. Tags with average rating 1.0 or above are considered to be relevant. * indicates statistical significance w.r.t. predicted tags (tf) with paired t-test $p<0.05$

Web dataset of the scale that search engines possess, even better results should be possible.

With regard to term weighting schemes, differences between tf and tf-idf were generally not statistically significant. However for Delicious tags, the default tf method seems to outperform the alternative tf-idf method, although in Table 5.4, a higher precision@5 was achieved using tf-idf. For the predicted tags it is not clear which ranking method is better. The tf method generally outperforms tf-idf, however in Tables 5.4 and 5.5, a higher precision@1 was achieved using tf-idf. It may be useful to experiment with additional weighting schemes to determine the optimal way of ranking tags generated from Web annotations.

An important aspect of tagging is its inherent subjectivity. Different users can have different opinions on whether or not a tag is relevant or not. Table 5.7 shows the breakdown of annotator agreement for the tag relevance assessment task. Evaluators agreed completely in only 35% of cases. However they at least almost agreed in nearly 85% of cases. In the remaining 15% of cases there was no clear consensus between evaluators. This suggests that sometimes it is impossible to precisely measure the true relevance of a tag, since each individual has their own personal idea of relevance.

Extent of agreement	Frequency
All 3 scores agree	35.0%
All 3 scores almost agree	49.9%
All other cases	15.1%

Table 5.7: Breakdown of evaluator agreement. "Almost agree" means that 2 scores are equal and the other score is just one point higher or lower

5.5 Conclusions

We have presented a comparison of the usage of anchortext and tags on the Web, and an approach for inferring tags from the Web annotations that already exists online. We observe that there is a substantial overlap between the tags assigned on a popular social bookmarking service and the anchortext for these pages. A human evaluation has confirmed that the relevance of tags predicted from anchortext is not far behind that of human-assigned tags. In particular

for webpages that have a large amount of incoming anchortext, the precision of our approach is comparable to that of the tags from Delicious. Our approach enables existing annotations on the Web to be reused as tag suggestions in order to reduce the effort involved in tagging a document collection.

Our experiments focused on tagging documents, since we had a large Web crawl available for evaluation purposes. The results show that anchortext can be a useful data source for predicting tags for social bookmarking sites. However the approach could also be applied to other social media items, in particular those types of items which most often attract hyperlinks, such as blog posts, videos (*e.g.*, from YouTube) and photos (*e.g.*, from Flickr). Tags from an external source are very useful because creators of social media items often do not go to great lengths to exhaustively annotate their own content, often focusing more on maintaining a consistent vocabulary and allowing easy navigation within their own archive. Thus the generation of tags from anchortext can potentially improve retrieval in social media by adding novel relevant terms to a post's annotations.

Chapter 6

Geolocation using Language Models from Geotags*

6.1 Introduction

In the previous chapter we studied tag prediction for resources, using only textual descriptions extracted from anchortexts on the Web. Now we investigate the problem of predicting the location of a social media post. Like Chapter 5, the approach presented in the current chapter makes use of texts generated by a community of Web users, but also requires the location data given by geotags. Therefore we again reuse existing, community-created Web data, but now also make use of additional information provided by Web 2.0 services.

Geographic information about social media items is important for location-related search tasks such as seeking updates on local news or conducting research in preparation for travel. Recently, the ability to add location metadata to user-generated content, and the proliferation of GPS-enabled devices such as smartphones, means that social media sites can often provide exactly this type of geographic data. Flickr, YouTube and Twitter all allow geotags to be added to items, either manually by the user or automatically by a GPS-enabled device. Users of these websites can then make use of the geographic metadata associated with resources to easily locate content originating from the place or region of interest to them.

*This chapter is partially based on (Kinsella, Murdock, & O'Hare, 2011)

However, while geotags are undoubtedly a useful source of information, there is still a long way to go in terms of user uptake. We report later in the chapter that in a 2010 sample of Twitter updates, fewer than 1% were geotagged. Alternative methods of identifying the location of origin are unreliable. Many social media sites, including Twitter, allow users to make their location publically available in their profile, but these self-reported locations are often inaccurate or imprecise. Cheng et al. (2010) report that they found only 21% of Twitter profiles reported a location as granular as a city name, and only 5% reported a location as granular as a latitude/longitude coordinate. Hecht et al. (2011) found that 66% of Twitter users provided any valid location information, and that of these, 64% provided a location as granular as a city name. Almost no users provided anything more precise than city-level location information. Many users (16%) entered non-geographical information, such as pop culture references and non-earth locations. Another finding was that the non-geographical or nonsensical information often entered in the Location field could fool reverse-geocoding systems into incorrectly assigning a location to a user. IP addresses can be used to estimate the location of a user, but these are only available to the social media sites to which content is submitted, so would not be useful for a scenario where data has been aggregated from various sources. Furthermore, it has been found by MaxMind, Inc. (2010) that on a local level, IP address is an unreliable indicator of true user location. For the United States, 83% of addresses can be accurately resolved to within 25 miles, and for Ireland, this figure drops to 61%. On a postal code level, *i.e.* within a mile or two, the accuracy is likely to be much lower. Therefore alternative methods of locating a user, especially at a very local level, are required.

In this chapter, we investigate the potential of using the text of geotagged Twitter posts to determine the location of other, non-geotagged updates, based only on the text that they contain. We use geotagged status updates from Twitter to build language models of geographic regions worldwide, from countries down to postal codes. Our method is not limited to explicitly mentioned geographic entities, but makes use of all text found in the tweets, which can include implicit geographic clues such as regional slang, business names, local specialities, and the language which is used. We investigate how our method compares to alternative approaches such as parsing user status updates for explicit locations, and parsing

the user's self-reported location. Our results show that using language models gives better results than relying on explicit geographic references, and at the postal code level the language models enable us to geolocate a post with even more success than using the user's self-reported location.

6.2 Approach

Our approach involves building language models of locations using a training set of geotagged tweets. In order to build the language models, we first need to reverse geocode each geotag: *i.e.*, map the latitude/longitude pair to some region such as a city or neighbourhood. The tweets originating from each location are aggregated to build corresponding language models, which can then be used to predict the location of new content which does not have geographic metadata. Figure 6.1 shows a flowchart for the processes involved in the location prediction of a social media item, which we describe in more detail in this section.

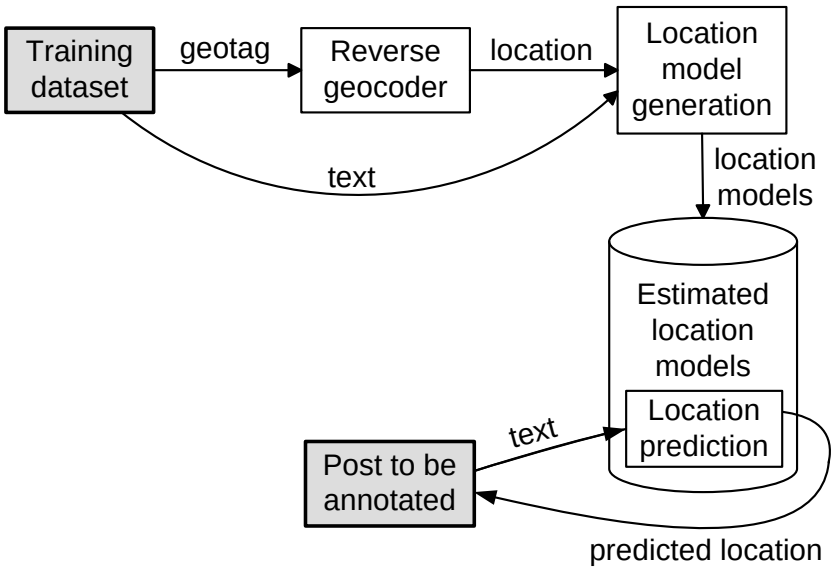

Figure 6.1: Flowchart for the location modelling approach

6.2.1 Reverse geocoding

The first step of the approach is to find the location associated with each tweet, which is done by reverse geocoding the coordinates specified in its geotag. Reverse geocoding involves mapping a point (latitude and longitude) to some defined region with a place name or identifier. It is the opposite of geocoding, which is the process of converting addresses to their associated geographic coordinates. Any reverse geocoding service could be used to perform this task, and different definitions of location could be considered, including very specific regions such as street addresses.

We used the Yahoo! GeoPlanet Web service[1] to transform a latitude/longitude pair into a spatial entity. GeoPlanet resolves the point to a unique place and also returns larger regions that contain the point. For example, a lookup for the coordinates (53.3,-9.1) will return the neighbourhood Bushypark, the city of Galway, the county of Galway and the country of Ireland. The place types that are used in the experiments later in this chapter are listed and defined in Table 6.1.

Place Type	Description	Synonyms
Country	A country (in ISO 3166-1 standard)	
State	A primary administrative area	province, region
City	A major populated place	town, village
Neighbourhood	A subdivision within a city	suburb, ward
Postal Code	Code assigned to address for sorting mail	zip code

Table 6.1: Place types included in the geolocation experiments

6.2.2 Language modelling and location prediction

The second step of the approach is to build a language model describing each location. A language model is a probability distribution over a set of terms. In natural language processing applications such as speech recognition and machine translation, language models are used to describe the characteristics of a language and thus aid in the prediction of the next word in a sequence. Language modelling is also commonly used in information retrieval, in which case a language model is

[1] `http://developer.yahoo.com/geo/geoplanet/`, accessed July 2011

assigned to each document in the collection. For information retrieval purposes, the order of words is not of vital importance, so unigram models are typically used. Each document is taken as a bag of words and the language model generated is a multinomial probability distribution over the vocabulary of terms in the document. The retrieval process is modelled as a random sampling of terms from some document – *i.e.*, the user has some ideal document in mind and they attempt to guess some terms from that document in order to formulate their query. Thus the language modelling approach does not explicitly consider the notion of relevance, but instead ranks documents by the probability that each document would generate a given query text under independent and identically distributed random sampling.

We use the language modelling approach of Ponte and Croft (1998) to build models of locations. Our experiments use two different but equivalent implementations for the location prediction problem, based on query likelihood and Kullback-Leibler divergence. The reason for using both of these approaches was that an implementation of Kullback-Leibler divergence was freely available as part of an open-source package and was suitable for small scale experiments, however in order to scale up to larger experiments, it was necessary to use an alternative implementation of query likelihood.

Query likelihood measures the probability that a document in a collection is relevant to a certain query. In our application, locations are treated as documents and posts are treated as queries. For each location l, we build an estimated model M_l based on the distribution of terms found in posts originating from that location. Then, given a post p, our goal is to rank the locations by $P(l|p)$, the probability that each location generated this post. Using Bayes' theorem we can express the probability that a location generated a post as:

$$P(l|p) = \frac{P(p|M_l)P(l)}{P(p)} \tag{6.1}$$

We assume that $P(l)$, the prior probability of location is distributed uniformly (though we later explain how the smoothing applied to our models considers document length and therefore the freqency of previous tweets from each location). We ignore P(p), the prior probability of the post, since it is the same for all locations. Therefore the model reduces to $P(p|M_l)$, which we estimate as the

probability that a random sampling of M_l would generate the post. We use a multinomial unigram language model to model the probability distribution over each term t:

$$P(p|M_l) = \prod_{t \in p} P(t|M_l) \tag{6.2}$$

The probability of a term for a location, $P(t|M_l)$ may be calculated simply as the term frequency for that location, $\mathrm{tf}_{t,l}$, divided by the total number of terms in that location, $|l|$, *i.e.*,

$$P(t|M_l) = \mathrm{tf}_{t,l}/|l| \tag{6.3}$$

However this means that if a post relevant to a certain location were to contain a previously unseen term, the probability for that term would be zero and hence the probability for the whole post would also be calculated as zero. Therefore we introduce a smoothing function in order avoid zero probabilities. We use Dirichlet smoothing as described in (Zhai & Lafferty, 2004). The resulting model combines the probability of a term occurring in a location with the probability of it occurring in the whole collection:

$$P(t|M_l) = \frac{\mathrm{tf}_{t,l} + \mu P(t|M_c)}{|l| + \mu} \tag{6.4}$$

where M_c is a language model built from the entire collection, and μ is a parameter whose optimum value should be determined experimentally. Dirichlet smoothing takes into account document length, with the result that longer documents will receive less smoothing from the collection model than shorter documents. Higher values of μ favour longer documents, which for our scenario means that locations which had more tweets will tend to be ranked higher. Therefore the μ parameter affects both how much collection smoothing is applied, and how much correction is applied for document length (Losada & Azzopardi, 2008), and thus implicitly determines how much influence the frequency of tweets per location will have on the relevance ranks.

Kullback-Leibler (KL) divergence measures the difference between two distributions, and allows the locations to be ranked in order of how divergent each location model M_l is from the post model M_p. KL-divergence is calculated as

follows:

$$KL(M_p|M_l) = \sum_{t \in V} P(t|M_p) log \frac{P(t|M_p)}{P(t|M_l)} \quad (6.5)$$

$$= \sum_{t \in V} P(t|M_p) log P(t|M_p) - \sum_{t \in V} P(t|M_p) log P(t|M_l) \quad (6.6)$$

$$= -\sum_{t \in V} P(t|M_p) log P(t|M_l) \quad (6.7)$$

where t is a term and V is the vocabulary. The $\sum_{t \in V} P(t|M_p) log P(t|M_p)$ component of the equation does not depend on the estimated location model, and is therefore ignored.

When computing KL-divergence, the post model $P(t|M_p)$ may be estimated in different ways. Our experiments use a maximum likelihood estimate which simply considers the frequency of term occurrences in the tweet *i.e.*, $P(t|M_p) = \text{tf}_{t,p}/|p|$, which has the result that the rankings generated are equivalent to those generated by the query likelihood approach. As with the query likelihood method, we apply Dirichlet smoothing (Zhai & Lafferty, 2004) to the location model.

6.2.3 Comparison with previous approaches

What distinguishes our approach from the other model-based (as opposed to gazetteer-based) geolocation studies described in Section 4.3.2 is that we apply our approach on a worldwide dataset, we model locations as semantically-meaningful regions, and predict locations to as precise a granularity as postal codes. Alternative approaches for tweet location were limited to the continental US (Cheng et al., 2010; Eisenstein et al., 2010), or only four countries (Hecht et al., 2011). Additionally, the studies of Cheng et al. (2010) and Hecht et al. (2011) reported results at the city level at best. There have also been studies on placing Flickr photos based on textual features, but they defined locations as arbitrary grid cells (Serdyukov et al., 2009), or focused on a limited number of famous landmarks (Crandall et al., 2009).

Another difference between our study and the other tweet-placing studies is that we make use of the individual geotags of tweets to learn language models, while the other approaches rely on Location field information or on one single geotag, which means that the resulting models are less accurate.

In addition, the three previously mentioned studies on location prediction in Twitter (Cheng et al., 2010; Eisenstein et al., 2010; Hecht et al., 2011) focus on placing users and restrict their experiments to users with a certain minimum number of tweets, while we focus on placing individual tweets as well as users. Placing tweets is more difficult than placing users because the amount of content is much less, but we still achieve promising results. We do additionally report user location prediction but did not fine-tune our method for this purpose, and we have no restrictions on the number of tweets per user. It is likely that our user placing results would improve if we did introduce such limitations.

6.3 Data Corpus

6.3.1 Dataset description

We use two Twitter datasets in our analysis. Since we do not limit our method to a particular language, we do not perform stemming or stopword removal. Usernames and hashtags are preserved in our language models as tokens which are distinct from the same term without a @ or # at the beginning. Any duplicates that occur in the status stream are removed. We also remove retweets from our dataset, since they are duplicates or near-duplicates of the original tweet. Specifically, we discard all tweets that have the API field `retweeted_status` set, and all tweets that contain the text "*RT @*" or "*rt @*". Finally, we remove all hyperlinks from tweets. Note that our usage of Twitter data, in particular geotagged data, means that the datasets will have a bias to a particular set of Web users. Geotags are usually generated by smartphones, which means that poorer regions or areas with low rates of smart phone usage are likely to have limited coverage in our data collection.

Spritzer. This dataset was obtained by consuming the public Twitter stream which is called the Spritzer. At the time of data collection, Twitter provided two types of generally available streams: a randomly sampled stream consisting of 5% of all tweets, and a filtered stream, consisting of all tweets that match specific criteria. One of the possible parameters for filtering is by location. The default level of access returns all tweets for up to 10 locations, specified by their bounding boxes of maximum one degree per side. The location filter operates

only on geotags, not on the self-reported location fields.

We used the location filter to retrieve all tweets from 10 cities with high Twitter usage. We chose the set of cities to be geographically and linguistically diverse. The cities were, in order of number of tweets: Jakarta (Indonesia), New York (USA), London (UK), Chicago (USA), San Francisco (USA), Houston (USA), Toronto (Canada), Amsterdam (The Netherlands), Sydney (Australia) and Santiago (Chile). This dataset covers the four week time period from May 25th to June 21st, 2010. We attempted to reverse-geocode each tweet to a city and neighbourhood. This dataset was used to perform initial exploratory experiments. It is a relatively small dataset and it would be feasible for anyone to gather in a month without special access to Twitter streams.

Firehose. This dataset is from the Twitter Firehose stream, which is the full stream of all public statuses. The original data consists of over 7.3 million tweets posted during Summer 2010. We attempted to reverse-geocode each tweet to a country, state, city and postal code. Table 6.2 shows the number of unique tweets that could be reverse geocoded to each place type. We were able to decode 7.26 million tweets to 222 different countries. There are a different number of tweets for each place type since not all tweets which correspond to a country also correspond to a city, for example.

Place Type	Num. Tweets	Distinct Places
Country	7,262,002	222
State	7,313,098	2,290
City	6,295,523	72,617
Postal Code	7,192,172	104,694

Table 6.2: Number of tweets that can be reverse geocoded to each place type

6.3.2 Data characteristics

Table 6.3 shows statistics for a sample of tweets, including properties of their content, frequency of re-tweeting, and prevalence of geotagging. These figures are based on a random sample of 5% of public tweets over a 24 hour period in Summer 2010. At that time, less than 1% of tweets were geotagged, while 80% had text of some sort provided in their user profile.

Average length (characters)	70.6 ± 40.4
Average length (words)	9.7 ± 6.8
% containing hashtag(s)	11.0%
% containing username(s)	52.0%
% with user location in profile	80.3%
% retweets (using official button)	5.4%
% retweets (using unofficial syntax)	14.0%
% containing any geotag	0.86%
% containing reverse geocoded place	0.61%
% containing geo coordinates	0.54%

Table 6.3: Properties of a random sample of tweets. The number after the ± indicates standard deviation

Table 6.4 lists the most commonly occurring source services of geotagged tweets in a random sample of 5% of geotagged tweets over a 24 hour period in Summer 2010. A source service is the website or application that the creator of the tweet used to send the tweet. This source of tweets is relevant because some services are geographically focused and are specifically intended to provide information about the place where the user is situated at the time of creating the tweet. For example, Foursquare, which accounted for 18% of geotagged tweets, is a service which allows users to "check-in" at a venue to win points. A check-in results in the creation of a tweet containing location information such as *"Safe travels! (@ LaGuardia Airport (LGA))"*. Foursquare is the most popular Twitter application of this type, but other location-based services such as Gowalla[2] provide similar functionality, allowing users to check-in to locations for rewards, and simultaneously updating their status. Tweets from these services are far more likely than average to contain geographic references.

Service	% of Tweets
UberTwitter	24.2%
Foursquare	18.3%
Twitter for iPhone	12.3%
Twitter for Android	6.3%
Echofon	5.7%

Table 6.4: Top 5 sources in a random sample of geotagged tweets

[2] `http://gowalla.com/`, accessed July 2011

6.4 Evaluation

6.4.1 Experimental setup

This section describes the prediction methods, baseline methods, and evaluation measures used in our experiments.

Prediction methods

We report results obtained using the following methods – three baseline methods, and two based on language models. Two of the baseline methods made use of Yahoo! Placemaker[3], a geoparsing service which identifies locations in unstructured text and returns the locations using the same identifiers as GeoPlanet. Thus Placemaker allows us to determine what percentage of tweets can be accurately placed based on explicit geographic information. We resolved locations identified by Placemaker to an ancestor of the appropriate place type – so a tweet mentioning LaGuardia Airport, for example, was resolved to New York City for city-level experiments and the USA for country-level experiments.

Trivial Classifier (TC). We computed the accuracy for a trivial classifier, where we simply selected the most commonly occurring place in the training set, and select this as the predicted location for every tweet in the test set.

Placemaker using Location field (PM-L). We extracted the user's self-reported location which appears in their profile, submitted this text to Placemaker, and identified the most probable candidate location. If necessary, we identified the ancestor of the appropriate place type and selected this as the predicted tweet location. This method allows the detection of explicit geographic references in a user's self-reported location.

Placemaker using Tweet content (PM-T). For each tweet, we submitted the content as a query to Placemaker and identified the most probable candidate location. If necessary, we identified the ancestor of the appropriate place type and selected this as the predicted tweet location. This method allows the detection of explicit geographic references in the tweet content.

Kullback-Leibler divergence (KL). Small-scale initial experiments were carried out on a single machine using the KL-divergence function (Equation 6.7)

[3] `http://developer.yahoo.com/geo/placemaker/`, accessed July 2011

implemented by the Lemur Toolkit for Language Modelling and Information Retrieval (Ogilvie & Callan, 2002) with Dirichlet smoothing. This method was chosen because it is included as part of a freely available open-source framework. For each tweet, locations are ranked according to the KL-divergence between the location model and the tweet model. The location whose model has lowest divergence is selected.

Query Likelihood (QL). For efficiency, it was necessary to perform large-scale experiments on a cluster using the Hadoop MapReduce[4] software, which Lemur does not support. Therefore we used a retrieval function previously implemented in Java which uses the Terrier toolkit (Ounis et al., 2006). The retrieval function uses a query likelihood language model with Dirichlet smoothing (Equations 6.2 and 6.4). For each tweet, locations are ranking according to their query likelihood and the location whose model ranks highest is selected.

Methods for user location prediction based on multiple tweets

For predicting the location of a user, we had to consider multiple tweets. This was not an issue for **TC**, because we simply chose the most commonly occurring location in the training set. For **PM-L**, the Location field generally did not vary, however in the rare cases when it did change, we queried Placemaker for the user's most common location. For **PM-T** and the two language model approaches we experimented with two options for dealing with multiple tweets:

Aggregation (agg.). All tweets were aggregated into one text which was used as the query text to determine the user location.

Majority Vote (m.v.). A location was predicted for each individual tweet and the most frequent one was taken as the user location.

Evaluation measures

The ground truth for each tweet was defined as the place from where it was posted, according to the geotag. We tuned parameters to maximise accuracy, but we additionally report accuracy within an extended geographic range as follows:

Accuracy (Acc). This indicates the percentage of correctly predicted locations over all test queries.

[4] `http://hadoop.apache.org/`, accessed July 2011

Accuracy within K hops (Acc@K). This indicates the percentage of predicted locations which lie within K hops of the correct location. For example, for postal code level prediction, Acc@1 measures the percentage of predicted postal codes that are correct, or are direct neighbours of the correct postal code. Acc@k enables us to measure the frequency of cases where a tweet has been placed in a region which is not the correct one, but is very close to the correct one. Neighbours of places were identified using the GeoPlanet Web service.

6.4.2 *Spritzer* experiments

Two sets of experiments were carried out using the *Spritzer* dataset – prediction of the city that each tweet originates from, and prediction of which neighbourhood each New York City tweet originates from. This evaluation was carried out using five-fold cross-validation, where in each round four parts of the data were indexed and a fifth was used as the test set. The dataset was partitioned by user in order to avoid similar tweets by the same user occurring in different partitions.

City-level tweet placing

We first investigated the performance of our method in classifying tweets on the city level. We built an index with a document corresponding to each of the ten cities in the dataset. On each round we performed a parameter sweep in the range [1, 5, 10, 50, 100, 500, 1,000, 5,000, 10,000] for the Dirichlet prior on a subset of the indexed data and on every round a parameter of 10,000 was found to be optimal.

The results are shown in Table 6.5. For over 10% of tweets, a geographic reference was detected and disambiguated by Placemaker (**PM-T**) and the correct parent city was reported. The high frequency of tweets from location-focused services is likely to have contributed to this score. The language model method (**KL**) correctly placed over 65% of tweets. In this small dataset, the success of the language model compared to Placemaker is largely due to the difference in languages between cities. Nevertheless the results show the advantages of using language information as well as placename knowledge for geographically placing text. For this dataset, due to the small number of regions involved (ten) and the skewedness of the tweet counts for each region, a trivial classifier (**TC**) also

performs well, correctly placing 40% of tweets.

Method	Acc
TC	0.403 ± 0.011
PM-T	0.108 ± 0.025
KL	0.657 ± 0.010

Table 6.5: Results of city prediction in the *Spritzer* dataset ($\pm$ 95% Confidence Interval)

Table 6.5 also shows the 95% confidence intervals determined using the t-test on the results of the five-fold cross-validation. All of the differences in acuuracy in this table are statistically significant. We do not report Acc@k for this experiment since none of the cities are neighbours.

Neighbourhood-level tweet placing

We next evaluated retrieval using the neighbourhood-level models. For this experiment we limited the geographic range to New York City. This subset of the dataset contained tweets from 502 New York City neighbourhoods indexed by GeoPlanet. We built an index with a document corresponding to each of the New York City neighbourhoods in the dataset. We performed the same parameter sweep for the Dirichlet prior as in the city prediction experiment and found 10,000 to be optimal for all splits.

The results for neighbourhood prediction are shown in Table 6.6. For retrieving the **PM-T** predictions in this experiment, we made use of Placemaker's focus location feature to inform Placemaker that places near New York City should be favoured in the location disambiguation process. Compared to the city-level experiments, **PM-T** performs poorly, correctly detecting explicit mentions of neighbourhoods in only 1.5% of tweets. The trivial classifier **TC** also performs poorly on the neighbourhood level. In this experiment we ran both the KL-divergence function provided by Lemur (**KL**) and the Terrier implemented query likelihood method (**QL**). Both language model methods achieved an accuracy of approximately 20%, far outperforming the trivial classifier and Placemaker's search for explicit geographic references. In this task, there may still be variation in the languages spoken in different regions, but to a lesser extent than for the city prediction task. Thus the good performance of the language modeling approach for

neighbourhood prediction shows that useful location clues can be found not only from the language used, but also from other regional variations such as mentions of venues or local slang.

Method	Acc
TC	0.034 ± 0.031
PM-T	0.015 ± 0.001
KL	0.209 ± 0.016
QL	0.203 ± 0.015

Table 6.6: Results of neighbourhood prediction for New York City in the *Spritzer* dataset (± 95% Confidence Interval)

Table 6.6 also shows the 95% confidence intervals for the results. The difference between **TC** and **PM-T** is not statistically significant, however the improvements of the language model methods over **TC** and **PM-T** are statistically significant. We do not report Acc@k for this experiment, because all of the neighbourhoods involved are located so close together that an improvement between Acc and Acc@1 is likely to be due to chance.

6.4.3 *Firehose* experiments

For the experiments using *Firehose*, the data was split to have roughly 80% of tweets for building models, 10% for tuning parameters, and 10% for testing. Thus for the country dataset, we reserved approximately 5.8 million tweets for index building, ∼700,000 for tuning, and ∼700,000 for testing. Again, the tweets were partitioned by user in order to avoid highly similar tweets by the same user occurring across different partitions. For each index we tuned the Dirichlet prior for retrieval to a parameter in the range $[1, 5, 10, 50...10^{10}]$. The wide range was chosen due to the exceptionally large sizes of some psuedo-documents, particularly those representing countries. When building the country index, the psuedo-document generated for the largest country (USA) was initially too large to index. Therefore a random sample of 80% of all tweets (approximately 4.6 million) was used to build the country index.

Due to the large size of the *Firehose* dataset we did not perform cross-validation in these experiments. Therefore we do not have standard deviation figures with which to determine statistical significance.

Tweets originating from location-focused services such as Foursquare comprise a significant component of the dataset, and are likely to have a major influence on the results, so we also report results when these tweets are omitted from the test sets. We manually inspected the sources, identified location-focused services which comprise more than 1% of total tweets, and removed these from the test set. Approximately 75% of tweets remain. The remaining tweets have less explicit references to locations and are more representative of tweets and social media items in general. We retained tweets from location-focused services in our language models, since we are interested in learning from these and using them to geolocate other tweets.

Geoplanet does not provide information about neighbours of countries, so we report only Acc for country-level experiments. For the other place types, we also report Acc@1 and Acc@2.

Tweet location prediction

The results for predicting the location of a tweet are shown in Table 6.7. The two best-performing methods are **PM-L**, which parses the user profile Location field for explicit geographic references, and **QL**, our language model method. Both methods achieve similar results on a country level, while **PM-L** performs better than **QL** at the state and city level. However at the very fine granularity of the postal code level, **QL** outperforms all other methods. The performance of both of the methods that operate on tweet content (**PM-T** and **QL**) decreases after omitting tweets that originate from location-focused services. Those tweets are very likely to have contained geographic references.

This experiment shows that the benefits of the language modeling approach are most clear at the zip code level. Considering the brevity of tweets and the lack of explicit geographic references in many tweets, the approach achieves promising results. Even though Placemaker could detect geographic entities in only 2.5% of all tweets, our language modeling approach could correctly place 13.9% of them. These results indicate that there are substantial benefits from text features other than explicit geographic references. The benefits of the language model approach are most obvious at the zip code level because users almost never provide their location to such a level of detail, and the tweets themselves are unlikely to mention official placenames or points of interest that would be found in a geographic

	All Tweets			w/o Location Services		
Method	Acc	Acc@1	Acc@2	Acc	Acc@1	Acc@2
Country						
TC	0.469	-	-	0.434	-	-
PM-L	0.528	-	-	**0.518**	-	-
PM-T	0.222	-	-	0.120	-	-
QL	**0.532**	-	-	0.514	-	-
State						
TC	0.063	0.082	0.101	0.060	0.076	0.091
PM-L	**0.407**	**0.449**	**0.462**	**0.401**	**0.440**	**0.450**
PM-T	0.160	0.170	0.173	0.076	0.081	0.084
QL	0.316	0.405	0.458	0.246	0.343	0.400
City						
TC	0.062	0.062	0.062	0.061	0.061	0.062
PM-L	0.269	**0.323**	**0.342**	**0.269**	**0.324**	**0.342**
PM-T	0.141	0.151	0.153	0.060	0.066	0.067
QL	**0.298**	0.317	0.326	0.217	0.234	0.244
Postal Code						
TC	0.004	0.004	0.004	0.005	0.005	0.005
PM-L	0.017	0.025	0.029	0.017	0.025	0.028
PM-T	0.025	0.034	0.034	0.018	0.025	0.026
QL	**0.139**	**0.166**	**0.188**	**0.052**	**0.073**	**0.094**

Table 6.7: Results for tweet location prediction in the *Firehose* dataset

database. Instead, the tweets may mention local venues, events or terms from local dialects that are not widely-known but can provide valuable information for a language model of that location. Therefore our approach would be most useful for hyperlocal applications.

Although the language model approach performed well for zip code prediction, we observe that for the country, state and city level, querying Placemaker for each user's self-reported location performed best. However this method requires that the user has provided a valid location. Our language model approach could still be useful for geolocating posts from the 20% of users who do not report a location (see Table 6.3), for tweets from users whose reported location does not resolve to an identifiable place, and for other applications where a user-provided location is unavailable.

User location prediction

The results for user location prediction are shown in Table 6.8. For these experiments, the ground truth was defined as the place from which the user most often posts. Again, we show results of both the entire test set, and the test set after tweets from location-focused services were removed. For **PM-T** and **QL**, we show the results of user location prediction based on both an aggregation of all their tweets into one text (**agg.**), and a majority vote from the results of placing each individual tweet (**m.v.**).

Comparing the different location prediction methods, we see a similar pattern to the tweet placing experiments. The language model method **QL** outperforms **TC** and **PM-T** in all cases. At a country level, **QL** and **PM-L** give similar results, for states and cities **PM-L** has the highest accuracy, and at a postal code level, **QL** gives the best results.

Results for country, state and city prediction improve on the user level compared to the tweet level, thanks to the additional information provided by multiple status updates. The accuracy for user country prediction after removing tweets from location-focused services is 71% compared to 52% for tweet country prediction. Results for postal code prediction are similar or slightly worse on a user level than on a tweet level. This could be because an individual's tweets are more likely to be dispersed across multiple postal codes than multiple states, for example.

	All Tweets			w/o Location Services		
Method	Acc	Acc@1	Acc@2	Acc	Acc@1	Acc@2
Country						
TC	0.446	-	-	0.426	-	-
PM-L	0.577	-	-	0.559	-	-
PM-T (agg.)	0.405	-	-	0.295	-	-
PM-T (m.v.)	0.418	-	-	0.295	-	-
QL (agg.)	**0.759**	-	-	**0.710**	-	-
QL (m.v.)	0.501	-	-	0.435	-	-
State						
TC	0.073	0.093	0.118	0.069	0.088	0.110
PM-L	**0.471**	0.507	0.518	**0.447**	**0.482**	**0.493**
PM-T (agg.)	0.276	0.326	0.350	0.185	0.226	0.247
PM-T (m.v.)	0.296	0.334	0.352	0.186	0.219	0.237
QL (agg.)	0.449	**0.541**	**0.589**	0.334	0.436	0.491
QL (m.v.)	0.343	0.412	0.453	0.238	0.313	0.356
City						
TC	0.034	0.034	0.035	0.031	0.032	0.033
PM-L	0.314	**0.361**	**0.379**	**0.292**	**0.339**	**0.356**
PM-T (agg.)	0.175	0.217	0.236	0.104	0.127	0.139
PM-T (m.v.)	0.218	0.244	0.256	0.114	0.132	0.142
QL (agg.)	**0.319**	0.346	0.362	0.215	0.234	0.249
QL (m.v.)	0.281	0.298	0.308	0.174	0.187	0.196
Postal Code						
TC	0.002	0.002	0.002	0.002	0.002	0.002
PM-L	0.023	0.034	0.038	0.023	0.033	0.037
PM-T (agg.)	0.025	0.045	0.058	0.012	0.020	0.026
PM-T (m.v.)	0.025	0.041	0.051	0.011	0.017	0.022
QL (agg.)	0.135	0.177	**0.213**	0.052	**0.080**	**0.106**
QL (m.v.)	**0.149**	**0.182**	0.207	**0.054**	0.077	0.099

Table 6.8: Results for user location prediction in the *Firehose* dataset

It is not clear from these results whether location prediction from a user's tweets should be calculated using aggregation or majority voting. For the language modeling approach, aggregation of tweets into one text gives better results at the city level and above, and when omitting location-based services. At the zip-code level, when location-based services are included, the majority-vote yields slightly better performance.

6.5 Conclusions

This chapter has presented a method for geolocating a Twitter status update or a Twitter user based on tweet content. Our approach involves using geotagged tweets to build language models corresponding to locations of varying granularity. These models then allow us to determine which is the most probable location for a tweet or user without geographic information. Our results show that using user-generated content as a source of information for creating language models of locations is promising, allowing us to pinpoint more than 70% of Twitter users to the correct country, and approximately 30% to the correct city. At the postal code level, our approach gives better results than all of the other alternatives that we investigated, including parsing the user's self-reported location for geographic information.

In a practical scenario, we could consider only adding a geographic annotation to those items for which confidence in location prediction is high, since many items do not have any specific geographic focus, especially at a very fine level of granularity. Since the score return by query likelihood is a probability, a threshold could be set which controls whether or not the predicted location is set for a certain item. This threshold would be applied at each level in the geographical hierarchy. As the prediction accuracy at the country level is reasonably good, many items could be tagged with their country, while fewer would be tagged with a city and less again with the zip code. This means that those minority of tweets which do have a strong local focus could be easily found by interested users from the relevant area, even if there are no mentions of geographic entities that would be recognised by a gazetteer.

The method presented in this chapter could also be applied to other geotagged social media items. For non-geotagged social media, it would be worth

investigating whether the models generated from tweet data could be used to make accurate predictions for non-Twitter posts. In conjunction with the other approaches for metadata augmentation presented in this thesis – tag prediction and topic classification – adding location information to social media items would enable users to more easily find geographic content on particular topics of interest. This would be very useful for local search, and could also be of interest for people who are planning a trip, or expatriates who want to keep up with events in their hometown.

Chapter 7

Topic Classification using Metadata from Hyperlinked Objects*

7.1 Introduction

The approach of Chapter 5 exploited anchortext on the Web for tag prediction, and Chapter 6 showed how geotagged text items can be used to predict the location relevant to a post. We now look at how hyperlinks to structured data sources can be used to classify the topic of a post. The motivation behind this approach is that posters often link to the objects which are the subject of the post, so the metadata of these objects should be useful for categorising the post. As in Chapter 6, this chapter relies on structured representations of data beyond the plain text of HTML documents. However while the last chapter made use of one specific type of metadata (geotags) for location prediction, the approach presented in this chapter compares any textual metadata types as data sources for text classification.

Interactions in social media are often based around objects, in a manifestation of the object-centred sociality theory introduced in Section 2.3. These objects are often identified via hyperlinks to relevant Web resources. Users share videos they have seen, point to products or movies they are interested in, and use external articles as references in discussions. These external resources can provide useful new data such as author information in the case of books, or genre information

*This chapter is partially based on (Kinsella et al., 2010a), (Kinsella et al., 2010b), (Kinsella, Passant, & Breslin, 2011) and (Kinsella, Wang, et al., 2011)

in the case of movies. Many of these hyperlinks are to websites that publish metadata about objects, such as videos (YouTube) or products (Amazon), and make this metadata available via an API or as Linked Data. Websites which provide structured data are particularly useful since they allow particular pieces of relevant data to be identified and extracted, along with their relationship to the hyperlinked resource. In some cases the metadata is published by external sources, *e.g.*, DBpedia provides a structured representation of Wikipedia.

In this chapter we investigate the potential of metadata from hyperlinked objects for improving topic classification in social media. Classifying social media items is challenging because posts are typically short and informal, and often rely on external hyperlinks to provide additional contextual information. In many cases, vital pieces of information are provided not within the text of the post, but behind hyperlinks that users post to refer to a relevant resource. For example, a poster may recommend a book by posting a hyperlink to a webpage where you can buy it, rather than providing the book title and name of the author. In the message board dataset described in Section 7.3, we found that 65% of posts that linked to books mentioned neither the complete title nor the complete author name, and at least 11% did not contain even a partial title or author name. Our approach exploits the fact that the subject of a post is often strongly related to objects to which the post is connected via hyperlinks, and the metadata from these objects can be of use to determine the topic of a post. Even if a post has no metadata such as tags or category information, the hyperlinked objects within the post may well have such metadata which is likely to also be relevant to the post itself. Figure 7.1 gives a graphical representation of how a social media dataset can be enhanced by integrating metadata retrieved from hyperlinks to various Web sources. The sources shown are those that will be used in experiments later in this chapter.

Figure 7.2 gives examples of two posts where additional useful text can be gained by considering the metadata of a hyperlinked object. In the first post, without the external metadata all we can conclude is that the poster is talking about a book, but by considering the external metadata we can infer that this post is on the topic of martial arts. In the second post, the artist name is redundant since it was already available from the URL, but the genre of the music artist is new and could be useful for a more accurate classification of this post. In the first

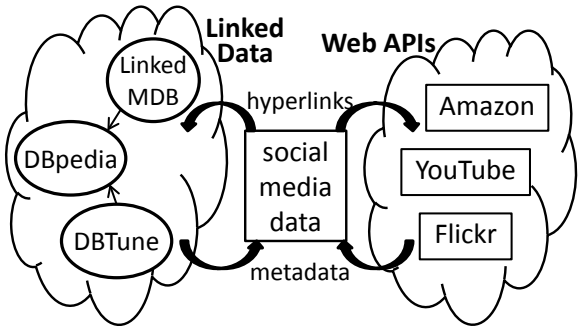

Figure 7.1: Web sources that were used to enrich social media data

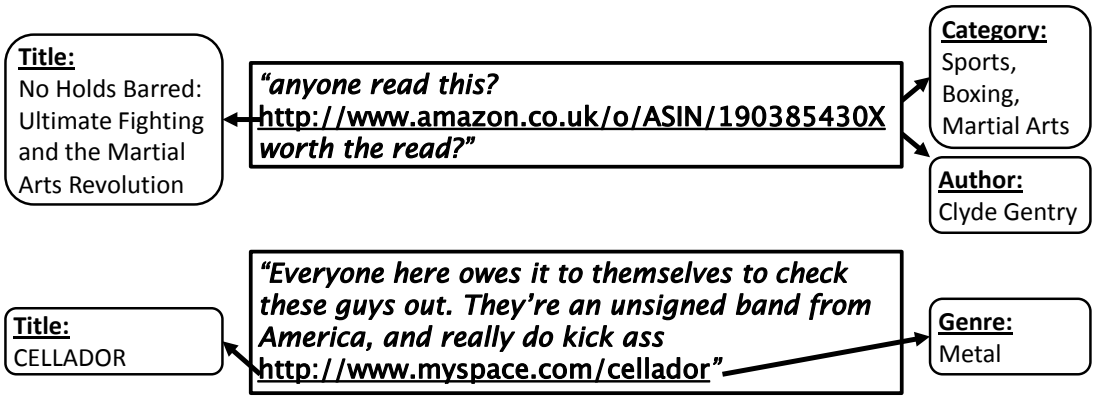

Figure 7.2: Examples of how social media posts can be augmented with relevant external structured data

case, the metadata of the book could also be useful for information retrieval, for example in a search scenario where a user queries for a book title. It is likely that certain metadata types will generally be more useful than others – for example, while the name of a book's publisher may sometimes indicate the topic of a book, in many cases it will not be helpful.

We focus on structured data sources in particular because they enable the identification of particular metadata items with specific semantic relationships to the relevant object. Thus it is possible to compare different metadata types and determine which types are most useful for post classification. Although many hyperlinks do not have a related structured data source, the percentage that do is quickly growing, as we show later in this chapter. Initiatives such as the Linking Open Data project and Facebook's Open Graph Protocol are

likely to ensure that the trend towards structured data continues. The growth of structured data is not only relevant for data organisation within those sources of structured data. This chapter shows how classification of unstructured social media data can benefit from the semantically-rich metadata of the hyperlinked objects within posts. Thanks to the linked nature of the Web, the increasing availability of structured data online can aid classification in non-structured Web data too.

7.2 Approach

Our approach involves using a Naïve Bayes text classifier to assign social media posts to one of a set of predefined topics or classes. In order to train a classifier, we require a dataset of posts which have already had topics assigned to them. As well as using the post content to build a model for each topic, we also experiment with various ways of incorporating metadata from hyperlinked objects. We identify hyperlinks to sources of structured data, retrieve the metadata of the corresponding objects, and include this text when training the classifier. When determining the most relevant class for a new post, we also include the structured metadata for any hyperlinks as part of the post representation. Figure 7.3 shows a flowchart for the processes involved in the topic classification of a social media item, which we now describe in more detail.

7.2.1 Dataset enrichment

Assuming that a social media dataset is already represented in the SIOC format for online community data, the data must be loaded into an RDF store so that SPARQL queries can be carried out against the dataset. External hyperlinks in the dataset are identified by executing a query using the sioc:links_to property, as shown in Listing 7.1.

From the hyperlinks identified in a dataset, those that point to resources for which there is available structured data are identified. We use regular expressions to determine which URLs are from a predefined list of domains that provide APIs or that are also available as Linked Data. For the Web API sources, we use Java wrappers to access the API, perform a query for the relevant metadata, convert

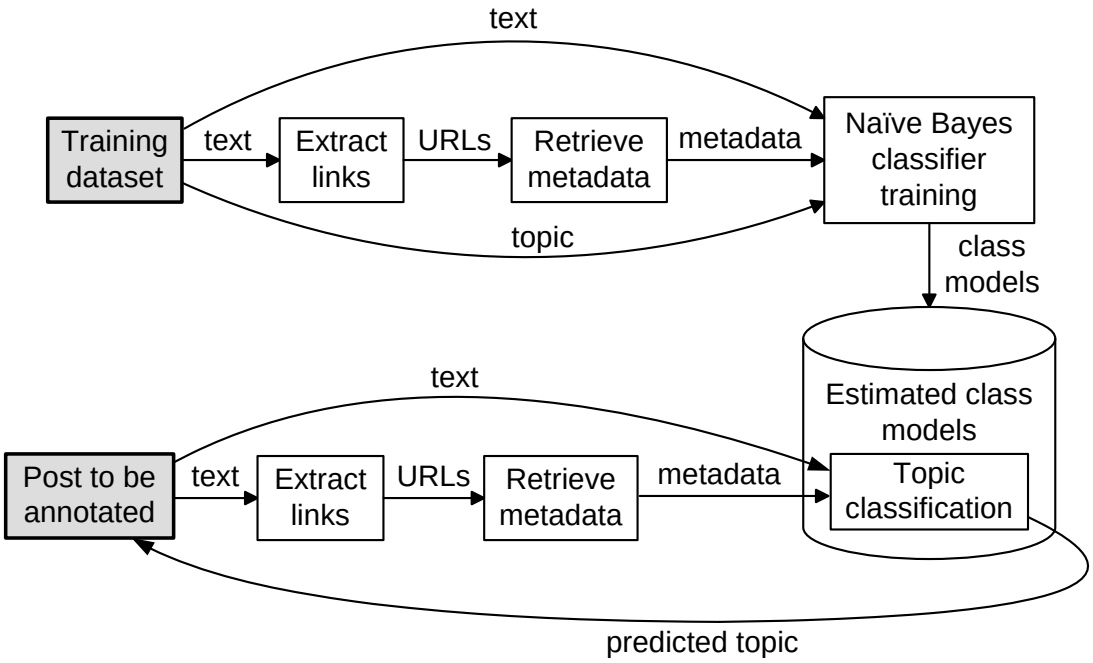

Figure 7.3: Flowchart for the topic classification approach

the data to RDF and add it to the dataset RDF store. For the Linked Data sources, we query the corresponding SPARQL endpoint for each URL, retrieve the related RDF data, and add it to the dataset RDF store. Listing 7.2 provides an example of how the LinkedMDB service[1] can be used to retrieve an RDF representation of the title and genre of a movie, based on the URL assigned to the movie in the well-known IMDb movie website.[2]

Storing the external data using RDF enables us to execute queries which involve both the original social media data and the external metadata. Listing 7.3 shows a query which returns all posts that link to tagged photos, and the tags of those photos. We used queries similar to this one to extract different types of metadata for the experiments reported later in this chapter.

[1] `http://linkedmdb.org/`, accessed July 2011
[2] `http://www.imdb.com/`, accessed July 2011

```
PREFIX rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .
PREFIX sioc: <http://rdfs.org/sioc/ns#> .

SELECT ?post ?resource WHERE {
  ?post rdf:type sioc:Post .
  ?post sioc:links_to ?resource .
}
```

Listing 7.1: SPARQL query for posts containing hyperlinks, and those hyperlinks

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .
PREFIX dc: <http://purl.org/dc/terms/> .
PREFIX movie: <http://data.linkedmdb.org/resource/movie/> .

CONSTRUCT {
  ?s foaf:page <http://www.imdb.com/title/tt0081505> .
  ?s dc:title ?title .
  ?s movie:genre ?genre .
} WHERE {
  ?s foaf:page <http://www.imdb.com/title/tt0081505> .
  ?s dc:title ?title .
  ?s movie:genre ?genre .
}
```

Listing 7.2: SPARQL query to retrieve the title and genre of an IMDb movie

```
PREFIX rdf: <http://www.w3.org/1999/02/22−rdf−syntax−ns#> .
PREFIX sioc: <http://rdfs.org/sioc/ns#> .
PREFIX dc: <http://purl.org/dc/terms/> .
PREFIX dcmit: <http://purl.org/dc/dcmitype/> .

SELECT ?post ?externaltags WHERE {
  ?post rdf:type sioc:Post .
  ?post sioc:links_to ?resource .
  ?resource rdf:type dcmit:StillImage .
  ?resource dc:subject ?externaltags .
}
```

Listing 7.3: SPARQL query for posts that link to tagged photos, and their tags

7.2.2 Topic classification

Each post is represented using the vector space model, the same model we used to describe resources in Chapter 5. A post p is represented as a vector of terms $p = (t_1, t_2 ... t_{n_p})$ where the entries in the vector are terms in the vocabulary, the value t_i denotes the weight of term i in document d, and n_p is the number of terms in p. This vector can be generated from the post content, from the external metadata, or from a combination of both. Functions such as tf-idf and document length normalisation can be applied to reduce the impact of variations in term frequency and document length.

A multinomial Naïve Bayes classifier is used to calculate class probabilities for each post. Each class corresponds to a topic or category. The classifier requires a set of training posts p, each of which has a class label c. The training set of posts and class labels $\langle p, c \rangle$ provides a sample of example posts for each category in the dataset. We use Bayes Theorem to determine the probability $P(c|p)$ that a post p belongs to a class c:

$$P(c|p) = \frac{P(c)P(p|c)}{P(p)} \tag{7.1}$$

$P(c)$ is the class prior probability. The post prior probability $P(p)$ is independent of the class, so can be ignored. $P(p|c)$ is the conditional probability of a post p occurring in a class c and is computed as the product of the probabilities of each term t in p, requiring the assumptions that the probability of each term is independent of all other terms in the document, and of its position in the document:

$$P(c|p) = P(c) \prod_{1 \leq k \leq n_p} P(t_k|c) \tag{7.2}$$

The class prior probability $P(c)$ is estimated from the training set as the number of posts N_c in c, divided by the total number of posts N, and weights the classes according to their relative frequency:

$$P(c) = \frac{N_c}{N} \tag{7.3}$$

The conditional probability $P(t|c)$ of term t occurring for class c is estimated from the training set as the relative frequency of term t in documents of class c,

i.e.,

$$P(t|c) = \frac{F_{tc}}{\sum_{x=1}^{N} F_{xc}} \tag{7.4}$$

where F_{tc} is the frequency of term t occurring in class c, and F_{xc} is the frequency of term x in the vocabulary occurring in class c. Smoothing is also performed in order to allow for very rare terms. Without smoothing, if any term t which occurs in p does not occur in the training set for class c, then $P(t|c)$ and therefore $P(c|p)$ will be computed as zero.

7.2.3 Comparison with previous approaches

In Section 4.3.3 we discussed related work on topic classification, of social media items, Web documents, and Web objects from specific domains. Much of the related work on categorisation in social media was based only on item content (Garcia Esparza et al., 2010; Sharifi, 2010; Rodrigues et al., 2008; Pal & Saha, 2010) or item metadata, for platforms where metadata is available such as blogs or content sharing sites (Berendt & Hanser, 2007; Sun et al., 2007; Huang et al., 2010). Other works (Genc et al., 2011; Firan et al., 2010) classified social media items by calculating the similarity between the items and objects from other sources (*e.g.*, Wikipedia, Upcoming). These approaches did not make use of hyperlinks occurring in posts, which we argue are a valuable source of topic information.

Exploiting hyperlinks is a common approach in Web document classification. Studies on categorising webpages show that the metadata of a webpage and its neighbours, especially citing pages, are useful text sources for topic classification (Attardi et al., 1999; Fürnkranz, 1999; Ghani et al., 2001; Glover et al., 2002; Sun et al., 2002; Lim et al., 2005; Qi & Davison, 2008). For social media posts however, this information is often not available. In many types of social media, such as social network sites, message boards and microblogs, posts have little or no textual metadata. As for citing pages, many types of social media items do not generally attract any inlinks since linking to other posts is not a standard practice to refer to another post (though there are exceptions, such as blogs and content sharing services). Instead, users of social media often quote the content of a previous post (*e.g.*, on message boards), or they quote the username of the post's creator (*e.g.*, on Twitter). However outgoing hyperlinks are a common feature of

these social media items, so we exploit such hyperlinks in our approach.

Another key difference between our work and the related work on Web document classification that the previously cited approaches operate on entire webpages, structural parts of webpages (headings) or very limited metadata (title and the meta tag). The related work on classifying social media items which considers information from hyperlinks similarly operates only on whole Web documents or their structural parts Antonelli and Sapino (2005); Irani et al. (2010). We on the other hand include a range of metadata items that have a semantic relation to the objects discussed in a post, such as movie directors and music genres. This approach enables us to compare the effectiveness of different metadata types for improving classification.

Other work on categorising Web objects such as music artists and videos (Figueiredo et al., 2009) made use of the object metadata, and compared the accuracy resulting from classification based on different metadata types. This approach is related to ours, however rather than using object metadata to categorise the objects, we investigate whether we can use it to categorise the posts which link to objects of various types.

7.3 Data Corpus

In this section we describe the social media datasets that we used in our experiments: one from a message board and one from a micro-blogging site. We also provide a detailed analysis of the message board dataset for which we have ten years worth of data, in order to gain insight into hyperlink posting patterns over time. Next, we describe the process of enriching the datasets with external structured data from the Web. Finally we present another detailed analysis of the message board dataset, this time focusing on the related structured data that we retrieved for the final year of the dataset.

7.3.1 Dataset description

General and *Music* datasets

We use the dataset from the `boards.ie` SIOC Data Competition[3] which was held in 2008. The data covers the first 10 years of existence of the message board, from February 1998 to February 2008. Over 130,000 users feature in the dataset, along with over 7 million posts which they have authored. The discussions are represented in the SIOC (Semantically-Interlinked Online Communities) format. The FOAF (Friend-of-a-Friend) and DC (Dublin Core) vocabularies were also used to represent elements of the dataset. The analysis that we present in Section 7.3.2 makes use of the entire `boards.ie` SIOC dataset.

For the classification experiments, we only consider only the posts contained in the final year of the dataset, as we found that links to structured data sources were far more likely to occur during this period. Forum titles were used as categories for the classification experiments, since authors generally choose a forum to post in according to the topic of the post. We choose two sets of forums for our classification experiments – ten forums based on quite high-level topics (*General*), and five forums related to specific genres of music (*Music*). The forums included in each dataset are listed in Table 7.1. These forums were chosen based on the criteria that they were among the most popular forums in the dataset and they each have a clear topic (as opposed to general "chat" forums). For the 23% of posts that have a title, this is included as part of the post content. Since discussion forums are typically already categorised, performing topic classification is not usually necessary. However, this data is representative of the short, informal discussion systems that are increasingly found on Web 2.0 sites, so the results obtained from utilising the class labels in this dataset should be applicable to similar uncategorised social media sites.

Twitter dataset

The Twitter dataset[4] comes from Yang and Leskovec (2011), and covers 476 million posts created by 17 million users between June 2009 and December 2009. Due to the post length restriction of Twitter, users make frequent use of URL

[3] `http://data.sioc-project.org/`, accessed July 2011
[4] `http://snap.stanford.edu/data/twitter7.html`, accessed July 2011

General	*Music*
Atheism	Alternative
Martial Arts	Electronic
Motors	Hip-Hop
Movies	Punk
Musicians	Rock
Photography	
Poker	
Politics	
Soccer	
Television	

Table 7.1: Forum titles in the *General* and *Music* datasets

shortening services such as bit.ly[5] to substantially shorten URLs in order to save space. Therefore for this dataset it was necessary to first decode short URLs via cURL.[6] Like many social media websites, but in contrast to the previous dataset, Twitter does not provide a formal method for categorising tweets. However, a convention has evolved among users to tag updates with topics using words or phrases prefixed by a hash symbol (#). We make use of these hashtags to create six categories for classification experiments. Our approach borrows the hashtag-to-category mappings method from Garcia Esparza et al. (2010) to identify tweets that relate to selected categories. We reuse and extend the hashtag categories of Garcia Esparza et al. (2010); Table 7.2 shows the mappings between hashtags and categories. These categories were chosen because they occur with a high frequency in the dataset and they have a concrete topic. Tweets belonging to more than one category were omitted, since our goal is to assign items to a single category. All hashtags were removed from tweets, including those that do not feature in Table 7.2, since they may also contain category information. Any URLs to websites other than the selected metadata sources were eliminated from tweets. Finally, to avoid repeated posts caused by users retweeting (resending another post), all retweets were omitted.

[5] `http://bit.ly/`, accessed July 2011
[6] `http://curl.haxx.se/`, accessed July 2011

Category	#hashtags
Books	book, books, comic, comics, bookreview, reading, readingnow, literature
Games	game, pcgames, videogames, gaming, gamer, xbox, psp, wii
Movies	movie, movies, film, films, cinema
Photography	photography, photo
Politics	politics
Sports	nfl, sports, sport, football, f1, fitness, nba, golf

Table 7.2: Categories and corresponding hashtags in the *Twitter* dataset

Enriching the datasets with external structured data

For *General* and *Music*, we identified Amazon, YouTube and Flickr as useful sources of metadata via APIs, based on the frequency of hyperlinks to each domain. We identified Myspace, IMDb and Wikipedia as good sources of Linked Data, via third-party data publishers. From the most common domains of *Twitter* we identified Amazon, YouTube and Flickr as sources of metadata via APIs. Hyperlinks to the Linked Data sources did not feature prominently in this dataset.

Amazon product, Flickr photo and YouTube video metadata was retrieved from the respective APIs. Myspace music artist information was obtained from DBTune[7] (an RDF wrapper of various musical sources which at the time of data collection included Myspace), IMDb movie information from LinkedMDB, (a movie dataset with links to IMDb) and Wikipedia article information from DBpedia.[8] The latter three services are part of the Linking Open Data project. Table 7.3 gives a summary of the domains for which we retrieved metadata, the sources of the structured metadata, and the metadata types that we retrieved. For our analysis, we selected only the most commonly available metadata types in order to make comparisons between them, but our method could be applied using arbitrary metadata. The metadata types that we chose were Title, Category (includes music/movie genre), Description (includes Wikipedia abstract), Tags and Author/Director (for Amazon books and IMDb movies only).

For *General*, the percentage of posts that have hyperlinks varies between forums, from 4% in Poker to 14% in Musicians, with an average of 8% across forums. These are a minority of posts; however, we believe they are worth focusing on

[7] `http://dbtune.org/`, accessed July 2011
[8] `http://dbpedia.org/`, accessed July 2011

Website	Object type	Structured Data Source	Title	Description	Category	Tags	Author/Director
Amazon	Product	Amazon API	x		x	x	x
Flickr	Photo	Flickr API	x	x		x	
IMDb	Movie	LinkedMDB	x		x		x
Myspace	Music artist	DBTune	x		x		
Wikipedia	Article	DBpedia	x	x	x		
YouTube	Video	YouTube API	x	x	x	x	

Table 7.3: External websites and the metadata types used in our experiments

because the presence of a hyperlink often indicates that the post is a useful source of information rather than just chat. Of the posts in *General* with hyperlinks, 23% link to one or more of the structured data sources listed previously.

The number of posts containing links to each type of object in *General*, *Music* and *Twitter* are shown in Figures 7.4, 7.5 and 7.6. For the *General* dataset, hyperlinks to music artists occur mainly in the Musicians forum, movies in the Films forum, and photos in the Photography forum. The other object types are spread more evenly between the remaining seven forums. In total, *General* contains 6,626 posts, *Music* contains 1,564 posts and *Twitter* contains 2,415 posts. Note that in rare cases in *General* and *Music*, a post contains links to multiple object types, in which case that post is included twice in a column. Therefore the total counts in Figure 7.4 and Figure 7.5 are inflated by approximately 1%.

7.3.2 Data characteristics: The `boards.ie` SIOC Data Competition

We carried out an in-depth analysis of the whole `boards.ie` SIOC Data Competition dataset. We focus on this dataset because it covers a ten year time period, allowing us to observe changes in user behaviour over time. The properties of the dataset are described in detail in Table 7.4. Note that some encoding issues were encountered while processing the data, resulting in the omission of some

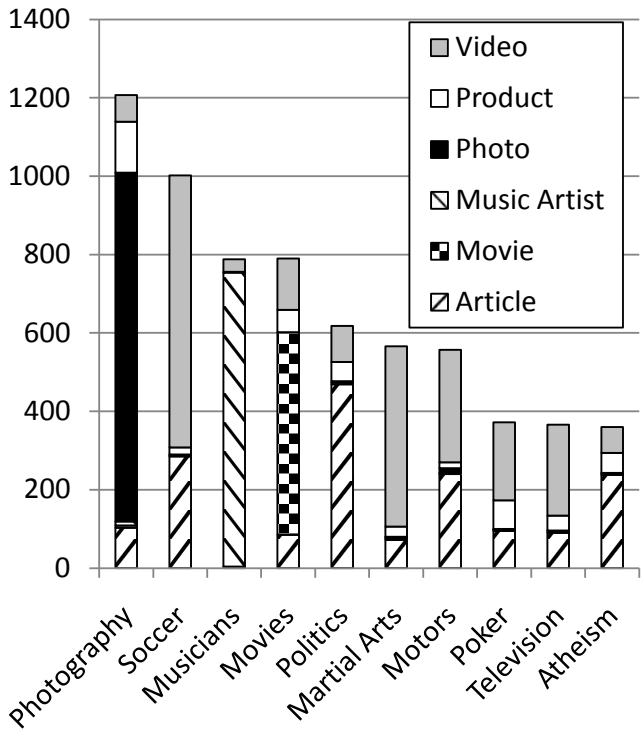

Figure 7.4: Number of posts containing links to each type of object for *General*

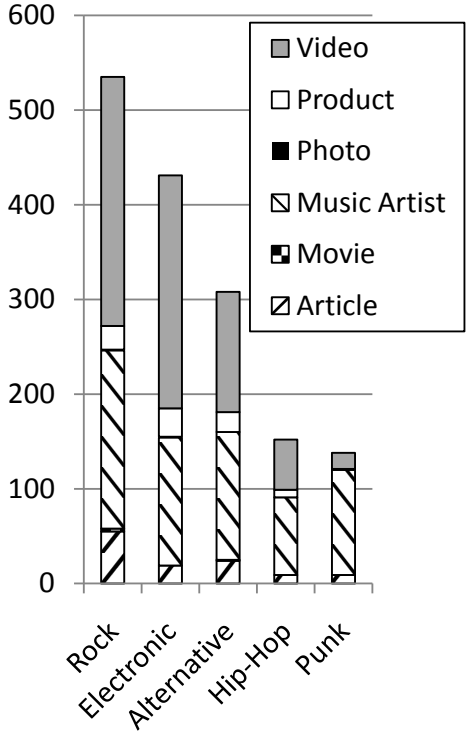

Figure 7.5: Number of posts containing links to each type of object for *Music*

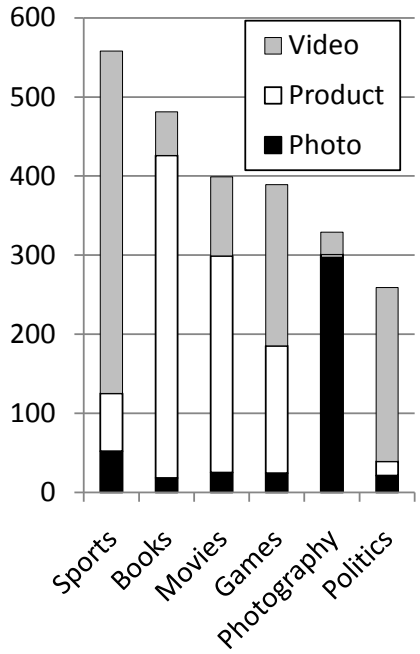

Figure 7.6: Number of posts containing links to each type of object for *Twitter*

posts (0.65% of total). The long time span covered by the dataset allows us to investigate the increase in links with related structured data available, and the size of the dataset enables us to extract enough links to study the usefulness of the data. The message boards are moderated so any spam or advertisement links have been removed from the messages, and consequently from the dataset used in our analysis.

Start Date	12/2/1998
End Date	13/2/2008
Users	138,139
Forums	967
Threads	657,169
Posts	7,794,495
External Links	625,723

Table 7.4: Basic properties of the SIOC Data Competition dataset

Figure 7.7 shows the growth of the message board in terms of posts, threads and forums. Note that Year 1 corresponds to the first twelve month period covered by the dataset, 1998/1999, and Year 10 corresponds to the last, 2007/2008. Usage of the message board has grown steadily, and in 2007/2008, approximately

2.2 million new posts were created. For the remainder of this section, we focus specifically on studying the properties of the links posted over the span of the dataset.

Frequency of hyperlinks posted

Figure 7.8 shows the growth of the number of external hyperlinks, in terms of total number of links, unique links, and domains linked to. We exclude syntactically invalid links and links internal to the domain of the message board. Note that we did not resolve all links, therefore multiple distinct hyperlinks may point to the same resource (or no valid resource). Figure 7.9 shows the percentage of posts containing links per year, and Figure 7.10 shows the average number of links in each of these posts. These figures show that over time, users are more frequently including links in their posts, and they are also including more links per post. Thus hyperlinks are becoming more prevalent in user conversations.

This increase in linking frequency shows a shift in user behaviours – in the initial stages of the message board, user attention was held almost entirely based on the text that participants created, whereas later users much more frequently made use of external resources to enrich their posts. One probable factor is that nowadays, there is simply more content available to link to. If a fan starts a thread about their favourite band, it requires very little effort to embed a video from YouTube, whereas finding a video online several years ago would have been much more difficult. A technical aspect that may have impacted linking habits is that message board software has improved over the years, and the task of inserting a link has become easier. The increasing practice of augmenting posts with links to external information suggests that perhaps an automatic method for doing this could be popular. An example of a service performing such a task is Zemanta,[9] an engine for blogs which analyses posts content and suggests related tags and webpages.

Locations of hyperlinks posted

It is not only the frequency of links that is relevant in online conversation, but also the stage at which they occur. Table 7.5 shows the percentage of posts

[9] `http://www.zemanta.com/`, accessed July 2011

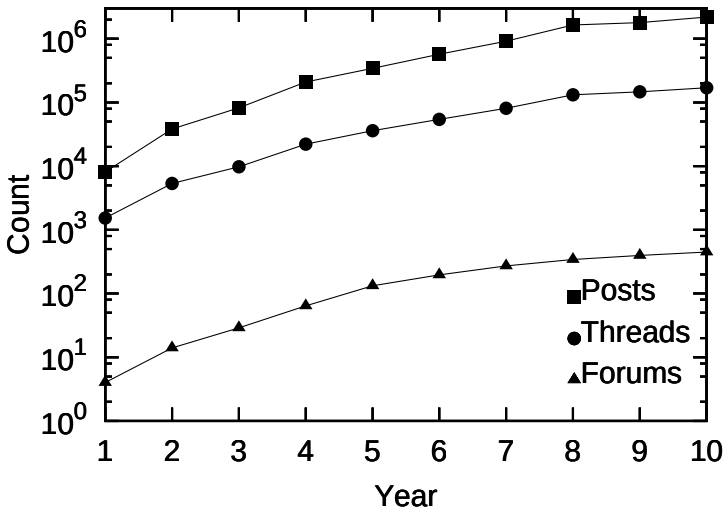

Figure 7.7: Posts, threads, and forums per year

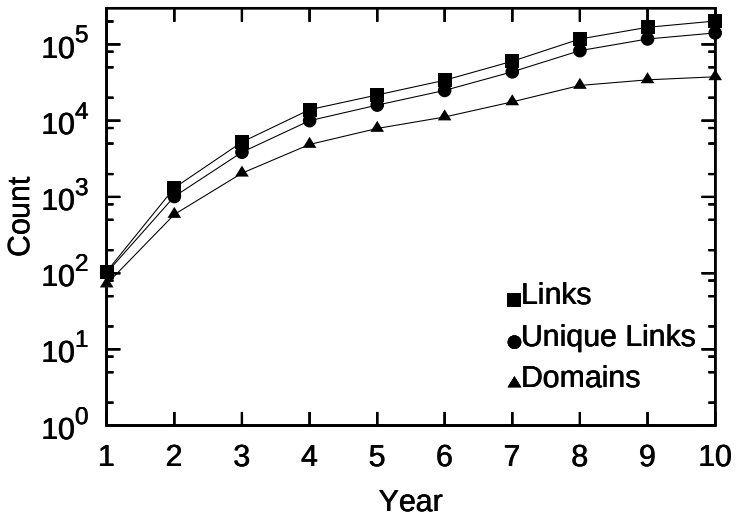

Figure 7.8: Links, unique links and domains linked to per year

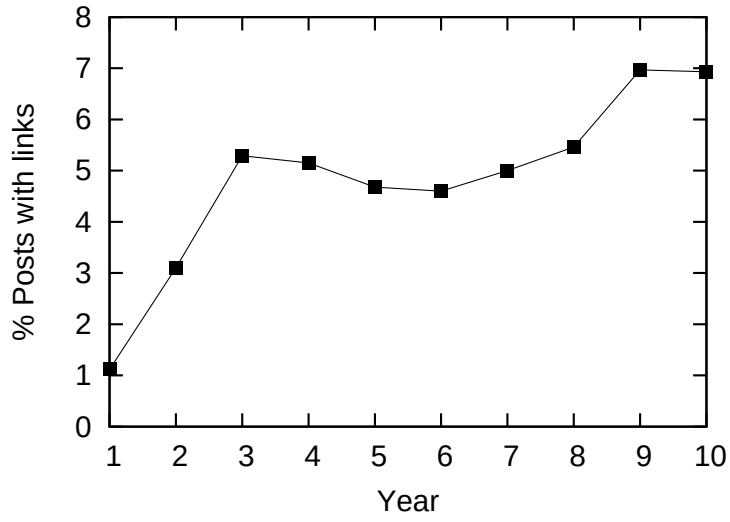

Figure 7.9: Percentage of posts containing links per year

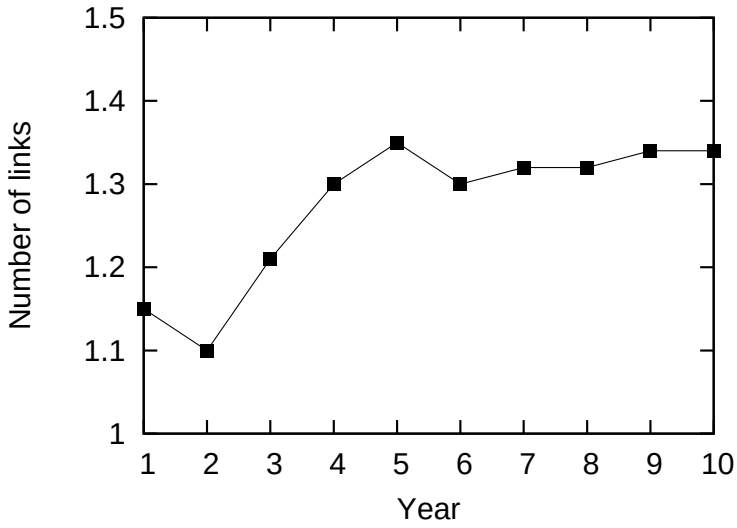

Figure 7.10: Average link count for all posts that contain links per year

with links, for all posts, for first posts of threads, and for subsequent posts. In first posts, where a conversation is being initiated, there is a far higher rate of link posting. This suggests that the link is important for the entire conversation that follows. It is likely that if we identify the object(s) described in that first link (movie, book, musicians, etc.), that would then enable us to have a better understanding of the focus of the conversation, and the main topic discussed in the rest of the thread.

Posts	% with link(s)
All posts	6.55%
First posts	14.61%
Subsequent posts	5.84%

Table 7.5: Percentage of posts containing hyperlinks, divided according to whether they are first in a thread

Targets of hyperlinks posted

We explored the domains that are most often linked to in the dataset and how their popularity changes over time. Table 7.6 shows the most popular sites linked to in 2002/2003, and Table 7.7 shows the most popular sites linked to in 2007/2008. We have labelled each site with a description of the type of content most commonly linked to on that site – note that this may not be the only type of content on the site, but it is the type that the message board's users most often linked to at this time.

The list of most popular domains during the early years of the message board is very different from the list of popular domains from recent years. Most notably, there has been a shift from unique, read-only sites created on Web hosting services, towards collaboratively-created read-write or content sharing sites with many contributors. These collaborative sites make it easier for users to put content online without requiring technical skills. Collaborative and sharing sites also typically feature an innate structure, and they often have associated APIs or Linked Data to enable data reuse, thereby yielding much more semantically rich data than traditional sites.

Rank and Domain	Dominant content type
1. bbc.co.uk	news media
2. komplett.ie	shop
3. ireland.com	news media
4. eircom.net	Web hosting
5. yahoo.com	news/discussion groups
6. rte.ie	news media
7. google.com	Web search
8. geocities.com	Web hosting
9. iol.ie	Web hosting
10. microsoft.com	technical support

Table 7.6: Top 10 domains linked to in 2002/2003

Rank and Domain	Dominant content type
1. youtube.com	video-sharing
2. wikipedia.org	collaborative encyclopedia
3. komplett.ie	shop
4. myspace.com	social networking/music
5. flickr.com	photo-sharing
6. bbc.co.uk	news media
7. rte.ie	news media
8. carzone.ie	shop
9. photobucket.com	media hosting
10. ebay.ie	shop

Table 7.7: Top 10 domains linked to in 2007/2008

Frequency of hyperlinks to structured datasets

Figure 7.11 shows the percentage of links in the data per year for which we could retrieve external data – *i.e.* where there is currently available a structured representation. We focus on the popular domains which we identified as being likely to have a useful structured data source i.e. `youtube.com`, `wikipedia.org`, `myspace.com`, `flickr.com`, `imdb.com` and `amazon.*`. Note that for the early years (1998-2005), very little data is available, but more recently the amount of structured data has grown rapidly. In total, we could retrieve structured data for links in 24,264 posts, the majority of which occurred in the last year of the dataset. For 2007/2008, 9% of links posted were to resources available as structured data. It is very likely that by now the percentage of links with

structured representations is much higher.

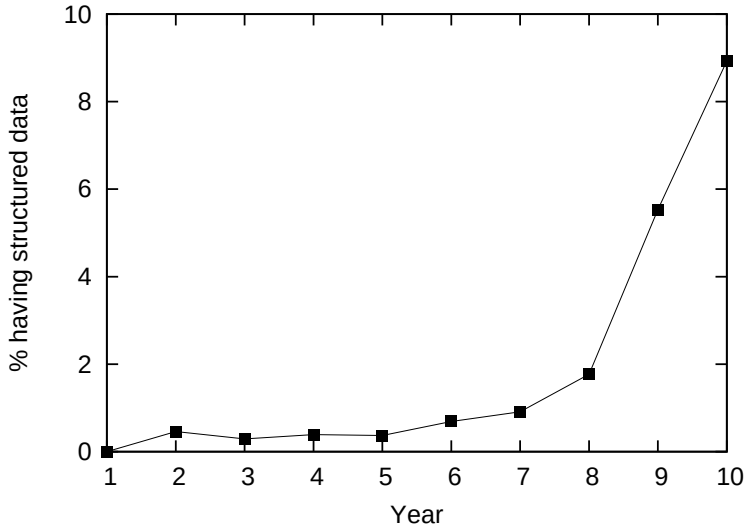

Figure 7.11: Percentage of hyperlinks posted for which there is structured data currently available

7.3.3 Data characteristics: External metadata for *General*

We now present a detailed analysis of the external metadata retrieved for *General*. The first section of Table 7.8 shows the percentage of non-empty metadata for each type of object. These figures are of interest since a metadata type that occurs rarely will have limited usefulness. Due to the unique features of each website, not every object type can have every metadata type. There are large variations in the percentage of non-empty features for different metadata types. Titles are typically essential to identify an object and categories are often required by a website's browsing interface, so these features are almost always present. For user-generated content, the frequency of non-empty fields is depends on whether the field is mandatory. For example, tags are often absent in Flickr because they are optional, while for videos they are almost always present because in the absence of user-provided tags, YouTube automatically assigns tags. For products, the author feature is often empty since this field is only available for books. For movies, the director feature is sometimes empty, which seems to be due to some inconsistencies in the various sources from which LinkedMDB integrates data.

	Title	Category	Description	Tags	Author/ Director
Average % of text features that are non-empty after preprocessing					
Article	100.0	100.0	99.7	-	-
Movie	100.0	100.0	-	-	39.9
Music artist	99.7	100.0	-	-	-
Photo	100.0	-	58.8	84.9	-
Product	100.0	100.0	-	75.2	65.4
Video	100.0	100.0	99.5	99.5	-
Average unique metadata tokens for non-empty fields (± standard deviation)					
Article	2.1 ± 0.9	13.6 ± 12.1	15.8 ± 8.3	-	-
Movie	1.7 ± 0.7	4.1 ± 1.8	-	-	2.2 ± 0.6
Music artist	1.8 ± 0.9	2.7 ± 0.9	-	-	-
Photo	2.0 ± 1.1	-	10.9 ± 17.2	6.5 ± 4.9	-
Product	5.2 ± 3.0	11.5 ± 7.8	-	5.7 ± 2.1	2.0 ± 0.4
Video	3.7 ± 1.6	1.0 ± 0.0	13.1 ± 26.3	7.2 ± 5.0	-
Average % of unique metadata tokens that do not occur in post content					
Article	0.0	78.5	68.4	-	-
Movie	17.4	76.2	-	-	43.3
Music artist	10.1	85.4	-	-	-
Photo	72.5	-	50.3	74.6	-
Product	39.5	81.0	-	51.1	32.2
Video	62.0	95.7	78.5	74.4	-

Table 7.8: Properties of external metadata content for *General*

The second section of Table 7.8 shows the average number of unique tokens found in non-empty metadata fields. These figures are an indicator of how much information each feature provides. In general, titles and authors/directors provide few tokens since they are quite short. For categories, the number of tokens depends on whether the website allows multiple categories (*e.g.*, Wikipedia) or single categories (*e.g.*, YouTube). The number of unique tokens obtained from descriptions and tags are quite similar across all object types studied.

The third section of Table 7.8 gives the average percentage of unique tokens from metadata that do not occur in post content. This section is important since it shows which features tend to provide novel information. Note that for article titles, the percentage is zero since all titles are contained within the article's URL. For music artist titles, the figure is low since bands often use their title as their Myspace username, which is contained within the artist's URL. All other object

types have URLs that are independent of the object properties. This section also allows us to see how users typically describe an object. For example, 40% of the tokens from product titles are novel, indicating that posters often do not precisely name the products that they link to. For the subset of products that are books, 23% of tokens from titles were novel. Approximately 32% of the tokens from book authors and 43% of the tokens from movie directors are novel, showing that posters often mention these names in their posts, but that in many other cases this is new information which can aid categorisation.

7.4 Evaluation

7.4.1 Experimental setup

For each post, the following representations were derived, in order to compare their usefulness as sources of features for topic classification:

Content (without URLs): Original post content with hyperlinks removed.

Content: The full original content with hyperlinks intact.

HTML: The text parsed from the HTML document(s) to which a post links. HTML tags were stripped out and only the visible text of the webpage was retained.

Metadata: The external metadata retrieved from the hyperlinks of the post.

For any cases where a HTML document was not retrievable, this object was removed from the dataset.

Classification experiments were performed using the Weka toolkit (Hall et al., 2009). Document length normalisation and tf-idf weighting were applied to each feature vector. We used the default Weka setting of a maximum 1,000 words per class. Aggregate feature vectors were generated for the combinations of Content+HTML and Content+Metadata. An aggregate vector for two text sources was obtained by adding their individual feature vectors, after document length normalisation and tf-idf weighting. Two methods were tested for combining different sources of textual information into a single vector:

Bag-of-words (BOW): The same term in different sources was represented by the same element in the document vector. For these experiments, we tested different weightings of the two sources, specifically {0.1:0.9, 0.2:0.8, ... , 0.9:0.1}. Two vectors v_1 and v_2 were combined into a single vector v where a term i in v is given by, for example, $v[i] = (v_1[i] \times 0.1) + (v_2[i] \times 0.9)$.

Concatenate (CON): The same term in different sources is represented by different elements in the feature vector – *i.e.*, "music" appearing in a post is distinct from "music" in a HTML page. Two vectors v_1 and v_2 were combined into a single vector v via concatenation, *i.e.*, $v = \langle v_1, v_2 \rangle$.

The text classifier used in our experiments was the Multinomial Naïve Bayes classifier implemented in Weka. A ten-fold cross-validation was used to assess the performance of the classifier on each type of document representation. K-fold cross-validation involves randomly splitting a dataset into K folds, and using one fold as test data and the remaining $K-1$ folds as training data. The process is repeated so that each of the K folds is used as a test set exactly once. Duplication of hyperlinks across splits was disallowed, so the metadata of a particular object cannot occur in multiple folds. In order to avoid duplication of post content due to one post quoting another, *General* and *Music* were split by thread so that multiple posts from one thread do not occur in separate folds. Duplication was not an issue in *Twitter* since retweets had been removed. These restrictions resulted in the omission of approximately 11% of the *General* and *Music* posts from any fold.

The accuracy of classification for each representation is measured using the F_1 measure, which takes into account both precision p and recall r and is defined as $F_1 = \frac{2.p.r}{p+r}$. Micro-averaged F_1 is calculated by averaging F_1 over each test instance and is therefore more heavily influenced by common categories, while macro-averaged F_1 is calculated by averaging F_1 over the result for each category and is therefore more heavily influenced by rare categories.

7.4.2 Experimental results

The results of the classification experiments for each post representation are shown in Table 7.9 with their 90% confidence intervals. The confidence intervals

Data Source	*General*	*Music*	*Twitter*
Content (without URLs)	0.745 ± 0.009	0.654 ± 0.021	0.722 ± 0.019
Content	0.811 ± 0.008	0.683 ± 0.023	0.759 ± 0.015
HTML	0.730 ± 0.007	0.548 ± 0.024	0.645 ± 0.020
Metadata	0.835 ± 0.009	0.766 ± 0.029	0.683 ± 0.018
Content+HTML (BOW)	0.832 ± 0.007	0.736 ± 0.022	0.784 ± 0.016
Content+HTML (CON)	0.795 ± 0.004	0.678 ± 0.024	0.728 ± 0.016
Content+Metadata (BOW)	0.899 ± 0.005	0.830 ± 0.025	0.820 ± 0.013
Content+Metadata (CON)	0.899 ± 0.005	0.820 ± 0.026	0.804 ± 0.018

Table 7.9: Micro-averaged F_1 for *General*, *Music* and *Twitter* (± 90% Confidence Interval)

were determined using the t-test on the results of the ten-fold cross-validation. For all datasets, classification results based on content improve when tokens from URLs within posts are included. Classification using only the HTML pages linked to by posts gives relatively poor results, while classification using only metadata from hyperlinked objects improves accuracy for both of the message board datasets, but decreases accuracy for *Twitter*. Those differences are all statistically significant. For the combined representations, the bag-of-words representation gives slightly better results than concatenation. For both HTML and Metadata, a bag-of-words combination with Content outperforms results for Content alone. The Content+Metadata approach significantly outperforms the Content+HTML approach, for both datasets.

The results reported in Table 7.9 for the bag-of-words representations are for the best-performing weightings. Figures 7.12 and 7.13 shows the performance of the weightings tested for Content+HTML (BOW) and Content+Metadata (BOW). A content weight of 1 is equivalent to classification based exclusively on content, while a weight of 0 is equivalent to classification based exclusively on HTML or metadata. The results show that combining content with either HTML or metadata proves beneficial in comparison to classification based on any single source. The Content+HTML weightings shown in Figure 7.12 follow similar patterns for all three datasets, with optimal performance achieved when content is relatively highly weighted. The results for the Content+Metadata weightings shown in 7.13 indicate that the two message board datasets (*General* and *Music*) behave in a similar manner, with the highest accuracy achieved when

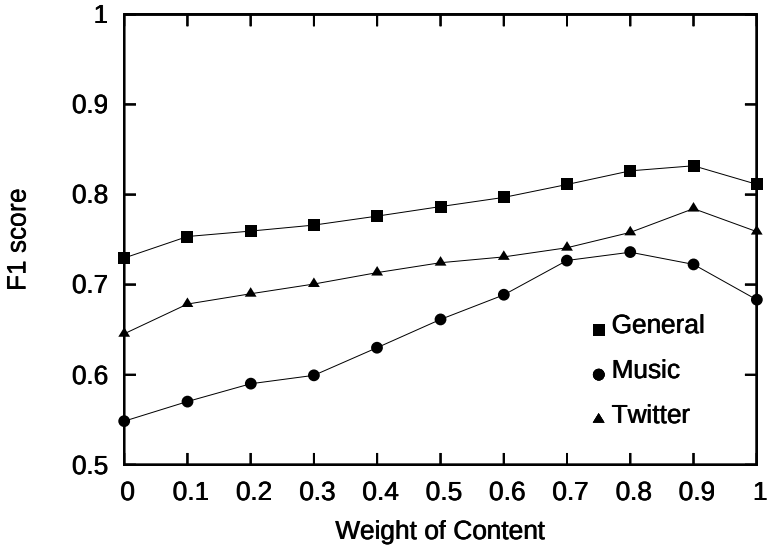

Figure 7.12: Performance of weightings tested for Content+HTML (BOW)

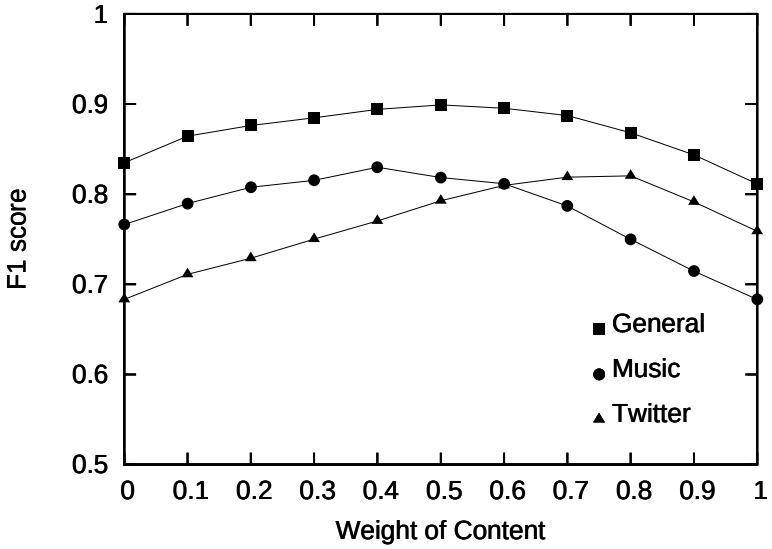

Figure 7.13: Performance of weightings tested for Content+Metadata (BOW)

there is a fairly even balance between content and metadata. However the *Twitter* dataset shows a different trend, with optimal performance achieved when content is emphasised.

Tables 7.10, 7.11 and 7.12 shows the detailed results for each category in each dataset, for Content, Metadata and Content+Metadata (using the bag-of-words weighting with the best performance). There is a large variation in classification results for different categories. Despite the variation between categories, Content+Metadata always results in the best performance. For the two single source representations, some categories obtain better results using Content and others using Metadata. The higher result between these two representations is highlighted with italics. For *Music*, Metadata always outperforms Content while for the other two datasets the better-performing data source varies. This variation could be partially due to the fact that in certain categories, there are good indicators of topic available from the URLs in the post content. For example, in *General*, links to Myspace music artists and Flickr photos are strongly associated with the Musicians and Photography categories respectively, which means that posts containing hyperlinks to these websites can be easily classified using post content. As a result, the accuracy within the Musicians and Photography categories shown in Table 7.10 is higher when using Content rather than Metadata, in contrast to most of the other categories. Another example from the *General* results is that the Politics and Atheism categories, which contain many links to Wikipedia articles (see Figure 7.4), both have higher accuracy for Content than Metadata. This is probably due the fact that Wikipedia URLs contain article titles, so the post content already contains some object metadata as well as the post text.

Table 7.13 shows the gains in accuracy achieved by performing classification based on various individual metadata types. The results shown are for Wikipedia articles and YouTube videos in the *General* dataset, and Myspace artists in the *Music* dataset. We limit our analysis to these object types because they have consistently good coverage across all of the categories, apart from the Musicians category in *General* which we hence excluded from this analysis. These results are based only on the posts with links to objects that have non-empty content for every metadata type and amount to 1,623 posts for Wikipedia articles, 2,027 posts for YouTube videos and 1,376 posts for Myspace artists. We compare the

Category	Content w/o URLs	Content	Metadata	Content+ Metadata
Musicians	0.921	*0.973*	0.911	0.981
Photography	0.769	*0.922*	0.844	0.953
Soccer	0.793	0.805	*0.902*	0.945
Martial Arts	0.755	0.788	*0.881*	0.917
Motors	0.712	0.740	*0.869*	0.911
Movies	0.705	0.825	*0.845*	0.881
Politics	0.773	*0.791*	0.776	0.846
Poker	0.602	0.646	*0.757*	0.823
Atheism	0.735	*0.756*	0.732	0.821
Television	0.524	0.559	*0.664*	0.716
Macro-Avgd	0.729	0.781	0.818	0.879

Table 7.10: F_1 of classifier for each category in *General*, ordered by performance

Category	Content w/o URLs	Content	Metadata	Content+ Metadata
Hip-Hop	0.603	0.690	*0.804*	0.876
Electronic	0.698	0.709	*0.788*	0.864
Rock	0.687	0.717	*0.806*	0.857
Punk	0.639	0.706	*0.785*	0.846
Alternative	0.573	0.583	*0.644*	0.719
Macro-Avgd	0.640	0.681	0.766	0.832

Table 7.11: F_1 of classifier for each category in *Music*, ordered by performance

Category	Content w/o URLs	Content	Metadata	Content+ Metadata
Books	0.788	0.804	*0.836*	0.877
Photography	0.648	*0.785*	0.728	0.842
Games	0.767	*0.772*	0.675	0.830
Movies	0.697	0.718	*0.777*	0.827
Sports	0.706	*0.744*	0.563	0.781
Politics	0.666	*0.685*	0.499	0.733
Macro-Avgd	0.712	0.751	0.680	0.815

Table 7.12: F_1 of classifier for each category in *Twitter*, ordered by performance

Metadata Type	Content w/o URLs	Metadata Only	Content+ Metadata
Wikipedia articles (*General*)			
Category	0.761 ± 0.014	0.811 ± 0.012	0.851 ± 0.009
Description		0.798 ± 0.016	0.850 ± 0.009
Title		0.685 ± 0.016	0.809 ± 0.011
YouTube videos (*General*)			
Tag	0.709 ±0.011	0.838 ± 0.019	0.864 ± 0.012
Title		0.773 ± 0.015	0.824 ± 0.013
Description		0.752 ± 0.010	0.810 ± 0.013
Category		0.514 ± 0.017	0.753 ± 0.014
MySpace music artists (*Music*)			
Category	0.649 ± 0.030	0.805 ± 0.027	0.798 ± 0.016
Title		0.679 ± 0.039	0.719 ± 0.027

Table 7.13: Micro-averaged F_1 for classification based on selected metadata types (± 90% Confidence Interval)

results against Content (without URLs), because Wikipedia URLs contain article titles and our aim is to measure the effects of the inclusion of titles and other metadata. The results for different metadata types vary considerably. For posts containing links to Wikipedia articles, the article categories alone result in a better classification of the post's topic than the original post content, with an F_1 of 0.811 compared to 0.761. Likewise, for posts that contain links to YouTube videos, the video tags provide a much better indicator of the post topic than the actual post text. For posts that link to Myspace music artists, the artist category (*i.e.* genre) is the source of textual features that results in highest accuracy. The Content+Metadata column shows results where each metadata type was combined with post content (without URLs), using a bag-of-words representation with 0.5:0.5 weightings. Every metadata type examined improved post classification relative to the post content alone. However some metadata types improve the results significantly more than others, with Content+Category achieving the best scores for articles, and Content+Tags achieving the best scores for videos. In addition to testing each metadata type individually, it is possible to perform a sweep of possible combinations of content and metadata types with various weightings, in order to identify the weightings necessary to create an optimal post representation for topic classification.

7.5 Conclusions

In this chapter, we have investigated the potential of using metadata from hyperlinked objects for classifying the topic of posts in online forums and microblogs. The approach could also be applied to other types of social media. Our experiments show that post categorisation based on a combination of content and object metadata gives significantly better results than categorisation based on either content alone or content and hyperlinked HTML documents. We observed that the significance of the improvement obtained from including external metadata varies by topic, depending on the properties of the URLs that tend to occur within that category. We also found that different metadata types vary in their usefulness for post classification, and some types of object metadata are even more useful for topic classification than the actual content of the post. We conclude that for posts which contain hyperlinks to structured data sources, the semantically-rich descriptions of these entities can be a valuable resource for post classification. For posts which contain hyperlinks to webpages that do not have a structured representation, we have shown that including the plain text from HTML documents also improves classification, to a lesser degree. For posts which do not contain hyperlinks, a simple content based approach can be applied (Garcia Esparza et al., 2010).

Although our evaluation was carried out on objects from specific domains, the approach is applicable to objects from arbitrary domains, as long as they have been represented in a structured format and using common vocabularies. For example, if multiple publishers of movie related data all use Facebook's Open Graph Protocol to represent their data, then a classifier system can automatically recognise movies from these publishers, and weight their metadata accordingly. Therefore our approach which considers RDF mark-up provides a feature which is not available to classifiers that only consider HTML documents – the ability to identify semantically-equivalent fields across websites which may have completely different HTML structures.

Our approach exploits the fact that online conversations often revolve around objects, in line with the theory of object-centred sociality, and that these objects often have structured representations. Participants in social media use hyperlinks to provide references to objects, so that their readers understand what

exactly they are talking about and can find further information. We have shown that these descriptions can also be used for automatically augmenting posts with additional relevant data for the purpose of topic classification. The approach demonstrates how structured data can be used to aid navigation even in unstructured Web content. The categories assigned by this approach would allow a user to browse social media posts with hyperlinks by topic, even if the text of the post itself is not sufficient for accurate automatic categorisation of the post.

Part IV

Conclusion

Chapter 8

Summary and Future Directions

Social media enables millions of members of online communities to hold conversations, share content, and work collaboratively. As a result of this activity, the Social Web is becoming an increasingly important and ubiquitous information resource. Compared to traditional media such as newspapers, books and magazines, social media offers a cheap, accessible and rapid way of accessing or publishing information across the globe. However the same qualities of freedom and openness that make it such a useful source of knowledge also result in issues in data organisation and retrieval. Posts are often short, informal, and lacking in metadata and contextual information.

The goal of this thesis was to find ways of enriching social media with metadata, by using the contextual information available from related data on the Web. We have presented approaches to infer three types of commonly used metadata: tags (social annotations, which can describe any aspect of a resource), location (geographic information, at varying levels of granularity), and topic (thematic classifications). The methods that we presented take advantage of a variety of Web data – plain HTML pages, structured data from APIs, and semantically rich Linked Data. Further, we demonstrate how the generated metadata can be represented using open standard formats so that distributed social media data from multiple formats can be integrated together and queried in a consistent way.

In this concluding chapter we describe how the approaches of Chapters 5, 6 and 7 can be applied together to a social media dataset in order to augment posts with tag, location and topic metadata. We summarise the contributions of this thesis and discuss potential future research directions.

8.1 Converging the Approaches

The core part of this thesis presented methods for automatically inferring tag, location and topic information for social media items. We now illustrate how these approaches can be combined to automatically annotate a post with structured metadata which can be exploited for enhanced search and navigation. Figure 8.1 shows a flowchart which summarises the processes involved in the metadata generation approaches presented in this thesis. It is assumed that the topic classifier and location models have already been trained, and hence the training process for these models is omitted from the flowchart.

Figure 8.1 also contains two simple possible extensions to the metadata generation process, shown as dotted lines, which reuse existing functionality from elsewhere in the process. Firstly, the metadata extracted from hyperlinked objects in preparation for topic classification could also be propagated to the location models in order to improve location prediction. The hyperlinks embedded in social media posts may contain geographic information which can help pinpoint the origin of those posts. Secondly, for hyperlinked objects which have missing or sparse metadata, the tag prediction approach could be used to infer annotations for these objects. These predicted tags could then be used in addition to existing metadata to augment the post representation for topic classification. These extensions make further use of the Web graph for metadata generation and may allow additional or more accurate annotations to be generated.

All of the methods shown in the flowchart, including both our implemented approaches and the proposed extensions, involve viewing a post not just as an isolated piece of data, but an object which exists in a rich network of interlinked data that provides vital context for the post. It is only by taking advantage of this background information that we can succeed in implementing accurate methods for automatically annotating or classifying social media items on the Web.

For a given social media item, we can apply some combination of the approaches presented in this thesis, depending on whether the item has incoming hyperlinks, outgoing hyperlinks, and textual content. In the absence of these features, other standard approaches from the state of the art, such as those outlined in Section 4.3, could be applied. In practice, it may not always be appropriate to infer all three types of metadata for every post. For example, a post may

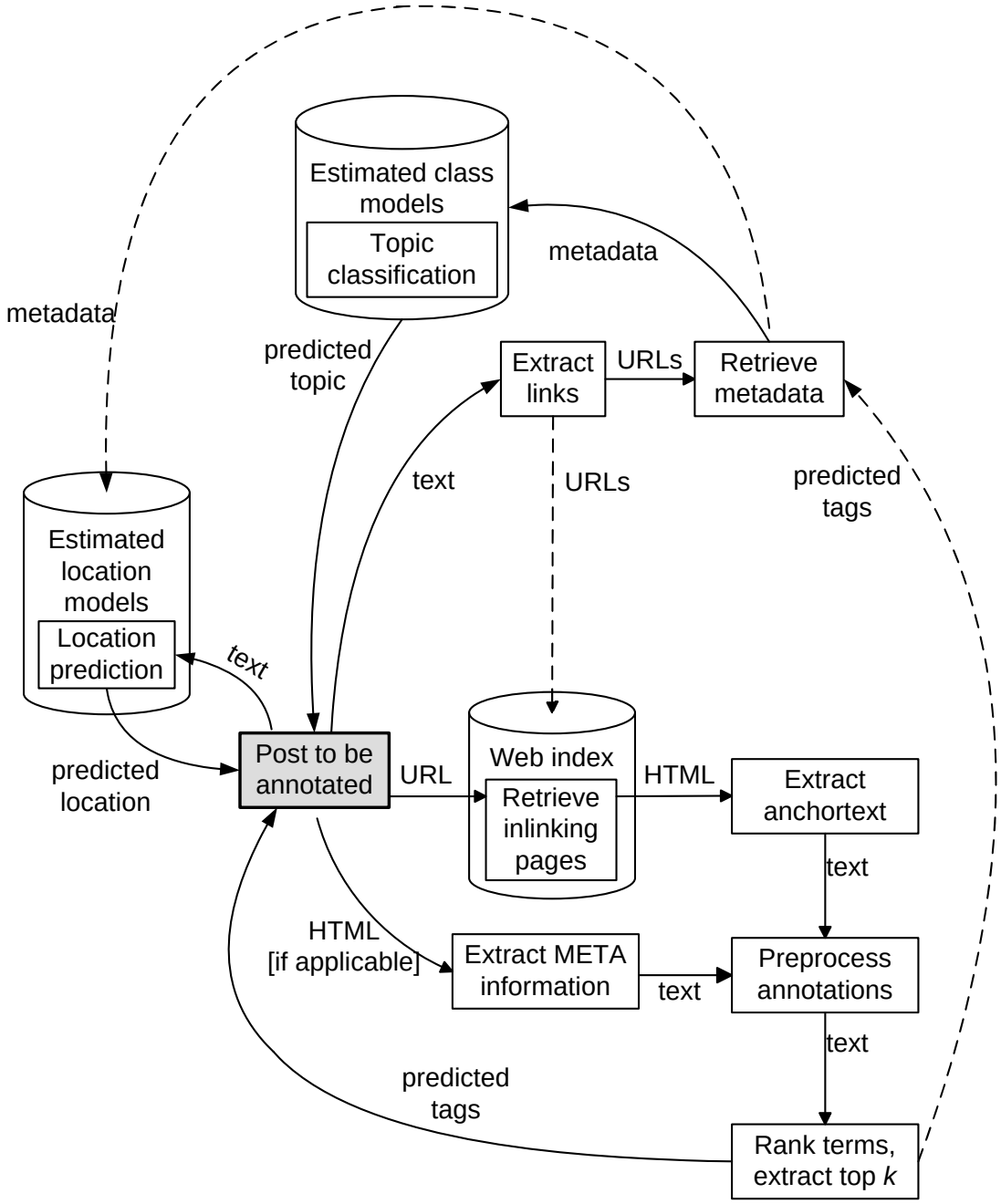

Figure 8.1: Flowchart for the metadata prediction approaches

have a strong topical focus but no geographic information. In these cases, we could introduce a probability threshold for each metadata item, and only assign predicted annotations for which there is sufficient evidence.

8.2 Contributions

This thesis has contributed to the state of the art by providing the following three approaches for automatic metadata generation in social media, as well as showing how they can be used in combination to annotate social media items from any source in a standard and uniform way, thus enabling enhanced search and browsing of cross-platform social media collections.

Tag prediction

We have proposed a method for generating tags for items from the anchortext of incoming hyperlinks. Our approach uses a vector space model to identify the top ranking tags for a given resource. A human evaluation comparing the tags predicted by our method to those assigned in Delicious showed that our method gave reasonable results (precision@5 of 0.66) compared to the user-assigned tags (precision@5 of 0.78). This approach allows preexisting social annotations in the form of anchortext to be re-used in automatic tag generation for untagged social media items. These tags help form a social index for an object, allowing new annotations to be created which may not have existed in the vocabulary of the original post.

Geolocation

We have presented a method for predicting the location of origin of a social media post or user. Our approach makes use of the content of geotagged items in order to build language models at different levels of geographic granularity. In a microblog dataset, after excluding tweets from location-focused services, we were able to correctly place 52% of tweets to the correct country, 22% to the correct city, and 5% to the correct postal code. At the postal code level, our approach outperformed all of the other alternative methods that we investigated, including geoparsing the user's self-reported location. This approach uses the implicit

location clues that occur in tweets to generate models based on their geotags, which can then be used to detect implicit location information in subsequent non-geotagged tweets. The predicted locations can be used to enable hyperlocal search and browsing of social media.

Topic classification

We have proposed a method for automatically classifying the topic of social media posts, which takes advantage of objects linked to within posts, and in particular makes use of their metadata in those cases where a structured representation is available. In a message board dataset, a classifier trained using our approach of combining post content with object metadata achieved an F_1 score of 0.90, significantly outperforming a classifier trained on post content (0.81) or post content and hyperlinked HTML pages (0.83). Experiments on a microblog dataset showed yielded similar results. We also show that our approach can be used to identify the most relevant types of object metadata for categorising posts which link to a certain type of object. This approach shows how structured data on the Web can be beneficial for generating new topic metadata in originally unstructured social media items. The thematic categories assigned to items can be used to enable users to browse conversations about an area of interest, or to filter social media search results to a desired topic.

8.3 Future Directions

This thesis has presented research at the intersection of the Social Web and the Semantic Web. The approaches we have proposed both produce and consume structured social data, with the goal of improving information-finding in online communities. Our work opens up new directions for future research on the automatic generation of semantically-rich metadata for social media items, and the utilisation of such metadata in novel tools for searching and exploring social media. In this section we discuss some of these potential research areas. We also examine the possibility that as these techniques become more widespread on the Web, they may become a target for spamming or other malicious activity, and we outline potential approaches for combating such attacks.

8.3.1 Further integration with the Web of Data

We now describe how the methods of this thesis can be more tightly integrated with structured data on the Web, by making more use of structured data in our metadata prediction, by adding more structure to the data we produce, and by using the generated metadata to bootstrap Semantic Web applications.

Exploiting the full potential of semantically-rich data

One possible enhancement to our work would be to make more use of semantically-rich data in the metadata prediction process. In our topic classification experiments, we used the structure of the external data to identify which types of metadata provide the most useful textual features for improving the accuracy of categorisation. However in addition to providing textual metadata, the Linked Data sources are also part of a rich interconnected graph with semantic links between related entities. We have shown that the textual information associated with resources can improve categorisation, and it would be interesting to also make use of the semantic links between concepts.

For example, imagine a post is created which contains a hyperlink to the comedy series dbpedia:Fawlty_Towers and is manually assigned to the Television category. A later post that links to the comedy series dbpedia:Mr_Bean could be automatically classified as belonging to the same category, due to the fact that the concepts are linked in several ways in DBpedia's RDF graph, including through their genres and the fact that they are both produced by British television channels. Just as we used machine-learning techniques to identify the most beneficial metadata types for topic classification of posts, we could also experimentally identify the property paths between entities that are most indicative of topical similarity.

It would also be possible to enhance our location prediction method to take advantage of the Semantic Web. The approach could be extended to extract structured data from hyperlinks in posts in order to infer semantic relationships between posts and locations via the hyperlinked objects. For example, posts which link to the concept dbpedia:Guinness are likely to have a heightened probability of originating from Ireland.

Producing more semantically-rich data

We could also further interlink the metadata we produce with the Web of Data. Although we have already provided a schema for expressing the metadata inferred by our methods in RDF, we could still add more semantic structure to the data. As we point out in Section 1.4, semantifying known locations and predefined topics is straightforward, but semantifying tags requires additional processing. Existing approaches for adding sementics to tags include Tesconi et al. (2008) and García-Silva et al. (2009), which both disambiguate tagsets to Wikipedia concepts based on the article content. By combining our tag prediction approach with one of these tag disambiguation approaches, we could automatically generate semantic tags for untagged social media items, thereby annotating posts with semantically-meaningful concepts rather than free text keywords. The Meaning Of A Tag (MOAT) vocabulary (Passant & Laublet, 2008) could be used to denote these concepts as semantic tags for a post.

The use of semantic tags could also make it unnecessary to consider tags and topics separately, since inference could be used to determine whether an item with certain tags is related to a topic. For example, consider the use-case of Section 1.5.2, where a user wishes to browse local sports information. A system augmented with semantic tag prediction could annotate a post whose anchortext includes *gymkhana* with the concept dbpedia:Gymkhana_(equestrian). Since this concept occurs under dbpedia:Sport in the DBpedia category hierarchy, the system can infer that a post annotated with this URI is relevant to the user's information need.

Bootstrapping Social Semantic Web applications

In Section 3.4.5 we described Semantic Web applications that make use of structured social media items integrated from multiple sources for applications including social recommendations, expert-finding, and enhanced visualisations. These applications depend on the presence of metadata in a structured format, including tag, location and topic information. Therefore the coverage of these applications is limited because much of the content on the Social Web is lacking such metadata. Many relevant items may be overlooked by an application because the location or topic is not explicitly stated. This lack of coverage means that the

full potential of Social Semantic Web applications is not realised.

The approaches that we presented in this thesis can support these applications by allowing systems that consume social media data to automatically infer missing metadata for social media items. By integrating automatic metadata generation into the data collection process, Social Semantic Web applications would be less dependent on data publishers to provide complete structured representations of posts and would be able to integrate content from a wider range of sources.

8.3.2 Combating malicious activity on the Social Semantic Web

In this thesis we have focused on the problem of making uncurated and poorly-annotated social media data easier to search and navigate. The approaches that we propose make use of structured and interlinked Web data to add metadata to social media posts and thus contribute to a Social Semantic Web of interconnected and semantically-rich user-generated content. An important factor to consider when working with user-generated content is the credibility of information on the Web. Any technique which makes use of data from the open Web is vulnerable to exploitation since the input data is unverified and unpredictable. Some malicious users publish false or misleading information in an attempt to gain advantages such as increased visibility in search results. Examples of such activity include link farms, where a group of websites all link to each other in order to exploit ranking algorithms that involve link analysis, or keyword stuffing, a search engine optimization technique which involves placing popular terms in the content or META element of a webpage, even if they have no connection to the document topic. These methods are well known and search engines have developed ways to detect and counteract such behaviour. Two common approaches used by search engines are analysis of the graph structure of the Web to identify and remove link farms from their index, and content analysis of webpages in order to learn the typical characteristics of spam documents and thus automatically detect them. It is likely that as the Semantic Web increases in popularity, similar malicious behaviour will emerge which exploits semantic data on the Web. If a system makes use of structured data from external sources, as the approaches that we propose do, then the reliability of the external data must be considered, just

as on the HTML Web. Some of the anti-spam solutions from traditional Web information retrieval may be applicable to structured data, but it may also be necessary to develop defenses against attacks specific to the Semantic Web.

Social media is particularly attractive to spammers and other malicious users as a way to entice Web users to their webpages by misleadingly advertising their content, and as a medium to virally spead hyperlinks or false information in order to attract more users. As a result, spam and other types of abuse are a very common problem in social media and user-generated content. A recent study by Grier et al. (2010) found that 8% of a large sample of URLs posted on Twitter pointed to malware, scams or phishing websites. Often, spammer's techniques involve creating misleading content or metadata. There has been significant research activity on detecting illegitimate usage of social media. As on the Web in general, graph analysis techniques have been applied to social media for the detection of malicious behaviour, using the social graph or content graph rather than Web documents. Content analysis is another valuable tool in detecting abuse in user-generated content. An assessment of the trustworthiness of an item can also be crowdsourced by examining quality ratings, comments or other attention measures. Existing systems to combat malicious activity in social media often use combinations of graph analysis, content analysis and crowdsourcing in order to verify the trustworthiness of information and of users. In general, these approaches are equally applicable to the Social Semantic Web. However integrating semantics into social media results in some new problems for combatting spam, and also enables novel potential solutions. We now discuss how algorithms which rely on structured user-generated content could be exploited and what measures could be considered in order to defend against malicious behaviour. We focus in particular on solutions which take advantage of Semantic Web techniques.

Graph analysis

A consequence of integrating semantics into the Social Web is that by providing common vocabularies, structurally rich user-generated content and social network data can be easily distributed across different interlinked sources. This has the advantage of enabling users to control their own data, to make claims about arbitrary Web resources, and to connect to friends who use different services without needing to create new accounts and port across existing data. However

there are also challenges which result from this approach. When data is not under the control of a centralised authority, it is possible to make false assertions, such as incorrectly stating that a video has a certain topic or creator, by reusing existing legitimate URIs. This presents a predicament – while there is now a rich network of decentralised and semantically rich social data, where people and objects can be referenced across the Web by their URIs, it is not reasonable to presume that all of these data publishers are reliable.

One possible solution for handling this issue is to automatically determine the reputation of data sources. Related work includes (Harth et al., 2009), where link analysis was performed on a source-level link graph in order to score data publishers in terms of authority and subsequently rank data items published by them. Using this method, sources whose URIs are often reused by other sources are ranked as being more authoritative. The same method could be used on the Social Semantic Web to determine whether a community recognises a publisher as being trustworthy based on the assumption that if other users reuse identifiers coined by a source, they consider this publisher legimate. Assertions made by a source with a high authority score could be accepted and presented to users or used in a reasoning engine, while assertions made by a source with low authority could be rejected or given a lower weighting.

Content analysis

Whether a system is handling social media from one source or from multiple distributed sources, assessing the accuracy of individual metadata items is important. Although the ability to annotate objects with well-defined concepts instead of free-text allows the assignment of metadata items which have unambiguous meaning and are linked to related concepts, it does not ensure that this structured metadata is actually correct. The accuracy of metadata on the Semantic Web is especially important because unlike on the HTML Web, this information is not just presented to the user who can its credibility for themselves – instead, the information may be integrated with more information and inferences made to generate new knowledge, or in the case of the methods presented in this thesis, infer new metadata. Therefore a system which relies on structured user-created metadata would benefit from methods to verify that individual metadata items are accurate.

An example of a potential attack involves the approach of Chapter 7, which uses external structured data from hyperlinks for topic classification. A system which implements such an approach would rely on the accuracy of the external structured data in order to correctly classify posts. Malicious data publishers could exploit this system to achieve attention from people who search for a particular topic unrelated to their content. A publisher could gain legitimate popularity with some content item, while attaching misleading metadata in RDF that indicates that the item belongs to an irrelevant topic of the publisher's choice. Hence documents which link to the malicious data publisher would be inappropriately classified, and would lead to users accessing the content under the impression that it is relevant to their topic of interest.

A potential solution for this scenario, and for the verification of metadata in general, would be to use content analysis to ensure that the metadata assigned to each object is in some way relevant to the content. Apart from simple text analysis, background knowledge from ontologies could also be exploited in order to ensure that the metadata assigned to objects is genuine. For example, a paper about SPARQL may be assigned a sioc:topic of dbpedia:Semantic_Web even though the document may not necessarily contain those words. By looking at concepts related to dbpedia:Semantic_Web in the RDF graph, it can be easily inferred that this concept is an appropriate and trustworthy annotation for the document – a fact that would be more difficult to infer from simple text analysis.

Crowdsourcing

Often, the quality of information on the Web can be judged by how the community reacts to it. On the Web 2.0, it is common for websites to allow users to rate, re-share, comment on, or link to content. The Semantic Web enables more meaning to be added to these actions, which can then be used as a factor in assessing the credibility of an item. For example, is a user commenting on a data item to criticise it or to praise it? A system could be developed which interprets the sentiment of typed links and thus determine whether a certain item is gaining good feedback or bad.

In addition, there may exist background information about users on the Web which could be used in order to determine their authority as a rater of items on a certain topic. If we can determine from a user's FOAF file and structured Web

sources such as DBLP that they are an expert in a certain area, their opinion on data relating to that topic is probably more trustworthy than that of an average user. This type of background information could similarly be used in order to detect conflicts of interest, for example where a Web user with a particular political allegiance disagrees with information provided by their opponents.

These semantics-aware crowdsourcing methods would take advantage of human judgement and structured data in order to automatically aggregate and assess community feedback. In combination with ontology-enhanced content analysis and graph analysis of structured data sources, the quality of structured user-generated metadata could be rated. These methods could be used to help ensure that any algorithms which learn from user-supplied data on the Social Semantic Web make use of only trustworthy input data and therefore produce high quality results.

Bibliography

Adida, B., & Birbeck, M. (2008). RDFa Primer `http://www.w3.org/TR/xhtml-rdfa-primer/`. URL accessed July 2011. W3C Working Group Note 14 October 2008.

Alani, H., Dasmahapatra, S., O'Hara, K., & Shadbolt, N. (2003). Identifying communities of practice through ontology network analysis. *IEEE Intelligent Systems*, *18*(2), 18–25. IEEE Computer Society.

Aleman-Meza, B., Nagarajan, M., Ramakrishnan, C., Ding, L., Kolari, P., Sheth, A. P., et al. (2006). Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web* (pp. 407–416). ACM.

Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In *CHI '07: Proceedings of the SIGCHI Conference on Human factors in Computing Systems* (pp. 971–980). ACM.

Amitay, E., Har'El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging Web content. In *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 273–280). ACM.

Antonelli, F., & Sapino, M. (2005). A rule based approach to message board topics classification. In *MIS '05: Proceedings of the 11th International Workshop on Multimedia Information Systems* (pp. 33–48). Springer-Verlag.

Attardi, G., Gullì, A., & Sebastiani, F. (1999). Automatic Web page categorization by link and context analysis. In *THAI-99: Proceedings of the European Symposium on Telematics, Hypermedia and Artificial Intelligence* (pp. 105–119).

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007).

Dbpedia: A nucleus for a web of open data. In *ISWC '07: Proceedings of the 6th International Semantic Web Conference* (pp. 722–735). Springer-Verlag.

Backstrom, L., Kleinberg, J., Kumar, R., & Novak, J. (2008). Spatial variation in search engine queries. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web* (pp. 357–366). ACM.

Baeza-Yates, R. A., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley Longman Publishing Co., Inc.

Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007). Optimizing Web search using social annotations. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (pp. 501–510). ACM.

Barrat, A., Cattuto, C., Szomszor, M., Broeck, W. Van den, & Alani, H. (2010). Social dynamics in conferences: Analyses of data from the Live Social Semantics application. In *ISWC '10: Proceedings of the 9th International Semantic Web Conference* (pp. 17–33). Springer-Verlag.

Beckett, D., & Berners-Lee, T. (2011). Turtle – Terse RDF Triple Language `http://www.w3.org/TeamSubmission/turtle/`. URL accessed July 2011. W3C Team Submission 28 March 2011.

Berendt, B., & Hanser, C. (2007). Tags are not metadata, but "just more content" – to some people. In *ICWSM '07: Proceedings of the 1st International Conference on Weblogs and Social Media.*

Berners-Lee, T. (1989). Information management: A proposal. CERN.

Berners-Lee, T. (2006). Design issues: Linked Data `http://www.w3.org/DesignIssues/LinkedData.html`. URL accessed July 2011.

Berners-Lee, T., Fielding, R., & Masinter, L. (2005). RFC 3986: Uniform Resource Identifier (URI): Generic syntax. `http://www.ietf.org/rfc/rfc3986.txt`. URL accessed July 2011.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 34–43.

Bian, J., Liu, Y., Agichtein, E., & Zha, H. (2008). Finding the right facts in the crowd: Factoid question answering over social media. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web* (pp. 467–476). ACM.

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used

for search? In *CIKM '08: Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 193–202). ACM.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data – The story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. IGI Global.

Bodoff, D. (2006). Relevance for browsing, relevance for searching. *Journal of the American Society for Information Science and Technology*, *57*(1), 69–86. John Wiley & Sons, Inc.

Bonatti, P. A., Hogan, A., Polleres, A., & Sauro, L. (2011). Robust and scalable Linked Data reasoning incorporating provenance and trust annotations. *Journal of Web Semantics*, *In Press, Corrected Proof*. Elsevier.

boyd, d., & Ellison, N. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230. Blackwell Publishing Inc.

Breslin, J., Harth, A., Bojars, U., & Decker, S. (2005). Towards semantically-interlinked online communities. In *ESWC '05: Proceedings of the 2nd European Semantic Web Conference* (pp. 71–83). Springer-Verlag.

Brickley, D., & Guha, R. (2004). RDF Vocabulary Description Language 1.0: RDF Schema `http://www.w3.org/TR/rdf-schema/`. URL accessed July 2011. W3C Recommendation 10 February 2004.

Brickley, D., & Miller, L. (2010). FOAF Vocabulary Specification 0.98 `http://xmlns.com/foaf/spec/`. URL accessed July 2011.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. Elsevier.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., et al. (2000). Graph structure in the Web. *Computer Networks*, *33*(1–6), 309–320. Elsevier.

Budura, A., Michel, S., Cudr-Mauroux, P., & Aberer, K. (2009). Neighborhood-based tag prediction. In *ESWC 2009: Proceedings of the 6th European Semantic Web Conference* (pp. 608–622). Springer-Verlag.

Business Insider Inc. (2011, January 5). Facebook has more than 600 million users, Goldman tells clients. `http://www.businessinsider.com/facebook-has-more-than-600-million-users-goldman-tells`

-clients-2011-1. URL accessed July 2011.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *WWW '11: Proceedings of the 20th International Conference on World Wide Web* (pp. 675–684). ACM.

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 759–768). ACM.

Chirita, P.-A., Costache, S., Nejdl, W., & Handschuh, S. (2007). P-TAG: Large scale automatic generation of personalized annotation tags for the Web. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (pp. 845–854). ACM.

comScore, Inc. (2011, February 7). The 2010 U.S. Digital Year in Review.

Connolly, D. (2007). Gleaning Resource Descriptions from Dialects of Languages (GRDDL) http://www.w3.org/TR/grddl/. URL accessed July 2011. W3C Recommendation 11 September 2007.

Connolly, D., & Masinter, L. (2000). RFC 2854: The 'text/html' Media Type. http://www.ietf.org/rfc/rfc2854.txt. URL accessed July 2011.

Crandall, D., Backstrom, L., Huttenlocher, D., & Kleinberg, J. (2009). Mapping the world's photos. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web* (pp. 761–770). ACM.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 250–257). ACM.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 708–716). Association for Computational Linguistics.

DCMI Usage Board. (2010). DCMI Metadata Terms http://dublincore.org/documents/dcmi-terms/. URL accessed July 2011.

Delicious Blog. (2008, November 6). Delicious is 5! http://blog.delicious.com/blog/2008/11/delicious-is-5.html. URL accessed July 2011.

Dividino, R., Sizov, S., Staab, S., & Schueler, B. (2009). Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Journal of Web Semantics*, *7*(3), 204–219.

Eck, D., Lamere, P., Bertin-Mahieux, T., & Green, S. (2007). Automatic generation of social tags for music recommendation. In *NIPS 2007: Proceedings of the 21st Annual Conference on Neural Information Processing Systems* (pp. 385–392). MIT Press.

Eiron, N., & McCurley, K. S. (2003). Analysis of anchor text for Web search. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 459–460). ACM.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. (2010). A latent variable model for geographic lexical variation. In *EMNLP '10: Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1277–1287). Association for Computational Linguistics.

Engeström, J. (2005, April 13). Why some social network services work and others don't – Or: The case for object-centered sociality `http://www.zengestrom.com/blog/2005/04/13`. URL accessed July 2011.

Erétéo, G., Buffa, M., Gandon, F., & Corby, O. (2009). Analysis of a real online social network using Semantic Web frameworks. In *ISWC '09: Proceedings of the 8th International Semantic Web Conference* (pp. 180–195). Springer-Verlag.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., et al. (1999). RFC 2616: Hypertext Transfer Protocol – HTTP/1.1. `http://www.ietf.org/rfc/rfc2616.txt`. URL accessed July 2011.

Figueiredo, F., Belém, F., Pinto, H., Almeida, J., Gonçalves, M., Fernandes, D., et al. (2009). Evidence of quality of textual features on the Web 2.0. In *CIKM '09: Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 909–918). ACM.

Finin, T., Ding, L., Zhou, L., & Joshi, A. (2005). Social networking on the Semantic Web. *The Learning Organization*, *12*(5), 418–435. Emerald Group Publishing Limited.

Fink, C., Piatko, C., Mayfield, J., Chou, D., Finin, T., & Martineau, J. (2009). The geolocation of Web logs from textual clues. In *CSE '09: Proceedings*

of the 12th International Conference on Computational Science and Engineering (pp. 1088–1092). IEEE Computer Society.

Firan, C. S., Georgescu, M., Nejdl, W., & Paiu, R. (2010). Bringing order to your photos: Event-driven classification of Flickr images based on social knowledge. In *CIKM '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (pp. 189–198). ACM.

Flickr Blog. (2010, September 19). 5,000,000,000. `http://blog.flickr.net/en/2010/09/19/5000000000/`. URL accessed July 2011.

Fürnkranz, J. (1999). Exploiting structural information for text classification on the WWW. In *IDA '99: Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis* (pp. 487–498). Springer-Verlag.

Garcia Esparza, S., O'Mahony, M. P., & Smyth, B. (2010). Towards tagging and categorization for micro-blogs. In *AICS '10: Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science.*

García-Silva, A., Szomszor, M., Alani, H., & Corcho, O. (2009). Preliminary results in tag disambiguation using DBpedia. In *CKCaR 2009: Proceedings of the 1st International Workshop on Collective Knowledge Capturing and Representation.*

Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: Classifying tweets through a semantic transform based on Wikipedia. In *HCI 2011: Proceedings of the 14th International Conference on Human-Computer Interaction.* Springer-Verlag.

Ghani, R., Slattery, S., & Yang, Y. (2001). Hypertext categorization using hyperlink patterns and meta data. In *ICML '01: Proceedings of the 18th International Conference on Machine Learning* (pp. 178–185). Morgan Kaufmann Publishers Inc.

Gilbert, E., & Karahalios, K. (2009). Predicting tie strength with social media. In *CHI '09: Proceedings of the 27th International Conference on Human factors in Computing Systems* (pp. 211–220). ACM.

Glover, E. J., Tsioutsiouliklis, K., Lawrence, S., Pennock, D. M., & Flake, G. W. (2002). Using Web structure for classifying and describing Web pages. In *WWW '02: Proceedings of the 11th International Conference on World Wide Web* (pp. 562–569). ACM.

Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative

tagging systems. *Journal of Information Science*, *32*(2), 198–208. Sage Publications, Inc.

Golub, K., & Ardö, A. (2005). Importance of HTML structural elements and metadata in automated subject classification. In *ECDL 2005: Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries* (pp. 368–378). Springer-Verlag.

Grant, J., & Beckett, D. (2004). RDF Test Cases `http://www.w3.org/TR/rdf-testcases/`. URL accessed July 2011. W3C Recommendation 10 February 2004.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @spam: the underground on 140 characters or less. In *CCS '10: Proceedings of the 17th ACM Conference on Computer and Communications Security* (pp. 27–37). ACM.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. Academic Press Ltd.

Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., & Sheth, A. (2009). Context and domain knowledge enhanced entity spotting in informal text. In *ISWC '09: Proceedings of the 8th International Semantic Web Conference* (pp. 260–276). Springer-Verlag.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: An update. *SIGKDD Explorations*, *11*(1). ACM.

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (pp. 211–220). ACM.

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005). Social bookmarking tools (I). *D-Lib Magazine*, *11*(4), 1082–9873.

Harris, S., & Seaborne, A. (2011). SPARQL 1.1 Query Language `http://www.w3.org/TR/sparql11-query/`. URL accessed July 2011.

Harth, A., Kinsella, S., & Decker, S. (2009). Using naming authority to rank data and ontologies for Web search. In *ISWC '09: Proceedings of the 8th International Semantic Web Conference* (pp. 277–292). Springer-Verlag.

Hassanali, K.-n., & Hatzivassiloglou, V. (2010). Automatic detection of tags for political blogs. In *WSA '10: Proceedings of the NAACL HLT 2010*

Workshop on Computational Linguistics in a World of Social Media (pp. 21–22). Association for Computational Linguistics.

Hays, J., & Efros, A. (2008). IM2GPS: Estimating geographic information from a single image. In *CVPR 2008: IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE Computer Society.

Hecht, B., Hong, L., Suh, B., & Chi, E. (2011). Tweets from Justin Bieber's heart: The dynamics of the "location" field in user profiles. In *CHI '11: Proceedings of the ACM CHI Conference on Human Factors in Computing Systems.* ACM.

Hepp, M. (2008). GoodRelations: An ontology for describing products and services offers on the Web. In *EKAW2008: Proceedings of the 16th International Conference on Knowledge Engineering and Knowledge Management* (pp. 332–347). Springer-Verlag.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve Web search? In *WSDM '08: Proceedings of the International Conference on Web Search and Web Data Mining* (pp. 195–206). ACM.

Heymann, P., Ramage, D., & Garcia-Molina, H. (2008). Social tag prediction. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 531–538). ACM.

Hogan, A., Harth, A., Passant, A., Decker, S., & Polleres, A. (2010). Weaving the Pedantic Web. In *LDOW2010: Proceedings of the 3rd International Workshop on Linked Data on the Web.* CEUR Workshop Proceedings, vol. 628, CEUR-ws.org.

Huang, C., Fu, T., & Chen, H. (2010). Text-based video content classification for online video-sharing sites. *Journal of the American Society for Information Science and Technology*, *61*(5), 891–906. John Wiley & Sons, Inc.

Irani, D., Webb, S., Pu, C., & Li, K. (2010). Study of trend-stuffing on Twitter through text classification. In *CEAS '10: Proceedings of the 7th Collaboration, Electronic messaging, Anti-Abuse and Spam Conference.*

Jadhav, A., Purohit, H., Kapanipathi, P., Ananthram, P., Ranabahu, A., Nguyen, V., et al. (2010). Twitris 2.0: Semantically empowered system for understanding perceptions from social data. In *ISWC '10: Proceedings of the 9th International Semantic Web Conference.*

Jones, R., Zhang, W., Rey, B., Jhala, P., & Stipp, E. (2008). Geographic intention and modification in Web search. *International Journal of Geographical Information Science*, *22*(3), 229–246. Taylor & Francis, Inc.

Kinsella, S., Bojars, U., Harth, A., Breslin, J. G., & Decker, S. (2008). An interactive map of Semantic Web ontology usage. In *IV '08: Proceedings of the 12th International Conference on Information Visualisation* (pp. 179 –184). IEEE Computer Society.

Kinsella, S., Breslin, J. G., Passant, A., & Decker, S. (2008). Applications of Semantic Web methodologies and techniques to social networks and social websites. In C. Baroglio, P. Bonatti, J. Maluszynski, M. Marchiori, A. Polleres, & S. Schaffert (Eds.), *Reasoning Web* (Vol. 5224, pp. 171–199). Springer-Verlag.

Kinsella, S., Budura, A., Skobeltsyn, G., Michel, S., Breslin, J. G., & Aberer, K. (2008). From Web 1.0 to Web 2.0 and back –: How did your grandma use to tag? In *WIDM '08: Proceedings of the 10th ACM Workshop on Web Information and Data Management* (pp. 79–86). ACM.

Kinsella, S., Harth, A., Troussov, A., Sogrin, M., Judge, J., Hayes, C., et al. (2008). Navigating and annotating semantically-enabled networks of people and associated objects. In T. N. Friemel (Ed.), *Why Context Matters* (pp. 79–96). VS Verlag für Sozialwissenschaften.

Kinsella, S., Murdock, V., & O'Hare, N. (2011). "I'm eating a sandwich in Glasgow": Modeling locations with tweets. In *SMUC '11: Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents* (pp. 61–68). ACM.

Kinsella, S., Passant, A., & Breslin, J. G. (2010a). Ten years of hyperlinks in online conversations. In *WebSci10: Proceedings of the Web Science Conference 2010: Extending the Frontiers of Society On-Line.*

Kinsella, S., Passant, A., & Breslin, J. G. (2010b). Using hyperlinks to enrich message board content with Linked Data. In *I-SEMANTICS 2010: Proceedings of the 6th International Conference on Semantic Systems.* ACM.

Kinsella, S., Passant, A., & Breslin, J. G. (2011). Topic classification in social media using metadata from hyperlinked objects. In *ECIR '11: Proceedings of the 33rd European Conference on Information Retrieval* (pp. 201–206). Springer-Verlag.

Kinsella, S., Passant, A., Breslin, J. G., Decker, S., & Jaokar, A. (2009). The future of Social Web sites: Sharing data and trusted applications with semantics. In M. V. Zelkowitz (Ed.), *Social Networking and The Web* (Vol. 76, pp. 121 – 175). Elsevier.

Kinsella, S., Wang, M., Breslin, J. G., & Hayes, C. (2011). Improving categorisation in social media using hyperlinks to structured data sources. In *ESWC '11: Proceedings of the 8th Extended Semantic Web Conference* (pp. 390–404). Springer-Verlag.

Knorr Cetina, K. (1997). Sociality with objects: Social relations in postsocial knowledge societies. *Theory, Culture & Society*, *14*(4), 1–30. Sage Publications, Inc.

Lim, C. S., Lee, K. J., & Kim, G. C. (2005). Multiple sets of features for automatic genre classification of Web documents. *Information Processing and Management*, *41*(5), 1263–1276. Pergamon Press, Inc.

Lipczak, M., & Milios, E. (2010). The impact of resource title on tags in collaborative tagging systems. In *HT '10: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (pp. 179–188). ACM.

Liu, Y., Kumar, R., & Lim, K. (2008). Taggers versus linkers: Comparing tags and anchor text of Web pages. Report 2008-020. `http://escholarship.org/uc/item/8b40q59k`. URL accessed July 2011.

Losada, D. E., & Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Journal of Information Retrieval*, *11*, 109–138. Kluwer Academic Publishers.

Lu, Y.-T., Yu, S.-I., Chang, T.-C., & Hsu, J. Y.-j. (2009). A content-based method to enhance tag recommendation. In *IJCAI-09: Proceedings of the 21st International Joint Conference on Artifical Intelligence* (pp. 2064–2069). Morgan Kaufmann Publishers Inc.

Manola, F., & Miller, E. (2008). RDF Primer `http://www.w3.org/TR/rdf-primer/`. URL accessed July 2011. W3C Recommendation 10 February 2004.

Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *HT '06: Proceedings of the 17th Conference on Hypertext and Hypermedia* (pp. 31–40). ACM.

Matsuo, Y., & Yamamoto, H. (2009). Community gravity: Measuring bidirec-

tional effects by trust and rating on online social networks. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web* (pp. 751–760). ACM.

MaxMind, Inc. (2010). GeoIP city accuracy for selected countries, 17th May 2010. `http://www.maxmind.com/app/city_accuracy`. URL accessed July 2011.

McGuinness, D. L., & Harmelen, F. van. (2004). OWL Web Ontology Language Overview `http://www.w3.org/TR/owl-features/`. URL accessed July 2011. W3C Recommendation 10 February 2004.

Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 1318–1327). Association for Computational Linguistics.

Mika, P. (2005a). Flink: Semantic Web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, *3*(2-3), 211–223. Elsevier.

Mika, P. (2005b). Ontologies are us: A unified model of social networks and semantics. In *ISWC '05: Proceedings of the 4th International Semantic Web Conference* (pp. 522–536). Springer-Verlag.

Miles, A., & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System Reference `http://www.w3.org/TR/skos-reference/`. URL accessed July 2011. W3C Recommendation 18 August 2009.

Miniwatts Marketing Group. (2009, January 5). Internet Growth Statistics – Today's road to eCommerce and global trade. `http://www.internetworldstats.com/emarketing.htm`. URL accessed July 2011.

Mishne, G. (2006). AutoTag: A collaborative approach to automated tag assignment for weblog posts. In *WWW '06: Proceedings of the 15th International Conference on World Wide Web* (pp. 953–954). ACM.

Mishne, G., & Glance, N. (2006). Leave a reply: An analysis of weblog comments. In *WWE 2006: Proceedings of the 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.*

Moxley, E., Kleban, J., & Manjunath, B. S. (2008). Spirittagger: A geo-aware tag suggestion tool mined from Flickr. In *MIR '08: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval* (pp. 24–30). ACM.

Nakamura, S., Shimizu, M., & Tanaka, K. (2008). Can social annotation support users in evaluating the trustworthiness of video clips? In *WICOW '08: Proceedings of the 2nd ACM Workshop on Information Credibility on the Web* (pp. 59–62). ACM.

Noll, M. G., & Meinel, C. (2007). Authors vs. readers: a comparative study of document metadata and content in the WWW. In *DocEng '07: Proceedings of the 2007 ACM Symposium on Document Engineering* (pp. 177–186). ACM.

Noll, M. G., & Meinel, C. (2008). The metadata triumvirate: Social annotations, anchor texts and search queries. In *WI-IAT '08: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 640–647). IEEE Computer Society.

Ogilvie, P., & Callan, J. (2002). Experiments using the Lemur toolkit. In *NIST special publication 500-250: Proceedings of the 10th Text REtrieval Conference (TREC 2001)* (pp. 103–108).

Oliveira, B., Calado, P., & Pinto, H. (2008). Automatic tag suggestion based on resource contents. In *EKAW '08: Proceedings of the 16th International Conference on Knowledge Engineering: Practice and Patterns* (pp. 255–264). Springer-Verlag.

Othman, M., Yusuf, L., & Salim, J. (2010). Features discovery for Web classification using Support Vector Machine. In *ICICCI 2010: Proceedings of the 2010 International Conference on Intelligent Computing and Cognitive Informatics* (pp. 36–40). IEEE Computer Society.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Lioma, C. (2006). Terrier: a high performance and scalable information retrieval platform. In *OSIR 2006: Proceedings of the ACM SIGIR'06 Workshop on Open Source Information Retrieval.*

Pal, J. K., & Saha, A. (2010). Identifying themes in social media and detecting sentiments. In *ASONAM 2010: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining* (pp. 452–457). IEEE Computer Society.

Passant, A., & Laublet, P. (2008). Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *LDOW 2008: Proceedings of the Linked Data on the Web Workshop.*

Passant, A., & Raimond, Y. (2008). Combining Social Music and Semantic Web for music-related recommender systems. In *SDoW2008: Proceedings of the Workshop on Social Data on the Web.* CEUR Workshop Proceedings, vol. 405, CEUR-ws.org.

Ponte, J. M., & Croft, W. B. (1998). A language modeling approach to information retrieval. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 275–281). ACM.

ProgrammableWeb Blog. (2011, January 3). API growth doubles in 2010, social and mobile are trends. `http://blog.programmableweb.com/2011/01/03/api-growth-doubles-in-2010-social-and-mobile-are-trends/`. URL accessed July 2011.

Prud'hommeaux, E., & Seaborne, A. (2008). SPARQL Query Language for RDF `http://www.w3.org/TR/rdf-sparql-query/`. URL accessed July 2011.

Pudota, N., Dattolo, A., Baruzzo, A., Ferrara, F., & Tasso, C. (2010). Automatic keyphrase extraction and ontology mining for content-based tag recommendation. *International Journal of Intelligent Systems*, *25*(12), 1158–1186. John Wiley & Sons, Inc.

Qi, X., & Davison, B. (2008). Classifiers without borders: Incorporating fielded text from neighboring Web pages. In *SIGIR '08: Proceedings of the 31st International SIGIR Conference on Research and Development in Information Retrieval* (pp. 643–650). ACM.

Rae, A., Sigurbjörnsson, B., & Zwol, R. van. (2010). Improving tag recommendation using social networks. In *RIAO 2010: Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information.* CID.

Riboni, D. (2002). Feature selection for Web page classification. In *Proceedings of the Workshop on Web Content Mapping: A Challenge to ICT at EURASIA-ICT 2002.*

Rodrigues, E. M., Milic-Frayling, N., & Fortuna, B. (2008). Social tagging behaviour in community-driven question answering. In *WI-IAT '08: Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (pp. 112–119). IEEE Computer Society.

Rowe, M., & Mazumdar, S. (2010). Interlinking and Interpreting Social Data from Heterogeneous Sources. In *LUPAS 2010: Proceedings of the International Workshop on Linking of User Profiles and Applications in the Social Semantic Web.* CEUR Workshop Proceedings, vol. 595, CEUR-ws.org.

San Martín, M., & Gutierrez, C. (2009). Representing, querying and transforming social networks with RDF/SPARQL. In *ESWC 2009: Proceedings of the 6th European Semantic Web Conference* (pp. 293–307). Springer-Verlag.

Schonhofen, P. (2006). Identifying document topics using the Wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 456–462). IEEE Computer Society.

Schopman, B., Brickly, D., Aroyo, L., Aart, C. van, Buser, V., Siebes, R., et al. (2010). NoTube: making the Web part of personalised TV. In *WebSci10: Proceedings of the Web Science Conference 2010: Extending the Frontiers of Society On-Line.*

Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., et al. (2006). tagging, communities, vocabulary, evolution. In *CSCW '06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (pp. 181–190). ACM.

Serdyukov, P., Murdock, V., & Zwol, R. van. (2009). Placing Flickr photos on a map. In *SIGIR 2009: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 484–491). ACM.

Sharifi, B. (2010). *Automatic Microblog Classification and Summarization.* Unpublished master's thesis, University of Colorado, Colorado Springs, CO, USA.

Siersdorfer, S., Chelaru, S., Nejdl, W., & San Pedro, J. (2010). How useful are your comments?: Analyzing and predicting YouTube comments and comment ratings. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web* (pp. 891–900). ACM.

Sigurbjörnsson, B., & Zwol, R. van. (2008). Flickr tag recommendation based on collective knowledge. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web* (pp. 327–336). ACM.

Sood, S., Owsley, S., Hammond, K., & Birnbaum, L. (2007). TagAssist: Auto-

matic tag suggestion for blog posts. In *ICWSM 2007: Proceedings of the 1st International Conference on Weblogs and Social Media.*

Stankovic, M., Wagner, C., Jovanovic, J., & Laublet, P. (2010). Looking for experts? What can Linked Data do for you? In *LDOW2010: Proceedings of the 3rd International Workshop on Linked Data on the Web.* CEUR Workshop Proceedings, vol. 628, CEUR-ws.org.

Strube, M., & Ponzetto, S. P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *AAAI '06: Proceedings of the 21st National Conference on Artificial Intelligence* (pp. 1419–1424). AAAI Press.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *WWW 2007: Proceedings of the 16th International Conference on World Wide Web* (pp. 697–706). ACM.

Suchanek, F. M., Vojnovic, M., & Gunawardena, D. (2008). Social tags: Meaning and suggestions. In *CIKM '08: Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 223–232). ACM.

Sun, A., Lim, E.-P., & Ng, W. K. (2002). Web classification using Support Vector Machine. In *WIDM 2002: Proceedings of the Fourth ACM CIKM International Workshop on Web Information Data Management* (pp. 96–99). ACM.

Sun, A., Suryanto, M., & Liu, Y. (2007). Blog classification using tags: An empirical study. In *ICADL '07: Proceedings of the 10th International Conference on Asian Digital Libraries* (pp. 307–316). Springer-Verlag.

Tesconi, M., Ronzano, F., Marchetti, A., & Minutoli, S. (2008). Semantify del.icio.us: Automatically turn your tags into senses. In *SDoW2008: Proceedings of the Workshop on Social Data on the Web.* CEUR Workshop Proceedings, vol. 405, CEUR-ws.org.

Twitter Blog. (2011, June 30). 200 million Tweets per day `http://blog.twitter.com/2011/06/200-million-tweets-per-day.html`. URL accessed July 2011.

Wang, J., & Davison, B. D. (2008). Explorations in tag suggestion and query expansion. In *SSM '08: Proceedings of the 2008 ACM Workshop on Search in Social Media* (pp. 43–50). ACM.

Wang, S., & Groth, P. (2010). Measuring the dynamic bi-directional influence between content and social networks. In *ISWC '10: Proceedings of the 9th*

International Semantic Web Conference (pp. 814–829). Springer-Verlag.

Yahoo! Research. (2007). Web Collection UK-2007. `http://research.yahoo.com/` Crawled by the Laboratory of Web Algorithmics, University of Milan, `http://law.dsi.unimi.it/`. URLs accessed July 2011.

Yanbe, Y., Jatowt, A., Nakamura, S., & Tanaka, K. (2007). Can social bookmarking enhance search in the Web? In *JCDL '07: Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries* (pp. 107–116). ACM.

Yang, J., & Leskovec, J. (2011). Patterns of temporal variation in online media. In *WSDM '11: Proceedings of the 4th International Conference on Web Search and Data Mining* (pp. 177–186). ACM.

Yee, W. G., Yates, A., Liu, S., & Frieder, O. (2009). Are Web user comments useful for search? In *LSDS-IR'09: Proceedings of the 7th Workshop on Large-Scale Distributed Systems for Information Retrieval* (pp. 63–70).

Yi, X., Raghavan, H., & Leggetter, C. (2009). Discovering users' specific geo intention in Web search. In *WWW '09: Proceedings of the 18th International Conference on World Wide Web* (pp. 481–490). ACM.

Yin, Z., Li, R., Mei, Q., & Han, J. (2009). Exploring social tagging graph for Web object classification. In *KDD '09: Proceedings of the 15th SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 957–966). ACM.

YouTube Blog. (2010, November 10). Great Scott! Over 35 hours of video uploaded every minute to YouTube. `http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html`. URL accessed July 2011.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, *22*(2), 179–214.

Zhou, D., Bian, J., Zheng, S., Zha, H., & Giles, C. L. (2008). Exploring social annotations for information retrieval. In *WWW '08: Proceedings of the 17th International Conference on World Wide Web* (pp. 715–724). ACM.