Realising Context-Oriented Information Filtering [1]

David Edward Webster
Department of Computer Science
The University of Hull, Scarborough Campus
YO11 3AZ, United Kingdom
D.E.Webster@leeds.ac.uk

May 22, 2010

[1] Submitted in accordance with the requirements for the degree of Doctor of Philosophy. The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

Abstract

The notion of *information overload* is an increasing factor in modern information service environments where information is ‘pushed’ to the user. As increasing volumes of information are presented to computing users in the form of email, web sites, instant messaging and news feeds, there is a growing need to filter and prioritise the importance of this information. ‘Information management’ needs to be undertaken in a manner that not only prioritises what information we do need, but to also dispose of information that is sent, which is of no (or little) use to us.

The development of a model to aid information filtering in a context-aware way is developed as an objective for this thesis. A key concern in the conceptualisation of a single concept is understanding the context under which that concept exists (or can exist). An example of a concept is a concrete object, for instance a book. This contextual understanding should provide us with clear conceptual identification of a concept including implicit situational information and detail of surrounding concepts.

Existing solutions to filtering information suffer from their own unique flaws: text-based filtering suffers from problems of inaccuracy; ontology-based solutions suffer from scalability challenges; taxonomies suffer from problems with collaboration. A major objective of this thesis is to explore the use of an evolving community maintained knowledge-base (that of Wikipedia) in order to populate the context model from prioritise concepts that are semantically relevant to the user’s interest space. Wikipedia can be classified as a weak knowledge-base due to its simple TBox schema and implicit predicates, therefore, part of this objective is to validate the claim that a weak knowledge-base is fit for this purpose. The proposed and developed solution, therefore, provides the benefits of high recall filtering with low fallout and a dependancy on a scalable and collaborative knowledge-base.

A simple web feed aggregator has been built using the Java programming language that we call DAVe’s Rss Organisation System (DAVROS-2) as a testbed environment to demonstrate specific tests used within this investigation. The motivation behind the experiments is to demonstrate that the combination of the concept frame-

work instantiated through Wikipedia can provide a framework to aid in concept comparison, and therefore be used in news filtering scenario as an example of information overload. In order to evaluate the effectiveness of the method well understood measures of information retrieval are used. This thesis demonstrates that the utilisation of the developed contextual concept expansion framework (instantiated using Wikipedia) improved the quality of concept filtering over a baseline based on string matching. This has been demonstrated through the analysis of recall and fallout measures.

Acknowledgments

I would like to take the time here to thank the numerous individuals who have provided kind assistance towards the completion of my doctoral thesis:

Darren Mundy, University of Hull. I would like to express my gratitude to Darren for taking on the difficult task of accepting me as a PhD student, after the departure of my original supervisor. I would like to thank Darren for his assistance and dedication over the years in keeping me straight on the process of academic rigour.

Paul Warren, University of Hull. I would like to express my heartfelt gratitude to Paul who has dedicated much time over the duration of my PhD. He has helped me widen my knowledge of the research world and has helped me through the sometime soul-crushing times of being a PhD student. Finally, I would like to thank Paul for his infinite supply of nuclear strength coffee and connoisseurship.

Weihong Huang, University of Hull, now at University of Kingston. I would like to thank Weihong, who acted as my original supervisor and gave me the inspiration to step on to the path of becoming a researcher.

I would like to thank all the staff at the Scarborough Campus and the University of Leeds for their professionalism and advice over the years.

Finally, I would like to thank my dear Alexandra and my parents for their support and encouragement to keep me going through the many years of PhD research and my taking of a full-time research position at the University of Leeds.

Contents

List of Figures

Chapter 1

Introduction

> ’Space’ it says ’is big. Really big. You just won’t believe how vastly, hugely, mindbogglingly big it is. I mean, you may think it’s a long way down the road to the chemist, but that’s just peanuts to space. Listen ...’ and so on. - *Douglas Adams* (Adams 1979)

The same could be said regarding the volume of information currently, and ever expanding, on the World Wide Web *(WWW)*. To give a tangible feel to this expansion, according to Zakon (2005), the size of the WWW had grown from approximately 100,000 websites to approximately 70,000,000 websites from January 1996 to September 2005. This expansion can be said to be due to the increase in non-expert computer users publishing information online, partly down to the tools now available to aid web-based publication (Liu, Petrovic and Jacobsen 2005).

On the WWW today, search engines, such as Google and Yahoo give the WWW user a convenient method of searching for information through these websites’ services. Whilst convenient for the user, the recall accuracy, quality and trustworthiness of web sites returned by web search queries leaves something to be desired. Such issues of quality and trust in searching the WWW lay not entirely within the WWW search engines, but mainly in the open nature of publishing and partially-standardised information formats that are commonly used on the WWW along with the human orientated creation of web sites. However, this criticism should be countered with the argument that "*[the internet] ...it is now, and will forever remain, credibility and value-neutral*" (Berghel 1997). One of the main features of the WWW is that it is designed to be read by humans (alongside machines) and thus WWW information formats are designed with that principle in mind.

Associated with the notion of querying web sites is the area of information retrieval, wherein much research has been conducted in the areas of indexing and text mining, with popular algorithms to mine web sites for information and authority of such information, for instance Google's PageRank, taking much of a spotlight in information retrieval research circles.

Maes and Sheth (1993) introduces the distinction between *information retrieval* and *information filtering*. The distinction between these terms is that information retrieval involves extracting a small amount of relevant information from a knowledge-base or largely irrelevant information. Information filtering, on the other-hand, involves separating largely relevant information from a stream against less relevant information relatively speaking.

The Semantic Web, proposed in the latter part of the 1990's by Tim Berners-Lee, is both an extension and an evolution of the current WWW (Berners-Lee 1999). Unlike the current WWW that is designed to be read by humans through a web browser, the Semantic Web is designed to be read and understood by machines. This is achieved by giving the information on the Semantic Web a *well defined meaning* (Berners-Lee and Miller 2002). At the present date, the Semantic Web is in a developmental phase and is not set to completely replace the current WWW in the near future (if at all). As a result of this scenario, it is likely that in the near future we will see a parallelism of both WWW and Semantic Web approaches to information representation co-existing on the Internet. Due to this duality of information sources, content publishers and agent designers will have to consider this duality, with the possibility of bridging the two webs preferable.

The proposed research in this thesis focuses on the exploration into and application of the notion of context to the quality of information retrieval and filtering in a WWW/Semantic Web environment. We focus further on the notion of context in **Chapter 3**. This area is not only limited to the context of the Semantic Web, but also ideally to other intelligent service domains featuring knowledge-awareness and agent-based information retrieval and filtering. The specific research problems addressed by this thesis are presented in the following section of this chapter.

1.1 Research Problems Analysis

This section will involve a descriptive account of the research problems addressed by this thesis; initially proposing a problem statement to identify and describe the problems that are addressed within. Following on from the problem description, a concise solution proposal will be presented, identifying a number of important topic areas that will be further expanded upon in subsequent chapters. Conclusion of this introduction chapter will involve the proposal of several research hypothesis along with criteria for success for the thesis.

1.1.1 Problem Research Areas

Information Overload

The notion of *information overload* (Berghel 1997) is an increasing factor in modern information service environments, particularly in push-based information access. As increasing volumes of information are presented to computing users in the form of email, web sites, instant messaging and news feeds, there is a growing need to prioritise and filter this information. Lang (1995) states that *"as the number of participants in Usenet continues to multiply, so does the spectrum of topics covered. Users find it increasingly difficult to locate useful or interesting information as this diversity expands"*. Whilst this statement is directed at the Usenet information source, one can apply this statement to other Internet-based information sources.

Information management needs to be undertaken in a manner that not only prioritises what information we do need, but to also dispose of information that is sent, which is of no (or little) use to us. An example of whether an item is relevant or not would include email spam, whereby large quantities of unsolicited email appears in a user's inbox representing no interest to them. Unfortunately (perhaps due to the popularity of email) the quantity of email being sent and unsolicited email spam are forces contributing to the attrition to its usefulness as a messaging medium.

Knowledge-base Aided Information Processing

Agent-based knowledge processing systems are currently being developed for Semantic Web resources (Fensel, Wahlster and Lieberman 2002, Passin 2004). Knowledge-bases are used for representing knowledge in a controlled manner and drawing inferences across this knowledge. The Semantic Web aims to provide a solid architecture for creating a rich and strongly formalised knowledge-base, the qualification for this statement is the Semantic Web is by definition a strong knowledge-base due to its rich description of knowledge concepts and the ability to create extensional knowledge through rich predicates. A current limitation is that of providing a world knowledge-base for agents to make heuristic decisions on everyday tasks. Due to the formal ontological nature of the Semantic Web, the provision of a strongly formalised *global knowledge-base* is unlikely to appear in the near future.

It should be noted at this point that wiki-based encyclopaedias (for instance Wikipedia) offer such a grand (although less formally defined and organised) source of world knowledge. It would, therefore, be unwise to completely discard this source as a source of world knowledge. This is because Wikipedia aims to apply a level of homogeneity and consistency to WWW-based evolving community maintained knowledge-base, albeit in a much weaker form than with Semantic Web logics.

1.1.2 Research Problem Statement

As discussed within this chapter, the quality of information retrieval and contextual operations on that information (for instance, filtering or prioritising) stands to benefit greatly from knowledge-bases of world knowledge. Existing solutions to filtering information suffer from their own unique flaws: text-based filtering suffers from problems of inaccuracy; ontology-based solutions suffer from scalability challenges; taxonomies suffer from problems with collaboration.

The area of ontology-based concept similarity determination is a relatively new area, but is based on existing schema/tree matching work. The main body of work currently relating to this area is in comparing and matching fragments of ontology graphs, not strictly measuring their similarity. There is a body of existing work in the area of semantic relatedness of concepts based upon a topic taxonomy (or semantic lexicon). The problem of this approach is the availability of topic taxonomy that contains a comprehensive world knowledge and omits the inclusion of concepts based on products, events and influential/popular people. Tagging, on the other hand to the taxonomy, provides the agility that taxonomies lack (Hepp, Siorpaes and Bachlechner 2007), but retains many of the inaccuracy weaknesses that keyword based concept matching carries.

Knowledge-bases, however, need to be populated and processed to be utilised by agents for information filtering. The research area proposed within this thesis is that of developing and implementing a framework to supplement the contextualised information filtering capability of agents by taking advantage of contextual knowledge about concepts represented within a knowledge-base.

1.1.3 Proposed Solution

In order to supplement today and tomorrow's Internet, a context-aware framework for use in information filtering and agent service management is proposed. Considering that *context* itself is a contextual and conceptual notion, research will focus on pinning down a definition as to notion of context-awareness. At the current point in time we are presented with a scenario whereby the Semantic Web and web agent development exist in such a (relatively speaking) early stage of development, thus there is a lack of agent-based Semantic Web services which can be used to demonstrate a reliable context-based system. The thesis will, therefore, focus on investigating the general application of context-awareness to informational filtering by developing a framework to utilise the contextual distance between concepts from a knowledge-base. Whilst not explicitly tied to a particular technology, it is intended that this framework be applied to the Semantic Web in the future. The creation of a prototype service will be considered as part of this research as a testbed for such a context framework. This prototype is a web feed aggregator that filters news items based on the developed contextual framework.

This thesis proposes the wiki as a potential bridge between the text-oriented WWW and the data-based Semantic Web as a candidate for a world knowledge-base. A specific quality of Wikipedia wiki as a knowledge-base is that it can be defined as a ‘weak’ knowledge-base due to its weak schema formalism - this qualification is explored later within the thesis. As a result of this point a major objective of this thesis is to explore the use of an evolving community maintained weakly formalised knowledge-base (that of Wikipedia) in order to populate the context model from prioritise concepts that are contextually relevant to the user’s interest space.

From this investigation discussion focuses on how concept topography in a wiki knowledge-base can be used as a basis for contextual concept comparisons to aid in information filtering. As is introduced in **Section 2.4.1**, wikis provide a consistent method for web users to collaboratively build up a text-based knowledge-base with relatively low entry requirements and little knowledge as to the underlying technology. One of the contributions of this thesis’ work involves the usage of a wiki (in our case Wikipedia) as a knowledge-base to achieve a tangible ideal of a global knowledge-base, whereby all concepts are linked together in a reasonably consistent manner under an overarching control authority. At the same time, the structure and knowledge needs to possess the agility to be able to evolve over time.

In this thesis we focus on the explicit knowledge provided by the user. This knowledge is presented in the form of concepts that the user is interested in, be it explicitly or implicitly collected. We propose and demonstrate that through the use of a wiki knowledge-base one can implicitly collect related concepts based on a user’s given concept. Through the community maintained wiki knowledge-base we can combine explicit knowledge with expanded implicit knowledge and provide a personalised solution rather than a one-size-fits-all one improving over those that are described later in **Section 3.2.1**.

1.2 Research Hypotheses

A number of hypotheses are presented in this section to inform research and validation into the areas explored in this thesis:

Automated context modelling and processing will form the major part of the research within this thesis. To justify the need for rich automated context modelling through the use of a knowledge-base and processing thereof, the following hypotheses are proposed:

> 1: *In order to perform contextual comparisons between concepts in a non-domain independent manner, there is the need to utilise a knowledge-base of community maintained world knowledge. The second part of this hypothesis is that a weak knowledge-base is a satisfactory knowledge-base for the task.*
>
> 2: *Large scale wikis (for instance Wikipedia) provide a weak knowledge-base structure that can be interrogated for contextual information in a manner that cannot easily be achieved through: an unstructured WWW; traditional thesauri; or through the Semantic Web that is richly structured but in a fragmented manner.*
>
> 3: Gabrilovich and Markovitch (2006) propose the following unverified hypothesis in the context of a wiki structure, that this thesis aims to verify: *"given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from*

> *a concept is ill-advised, as this would collect a lot of weakly related material.*"

1.3 Research Aims and Objectives

The primary aim of this research is to investigate how a context framework supported by a knowledge-base can be developed to allow for human-like contextual comparison between concepts and expansion of concepts. In order for this to be realised, a suitable context framework for this purpose needs to be developed. One key concern of this context framework is in the conceptualisation of a single object, specifically understanding the context under which that object exists (or can exist). This contextual understanding should provide us with clear conceptual identification of an object including implicit situational information and detail of surrounding objects. From the initial framework, a method of contextual comparison between and expansion of concepts will be developed.

The objectives to achieve the above aims are as follows:

- Critique current literature in the field to understand past and current practice in the areas of the conceptualisation of a single object and understanding the context under which that object exists (or can exist). Furthermore this critique will focus on the topics of information filtering and concept comparison based on a knowledge-base.

- Establish the current limitations of the field in areas described in the previous point including identifying shortfalls of existing knowledge-bases for concept comparison.

- Develop a solution that provides a clear conceptual identification of an object including implicit situational information and detail of surrounding objects.

- Verify if a weakly formalised knowledge-base (by implementation, Wikipedia) is a suitable a knowledge-base to base the above concept expansion and comparison framework upon.

- Utilise the weakly formalised Wikipedia knowledge-base for the basis of a framework to expand out a given concept using contextual information provided by the knowledge-base. In order to achieve this, the derivation of contextual concept information is based on the topological structure of the wiki knowledge-base.
- Develop a web feed aggregator to utilise the developed concept expansion framework. This will allow the experimental testing of the developed concept expansion technique and web feed filters using it.
- Develop a testbed environment to demonstrate specific tests used within the experimental aspects of this investigation. Additionally determine measures to evaluate the effectiveness of the web feed filtering solution.

1.4 Research Methodology

As has been discussed previously within this chapter, this thesis' research will revolve around the development of a contextual concept comparison and expansion framework through the use of a knowledge-base. Due to the requirement of a knowledge-base containing a comprehensive representation of world concepts, for instance events, influential people and products to reflect the variability of concepts that web users search for, Wikipedia has been selected as an appropriate source for representation of these concepts due to its ease of user participation and evolvability. The knock-on effect of this is that the contextual comparison and evaluation framework will be based on a weakly formalised knowledge-base. This research will begin with familiarisation with current and proposed developments in the area of the information and service environments in respect to knowledge representation and processing. In this area, in addition to determining the state of the art in service environments and architectures, the work of Tim Berners-Lee and the W3C will be explored, if only at least to understand the mindset and direction that major players are planning for the Semantic Web and related information representation architectures for the Internet. Attention will also be importantly focused upon the current WWW, in order for a perspective to be given on current issues which this thesis's proposed framework will be able to alleviate.

Issues relating to the context-aware approach objective include a review of current research in order to aid the process of defining a context framework and to look at their current uses and how they deal with world knowledge. Methods of how this context framework can be represented and populated in relation to the WWW and Semantic Web will be investigated.

In relation to the life-cycle of the research, this research's methodology will follow a similar path to that proposed by Boehm's Spiral Model (Pressman 2001). Therefore research will initially begin with a broad, but shallow sweep of the subject areas to identify the research problem; this phase will be presented in **Chapter 2**. Following this stage a second research investigation will be used to identify the major research areas required to fully understand and resolve the stated research problem. Such research is detailed in subsequent chapters. Following this literature survey and review, research will progress deeper into each of these major topics with additional focus placed on major developments in these areas.

Following this investigation a context framework will be developed in order to compare and expand concepts and, in addition, methods will be created to utilise a knowledge-base in order to enrich the framework with a world knowledge-base. This contextual understanding should provide us with clear conceptual identification of an object including implicit situational information and detail of surrounding objects.

The next part of the methodology is to explore and verify the use of Wikipedia as an evolving community maintained knowledge-base as one that is suitable to base the above concept expansion and comparison framework upon.

Following the development of the context framework and verification of the Wikipedia knowledge-base, the next part of the methodology is to develop a testbed to test and evaluate the framework through the use of real-world data.

1.4.1 Evaluation of Framework

In order to empirically judge the success of this research, a number of evaluations need to be performed. An implementation of the proposed framework will be benchmarked using a news headline aggregation system providing information items relating to topic information. This benchmark will provide a quantitative method of evaluating the feasibility of such an approach to contextual concept comparison. Such comparisons and review will form a major part of the evaluation into the feasibility and success of the approach chosen for the proposed framework. In the event of success, it will be deemed suitable to apply this approach and framework to other related areas, such as e-commerce in order to evaluate the scalability of the approach.

The development of a contextual news filtering system will allow the confirming or disproving of our research hypotheses. This evaluation can be conducted through tabulation and graphing the number of relevant search results as a result of the query in addition to the investigation as to the signal-to-noise ratio. Issues relating to accuracy of recall, such as false positives and false negatives, need to be considered in relation to this approach.

In the case that the hypotheses are proven false, an interesting situation will be presented. It may be that, for example, a context-based concept comparison and expansion method for a filtering system is found to provide no additional benefits over a non-contextual method. It is, therefore, important to consider and demonstrate why such an approach will fail. Current research in this thesis will proceed with an optimist view.

1.5 Research Contributions

For this thesis, a number of research contributions can be extracted as the main drivers.

- The first contribution is the context comparison framework and the method to perform the comparisons between modelled concepts in order to calculate semantic distance between them. As an integrated part of this framework, context-chains are introduced, not only to aid in calculating the contextual distance between objects, but to aid in the provenance of this comparison.

- The second contribution utilises the weak Wikipedia knowledge-base as the basis of a method to expand out a given concept using contextual information provided by the knowledge-base. In order to achieve this, the derivation of contextual concept information is based on the structure of the Wikipedia knowledge-base with a simple TBox schema.

As a direct successor to the proposed contextual modelling methodology, a case study is proposed in order to demonstrate the framework to assign a contextual distance value to a pair of scenario concepts and to allow the matching of related concepts through the expansion of the first object. The originality of this work is to demonstrate the use of contextual modelling in information filtering and to aim towards a mechanism of concept management that is more representative of a human than a first order logic approach, for instance. This mechanism will be realised through the introduction of context-chains provided by the context framework proposed in this work. These context-chains will provide a contextual linkage between given concepts with a number of paths, in order for semantic distance to be measured in the correct context between respective concepts.

1.6 Research Publications in Relation to this Thesis

Output from this thesis has been published in a number of academic publications from 2004 to 2008. The list of current publications is presented as follows:

- D.Webster, J.Xu, D.Mundy and P.Warren. A Practical Model For Conceptual Comparison Using A Wiki. Fourth International M3C Workshop on Interaction with Content around Curriculum Lifecycle. The 9th IEEE International Conference on Advanced Learning Technologies ICALT2009. July 14-18, 2009. Riga, Latvia.
- D. Webster, W. Huang, D.Mundy and P.Warren Context-Orientated News Filtering for Web 2.0 and Beyond , 15th International World Wide Web Conference. May 2006.
- W.Huang, D.Webster, D.Wood and T.Ishaya, "An Intelligent Semantic e-Learning Framework Using Context-Aware Semantic Web Technologies" in 'British Journal of Educational Technology', Special Issue on Semantic Web for e-Learning Blackwell Synergy, Feb, (2006)
- W. Huang, E. Eze and D Webster, "Towards integrating semantics of multimedia resources and processes in e-Learning" in 'ACM/Springer Journal on Multimedia Systems', Special Issue on Educational Multimedia Systems Springer, Feb, (2006)
- W. Huang, D. Webster, and E. Eze, LILAC: Learner-centric Intelligent e-Learning with Awareness of Contexts, (Submitted) Book Chapter in E-Service Intelligence - Methodologies, Technologies and Applications. Editors J. Lu, D. Ruan, G. Zhang, to be published by Springer in 2006.
- W. Huang and D. Webster, Enabling Context-aware Agents to Understand Semantic Resources on the WWW and the Semantic Web. 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI04), Beijing, China, September 2004

- W. Huang and D. Webster, Intelligent RSS News Aggregation Based on Semantic Contexts. ACM SIGIR 2004 Workshop on Information Retrieval in Context

1.7 Structure of Thesis

As previously discussed regarding the life-cycle of the research, the research methodology will follow a similar path to that proposed by Boehm's Spiral Model (Pressman 2001) for the first two-thirds of this thesis. We present the outputs from this research process in the following sequence of chapters.

- Chapter 2) **Background Literature Survey** - Chapter 2 begins with a survey of the area of information consumption within service based environments. In this section the notion of information overload and how this is a problem that we are increasingly encountering within information environments is presented.

  Limitations of information filtering in the WWW and the Semantic Web are presented with a number of cited issues that are present in the implementation of the Semantic Web vision including an overview of the differences between statistical and logical processing of the information present on them. Finally the Web 2.0 and associated developments are surveyed as a bottom-up social alternative to the formal Semantic Web. We discuss a number of benefits that Web 2.0 presents over the Semantic Web and the limitations of Web 2.0

  Concept representation in information systems is then explored. This exploration includes looking at existing approaches to determining similarity between concepts. In this chapter, the Ogden Meaning Triangle is surveyed to enable us to better understand how concepts relate to words and the world. Focus then proceeds to survey current work in knowledge-bases and present a number of relevant internet-based options. For each of these knowledge-bases the area of concept similarity and present current research pertaining to each type of knowledge-base is explored.

- Chapter 3) **Context and Information Filtering** - In Chapter 3 a survey of the notion of context is introduced for information retrieval. This begins with an understanding of the notion of context. Existing definitions of context are explored, which permits an understanding of not only dictionary definitions of context, but also semantically what context means regarding related objects. A definition of context is presented for this thesis before giving an overview as to how context helps us to understand the relationship between an object and its world. Focus is then passed on to what context means for information search including the citing of a number of properties of contextual information search including personalised filtering and historical context.

- Chapter 4) **Initial Context Modelling and Practical Experience** - This chapter presents a discussion on some of the early context modelling work conducted for this thesis. The FPO model is presented and describes how this helps us to represent context in a lightweight manner through relaxing the predicate semantics. This model is extended and then the limitations of this extension is presented. The latter half of Chapter 5 focuses on our experience with using a taxonomy to determine contextual similarity between concepts and introduces our method of contextualisation of similarity calculation for concepts. Throughout the use of prototyping, we uncovered a number of problems inherent with a rigid taxonomy and provide motivation that would ultimately lead us to favour the wiki structure over the taxonomy structure.

- Chapter 5) **Wiki-based Knowledge Processing** - Wiki-based Knowledge Processing is presented in Chapter 6. This chapter begins by exploring existing methods of determining semantic similarity within Wikipedia. These methods are reviewed and a new novel method is proposed to allow the semantic similarity between and expansion of concepts to be achieved.

- Chapter 6) **Experimentation and Evaluation: Context-Orientated News Feed Filtering Case Study** - In Chapter 7 the experimental method is presented for the proposed concept expansion method. This includes the introduction of a news feed filtering case study utilising source news items from the OSNews technology news website. From this case study it is possible to design a number of experiments along with an associated baseline experiment to allow exploitation of the contextual concept expansion method through the development of news item filters.

- Chapter 7) **Analysis and Evaluation: Context-Orientated News Feed Filtering Case Study** - Experimental analysis is presented in Chapter 8 where the effectiveness of the proposed experimental filter methods across our chosen OSNews topics is evaluated. For each filter the average of the recall and fallout measures across all topics is measured and analysed.

- Chapter 8) **Conclusion and Reflection** - This chapter concludes this thesis and outlines future work.

Chapter 2

Information Filtering

This chapter consists of an initial review of the foundation for this thesis' research and focuses on a number of key topic areas associated with information overload. The areas that we focus on are; *service orientation*, the *Semantic Web*, *information retrieval*, and *Web 2.0*.

2.1 Overview of Information Consumption within Service-Oriented Computing

In this section, the area of service-orientated computing is briefly discussed, primarily from an information consumer's perspective in order to highlight a number of research issues and to give a historical context to information systems and the Semantic Web.

In the subsequent sections we start with an overview of the problem of information overload in modern information service environments. Following this overview we present the progression of the development of service oriented computing over the last quarter of a century with particular emphasis on the volume of information provided by the evolution of the mediums. In addition with this evolution comes the reliance of modern society in the medium.

2.1.1 Information Overload

The notion of *information overload* is an increasing factor in modern information service environments (Berghel 1997). As an increasing volume of information is presented to computing users in the form of email, web sites, instant messaging and news feeds, there is a growing need to prioritise and filter this information. As Liu et al. (2005) states *“in addition, a user does not have the resources to monitor large number of feeds and hence the user can easily miss information of interest".* This volume of information from various sources is a problem in modern knowledge working environments (Allen 2003) and can seriously affect productivity. An article published in New Scientist (Knight 2005) to mention here if only for anecdotal reasons, claims that, *“the relentless influx of emails, cellphone calls and instant messages received by modern workers can reduce their IQ by more than smoking marijuana, suggests UK research.".*

In a complimentary argument, 'productivity guru' David Allen proposes in the book "Getting Things Done" (Allen 2003) that, *“you don't manage information overload*

– otherwise you'd walk into a library and die, or the first time you connected to the Web, or even opened a phone book, you'd blow up". The main motivation behind the Getting Things Done (GTD) method is that the user should not manage *'stuff'*, with *'stuff* referring to unorganised information, but rather manage actions based upon the collected, organised then reviewed information. The key here is that information in itself cannot be managed, but actions are managed in relation to a goal, with a project being a good example. To take this further, we can say that we are managing information in relation to a context.

Such management needs to be undertaken in a manner that not only prioritises what information we do need, but also disposes of information that is sent which is of no use to us. An example of whether an item is relevant or not would include email spam, where large quantities of unsolicited email appears in a user's inbox representing no interest to them. Unfortunately, the popularity of email, the quantity of email being sent and unsolicited email spam are forces contributing to the attrition to its usefulness as a messaging medium.

Email is a common example of a domain requiring much overhead in filtering on a user's part. During 2007 a presentation was given at the Google Campus by Merlin Mann titled Inbox Zero (Mann 2007). During the talk, the question of the number of company emails that an employee received was asked to the audience. The personal responses ranged from 100 to 500 per day. From this, 10 percent included actionable items. It was also noted that people often use their email inbox to manage particular aspects of their lives.

Web feeds present us with an example of an attempt to reduce the information retrieval required from web users by allowing web users to subscribe to lists of items (for example news items) from websites of their choice. In this manner there is a loose coupling between the website web feed service and the user aggregating the items.

Email and web feeds can be abstracted to a list of items that are subscribed to (or sent) by the user. In both cases, prioritisation is applied by the user to both not only whether an item is relevant, but also how relevant at a given time. Due to this similarity, it is common for modern email software to double as web feed aggregators, such as Mozilla Thunderbird (moz 2007).

One distinction to note is that of whether information is pulled by the user or pushed to the user. In both cases, the user will be presented with potentially relevant information that may need to be prioritised. This presents us with the question, how relevant is each item that is within the return set. This question brings us into the topic of information filtering, which is expanded upon later in **Section 2.2**.

2.1.2 UNIX Systems

Timesharing systems, such as UNIX, allowed the promotion of network services, such as remote login through a local network with the use of remote login protocols, such as TELNET gaining widespread use. UNIX also included utilities that could be combined together using pipes to create workflows (Robbins and Beebe 2005). In many ways this pattern can be seen as a precursor to modern web service based workflows.

In the early days of the Internet DARPANET network was developed. One of the first uses of such a network was to allow remote login into a system. Whilst this epoch was limited to very few academic partners, such a pattern would ultimately gain popularity, with public adoption in the late 1970's through BBS services and later the uptake of internet connections from the home.

2.1.3 Bulletin Board Systems: A precursor to the WWW

In a search for a precursor to the WWW, one can look to Bulletin Board Systems (*BBS*). The first BBS was developed by Ward Christensen in January 1978. For anecdotal value, Christensen built the implementation with Randy Suess whilst being snowed-in in the Great Chicago Snowstorm of 1979 as a result of a telephone conversation (Christensen 1992). The system implementation, called CBBS, was in use 30 days after its conception. In many ways this simple system can be seen as the genesis of service-oriented computing in the home. Now mainly relegated to history, a typical BBS would offer text based services through a standard telephone dial-up modem. The BBS system did not require a specific hardware or software platform to achieve inter-connection, which in an era of heterogeneous home hardware platforms and storage mediums provided a way forward for levels of interoperability and integration that we take for granted today.

The use and culture of BBS services, as documented by Sadofsky (2005), highlight a number of trends and problems that existed in BBSs that are still part of the current internet. Such problems include widespread intellectual property infringement and piracy, particularly with software. Other areas that were present at the time that BBS were popular include offensive and illegal content, which still exist on the internet today.

2.1.4 Current World Wide Web

The original WWW was proposed by Tim Berners-Lee in the early 1990's (Berners-Lee 1999). The W3C (W3C n.d.b) describes the WWW as:

> "*the universe of network-accessible information, the embodiment of human knowledge.*"

The WWW is an information space composed of three protocols, *URL*, *HTTP* and *HTML* (w3c n.d.a). URL (Uniform Resource Locator) provides a mechanism for identifying a resource on the network along with associated protocol information. The URL (*http://www.yahoo.com/index.html*), for example identifies the protocol as *http*, the network server as *www.yahoo.com* and the document being requested from that server being *index.html*. HTTP (Hypertext Transfer Protocol) is the primary protocol used for sending documents over the WWW network (Passin 2004). HTML (HyperText Markup Language) is a simplified version of Standard generalised Markup Language (SGML) document type for the WWW (Fensel et al. 2002). An HTML document is composed of tags, used for not only information presentation, but for URL-based links between HTML documents, otherwise known as hyperlinks.

One of the problems with HTML is that it was designed to be network-accessible, but not necessarily 'machine-understandable', meaning that HTML documents published on the WWW were only intended for human readership (Fensel et al. 2002). As WWW use evolved it was seen by some developers that machine processing of HTML pages would be desirable. To enable this, a technique called *web scraping* was developed to analyse patterns of key words in a HTML document to look for specific items such as book prices (Berners-Lee 1999). A problem with web scraping was that it was based on an ad-hoc analysis of web page and prone to failing if the web site's layout changed.

In the late 1990's the W3C, set about creating a new markup language to aid in machine parsing and understanding. This language is XML (eXtensible Markup Language. Concurrently to the development of XML a stack of languages is being developed by the W3C, with XML forming one of the base layers (Passin 2004).

2.2 Information Retrieval on the Web

To return from an informational overview of the current state of information representation on the WWW and Semantic web for this thesis, this section will cover the area of information retrieval, discussed in part in previous sections, in depth and relate this area back to the problem on information overload. Information retrieval is an area deeply intertwined with the WWW and drives many of the uses of the web today. To explain information retrieval in human terms, when a WWW user wants to find information on the web they can utilise a service, such as a WWW search engine, to help retrieve documents from the web. In order to utilise a search engine, a user can specify search criteria as a basic text string and then the search engine will attempt to match the search terms against pages that it has indexed using a term ranking algorithm. The search engine will typically present ranked pages to the user in the form of a dynamically generated WWW page.

This section will explore the fundamentals of information retrieval, followed by the notion of information overload and information filtering. The section concludes with a look into statistical processing versus logical processing of information associated with the WWW and the Semantic Web.

2.2.1 Overview of Information Retrieval Measures

Fundamental to the information retrieval domain are a number of measurable quantities reported by (Van Rijsbergen 1979) originally given by Cyril Cleverdon in 1966. Among these quantities are the quantities of effort, precision and recall.

Effort is described as:

> "*the effort involved on the part of the user in obtaining answers to his search requests*"

Precision is described as

> "*the precision of the system, that is, the proportion of retrieved material that is actually relevant*"

Precision can also be described as the *“fraction of the retrieved documents that are relevant".* (Egghe 2008)

$$Precision = \frac{|RetrievedDocuments \cap RelevantDocuments|}{|RetrievedDocuments|} \qquad (Egghe\ 2008) \quad (2.1)$$

Recall is described as:

> "*the recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request*"

Recall can also be described as the *“fraction of relevant documents that are retrieved".* (Egghe 2008)

$$Recall = \frac{|RetrievedDocuments \cap RelevantDocuments|}{|RelevantDocuments|} \qquad (Egghe\ 2008) \quad (2.2)$$

In addition, the fallout quantity, described in the Swets model (Van Rijsbergen 1979), is described as:

> "*an estimate of the conditional probability that an item will be retrieved given that it is non-relevant*"

Fallout can also be described as the *“fraction of non-relevant documents that are retrieved".* (Egghe 2008)

$$Fallout = \frac{|RetrievedDocuments \cap \neg RelevantDocuments|}{|\neg RelevantDocuments|} \qquad (Egghe\ 2008) \quad (2.3)$$

Connected to the area of information retrieval is that of similar objects determination and matching of such objects. This will be further explored in **Chapter 2.5**.

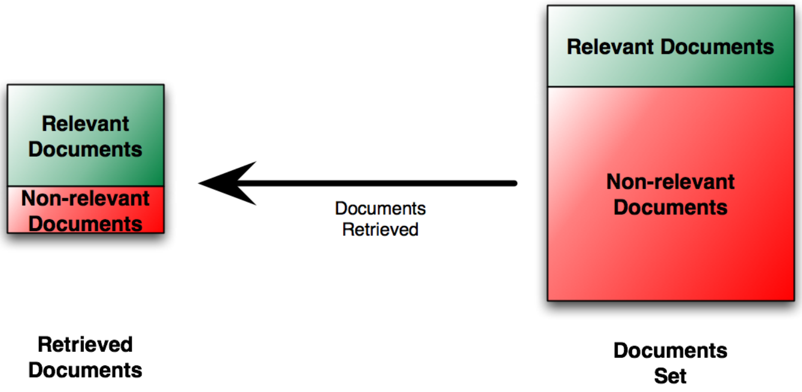
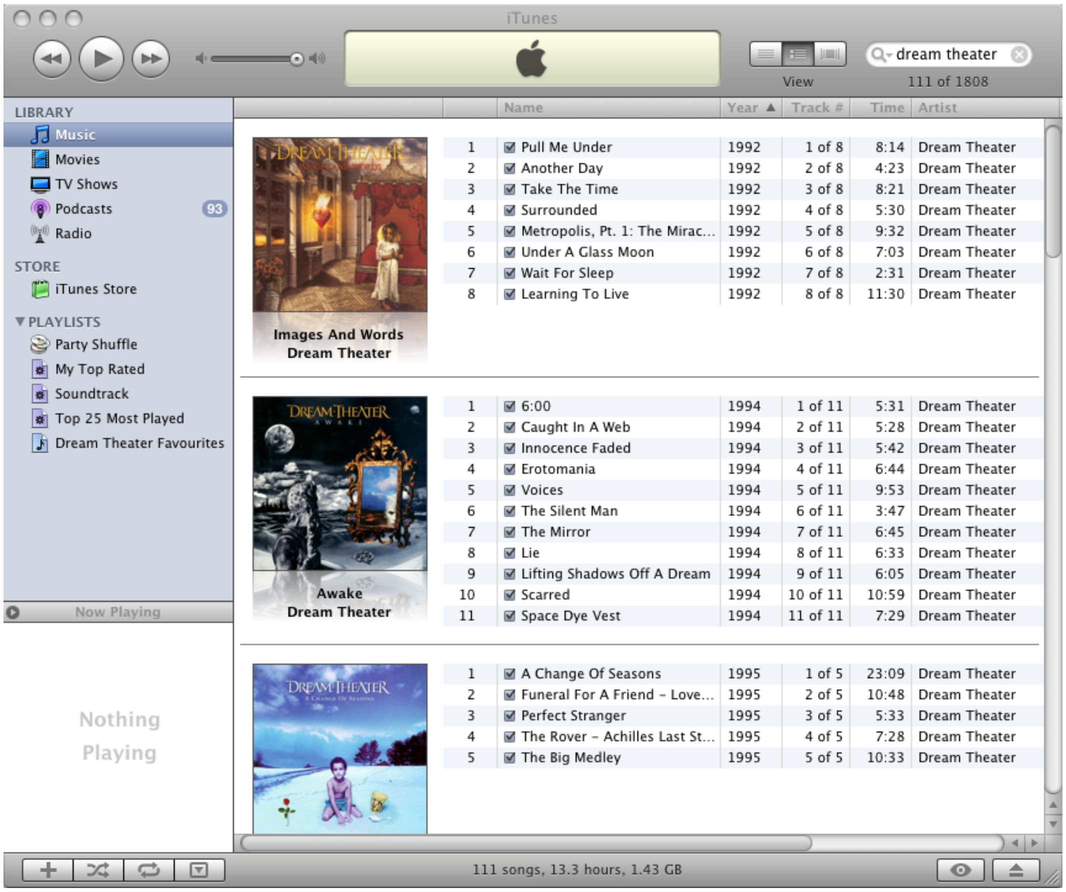


Figure 2.1: Illustration of Information Retrieval

2.2.2 Information Filtering

Maes and Sheth (1993) introduces the distinction between *information retrieval* and *information filtering*. The distinction between these terms is that information retrieval involves extracting a small amount of relevant information from a repository of largely irrelevant information. Information filtering, on the other-hand, involves separating largely relevant information from a stream against less relevant information. One driver for this differentiation concerns soft issues relating to the filter owner's current context for requiring information.

Two current approaches to information filtering include searching with keywords to match metadata attached to information items; a second approach is to browse through the metadata (or index terms) extracted from information items. An example of this would be to select a band name in music software from a list of extracted band names, filtering out all other music items that do not have the selected band name as a metadata attribute. **Figure 2.2** shows an example of filtering through a music collection through simple keyword filtering with a filter box in the top right of the interface.



	Name	Year ▲	Track #	Time	Artist
1	☑ Pull Me Under	1992	1 of 8	8:14	Dream Theater
2	☑ Another Day	1992	2 of 8	4:23	Dream Theater
3	☑ Take The Time	1992	3 of 8	8:21	Dream Theater
4	☑ Surrounded	1992	4 of 8	5:30	Dream Theater
5	☑ Metropolis, Pt. 1: The Mirac...	1992	5 of 8	9:32	Dream Theater
6	☑ Under A Glass Moon	1992	6 of 8	7:03	Dream Theater
7	☑ Wait For Sleep	1992	7 of 8	2:31	Dream Theater
8	☑ Learning To Live	1992	8 of 8	11:30	Dream Theater
1	☑ 6:00	1994	1 of 11	5:31	Dream Theater
2	☑ Caught In A Web	1994	2 of 11	5:28	Dream Theater
3	☑ Innocence Faded	1994	3 of 11	5:42	Dream Theater
4	☑ Erotomania	1994	4 of 11	6:44	Dream Theater
5	☑ Voices	1994	5 of 11	9:53	Dream Theater
6	☑ The Silent Man	1994	6 of 11	3:47	Dream Theater
7	☑ The Mirror	1994	7 of 11	6:45	Dream Theater
8	☑ Lie	1994	8 of 11	6:33	Dream Theater
9	☑ Lifting Shadows Off A Dream	1994	9 of 11	6:05	Dream Theater
10	☑ Scarred	1994	10 of 11	10:59	Dream Theater
11	☑ Space Dye Vest	1994	11 of 11	7:29	Dream Theater
1	☑ A Change Of Seasons	1995	1 of 5	23:09	Dream Theater
2	☑ Funeral For A Friend – Love...	1995	2 of 5	10:48	Dream Theater
3	☑ Perfect Stranger	1995	3 of 5	5:33	Dream Theater
4	☑ The Rover – Achilles Last St...	1995	4 of 5	7:28	Dream Theater
5	☑ The Big Medley	1995	5 of 5	10:33	Dream Theater

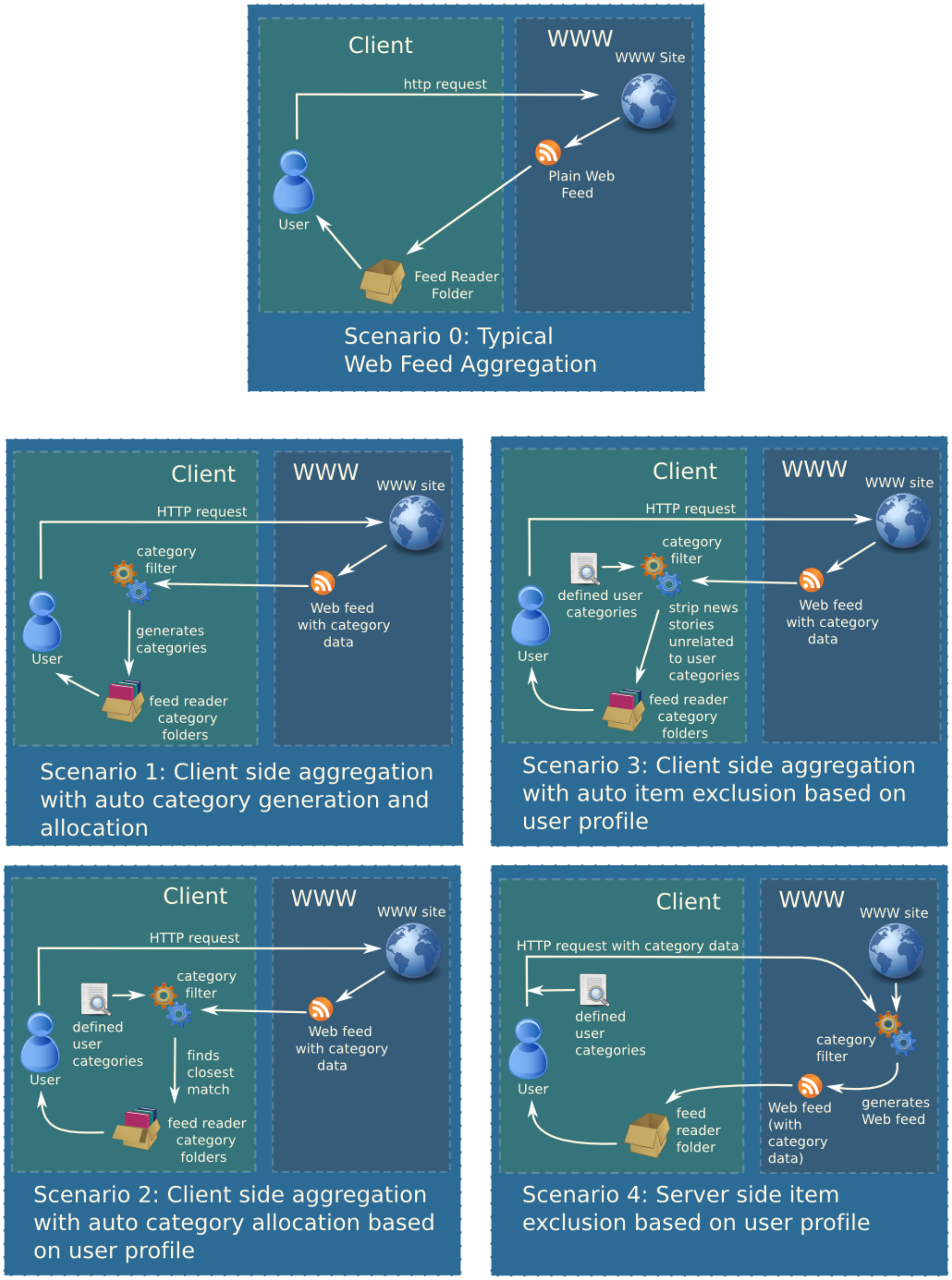
Figure 2.2: Filtering songs with iTunes

2.2.3 Publish/Subscribe Pattern

The Publish/Subscribe pattern has been used by related work, for instance as with Liu et al. (2005), to provide topic based brokering to reduce information overload overhead on the subscriber to web feed content. This paradigm allows loose coupling between publisher and subscriber. A further example of this pattern has been applied to web feeds in a poster paper at WWW06 as part of this thesis' work (Webster, Huang, Mundy and Warren 2006) as summarised in **Figure 2.3**. In this approach, we considered the processing of context information in web feed aggregation from two perspectives; the client side and the service provision side. As with traditional web feed aggregators where filtering happens on the client side, this filtering could easily be transferred to a WWW server-based solution (see Scenario 4). Just as a user can create a set of category/smart folders in his or her aggregation client, these preferences could be saved as a user profile and uploaded to a server. The server-based approach has the advantage of being usable with a dumb client, which can be used to subscribe to a single aggregated and filtered RSS feed. Through this pattern, users can subscribe to an archive of web feed items through keywords present in the item metadata.

Recent examples of a large scale web feed based publish/subscribe patterns include Rojo, Del.icio.us and eBay. In the case of Del.icio.us, a user can subscribe to a web feed of all new bookmarks that have been filed with the keyword specified in the subscription. With eBay, a user can subscribe to a web feed containing recent items listed under an item category.

The advantage of the approach taken by Liu et al. (2005) is that the architecture allows simple RDF graphs to be matched quickly, rather than just topic keywords. With a simple ontology and directed acyclic RDF graphs, a simple RDQL (RDF Query Language) query can be used to match subgraphs consisting of lexicographically related terms and structures from the subscription, however, problems arise regarding complexity with graphs richer than simple RDF graphs, particularly when these are cyclic.



Scenario 0: Typical Web Feed Aggregation

Scenario 1: Client side aggregation with auto category generation and allocation

Client
WWW
WWW site
HTTP request
category filter
defined user categories
Web feed with category data
strip news stories unrelated to user categories
User
feed reader category folders

Scenario 3: Client side aggregation with auto item exclusion based on user profile

Scenario 2: Client side aggregation with auto category allocation based on user profile

Scenario 4: Server side item exclusion based on user profile

Figure 2.3: Patterns of Web Feed Subscription. Presented in poster form at the International World Wide Web Conference in 2006.

2.3 The Semantic Web

At the first WWW conference in 1994, Tim Berners-Lee (the inventor of the WWW) discussed a notion that would lead to the development of what would become the Semantic Web; a logical connection of terms that would aid in interoperability between systems through semantic theory (Shadbolt 2006). As Greaves and Mika (2008) describe, the Semantic Web is a network connecting data through semantic relations.

In his book ’Weaving the Web’ (Berners-Lee 1999), Tim Berners-Lee’s proposed the second part of his Web vision as what would eventually become the Semantic Web. Whilst the WWW satisfied it’s demands in early years of infancy, recent information demands are revealing problems in the WWW that we have become accustomed to. A major problem with the current WWW is that it is ill suited to automated processing by machines due to unstructured information present within it (Flake, Pennock and Fain 2003).

The Semantic Web is an extension of the current World Wide Web *(WWW)*. Unlike the current HTML-derived WWW, which is designed to be read by humans through web browsers, the Semantic Web, is designed to be read and understood by machines. This is achieved by giving the information on the Semantic Web, a *’well defined meaning’* (Berners-Lee and Miller 2002, Greaves and Mika 2008) using explicit metadata so that machines may manipulate the semantic information automatically. Ideally the Semantic Web is envisioned as enabling seamless interactions among agents and enabling automatic discovery and integration (Bussler 2008).

The Semantic Web heavily relies upon the concept of *ontology*, which is used to formally describe a *domain of discourse* (Antoniou and Harmelen 2008). Semantic Web ontologies, therefore, allow for inferences to be performed on relations defined using the ontology and subsumption hierarchies of classes (Paolucci, Kawamura, Payne and Sycara 2002, Greaves and Mika 2008).

It is important to note here that the Semantic Web was proposed before *Web 2.0* for managing socially generated content (Greaves and Mika 2008). Both of these approaches will be discussed in the chapter.

2.3.1 Ontology

Fundamental to the realisation of the Semantic Web is the concept of ontology. Concisely, in the context of the Semantic Web, an ontology is used to create a *shared understanding* between two systems (Davies, Fensel and van Harmelen 2003) that they can reason about (Passin 2004). This section will now look at a number of definitions as to what defines an ontology including:

> 1: *"an ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary."* (Neches, Fikes, Finin, Gruber, Patil, Senator and Swartout 1991) as cited by (Corcho, Fernandez-Lopez and Gomez-Perez 2003)
>
> 2: *"an ontology is a hierarchically structured set of terms for describing a domain that can be used as a skeletal foundation for a knowledge-base."* (Swartout, Ramesh, Knight and Russ 1997) as cited by (Corcho et al. 2003)

These definitions, although not exhaustive, demonstrate that an ontology consists of a lexicon of terms with inference rules between them. A flat tree-based taxonomy of concepts (for instance that provided by the Mozilla Directory Project (DMOZ)), whilst useful for identifying objects, cannot provide a comprehensive representation of the intricacies of relations between these objects without further inference rules as would be provided by an ontology.

Ontologies in the Semantic Web are used to define a domain of interest through the definition of class hierarchies and properties between them, with axioms defining intensional knowledge (Deng, Haarslev and Shiri 2007).

2.3.2 Semantic Web Ontology and Logic

Logic is used to express a subset of knowledge that can be described with natural language (Passin 2004). Logic allows for the expression of truth conditions and rules of inference (Brachman and Levesque 2004) and these are expressed in an unambiguous manner (Antoniou and Harmelen 2008).

Ontologies and the Semantic Web are associated with logics (Fensel et al. 2002). A logic used with the Semantic Web is *FOL (First Order Logic) (otherwise known as predicate logic (Antoniou and Harmelen 2008) or predicate calculus).* FOL can be used to *"make statements about things and collections of things, their types and properties"*, however it cannot be used to describe the properties (predicates) (Passin 2004).

Ideally, as stated by (Sycara, Paolucci, Ankolekar and Srinivasan 2003), *"the Semantic Web provides a set of languages with well-defined semantics and a proof theory that allows agents to draw inferences over the statements in the language."*. FOL, however, is found to not scale too well over large quantities of data, which will prove problematic with the quantity of data that will be present in the Semantic Web (swi n.d.). Unfortunately, with an environment like the WWW, there is much contradictory information present. Knowledge in the Semantic Web will be patched together piecemeal and the problem with this that there is likely to be contradictory information present (Passin 2004). As Deng et al. (2007) states *"however, as shown in real world applications, it is possible for an ontology to contain two or more sources of inconsistencies and they may have different impact towards the inconsistencies"*. **Figure 2.4** demonstrates an example of how information encoded in the Semantic Web can create contradictions for inference engines. In this example, the semantic graph explicitly states that all Quakers *are* Pacifists and that all Republicans *are not* Pacifists. If Nixon is both a Quaker and a Republican, then there is a contradiction in that he by inference *is* and *is not* a Pacifist. Building on this argument that contradictory information in semantic graphs can create problems, Sirin, Parsia and Hendler (2004) state that there is a dependency between the detail of ontologies and the accuracy of matches found by an inference engine. As Brachman and Levesque (2004) state, *“no automated reasoning process for FOL can be both sound and complete in general"*.

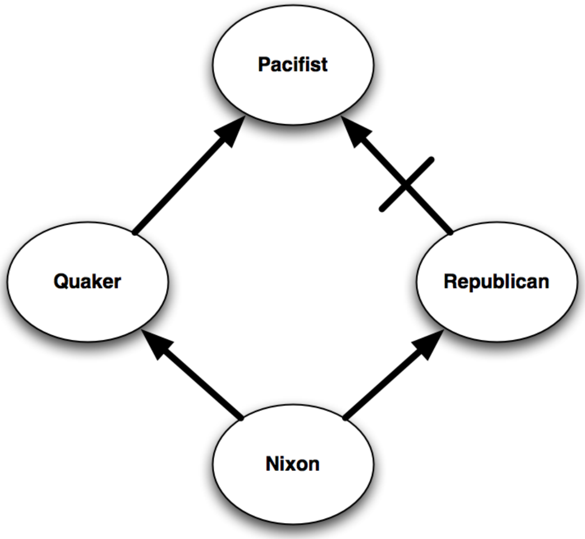


Figure 2.4: Example of contradictory information. Adapted from (Brachman and Levesque 2004)

In the Semantic Web - RDF (Resource Description Framework) is used for data representation. RDF forms the basis of the expression of inference rules for RDF parsing agents. RDF is composed of a nested graph of RDF-Triple statements in the form **(object-attribute-value)** (Davies et al. 2003), also known as **(subject-predicate-object)** . Such a collection of statements in a simplified non-XML syntax can read:

hasName('http://cic.hull.ac.uk/student/10011', 'Dave Webster')

hasWebpage('http://cic.hull.ac.uk/student/10011',
'http://dave-webster.com/')

hasTitle('http://dave-webster.com',
'Dave Webster's Webpage')

In the Semantic Web, ontologies build upon the RDF model and allow for standardisation of information representations (Lacy 2005).

Throughout the development of the Semantic Web there have initially been a number of evolving and competing ontology languages developed. From 2003 to the current date, the Semantic Web community has standardised around the W3C's *OWL* (Web Ontology Language) (owl 2004). OWL has evolved from predating ontology languages such as DAML + OIL (Passin 2004). Bussler (2008) notes that in many recent publications pertaining to the Semantic Web authors state that a homogeneous environment is used with respect to ontology and language used. Examples Bussler gives include: *“we developed an ontology that we use exclusively"*; *“we assume OWL-S as the SWS language"*; and *“all data is stored as triples in an RDF store"*.

2.3.3 Challenges to Semantic Web Ontology Construction

Despite a number of Semantic Web application implementations being developed and deployed in recent years, with drivers from business and defence, there are views that for the most part the dream of the Semantic Web is largely unrealised (Shadbolt 2006). There are opinions that the scale of the Semantic Web that we currently have is somewhat short of the ideal, as we explore in this section. The initial premise presented in the original Scientific American paper (Berners-Lee, Hendler and Lassila 2001) is that the 'Web of data' should *“routinely let us recruit the right data to a particular context"*, however, this task has proven less straight forward than originally presented.

Challenges concerned with the current WWW are identified by Goble (2003) who proposes a number of important identified questions concerning the development of the Semantic Web and Semantic Web ontologies. The statements from Goble concerning the WWW consist of:

1: *The Web grows from the bottom*,
2: *The Web is volatile and changeable*,
3: *The Web is heterogeneous*,

If these issues are to be inherited by the Semantic Web then we will encounter many issues when it comes to ontology development and use. The bottom-up approach to web content creation will mean that in some cases people will create their own ontologies without reference to public base ontologies, meaning that there will be many different representations of similar or identical terms or objects. This problem with the Semantic Web is that authors are not prevented from using different semantics to describe the same thing (Passin 2004). As Russomanno (2006) states, reasoning over ontology graphs suffers from sensitivity regarding *“subjective assertions of the ontological engineer"*. Echarte, Astrain, Cordoba and Villadangos (2007) discuss a number of challenges with the social and engineering process of building ontologies. These are summarised as:

> 1: *“Ontologies have some drawbacks due to their difficulty to be built and used".*
>
> 2: It is difficult to *“anticipate the whole domain and the complete set of point of views of all possible users".*
>
> 3: *“Ontologies are difficult to maintain, mainly in changing environments".*

This heterogeneity will present an issue and a need for creating inferences between web-like ontologies and identifying overlaps (Davies et al. 2003). Whilst this ontology mediation is possible, albeit with effort required, Bussler (2008) proposes that data and process heterogeneity needs to be addressed between standard ontologies and their layers.

These challenges lead into the volatile nature of WWW resources. In the WWW it is common to encounter a *404* (page not found) error when a HTML page has been removed from a web server. In the case of ontologies, if a personal ontology is referencing a parent ontology and the processing engine finds that its parent ontology disappears from the Semantic Web or is drastically changed in some way then problems will be encountered.

Flake et al. (2003) neatly summarises a fundamental challenge faced when attempting to annotate current WWW documents with Semantic Web markup. Having web content annotators use the same (or similar) annotation language will be a best-case scenario with the annotation of long passages or pages becoming a daunting task.

Jones (2004) discusses a **high end** Semantic Web in which reasoning over a universal world model with a controlled formal language occurs, which Jones notes cannot be achieved through surface word matching and statistical operations on words. This version of the Semantic Web will assist humans with decisions, rather than enabling them. The barrier to realising this tier of the Semantic Web is that a world model, or encyclopaedia, is needed to enable this seamless connectivity. The

middle end Semantic Web implies that an upper ontology of 'ordinary' knowledge is shared amongst applications, however the paper's author states that these types of models can quickly become over complicated. Finally, the **low end** version of the Semantic Web suggests lightweight statistical processing strategies over text with field tagging and lightweight natural language processing. This final version can be viewed as equivalent to Web 2.0. Jones states that a lesson learned from the field of natural language processing here is that shallow text operations still have a high degree of usefulness with minimal semantic tagging recommended.

2.3.4 Summary

At this point it is prudent to reiterate that there is a distinct divide between the WWW and the Semantic Web, both in the manner in which they are processed by machines (or agents) and the manner in which users create content within them.

One fundamental difference between the current text-based WWW and the Semantic Web is that of processing in terms of information retrieval and document similarity determination. Whereas the WWW relies upon natural language processing and statistical methods, the Semantic Web relies upon logic and directed graph matching algorithms (with limitations discussed in this chapter) and (at the current time) does not substantially make use of statistical methods (Berners-Lee, Hall, Hendler, O'Hara, Shadbolt and Weitzner 2006).

2.4 Web 2.0 as a Pragmatic Bottom-Up Alternative to the Semantic Web

Although a contentious topic to pin down to a definition, the Web 2.0 definition originally defined by Tim O'Reilly (O'Reilly 2006) is:

> "*Web 2.0 is the business revolution in the computer industry caused by the move to the internet as platform, and an attempt to understand the rules for success on that new platform. Chief among those rules is this: Build applications that harness network effects to get better the more people use them. (This is what I've elsewhere called "harnessing collective intelligence.")*"

Web 2.0 is developed to intelligently interoperate and integrate information systems on the WWW in addition to being a pragmatic method to build upon the current WWW to support services. Web 2.0 has been promoted as a simpler paradigm to that inherent in the Semantic Web. Web 2.0 aims to utilise current technology and application areas to provide such intelligent services without the need for overly complex stacks of standards (Ankolekar, Krötzsch, Tran and Vrandecic 2007), such as has been observed with the Web Services and Semantic Web approaches. Web 2.0, along with related components, such as tagging, folksonomies and wikis, can be seen as an alternative to deep ontologies (Shadbolt 2006), with such structures evolving organically to suit the individual or community's changing requirements.

Web 2.0 takes a **'worse-is-better'** approach, originally described by Richard P Gabriel, albeit in application to C and Unix systems. To better the worse-is-better approach to pragmatism, a quotation from Gabariel (1989) is presented as follows:

> "*The **big complex system scenario** goes like this:*
> *First, the right thing needs to be designed. Then its implementation needs to be designed. Finally it is implemented. Because it is the right thing, it has nearly 100% of desired functionality, and implementation*

> *simplicity was never a concern so it takes a long time to implement. It is large and complex. It requires complex tools to use properly. The last 20% takes 80% of the effort, and so the right thing takes a long time to get out, and it only runs satisfactorily on the most sophisticated hardware.*"

One can see parallels between the pragmatic (if not the most elegant) solution to information integration on the web provided by Web 2.0 versus the well engineered (although complex to implement) Semantic Web.

Additionally, Gabariel (1989) goes on to say:

> "*In the worse-is-better world, integration is linking your .o files together, freely intercalling functions, and using the same basic data representations. You don't have a foreign loader, you don't coerce types across function-call boundaries, you don't make one language dominant, and you don't make the woes of your implementation technology impact the entire system.*"

Although this approach to technology success has been disputed (Bourbaki 1992), one can see a pattern between the success of 'right thing' languages (for example, LISP and pragmatic languages, such as C, which are designed to 'get the job done' on as many platforms as possible) applied to the Semantic Web versus Web 2.0. This simplification of the interfaces of Web 2.0 applications and APIs is a major contributor to the trend in aggregating and remixing Web 2.0 content, otherwise known as a *“mash-up"* of content. The quality of human judgements in Wikipedia will be revisited in **Chapter 5**

As an aside, it is worth considering the quotation attributed to Albert Einstein at this point:

> "*Make everything as simple as possible, but not simpler.*"

2.4.1 A Web of Blogging, Tagging and Folksonomies

As part of the Web 2.0 trend, there are a number of technologies and social network interaction patterns associated with the trend. Such examples include weblogs, tagging, social networking and wikis.

Weblogs (blogs) and Web Feeds

Web feeds, introduced earlier in this thesis, provide a method for serialising and syndicating information items on the web. A common use of web feeds is in syndicating news items in what is more commonly known as a news feed.

Weblogs, or blogs, as they are more commonly known, are a web site pattern, whereby news items (for example from a diary or journal) are presented in a chronological list in a repeating pattern.

Weblogs are typically presented and distributed in two forms. The first form being standard HTML and the second a web feed. It is important to note that a large proportion of web feeds on the web source from blogs and that a large proportion of recent blogging software such as Wordpress and Typepad provide web feed functionality. More recent web feed formats, such as RSS 2.0 and ATOM 1.0 allow the publisher to include multiple category tags for each web feed item and for the web feed itself; this is a feature supported by modern blogging software.

Tagging and Popularity

Tagging, common to Web 2.0, involves attaching a keyword tag to an information item or resource on the web. Example uses include attaching keyword tags to web site URLs in order to provide index terms for later browsing.

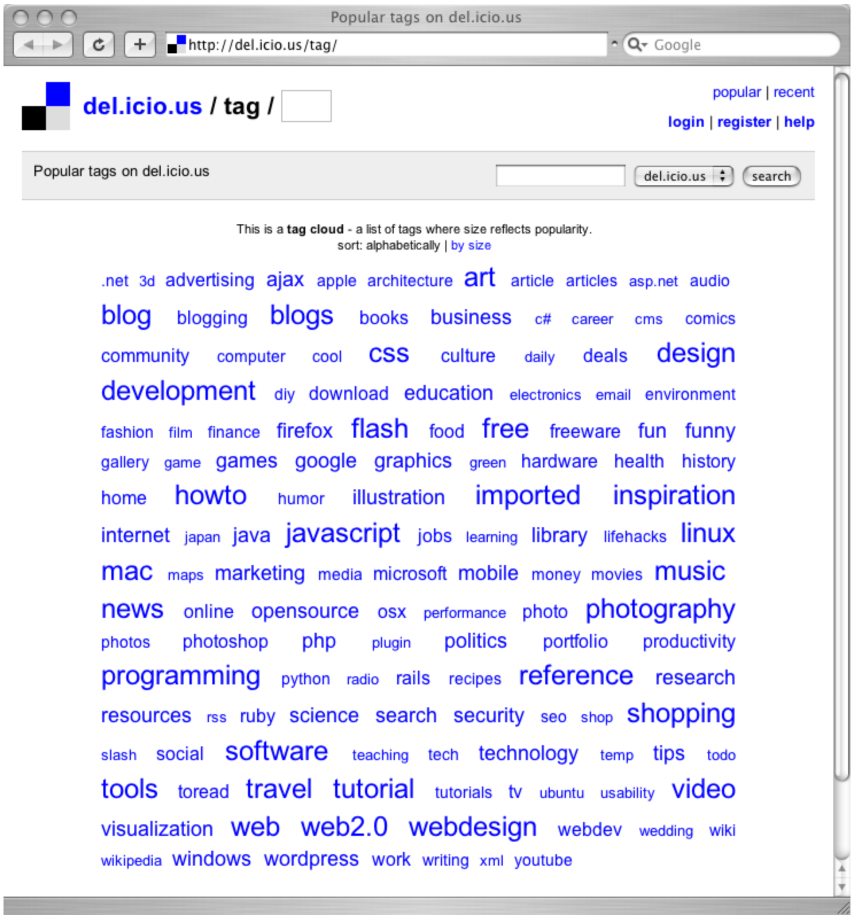
The current generation of blogging software allows authors to add free-text tags to blog posts. Such software will allow a blog author to create a list of internal text category labels, otherwise known as *“tags"*, of which multiple categories can be assigned to each blog post item. Categories from blogging software can be encoded in blog web feed export formats, such as RSS 2.0 and ATOM. This inclusion of categories in web feeds allows blog search engines, for instance Technorati, to index blog posts based on topic. The practical benefit of this method is that users can search for blogs (or blog items) based on a specific tag or tags.

By providing tags, blog authors can help reduce the *information overload* problem and aid in both information retrieval and filtering, by providing index terms (tags) for search engines to match, rather than relying on the search engines indexing the blog content by the text.

As has been highlighted in this chapter, there are a number of barriers to the realisation of annotating the WWW through the use of Semantic Web ontology-based technologies. Web 2.0, on the other hand, lowers the 'barriers to entry' in terms of attaching metadata to resources in the same way that the informality of tagging has the advantage of a low barrier of entry and can lead to high user base of tagging on the WWW (Berners-Lee et al. 2006, Ankolekar et al. 2007).

Folksonomy and Controlled Vocabulary

A driver behind the strength of Web 2.0 is within its associated social processes (Lerman 2007). Web 2.0 relies upon its user to tag and contribute content, whereby such popularity is used as a method of prioritising the display and indexing such information (Berners-Lee et al. 2006). In Web 2.0, it is common to use a socially managed controlled vocabulary, otherwise known as a folksonomy (Gruber 2005). In a folksonomy a category tag from a group of tags can be used within a website to annotate a particular item. This grouping of tags is maintained through a popularity measure and the encouragement for 'taggers' to use similar tags to those of their peers. A method of viewing this tag usage is to use a 'tag cloud', whereby the size of tags are determined by their popularity. This tag cloud allows the viewer to quickly see which tags are popular at the current time. **Figure 2.5** shows an example of a tag cloud for the social bookmarking website Del.icio.us.



del.icio.us / tag /

Popular tags on del.icio.us

This is a **tag cloud** - a list of tags where size reflects popularity.
sort: alphabetically | by size

.net 3d advertising ajax apple architecture art article articles asp.net audio
blog blogging blogs books business c# career cms comics
community computer cool css culture daily deals design
development diy download education electronics email environment
fashion film finance firefox flash food free freeware fun funny
gallery game games google graphics green hardware health history
home howto humor illustration imported inspiration
internet japan java javascript jobs learning library lifehacks linux
mac maps marketing media microsoft mobile money movies music
news online opensource osx performance photo photography
photos photoshop php plugin politics portfolio productivity
programming python radio rails recipes reference research
resources rss ruby science search security seo shop shopping
slash social software teaching tech technology temp tips todo
tools toread travel tutorial tutorials tv ubuntu usability video
visualization web web2.0 webdesign webdev wedding wiki
wikipedia windows wordpress work writing xml youtube

Figure 2.5: Tag cloud from Del.icio.us

Del.icio.us is a social web bookmarking site, whereby users can assign tags to their bookmarked web links. Del.icio.us attempts to control it's tag vocabulary through recommendations. When a user attempts to bookmark the website, the Del.icio.us interface will present a number of popular tags that other users have used to tag that website. In this manner, Del.icio.us encourages users to tag websites using similar tags and encourages users to reuse tags, rather than create a new variation on keywords.

Connected to the notion of a folksonomy, tags from a common external folksonomy are often used, such as Technorati, Del.icio.us or Flickr.

Wikis

Although one could argue, not strictly Web 2.0, Wikis have grown to become a popular tool for collaborative working and knowledge building. A wiki allows external users to create pages in a web site and edit them collaboratively. Users can then create links between these ad-hoc pages. Perhaps the most well known wiki today on the web is Wikipedia, an online encyclopaedia, whereby web users can contribute through adding and modifying existing information. Wikis build upon a principle similar to the Web 2.0 principle of user edited content and consensus through popularity.

2.4.2 Selection and Filtering

Popular with the Web 2.0 trend is the concept of mash-ups, whereby content from multiple sources is aggregated together and combined. Information can be aggregated from an external source through a public API or a web feed. In Web 2.0, XML is primarily used as the serialisation syntax, with many applications utilising XML based web feeds as a means of serialising lists of items. Such a pattern eases the loosely coupled interaction of these information services.

Applying this pattern to the UNIX pipes pattern, Yahoo Pipes (Yahoo! 2007) has provided a web based interface to 'pipe' information from a number of information services through a number of transform operations to outputs such as HTML or mapping services, for instance Yahoo Maps (Yahoo! 2007) or Google Maps (Google 2007). Such piping tools allow a structured approach to integrating loosely coupled Web 2.0 resources into a workflow. One aim of Web 2.0 is to aid in selection and filtering of information on the web. Such ability to create workflows along with intelligent filters allows for information to be filtered and 'noise' removed. However, these 'intelligent' filters need to be developed in order to be used in workflows.

2.4.3 Summary on Resource Authority

As presented in this chapter, one of the problems faced by the Semantic Web is that of the notion of a global ontology. The Web 2.0 style, on one hand, has demonstrated that social knowledge-bases (albeit of a simple, shallow nature), can evolve organically with no single authority of control. With this style we can see folksonomies present within such community-driven Web 2.0 applications such as Del.icio.us and Flickr.

The next chapter will consist of an investigation into the notion of context. Part of this investigation will involve gaining an understanding of this abstract notion and pinning it down to a definition. The relationship between context and information filtering will be explored.

2.5 Concept Representation and Concept Similarity

This section focuses on the representation of concepts in a knowledge-base before discussing semantic similarity between concepts based upon the characteristics of the knowledge-base. To expand this, we initially concern ourselves with the representation of concepts in a knowledge-base through the investigation of current Web-based knowledge-bases. Secondly, we then follow up with a summary of methods for measuring the similarity between concepts through the use of these knowledge-bases.

In this thesis we focus on the explicit knowledge provided by the user. This knowledge is presented in the form of concepts that they are interested in. We propose that through the use of a knowledge-base we can implicitly collect related concepts based on a user's given concept. Through a community maintained knowledge-base we can combine explicit knowledge and expand it through implicit knowledge and provide a personalised solution rather than a one-size-fits-all approach improving over those that are described later on in **Section 3.2.1**.

2.5.1 Understanding Words, Concepts and Physical Objects

This section gives an overview of how concepts can be represented using web based information systems and knowledge-bases. Through learning how concepts are represented in a knowledge-base and more importantly how they are interrelated we can investigate how to expand out explicit concepts using related neighbouring concepts - thus satisfying one of the objectives of this thesis.

To begin with we introduce the Ogden Meaning Triangle (Ogden and Richards 1923), described in (Fensel et al. 2002)) as illustrated in **Figure 2.6**.

The Ogden Meaning Triangle describes the relationships between: objects in the world (*Things*); *Concepts*; and words (*Symbols*). The important point of the diagram is to illustrate that a Word/Symbol cannot fully capture the essence of Concepts or Things. The connection between a Symbol/Word and a Thing is implicit and is achieved through the invoking of a Concept (Fensel et al. 2002). For the rest of this thesis *word* and *symbol* will be used interchangeably. Also *Thing* and *physical object* will be used interchangeably.

The understanding of the term *'concept'* is important for the work within this thesis. We understand that the concept that a word refers to may only exist as an abstract concept - this concept buffer between a symbol and a thing is fundamental as we are not necessarily talking about a physical instance of a real-world object, but rather to an abstract definition of a concept, which can be equated to a mental model. Furthermore, the concept entity helps anchor a specific word sense of a word to the concept entity.

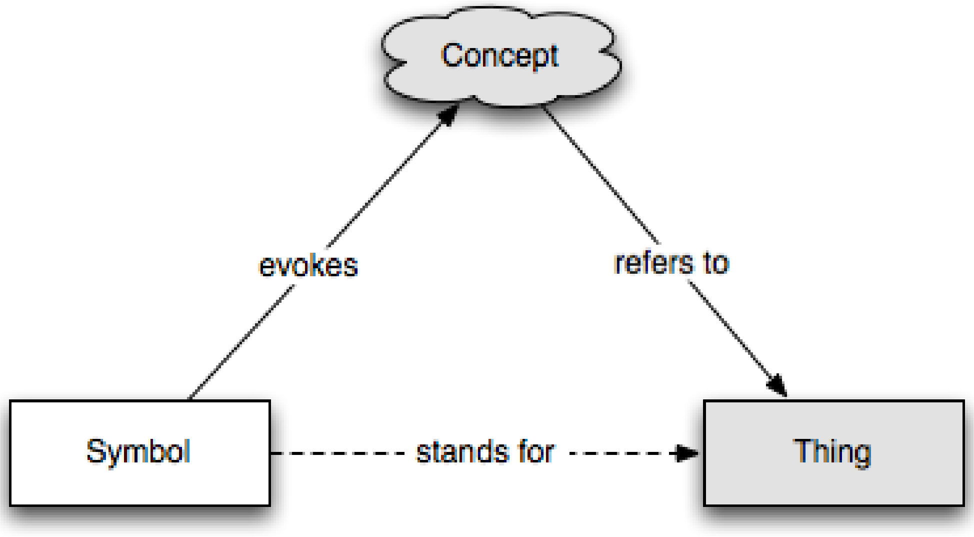


Figure 2.6: Ogden Meaning Triangle, redrawn from illustration from (Fensel et al. 2002)

2.5.2 Word Sense Disambiguation and Web 2.0 Tagging

Whilst the Web 2.0 method of tagging, based upon shared folksonomies, may seem like an ideal way to add community popular index terms to resources, the problem of word sense disambiguation still remains. A quick search on sites that provide a tag folksonomy, for example Technorati and Del.icio.us, reveal that there is no disambiguation method applied to the keywords themselves. Instead of adding disambiguation information to the category keywords, for example *“java (programminglanguage)"*, the shared folksonomies rely upon multiple tags being used when tagging items (for instance java programminglanguage), with the connection between the tags being implicit.

2.5.3 Concepts and Word Sense Disambiguation

Word sense disambiguation occurs when one tries to map a textual keyword to that of a concept. This has been introduced in the context of the Ogden Meaning Triangle earlier within this chapter. When a user desires to create a keyword annotation or tag for a concept, the issue of word sense disambiguation will become a concern (Murthy and Keerthi 1999). One example to explain this idea is to look at the keyword **Java**. One could easily think of three senses of the word **Java**; programming language, coffee or island. When a user enters the keyword into a software tagging system, the system should ideally identify and understand the word sense of the word. In order to aid in this process a knowledge-base or text corpus can be used.

2.5.4 Basic Level Variation

In addition to word sense disambiguation, there is also the issue of basic level variation, polysemy and synonymy. Similar to the problem of disambiguation in the previous section, variations in keyword representations of concepts are entirely possible in tagging systems (Gruber 2005).

In addition to these well understood typographic variations to human writing there are problems of errors present in human text, particularly on the WWW. A fitting comment from Dave Winer, the creator of RSS, (Winer 2006), states:

“You guys want users to enter metadata, I’m looking for ways to get around that,

because I have found that people don't even spell things right, much less label things."

Through the use of a well designed controlled vocabulary, as discussed earlier in this thesis, the user is encouraged to use a process to reduce typographical errors. An example of this in practice is the Del.icio.us bookmark tagging system, whereby the user is encouraged to reuse existing tags.

2.6 Knowledge-Bases

The definition of a *'knowledge-base'* is of value for this thesis as one of its contributions is to use Wikipedia as a knowledge-base. In order to validate that Wikipedia is a knowledge-base the terms *'knowledge'* and *'knowledge-base'* need to be defined as a goal of this task is to understand how knowledge is stored, represented and linked together within a knowledge-base. In research literature the grounding of the terms *'knowledge'* and *'knowledge-base'* are often loosely defined. Furthermore, where a term is defined in one source it will differ with a definition by another source.

2.6.1 Knowledge

Before we ground the term *'knowledge'*, we must understand the supporting terms *'data'* and *'information'*. Fortunately these terms are well understood and are summarised by Bergmann (2002) as follows:

- *'Data'* are syntactic patterns that contain no meaning in themselves and are subject to interpretation.
- *'Information'* is data with meaning which has already been interpreted.

In order to obtain *'information'*, *'data'* needs to be processed in some manner; this could be a mathematical process or a human process.

'Knowledge' is a term that is less consistently grounded than data and information, however. On one hand Bergmann (2002) again states that knowledge is information that is put into the context of a task or goal, described as *"a set of related information with pragmatics"*. On the other hand, Lehmann (1989) states in the context of humans that they can infer information from knowledge. Furthermore Mylopoulos, Borgida, Jarke and Koubarakis (1990) tells us that knowledge can be represented through language using propositions.

Nonaka (1994) discusses the concept of knowledge creation in the context of organisations where knowledge is created to help to solve defined problems. In this argument knowledge is defined as *'justified true belief'* which can be expressed in propositional form and formal logic. As with the previous definitions, the processing of information can lead to the creation of both information and knowledge; this implies that the process can work both ways or that information and knowledge can work on multiple levels. Knowledge is viewed as being *"created and organised by the very flow of information"* which is the necessary medium for initiating and formalising knowledge.

For this thesis the creation of knowledge is viewed from the perspective of a human creating and distilling this knowledge. Furthermore this creation process of knowledge is viewed as a human willingly externalising this knowledge for a goal in an explicit form. Knowledge is not just a random 'information dump', but rather is directed for a purpose and is reusable. In order for a human to produce this knowledge both information and knowledge are processed and these may come from new sources and/or existing sources within the human's mind. To give an example of this the human will produce an academic paper which will itself contribute to knowledge within the research field; this output will come from both the synthesis of knowledge from other academic papers and the processing of information from experiments, etc. The importance of knowledge, therefore, it that it is tangible, explicit and created for a purpose.

2.6.2 Knowledge-base

Moving on to the term *'knowledge-base'*, from a pragmatic perspective a common understanding in the literature is based on the use of a knowledge-base for drawing inference from, with Nardi and Brachman (2003) clarifying the characterisation of knowledge-based systems as the *"ability of a system to find implicit consequences of its explicitly represented knowledge"*. An example of this grounding can be seen in Fensel et al. (2002)'s definition of a knowledge-base as:

> *"a knowledge-base provides permanent storage for information and the ability to use knowledge to draw inferences from the information that has been explicitly stored in it."*

Taking this statement, there is a contradiction present in research literature whether a knowledge-base stores information or knowledge. Correlating with the above quotation, Buitelaar, Cimiano, Frank, Hartung and Racioppa (2008) describes extracted information from football matches being stored within a knowledge-base and using an ontology to allow for linking in a well founded manner to provide a background for the interpretation of new information. A correction to this definition proposed here is that a knowledge-base stores *'knowledge'* to allow it to be processed, not just random information, as correlates with Akerkar and Sajja (2009), Haarslev and Möller (2001) and Guo, Qasem, Pan and Heflin (2007).

When discussing the contents of knowledge-bases, literature generally agrees that contents are propositions (Lehmann 1989, Guarino 1995) or propositional sentences (Katsuno and Mendelzon 1991). Nardi and Brachman (2003) informs us that early developments within the field in the 1970's divided up these representations into two categories: logic based formalisms; and non logic-based representations for example ad-hoc data structures. As has been discussed in **Section 2.3.1** Description Logics can be used to construct formal ontologies to use as the TBox part of the knowledge-base. In the case of non Description Logic based knowledge-base work (Nakai and Kanehisa 1992), this may be manifested as a collection of "if-then" rules. Regardless of representation logic or formats, the purpose of a knowledge-base is to represent, as Guarino (1995) describes, *“objective reality instead of an agent’s mind”* or as Lehmann (1989) describes, *“A set (finite or infinite) of conditional assertions”* representing *“defeasible knowledge an agent may have.”*

A further facet to the knowledge-base that has been briefly introduced is its structure. As has been established, knowledge is produced for a purpose and a knowledge-base stores externalised knowledge, but is a collection of academic papers a knowledge-base? The answer here is no, as a knowledge-base needs to contain a structure rather than be just a collection of random pieces of knowledge. The knowledge-base needs to possess a logical schema to link knowledge together. In order for machine processing of a knowledge-base the knowledge fragments need to be connected in a homogeneous manner. This can be achieved through the use of a schema.

Knowledge-bases use schemas in a similar manner to databases and knowledge-bases are sometimes compared to databases in the literature, with a statement from Mylopoulos, Chaudhri, Plexousakis, Shrufi and Topologlou (1996) that, *“there are no technical grounds for distinguishing between the two terms”*. Not all sources agree on this however, as the two can be distinguished by the Description Logic and ontology research community which defines the terms *‘TBox’* and *‘ABox’*. A Description Logic defines a knowledge-base as consisting of both TBox and ABox components. TBox defines terminology/vocabulary which consists of concepts as can be equated to a database schema, whereas ABox defines assertions over that concepts. TBox describes *‘intensional knowledge’* and describes general properties of concepts. The ABox defines *‘extensional knowledge’* which is knowledge specific to individuals (Baader, Calvanese, McGuinness, Nardi and Patel-Schneider 2003, Guo et al. 2007). Furthermore, intensional knowledge is described by Baader as being relatively stable over time, whereas extensional knowledge will be subject to constant change and dependent on ongoing circumstances and can be related back to Guarino (1995)’s definition of objective reality instead of an agent’s mind. One further distinction is provided by Guo et al. (2007) who states that when evaluating knowledge-bases and databases, the former is evaluated based on completeness of their reasoning, whereas with the latter the evaluation is based on retrieval speed. Therefore, for this thesis, when compared to a knowledge-base a database is viewed as just structured information rather than structured knowledge and, therefore, a knowledge-base may be implemented on-top of a database.

Following the thread on the knowledge-base schema a key point of our knowledge-base definition is that the practical use of the knowledge-base is through the inferences that can be carried out over it. Knowledge is the primary source of any reasoning process, therefore, knowledge-bases can be used as a basis for inferences for knowledge processing. This knowledge reuse helps humans and computers make better use of experience, individually and through community contributions (Bergmann 2002). However, in order for a computer to be able to reason over this knowledge (Shipman and McCall 1994), a formalism is required, which is the identification of the machine-processable parts of information. A homogeneous schema contributes towards this formalism providing a representation of knowledge and the interrelations between pieces of knowledge to aid the use of First-order Logic or a Description Logic (Guarino 1995). We can, therefore, assert for our definition that the structure of a knowledge-base provides context for inference through propositions with underlying logic. To contrast to a random collection of information, the structure of a knowledge-base is based on logical propositions to create structured knowledge from unstructured knowledge.

2.6.3 Developing a Knowledge-base

In order for a knowledge-base to be useful it needs to be told facts, described by Brachman and Levesque (2004) as *“the beliefs of the system that are explicitly given"* with *“the entailments of that knowledge-base as the beliefs that are only implicitly given"*. From a practical use, the knowledge-base adjusts its behaviour based on the facts it has been told. Fellbaum (1998) tells us that *“common-sense reasoning requires extensive knowledge"*, therefore, millions of concepts are required for a knowledge-base that is useful for practical reasoning problems. Such a rich and comprehensive knowledge-base will take a large effort to build. Fellbaum also tells us that another desirable feature of a knowledge-base is *“to have a rich concept connectivity while using only a small set of relation types."*, he continues to state that inference is aided by rich connectivity in the knowledge-base. Just as with the problems discussed with FOL in the Semantic Web, more expressive representation languages become more difficult to reason across, leading to a tradeoff between formalism and scalability (Baader et al. 2003).

Shipman and McCall (1994) discussed the problems associated with formalism. There is a tradeoff here as on one hand it is proposed that greater levels of functionality require greater levels of formalism. On the other-hand, if the knowledge-base is contributed to by humans then strong formalisms are often time consuming for a human to provide and can dissuade them from contributing to the knowledge in the knowledge-base.

An important addition for our definition of a knowledge-base is that when concepts are defined in a knowledge-base, only one definition for a concept name is typically allowed. This allows for a unique definition of terms/concepts (De Bruijn, Eiter, Polleres and Tompits 2007, Nardi and Brachman 2003) to allow reasoning over and ensures that when multiple parties are reasoning over the knowledge-base they are reasoning about the same concept. Additionally, these definitions must not be cyclic, for example, definitions must not refer to themselves or be defined in terms of other concepts that refer to them Baader et al. (2003).

This discussion is continued in the context of a knowledge-base in the form of an encyclopaedia-like structure later on in **Section 5.5**.

Within the following sub-sections the methods and conventions used to represent concepts within web based information systems are discussed. Systems that are represented here include WWW, Semantic Web, Web 2.0 tagging and Semantic Lexicons.

2.6.4 WWW Representation of Concepts

The WWW provides content authors with the ability to represent ideas and concepts in a flexible manner. As described in **Section 2.1.4**, one of the problems with HTML is that it was designed to be network-accessible, but not necessarily ’machine-understandable’. Consequently, HTML documents published on the WWW were only intended for human readership. As a result of this situation, HTML provides no formal manner to identify and represent concepts over that of a human written document whereby concept identification is achieved through human comprehension of the page text.

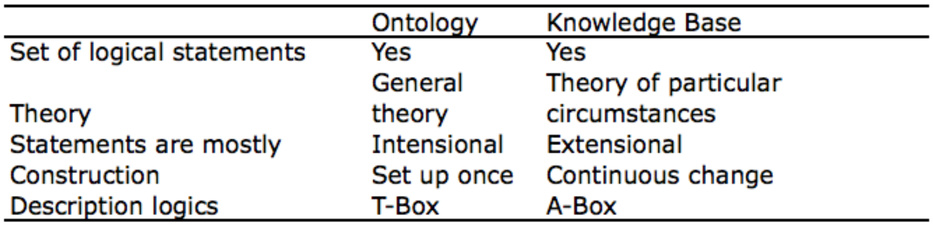
2.6.5 Semantic Web and Ontology

As previously discussed - in the Semantic Web, resources (including concepts) are identified (addressed) through the use of URIs (Passin 2004), therefore, the naming of RDF resources is achieved through the use of the URI. We note that whilst URIs are used to identify resources, the URI does not have to explicitly point to the RDF document (Powers 2003, Passin 2004). The RDF data model allows named resources to contain relations to other resources using the *Subject, Predicate, Object* model as we have covered in **Section 2.3.2**.

It is important to note that there are a number of distinguishing attributes between an ontology and a knowledge-base, despite in some literature a Description Logic Knowledge-base being equated to an ontology (Baader, Ganter, Sertkaya and Sattler 2007). These differences are summarised by (Fensel et al. 2002) in **Figure 2.7**, who states:

> *“An ontology constitutes a general logic theory, whereas a knowledge-base describes particular circumstances pertaining to such a theory ... an ontology is (mostly) constituted by intentional logic definitions, whereas a knowledge-base comprises (mostly) the extensional parts."*

Referring back to **Section 2.6**, the concept of ontology fits with the TBox component of a knowledge-base, whereby the vocabulary/concepts and formal relationships between the concepts are defined. The use of RDF is where the ABox extensional knowledge that provides the knowledge-base with the power to be used



	Ontology	Knowledge Base
Set of logical statements	Yes	Yes
Theory	General theory	Theory of particular circumstances
Statements are mostly	Intensional	Extensional
Construction	Set up once	Continuous change
Description logics	T-Box	A-Box

Figure 2.7: Distinguishing attributes between an ontology and a knowledge-base. Recreated from (Fensel et al. 2002)

.

for common-sense reasoning applications. Regarding a focus for this thesis, the stated differences between an ontology and a knowledge-base help us to focus on the knowledge-base over just the formal ontology as a source of contextual information in order to alleviate the information overload problem.

2.6.6 Semantic Lexicon

A semantic lexicon is a structure of concepts with defined associations between the concepts. A common type of relationship is the *is-a* relationship. An example of this relationship is presented in **Figure 2.8** from (Varelas, Voutsakis, Raftopoulou, Petrakis and Milios 2005). Instances of semantic lexicons used in similarity research are the DMOZ and WordNet lexicons (Chakrabarti, Punera and Subramanyam 2002). The DMOZ tree provides a topic hierarchy of concepts with web links attached to topics. WordNet expands on the notion of a topic hierarchy by providing hyponyms and synonyms of around 100,000 English concepts (Qu, Hu and Cheng 2006, Pedersen, Pakhomov, Patwardhan and Chute 2007). In addition to the discussed relationships, WordNet provides a short text gloss for each concept, consisting of a short sentence to describe the concept (Patwardhan, Banerjee and Pedersen 2003).

Berners-Lee et al. (2006) distinguishes WordNet from an ontology, describing WordNet as a collection of lexical items with different senses. He distinguishes this from

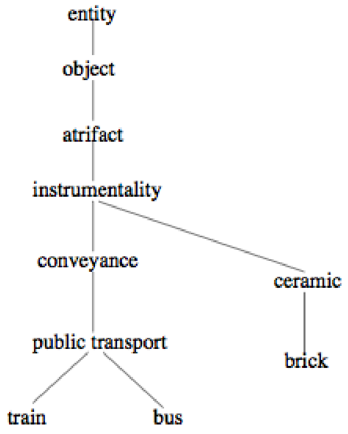


Figure 2.8: Is-A relationship from WordNet. Sourced from (Varelas et al. 2005)

.

an ontology, as an ontology will try to ensure a unique interpretation of the terms it uses. Unlike a basic topic hierarchy, WordNet provides facilities to represent synonyms, hyponyms and hypernyms of words in order to assist in the detection of word variations. To clarify, synonyms are words with similar meanings; hyponyms represent words of a specific category (downwards in the is-a relationship) and hypernyms represent words more general to the word (upwards in the is-a relationship) (Benatallah, Hacid, young Paik, Rey and Toumani 2006, Fileto, Liu, Pu, Assad and Medeiros 2003). Following this idea, the work by Varelas et al. (2005) describes an approach to expand out queries by aggregating synonyms, as can be shown in **Figure 2.9**. These variations prove useful in semantic matching (Qu et al. 2006).

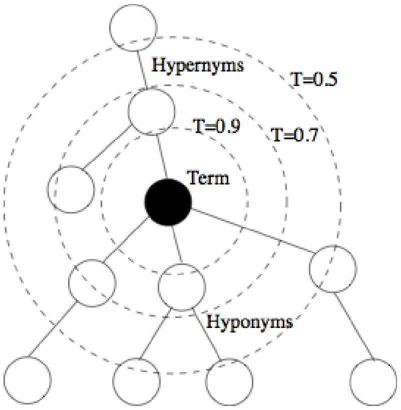


Figure 2.9: WordNet term expansion. Sourced from (Varelas et al. 2005)

.

2.6.7 Summary of Knowledge-bases

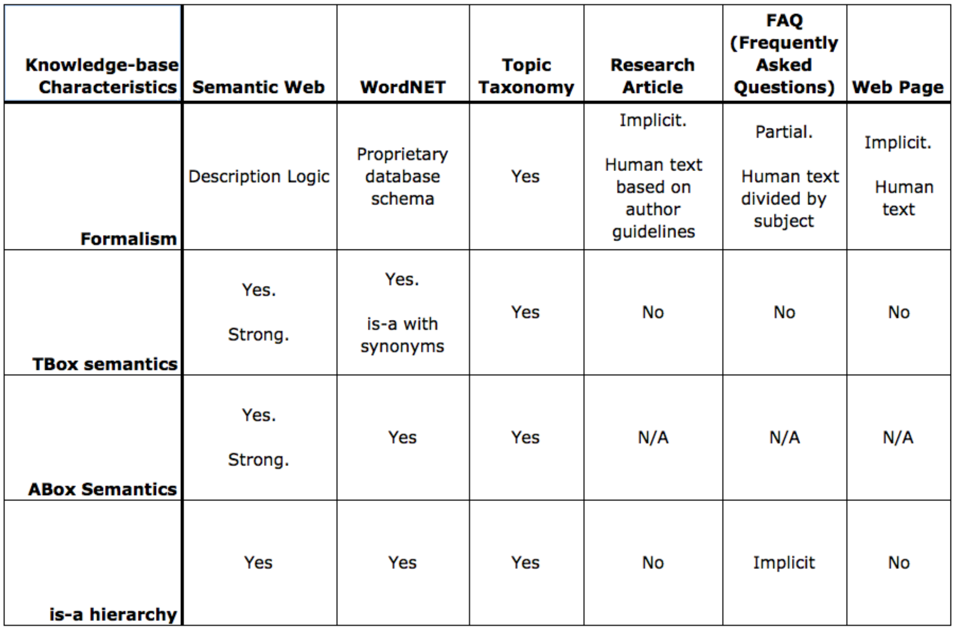
The definition of a knowledge-base can be summarised from **Section 2.6**. A knowledge-base contains knowledge which is represented and linked through propositional sentences. This is provided through a formalism; for instance a homogeneous schema that links knowledge together. A strong representation of this is through a Description Logic description of TBox semantics where concepts are defined - as can be equated to an upper ontology. ABox semantics describe intensional knowledge based on the TBox concepts. Both of these complete the knowledge-base.

Since the Semantic Web is founded on Description Logic ontologies and RDF ABox assertions, the Semantic Web is by definition a strong knowledge-base due to its rich description of concepts and the ability to create extensional knowledge through rich predicates.

A semantic lexicon is a weaker version of a knowledge-base, but still possesses strong formalisms. To qualify this statement, whilst a semantic lexicon (such as WordNet) does provide a homogeneous schema for defining and linking concepts, by definition of it's application domain it is a weaker knowledge-base due to the limited predicate semantics used to link concepts together compared to a full Description Logic knowledge-base. The main semantic being the *'is-a'* relationship forming a hierarchy of concepts and the linking to synonyms. A further difference to an ontology is that an ontology will ensure a unique interpretation of concepts where as the aim of a semantic lexicon is to describe *'words'* that in-turn map on to unique concepts.

Returning to the representation of concepts (and in turn knowledge) on the WWW, it is difficult to classify the WWW as a knowledge-base, even as a weak knowledge-base. The motivation behind this statement is that the WWW can be compared to the previously described *"collection of academic papers"*. Whilst the web provides the ability to hyperlink between resources, there is no schema assigned to which type of resources are linked. To contrast to a random collection of information, the structure of a knowledge-base is based on logical propositions to create structured knowledge from unstructured knowledge.

This can be further summarised in tabulated form in **Figure 2.10** below.



Knowledge-base Characteristics	Semantic Web	WordNET	Topic Taxonomy	Research Article	FAQ (Frequently Asked Questions)	Web Page
Formalism	Description Logic	Proprietary database schema	Yes	Implicit. Human text based on author guidelines	Partial. Human text divided by subject	Implicit. Human text
TBox semantics	Yes. Strong.	Yes. is-a with synonyms	Yes	No	No	No
ABox Semantics	Yes. Strong.	Yes	Yes	N/A	N/A	N/A
is-a hierarchy	Yes	Yes	Yes	No	Implicit	No

Figure 2.10: Summary for knowledge-base characteristics against knowledge-base candidates

2.7 Concept Similarity

In this section we will give a general overview of semantic similarity between concepts. To be specific, there is a defined difference between semantic similarity and relatedness. Semantic similarity is defined by Patwardhan et al. (2003) as:

> *"Semantic similarity is a kind of relatedness between two words that defines a resemblance."*

Likewise, Pedersen et al. (2007) views similarity of words as how they are deemed to be related based on their likeness. Semantic relatedness can be viewed as a more broad concept, which covers all kinds of relations (Strube and Ponzetto 2006), with semantic similarity being a special case of semantic relatedness (Pedersen et al. 2007).

Despite the simplicity of informal and simple concept representation systems with low barriers to entry (for instance, tagging) and the associated advantages they provide, formally engineered concept representation structures provide additional information regarding the concepts placed into them (for example, WordNet and the Semantic Web). Tree structures and directed graphs are typical examples of these structures; taxonomies and RDF/ontologies representing these structures respectively.

One disadvantage of using a tag-based approach, as opposed to using a formally defined knowledge-base to annotate a web resource, is that of word-sense disambiguation, briefly mentioned in **Section 2.4**. This notion is discussed further in **Section 2.5.3**. As has been introduced, the tagging of web resources can be a seen as a more pragmatic method for providing information than relying upon formal ontologies provided by a central control (Gruber 2005) in order to infer relatedness between resources. The informality of tagging has the advantage of a low barrier of entry and can lead to large user base to tagging as has been shown with the Web 2.0 trend (Berners-Lee et al. 2006, Ankolekar et al. 2007).

2.7.1 Taxonomy Approach to Similarity

The idea of arranging concepts in a tree-based taxonomy is not a recent one. For many years humans have arranged objects in the world into distinct classifications, for instance, arranging species in biological sciences. Due to the maturity of this discipline and commonality of taxonomy structures, there is a large body of work on calculating the similarity between concepts in a taxonomy tree.

In recent years for practical implementation of tree-based similarity algorithms, a topic hierarchy is primarily used (Chakrabarti et al. 2002, Pedersen et al. 2007, Patwardhan et al. 2003). The source for this taxonomy used in many of these publications is the DMOZ topic tree and the WordNet semantic lexicon.

Concept Distance Calculation

Both DMOZ and WordNet include an 'is-a' relationship tree between each concept. This means that all concepts are linked back to a common ancestor. For example let us take the word senses of programming and coffee. Taking a path through the DMOZ tree would give us:

Top/Computers/Programming
Top/Shopping/Food/Beverages/Coffee_and_Tea/Coffee

A limitation here with a taxonomy using only 'is-a' relationships is that one cannot state or infer that computer programmers are stereotyped as coffee drinkers, for instance.

To date there are number of different methods of measuring the semantic similarity of concepts in a topic hierarchy. A representative sample of these measures include:

- Simple edge counting scheme

- Resnik (Resnik 1995)
- Lin (Lin 1997)
- Jiang and Conrath (Jiang and Conrath 1997)
- Leaccock and Chodrow (Leacock and Chodorow 1998)
- Hirst and St-Onge (Hirst and St-Onge 1997)

The algorithms listed above can be loosely divided up into *information content approaches* and *edge-based (distance) approaches* (Jiang and Conrath 1997). Information content approaches, for example *Resnik* and *Jiang and Conrath* and *Lin*, concern themselves with the information that both compared concepts share in common. Edge based approaches are concerned with the geometric distance between concepts in the tree - this can be viewed a more natural way to compare concepts (Jiang and Conrath 1997).

2.7.2 Ontology Approach to Similarity

The area of ontology-based similarity determination is a relatively new area, but is based on existing schema/tree matching work. The main body of work currently relating to this area is in comparing and matching fragments of ontology graphs, not strictly measuring their similarity. A basic example of this in practical use is in the Semantic Web query languages, for instance SPARQL (Prud'hommeaux and Seaborne 2007), that formally matches portions of RDF graphs against each other and merges RDF graphs.

Ontologies are commonly matched using elements (Doan, Madhavan, Dhamankar, Domingos and Halevy 2003) and sub-graphs based upon formal semantics, for example exact matching, subsumption and intersection (Pan 2007). Matching can be described as the process of producing a mapping between two graph elements that correspond semantically (Giunchiglia and Shvaiko 2003). Semantic matching occurs at both the element level, as described above, but also applies to the structural level. An example of this occurring at both the structural level and element is provided by Embley, Xu and Ding (2004), who finds that up to 50 percent of matches between schemas are indirect, whereby elements in graphs are matched, but not necessarily in the same structural schema.

Existing work relating to ontology matching is based on schema matching deriving from the stable marriage problem, for instance, similarity flooding (Melnik, Garcia-Molina and Rahm 2002). To clarify, the *stable marriage* problem aims to best match couples from lists of men and women according to their preferences (Melnik et al. 2002). These techniques date back to the task of matching elements between databases. In another example, work by Qu et al. (2006) has applied schema matching techniques to ontologies. Qu et al. (2006) proposes the use of virtual documents and calculates a similarity value between these documents through the use of a vector space model. It should be noted at this point that the choice for concept similarity for that work was based upon string comparison of element labels. GLUE (Doan et al. 2003) uses machine learning to match elements between ontologies. When computing similarities, however, GLUE utilised existing taxonomy based similarity measures in order to calculate the similarity between concepts that the ontology elements represent.

As we introduced in **Section 2.2.3**, the Publish/Subscribe pattern relies upon semantic matching based on keywords and simple graph patterns. Complexity increases with graphs larger than simple RDF graphs, particularly when these graphs are cyclic. Consequently the time taken to perform the match increases accordingly (Liu et al. 2005).

2.7.3 Summary of Direction for Concept Similarity

For our use within this thesis, we will focus on the matching of single elements using our proposed context methods using a knowledge-base and building upon the work of the taxonomy research community. In our case, we will use the Wikipedia knowledge-base and develop a method of expanding out contextually similar concepts and measuring the semantic distance between concepts. In future work, however, we will endeavour to combine this element level matching with the work from the ontology community to expand the matching to cover the structure of elements within a directed graph.

2.8 Summary

As has been demonstrated in this chapter, there is a large body of existing work in the area of semantic relatedness based upon a topic tree or taxonomy. Despite the usefulness of taxonomies, they can be viewed as too rigid for social web use (Gruber 2005), but on the other hand, text is too weak to use for semantic information matching without corpora to utilise for statistical heuristic calculations. Tagging, on the other hand to the taxonomy, provides the agility that taxonomies lack (Hepp et al. 2007), but retains many of the weaknesses that keyword based matching carries.

Current thinking in web-based knowledge-base management has worked towards relating tagging folksonomies and wikis, based upon the grounding of taxonomy based similarity measures (Hepp et al. 2007, Strube and Ponzetto 2006). A place where this notion of word sense disambiguation through shared folksonomies (a controlled vocabulary) has occurred within Wikipedia page naming system, whereby, the title of a page will contain disambiguation information, for example *'Java (programming language)'*. This disambiguation occurs in an evolutionary manner, as multiple senses for a word emerge (Hepp et al. 2007). In the case that there are a number of senses for a word, the custom for Wikipedia authors is to create a disambiguation page.

The use of wiki technology to aid in knowledge-base information representation and ultimately contributing to information filtering will be explored and developed later in **Chapter 5**.

Chapter 3

Context and Information Filtering

In **Chapter 2**, we explored information retrieval and filtering as a way of retrieving information based upon a user's textual query. We also explored how the Web 2.0 pattern of retrieval through tagging allows a light-weight and simple way to filter resources on the WWW. For all the pragmatism of tagging, however, it does not take into consideration the user's information retrieval context and is limited to the textual keyword information present in the tags.

In this chapter we focus on two areas. The first area is to better understand the notion of context and to narrow this understanding down to a definition to be used for the remainder of this thesis. The second area is that of what it means to *understand the context* of the user's information retrieval or filtering task and how that context relates to wider world knowledge.

3.1 Context in Information Filtering

In this section we look at the application of the notion of context to improve the quality of information filtering. The scope of this thesis is that of filtering based on Web 2.0 style keywords (tags); this will inform the arguments presented within this chapter. To begin with we introduce what is meant by context and then look at how context is applied to information filtering and search.

3.1.1 What do we mean by Context?

Existing Definitions of Context

As a starting point for articulating the somewhat nebulous notion of context, we look towards the dictionary definitions of context in a general sense.

The Oxford English Dictionary (OED 2002) defines *context* in two ways:

> *“The circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood."*

and;

> *“The parts that immediately precede and follow a word or passage and clarify its meaning."*

These given definitions help us shed light on understanding the notion of context at a general level and help introduce the areas of context in relationship to the world. From this understanding one can generalise context as the relationship between a concept (or event) and its world.

Looking towards supporting definitions of context from literature, we extract some key points from a discussion of context presented by Dey (2001) described as:

> *“implicit situational information"*

and;

> a *“neighbourhood containing related or associated objects"*

Later in the paper, Dey (2001) proposes a definition of context:

> *“Context is any information that can be used to characterise the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves."*

From these points we can learn two things; the first is that context is provided through objects that surround a specified object in an implicit manner; secondly that context can be the situation that an object is currently in along with associated interactions. We can summarise this relationship in a simple diagram, as shown in **Figure 3.1**.

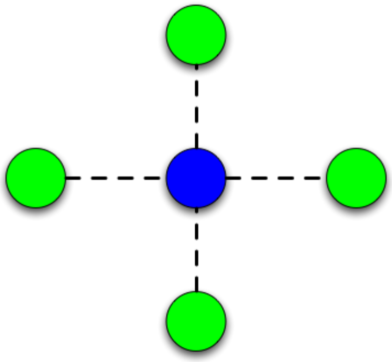


Figure 3.1: Inferred context between an object and its surrounding objects

Focusing on this we can see a similarity between the implicit surrounding objects and the method of tagging resources as described in **Section 2.4.1** called a tag cloud. Popular concepts relating to resources are determined at a point in time through the use of aggregation and ranking tags assigned to the resource by a social community. The tag cloud provides us with a collection of objects contextually related to the resource with a given rating of popularity. As a side-note, it is prudent to note that this popularity ranking will change through time.

To look at another approach to modelling context; the context model by Schmidt, Beigl and Gellersen (1999) is modelled as a set of statements, as follows:

- *A context describes a situation and the environment a device or user is in.*
- *A context is identified by a unique name.*
- *For each context a set of features is relevant.*
- *For each relevant feature a range of values is determined (implicit or explicit) by the context.*

Likewise Becker and Nicklas (2004) defines context as, *"the information which can be used to characterise the situation of an entity"* and a context aware application is described as so if it *"adapts its behaviour depending on the context."*.

Again, as with Dey (2001)'s definitions, we have a unique object with a surrounding group of objects or features. The second similarity of note is that of the environment that the object is within. It should be noted that in Schmidt et al. (1999)'s definition there is a notion of relevance is introduced. Rather than all surrounding objects to a defined (or unique) object contributing to its context equally, the context can come from a particular subset of surrounding objects and associated interactions.

To further our understanding, Gong (2005) presents two paradigms of context. These paradigms are *"in the context* **of** *X"* and *"the context* **for** *X"*. With the first paradigm we can constrain the events of concepts that we are talking about to the particular space of the object we are talking about. The second paradigm provides us with a *"collection of things and events externally related to"* the concept we are talking about. To illustrate these points diagrammatically, we provide the following illustrations of *"in the context of X"* and *"the context for X"* in **Figure 3.2** and **Figure 3.3** respectively:

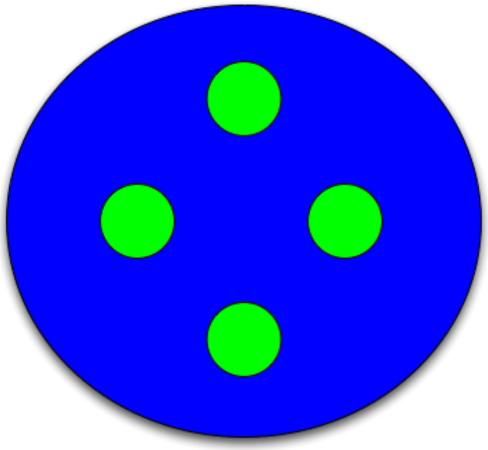


Figure 3.2: In the context **of** X

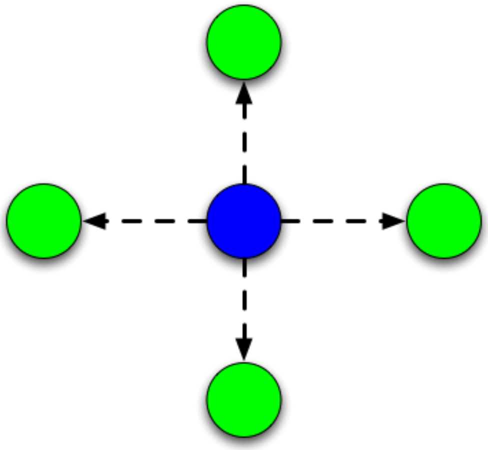


Figure 3.3: In the context **for** X

To summarise, Gong (2005) explicitly expresses the distinction between the environment that an object is placed within and the environment defined by (or within) the object.

3.1.2 A Definition of Context for this Thesis

Whilst it may never be possible to conclusively define context, an informed definition based upon existing consensus will be given for use during the remainder of this thesis in order to frame the domain of the concept of context. For the remainder of this thesis we define context as:

“the information which can be used to characterise the situation of a concept derived from a neighbourhood containing related or associated objects”

3.1.3 How does a Word relate to its World?

As with our definition of context, we can say that an object is related to the world, but furthermore we can say *how* an object relates to the world.

Regarding how humans convey ideas and react with each other Dey (2001) states, *“this is due to many factors: the richness of the language they share, the common understanding of how the world works, and an implicit understanding of everyday situations"*. The world knowledge is the key point to pick up on here which is again confirmed by Gabrilovich and Markovitch (2006), who stresses that world knowledge is important for tasks that require *“human level intelligence"*, for example human text classification.

As an example of how a word can link to its world, we look to artificial intelligence literature. *’Words and Worlds’* (Amsler 1987) discusses the notion of world knowledge in the context of artificial intelligence. He argues that humans have a shared knowledge of which part of that knowledge consists of knowledge of the outside world that the human has never experienced themselves.

To mention a related point proposed by Amsler (1987); he argues that there are many obstacles to acquiring world knowledge from text. The debate proposed in the work is whether this world knowledge needs to be hand coded and not derived from texts in order to be usable by machines. The work briefly discusses the large body of world knowledge that is present in printed literature and notes that a lexicon for machine understanding of such material is still not with us. This topic (whilst briefly introduced here) is further explored in **Section 5.4**.

3.2 Application of Context in Information Filtering

In recent years, the notion of context has been introduced into information search and taken seriously as a method of improving search results (Jones and Brown 2004). In the past, context and a user's search history has not been taken into consideration as part of information retrieval. This has been due to a simplification of the process (Ruthven 2004); this is also true in web search engines (Lawrence 2000) as they *"treat search requests in isolation"*. As a result, these traditional systems have been described as *"quite brittle"* (Gabrilovich and Markovitch 2006). Efforts towards improvement to the process are focussed towards better managing complexity (iri 2004) and reducing ambiguity (Murthy and Keerthi 1999) in information search.

As the focus for this thesis, we need to be able to alleviate information overload through the use of contextualised information filtering. Before focussing on information filtering, we first briefly look at information search, and in particular some attributes associated with context.

Kraft, Chang, Maghoul and Kumar (2006) defines contextual search as:

> *Contextual search refers to proactively capturing the information need of a user by automatically augmenting the user query with information extracted from the search context; for example, by using terms from the web page the user is currently browsing or a file the user is currently*

editing.

Points to highlight from this is that whilst traditional search does not explicitly define contextual information to aid in the search, there is information that can be inferred from the user's current search context, for example through past interactions with a web browser (Lieberman 1995) . This is useful, as it means that the user need not manually define a context for the search.

3.2.1 Contextual Information to Support Search

The current generation of research in contextual information (both explicit and implicit) search include a number of contextual factors relating to the user's search:

- Interests of the user (M. Marinilli and Sciarrone 1999)
- Past Searches (Kraft et al. 2006)
- Stereotypes of users (Shepherd, Watters and Marath 2002). Described by Ambrosini, Cirillo and Micarelli (1997) as *'A stereotype is a description of a prototypical user of a given class"*
- Community Based Popularity / Collaborative Filtering (Lerman 2007)

Each of these factors brings with it techniques of acquiring context from explicit and implicit sources of evidence (Belkin, Muresan and Zhang 2004, Lawrence 2000) which contribute to increasing the quality and personalisation of information filtering. In the following sections, we explore these topics closer and unpack their intentions.

Personalised Filtering

In this thesis we will focus on filtering based upon a personalised user profile; we focus on the interests of the user and expand out this profile using a world knowledge-base. For our understanding of this area, we need to look at existing work in the area of personalised filtering.

A body of work has gone into developing personalised information filtering agents (for example, news filtering) over the years based upon an existing user profile (Lang 1995, M. Marinilli and Sciarrone 1999, Murthy and Keerthi 1999). Ideally, a news filtering agent should block news items that a consumer is not interested in (Sebastiani 2002). A definition of personalised filtering from Yahoo Contextual Search (Kraft, Maghoul and Chang 2005) helps us understand the principle of personalised filtering and provides the following statement:

> *“A regular web search only uses your search term or "query" to find what you’re looking for. Y!Q uses both your query and information you highlight on a webpage to determine the "context" of your search. Using this "context," Y!Q provides exactly the type of information you’re looking for. A traditional search does not utilise context and isn’t a contextual search."*

The simplest method of personalised filtering is to use a list of words to represent the concepts that a user is interested in. This method of comparing keywords in a user profile to words present in a document is called the *‘bag of words’* (BOW) technique, with optional weights attached to the words (Sebastiani 2002). The BOW technique has the limitation that the word from the profile needs to be explicitly mentioned within the document. Additionally, this technique does not make use of world knowledge (Gabrilovich and Markovitch 2005).

A problem noted by (Lang 1995) is that a personalised news filtering system required not only the creation of a user profile representing their interests, but also the maintenance of that user profile over time. The technique he proposes to alleviate this problem is for a user to rate a news article on a rating of (1-5) and let the system learn from these ratings. A similar model is proposed by M. Marinilli and Sciarrone (1999), whereby a user model is built up through long-term interaction with the system. Whilst learning-based approaches to information filtering help reduce the user load on maintaining the interest profile, they do not eliminate the maintenance required by the user.

From this section, we have learned that a user's profile is a crucial element in order to perform personalised filtering. A decision relating to this profile concerns whether information is collected explicitly from the user or whether the user's interactions with the information system can generate background context data to be fed into the profile.

Historical Context

Relating to the above notion of maintaining a user's interests profile, historical context helps provide an augmentation to a user profile of the user's interests over time. As has been discussed, context data can come from interactions with a system over time. We present a short example here:

For the first example, in early agent work by Lieberman (1995) he states that the agent "*generates knowledge through a user's concrete actions through a web browser, rather than relying on a pre-existing knowledge-base.*" His tool (Letizia) generates a list of keywords from visited web pages. In that work, the user's browsing history is used to build up knowledge to help contextualise the information retrieval task.

A similar technique is also noted by Kraft et al. (2006) and Sebastiani (2002) who both propose the recording of a user's search history and usage patterns in order to generate a user profile as an aid in information retrieval. Whilst, we do not explicitly look at the user's search history in the framework developed for this thesis, instead modelling a user's profile explicitly, it is important to note that there is a body of work looking into this.

Information Retrieval as Part of a Larger Task

As discussed, the user's search history contains provenance information relating to that search history. Building on this, existing work factors in that a user's search queries will form part of a larger information retrieval task (Kraft et al. 2006). Whilst not focussing on a user's history in our framework, we appreciate that user's searches and filtering activities do not always occur in isolation. An area of contextual information retrieval is that of modelling the wider context of the user search (Ingwersen and Jarvelin 2004, Ruthven 2004). In other words; the search in context. Work by (Ingwersen and Jarvelin 2004) models information retrieval in the context of tasks. Tasks are broken down into job work tasks and non-job related tasks. This model can be illustrated in **Figure 3.4**. The work goes on to detail the socio-cultural background, such as goals, motivation and preferences behind the work task.

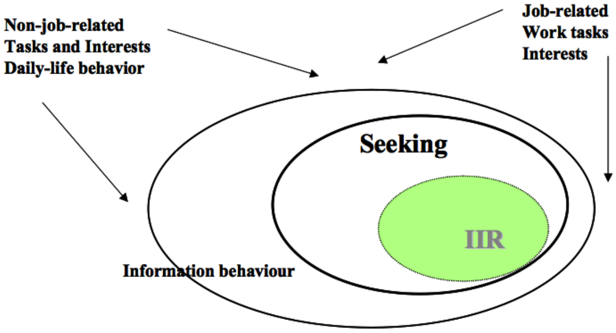


Figure 3.4: Information retrieval embedded in interactive information retrieval (Ingwersen and Jarvelin 2004).

We can see a correlation between this idea and Dey (2001)'s further definition of context in the domain of context-aware computing:

> *"A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task."*

Looking at the above examples in relation to the previously discussed *'Words and Worlds'* (Amsler 1987), a user will have a work task that they are currently engaged with. They will in addition possess a world knowledge that will likely inform or influence their work and searching pattern.

As Pitkow, Schutze, Cass, Cooley, Turnbull, Edmonds, Adar and Breuel (2002) discuss, the point of contextual computing is that it is not just about modelling the user behaviour, but rather the user's environment. This statement correlates well with the work by Ingwersen and Jarvelin (2004) and relates back to the definitions of context discussed in **Section 3.1.1**.

Community Based Filtering

As introduced in **Section 2.4.1**, community based filtering through the use of tagging (in the case of Web 2.0) can assist the web user in determining what resources are popular with the participating community. Through the use of this popularity data, we can apply context to the user's information search or filtering in order to augment it. One can see the connection here between this type of augmentation and the augmentation of a user's profile through the analysis of his or her search personal context. Lerman (2007) describes four characteristics that are shared by social media sites:

- *Users create or contribute content in various media types;*
- *Users annotate content with tags;*
- *Users evaluate content, either actively by voting or passively by using it;*
- *Users create social networks by designating other users with similar interests as contacts or friends.*

Contextual search can be viewed as moving us away from consensus relevancy to that which is more personal to the user (Pitkow et al. 2002). At this point it should be pointed out, however, that there appears to be a dissonance between the 'sheep' attitude of Web 2.0 users (based on the *'wisdom of crowds' (Lerman 2007)*) and that proposed by contextual search, which is finely personalised to the user.

Social media sites suffer from a *'short lifespan'* (Cho and Tomkins 2007), as contributors will actively tag content based on the current hot topic. A further complication to this point is that of search ranking based community values (Pitkow et al. 2002), which may influence (or differ from) that of the general population. As Cho and Tomkins (2007) describes, social media sites suffer from a *“locality of interest"*, whereby the contributions of users will have little significance to the general population. Gruhl, Meredith, Pieper, Cozzi and Dill (2006) describes this phenomenon as the creation of *‘icebergs’* around topics with only a minority of information above the *’popularity waterline’*.

Following on with this point, part of Google’s success has been due to the PageRank algorithm. With PageRank, web authors effectively vote for web sites through the creation of hyperlinks to them (Gruber 2005), therefore users are implicitly creating community data to influence all other users’ searches. A current problem with search engines is that they use a *’one size fits all’* (Lawrence 2000) model and each user gets presented with the same response to an identical query, which goes somewhat against the personalised proposed search solutions.

3.3 Summary and Direction for Contextual Filtering within this Thesis

In this thesis we focus on the explicit knowledge provided by the user to represent their interests in the form of concepts in a simple profile. We propose that through the use of a knowledge-base we can implicitly collect related concepts based on a user’s given concepts within that profile. This proposal is based upon existing research into mapping a user’s query to multiple formulations using semantic similarity (Pedersen et al. 2007). Through a community maintained knowledge-base we can combine explicit knowledge from the user’s profile and expand it through implicit knowledge from the knowledge-base and provide a personalised solution rather than a one-size-fits-all solution.

In the area of information retrieval we can say that *context* is used to expand information to enrich it in some meaningful way. This notion is not that different from providing *metadata* about information in an information environment in order to enrich its meaning. We can therefore make the connection between context and metadata and infer that metadata can be used to provide additional context for an entity. If context is used to provide additional information about an entity, we can propose that providing additional context can be used to strengthen the quality of knowledge as to that entity. Taking the word and world approach to context, one can see that having access to a world knowledge-base in order to relate the word (an object, etc) to the world will be of great use in context modelling.

We note at this point that the main bulk of research in contextual information retrieval operates with a model of information retrieval that compliments current search engines. Techniques such as query rewriting (Kraft et al. 2006) utilise a vector of context terms in addition to boolean semantics, which are augmented on top of current search engines. These techniques allow a degree of interaction between context modelling and current information retrieval engines. For our proposed research we abstract out to a generic solution that is not constrained to a specific search technology.

In the following chapter we focus on the issue or determining whether two concepts are the same as this is a relevant issue within information filtering. Not only do we want to say that two concepts are the same, but rather we want to say *how* similar the concepts are. In addition we separate out the concerns of variation referring to a distinct object (Berners-Lee et al. 2006), such as misspellings, against measuring the distance or distance between distinct objects. Summarising this point we refer the reader back to **Section 2.5** regarding approaches to determining the semantic relatedness between concepts.

Chapter 4

Initial Context Modelling and Practical Experience

This chapter details the initial modelling work on context for information filtering as part of this thesis' work. The section covers the progression of the modelling work, concluding with a final model that will be used as the basis of modelling for the remainder of this thesis. The motivation for this chapter is to document the 'lessons learned' from explorations into using a concept taxonomy as the basis for contextual concept distance measurement.

4.1 Building on Related Work in Context Modelling

As previously stated, context is a somewhat nebulous notion and, therefore, when surveying the area we discovered that existing work in context modelling suffers from overloading of the term *'context'*. For example a number of *'context models'* are domain specific and involve the modelling of specific classes of objects based on scenarios. Example domains include: spatio-temporal modelling geographic information systems (GIS) (Becker and Nicklas 2004); personal information management (PIM) in the mobile workspace (Tazari, Grimm and Finke 2003) - modelling a number of specific classes, such as documents, terminals, locations, users, times and tasks with dependencies between these aiding calendaring and address-book management. Finally in this list are context models for the integration of metadata with attributes such as reliability, precision, consistency and age (Honle, Kappeler, Nicklas, Schwarz and Grossmann 2005). As the authors state (for example Tazari et al. (2003)) these models become large very quickly with associated complexity problems.

We can summarise that these approaches to context modelling do not help us with being able to expand out a concept through the use of neighbouring concepts based on semantic relatedness derived from the Wikipedia knowledge-base - this is due to their domain specificity. Furthermore, the models surveyed do not feature a workflow for population or processing and suffer from a lack of ‘fit for purpose’ evaluation. For our context model, therefore, we pull from a number of modelling areas that are relevant to our generic domain of textual information filtering in order to develop a generic framework that can be applied to a range of disciplines.

There are a number of research papers that discuss context modelling in an abstract sense, but not in a manner that can be reused or implemented. There are elements of these models that help us to define our context model for use within this thesis; specific examples include:

- Dey (2001) who talks about a definition of context that involves surrounding concepts of an entity that are inferred.
- Schmidt et al. (1999)’s model includes the following properties: a set of relevant features; unique name.
- Gong (2005) discusses context ‘*for*’ X. As has been previously discussed, this paradigm provides us with a *“collection of things and events externally related to"* the thing we are talking about. This diagram is illustrated again here in **Figure 4.1**.

Knowledge gained from a study of these works has been helpful in the development

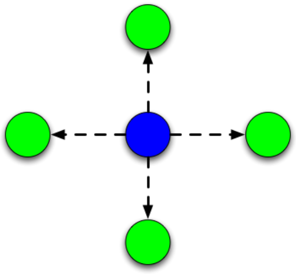


Figure 4.1: In the context **for** X

of our model's context cloud.

A family of related work that will help in defining our context model is the work on concept similarity through a concept taxonomy, as discussed in **Section 2.7.1**. Whilst this body of work does not explicitly describe itself as modelling *'context'*, it is nevertheless a useful and relevant body of work to draw upon. From these approaches, we can use associative reasoning to help us trace a path from one concept object to another; based on the *'shortest path heuristic'* (Brachman and Levesque 2004) whereby conclusions are preferred from shorter paths in the network. We can use the supposition to inform our context model development, whereby if concepts can be connected through a small number of edges, then they are in some way more strongly related than those with a larger (or no) edge path. In this manner, we can say that such concepts are contextually related in a stronger manner. Using a graph-based approach, we can calculate concept distances through the use of an Erdös number to determine the distance between nodes in a linked graph - this is expanded upon in **Section 5.7.2**.

4.1.1 Structural Modelling of Context

Based on our initial review of context and concept similarity, the initial design of our context model is intended to be used as a way of abstractly representing a concept and its related concepts in a simple form. In order to achieve this, the model needs to be able to represent the topology of concepts based on the wikipedia concept graph.

Another form of existing work in context modelling is through meta-data and ontologies. As previously discussed in **Section 3.3** the notion of context can be compared to metadata, whereby it is used to expand information to enrich it in some meaningful way. We can therefore make the connection between context and metadata and infer that metadata can be used to provide additional context for an entity. As previously discussed taking the word and world approach to context, one can see that having access to a world knowledge-base in order to relate the word (concept) to the world will be of great use in context modelling.

Dublin Core

A common example of RDF usage on the WWW is the Dublin Core RDF schema (DCMI 2008). Dublin Core consists of the the following elements:

contributor	coverage	creator
date	description	format
identifier	language	publisher
relation	rights	source
subject	title	type

Table 4.1: Dublin Core elements.

The aim of Dublin Core is to provide a simple way of describing web resources. Whilst the model does provide the relations shown in **Table 4.1**, many of these are specific to the published nature of web resources, for example date, format and rights. This model, however, provides a topology of a focus resource and a ring of surrounding resources that are bound to it via a number of pre-defined predicates. Whilst these relations do provide a useful starting point for contextual relatedness, we are more concerned with whether concepts *"are"* related, not limited by any one particular schema. Therefore, we utilise the topological structure of Dublin

Core without being restricted by its domain specific set of predicates.

4.1.2 Context Modelling - Relaxing the Semantics

In our proposed context model, we maintain the linked structure of the Dublin Core model, however, we relax the relation semantics to not be restricted to a single schema. We propose an initial model to describe the context of a focus concept item through what we call the *“Focus–>Predicate->Object"* (FPO) model. As indicated by the name, the model includes a focus concept surrounded by objects, connected through predicates. In an abstract sense, this model is an implementation of a simple single layer RDF graph as applied to a focus object. A visual representation of this model can be seen in **Figure 4.2**.

The FPO model uses a similar basis to RDF, but is intended to be abstract and, therefore does not require a predefined vocabulary for the predicates. A Dublin Core metadata description of resource, therefore, would easily be mappable to the FPO model. In FPO, the outer ring of concepts linked through the predicates is named as the contextual cloud.

The importance in the FPO model is the concentric rings of concepts around the focus concept. We name this the *‘concept cloud’*. It is these concepts that we are interested in, as these will be mined from a wiki knowledge-base when used loosely and contextually by the editorial community, rather than specified from the limited predicate set of Dublin Core.

To give an example of this model, represented in a rather simple arbitrary notation, we can describe an object in relation to it’s world, for instance:

Scarborough (locatedIn) UK
Scarborough (visitedBy) Tourists
Scarborough (population) 500,000

As can be demonstrated by this model, simple semantics can be presented whereby

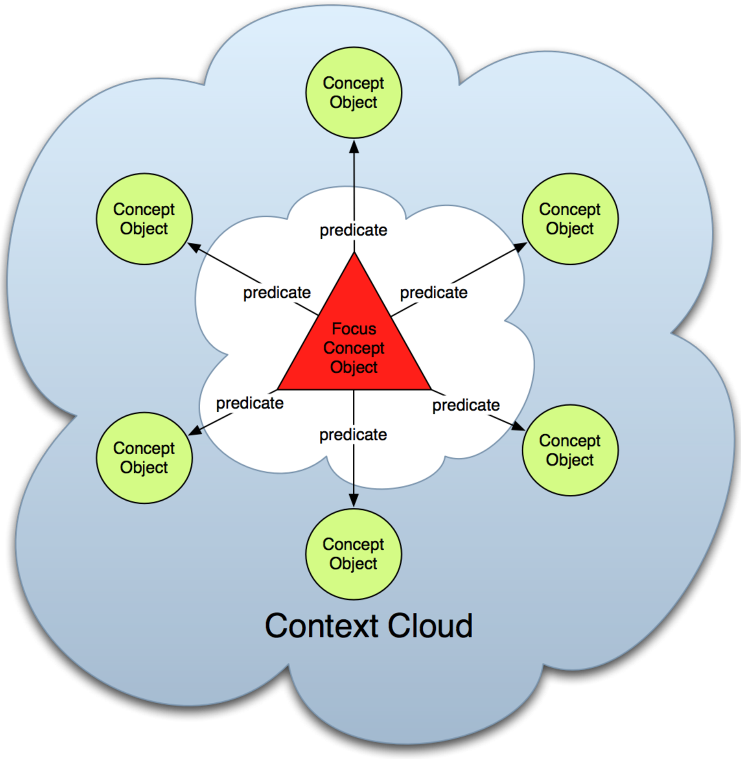


Figure 4.2: FPO model

concepts can be attached to a focus concept along with a single predicate. One limitation, however, is that the only predicate semantics that can be applied exist between the focus and the concepts in the cloud due to the graph being non-cyclic.

4.1.3 Extended FPO and Associated Problems

In the extended FPO model, the pattern of chaining together predicates and concept objects is applied. This chaining is achieved through the use of a preposition to denote context between the two attached concepts. An advantage of using such an extended model is of gaining the ability to express richer semantics through attaching prepositions. An exemplar is expressed below as before:

Scarborough (locatedIn) UK (in) Europe
Scarborough (locatedIn) UK (near) Sea
Scarborough (visitedBy) Tourists (in) Summer

A visual representation of this model can be seen in **Figure 4.3**.

This work has exposed the issue of extending the number of types of predicate types (for example prepositions) to expand on the semantics and has highlighted where the complexity threshold lies for the model's development for this thesis. As discussed in **Section 2.7.2** when working towards a method of determining the similarity of concepts, then we have structural issues to consider in addition to semantics attached to the graph edges through the use of predicates. As has been discussed in **Section 2.3.2**, the processing of RDF Semantic Web data requires the use of a logic, for instance predicate logic (aka FOL). FOL is found to not scale too well over large quantities of data, which will prove problematic with the quantity of data that will be present in a community maintained world knowledge-base. Unfortunately, with this type of knowledge-base there will be contradictory information and inconsistencies present that will limit the scalability of using this type of approach to semantic data processing. As Brachman and Levesque (2004) states, *"no*

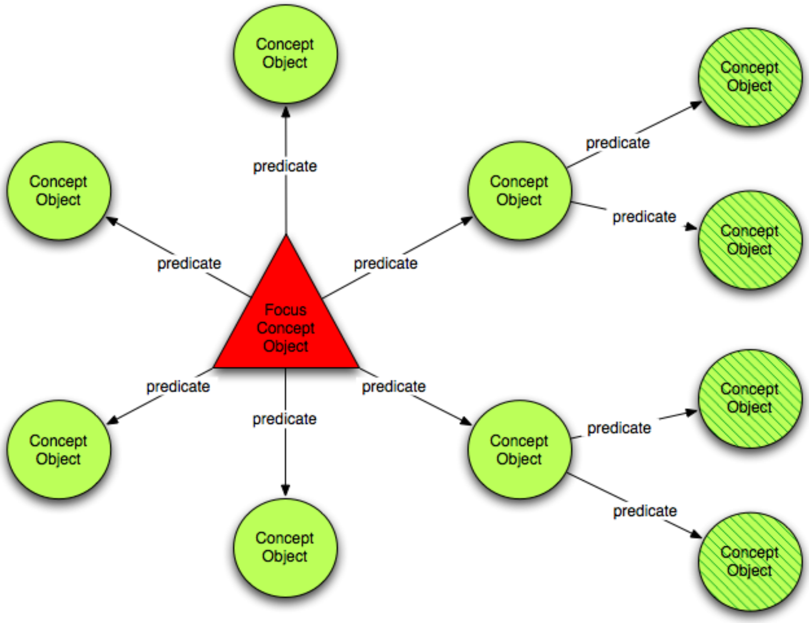


Figure 4.3: Extended FPO model

automated reasoning process for FOL can be both sound and complete in general".

A problem that is compounded by the Extended FPO model is that of aligning concepts when comparing two models. This occurs when attempting to determine the similarity between a pair of focus concepts. With the basic model, the sets of concept objects in the clouds have to be compared and aligned so that each concept is assigned to the best match. With the extended model, the problem is exacerbated, whereby the concept from a level 1 cloud may bind most closely with a concept from a level 2 cloud. In that case there exist problems with processing the semantics of the chains in a simple manner.

This alignment problem can be illustrated in **Figure 4.4**. In this example there is an uncertainty between mapping the objects within the chains. Whilst all concepts from the first model are present in the correct levels in the second model, the semantics of the chains are different.

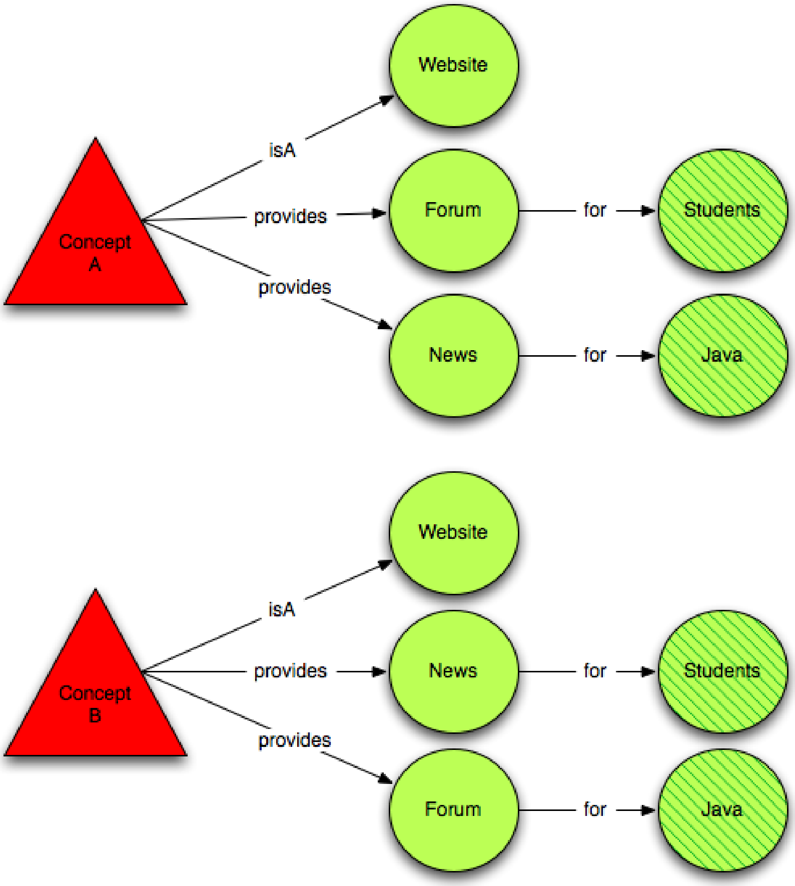


Figure 4.4: Model Alignment Problem

4.2 Summary of the Initial Context Model - F-O Model

Taking into consideration the motivation behind and scope of context modelling within this thesis, a further simplification towards the initial FPO model is made. As with the problems associated with the extended model, the explicit constraints imposed by the two layers of predicates and concepts can cause complications when aligning non-homogeneous semantics. It should be noted, however, that even at the level of a single layer in a contextual cloud, there may be unnecessary rigour imposed by forcing a child concept to be bound to a focus concept through an explicit predicate.

As has been demonstrated, it is possible to use predicates as a vehicle to relate the focus concept to the world using a contextual concept cloud. This approach can be taken further - the contextual concept cloud is used as a vehicle to relate the focus concept with the world associated with the focus concept. Predicates are no longer modelled, only an unnamed connection between concepts will be used. To describe this in another way, the concept becomes the semantic predicate in a chain of concepts.

For this current research, the simpler proposed F-O model will be used in order to demonstrate the basic modelling and contextual comparison framework without introducing unnecessary complexity through structural modelling. In this thesis, for the methods that we propose, we are focussing on dealing with the semantic distances of concepts, so we concern ourselves with the concepts that are associated with a focus concept and not at this stage with the semantics attaching them to the focus.

As previously discussed, one of Wikipedia's strengths is in that it has a very simple TBox model with relaxed semantics that are inferred through the user's reading context. In addition to this, the rich concept connectivity comes through the ability to link key terms/concepts in the page text of a concept to other concepts in the Wikipedia knowledge-base. A Wikipedia concept page is typically linked against multiple related concept pages. Regarding a focus for this thesis, the stated differences between an ontology and a knowledge-base help us to focus on the knowledge-base over just the formal ontology as a source of contextual information to aid in information filtering in order to alleviate the information overload problem.

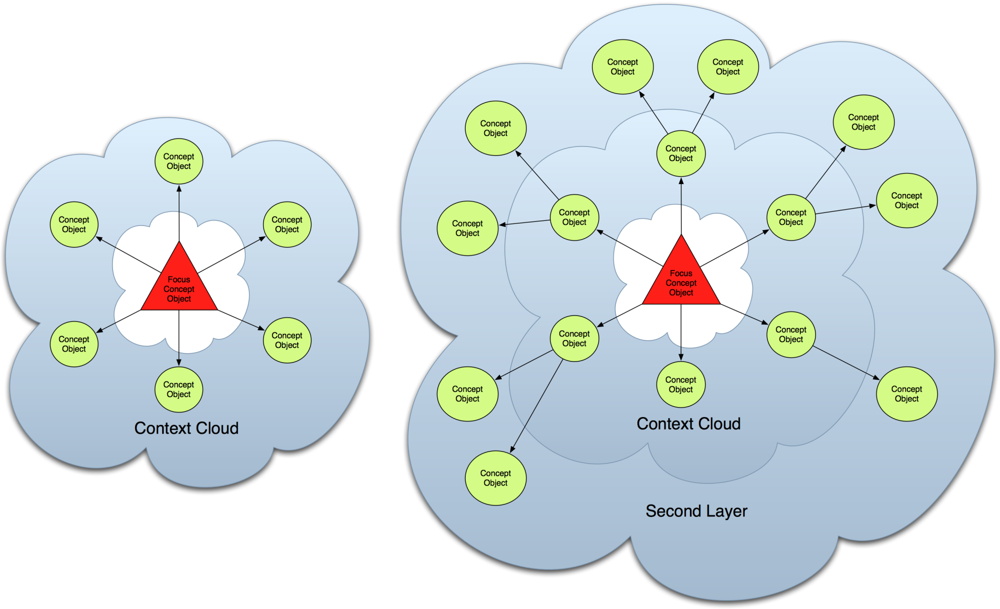


Figure 4.5: Focus-Object (F-O) model

Figure 4.5 (left) illustrates the basic structure with relaxed semantics of the contextual concept cloud. **Figure 4.5 (right)** illustrates the expansion of the F-O model to include multiple depth layers of the contextual concept cloud.

A limitation of our proposed context model is that we do not take into consideration the user's work tasks that the user's search context is related to as suggested by a number of authors (Kraft et al. 2006, Ingwersen and Jarvelin 2004, Ruthven 2004). This is acknowledged in this thesis as a possible expansion for future work.

4.3 Experience in Using a Taxonomy-based Knowledge-Base

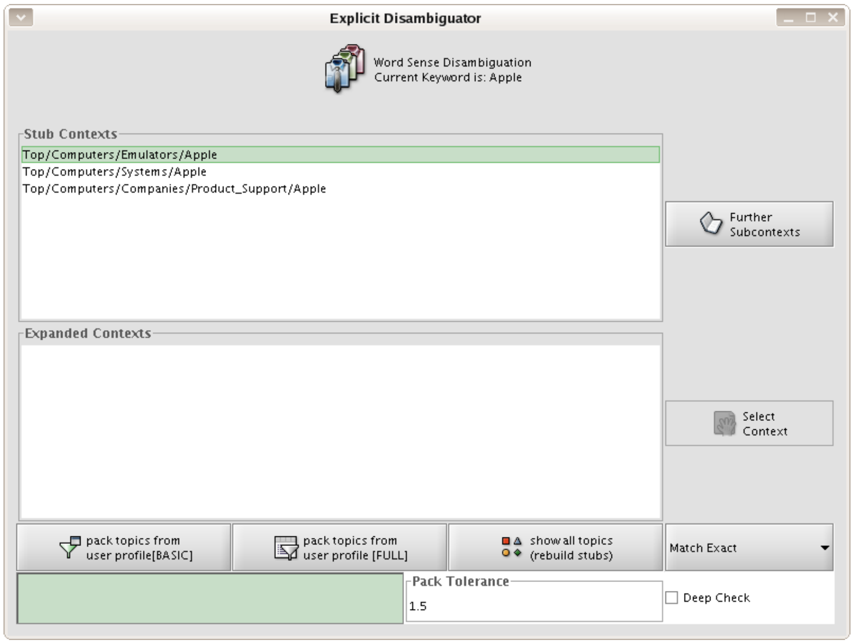
In this section we will discuss the exploratory practical work that had been conducted when working with a topic taxonomy as a knowledge-base to aid concept similarity determination.

In order to begin to compare concepts, keywords from a human input need to be mapped to a concept from within the taxonomy. At this stage we assume that a user will start to identify a concept by entering a keyword symbol for the concept in a manner relating back to the Ogden Meaning Triangle introduced in **Section 2.5.1**. Ideally, by explicitly identifying the word sense of a word, the comparison system does not have to determine the word sense of the user's annotation or query. A problem that still remains is that of being able to communicate this with the user. It is unlikely that a user will want to search a taxonomy for the word sense by hand, as this is a time-consuming process, and enter the path by hand for every noun (concept) in their query.

Taking the taxonomy approach, a simple software application called *ExplicitDisambiguator* was developed as part of the early development work within this thesis. In order to demonstrate the issue of word sense disambiguation from a user query we have developed a graphical interface between the user and the DMOZ taxonomy knowledge-base in *ExplicitDisambiguator*. The following interface can be shown below in **Figure 4.6**.

As can be seen in **Figure 4.6**, the user has initially entered a keyword of *Apple*. This keyword has produced three matches from the DMOZ knowledge-base. The matching between the keyword and the knowledge-base is achieved by matching the keyword to the labels of nodes in the tree. The path through the tree in relation to this match is represented in the software tool as a path of words and slashes (**'/'**) as delimiters. For this system we define currently three keyword match cases:

1) Match the keyword to the DMOZ word exactly.



Explicit Disambiguator

Word Sense Disambiguation
Current Keyword is: Apple

Stub Contexts

Top/Computers/Emulators/Apple
Top/Computers/Systems/Apple
Top/Computers/Companies/Product_Support/Apple

Expanded Contexts

Pack Tolerance: 1.5

Match Exact

□ Deep Check

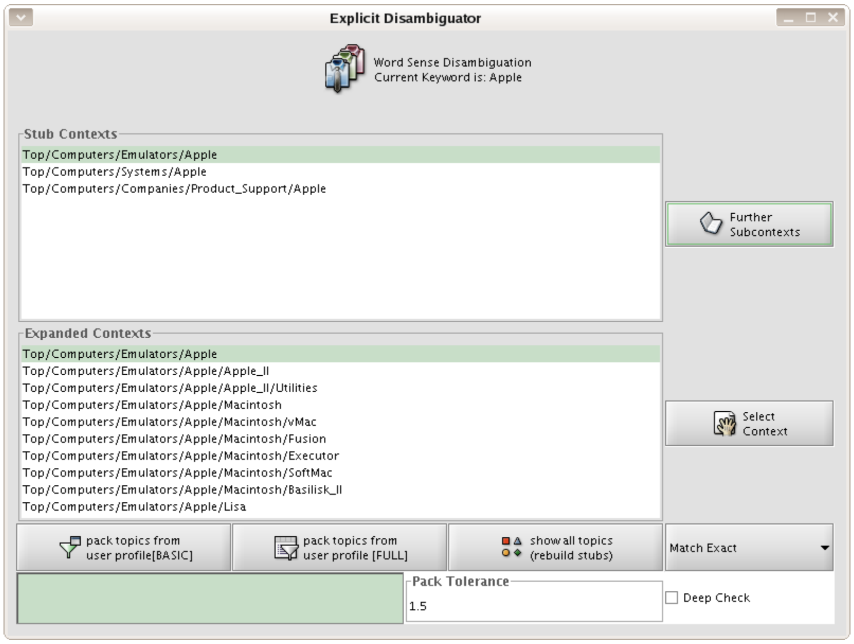
Further Subcontexts
Select Context
pack topics from user profile[BASIC]
pack topics from user profile [FULL]
show all topics (rebuild stubs)

Figure 4.6: Chosing a Word Sense Stub

2) Match the keyword to the beginning of the DMOZ word.
3) Match the keyword anywhere in the DMOZ word.

Once the keyword has been matched to a node in the DMOZ tree, a stub is then stored. Any deeper nodes from that stub node are spidered and stored.

The user will then select one of these word senses and then click the *'Further Subcontexts'* button. This will expand the stubbed node into the lower list by populating it with the spidered data, see **Figure 4.7**. Upon selecting a word sense from the lower list and clicking the *'Select Context'* button, the user will then be presented with a dialogue box showing the word sense and the path to that word sense, see **Figure 4.8**.


Explicit Disambiguator
Further Subcontexts
Select Context
pack topics from user profile[BASIC]
pack topics from user profile [FULL]
show all topics (rebuild stubs)

Word Sense Disambiguation
Current Keyword is: Apple

Stub Contexts

Top/Computers/Emulators/Apple
Top/Computers/Systems/Apple
Top/Computers/Companies/Product_Support/Apple

Expanded Contexts

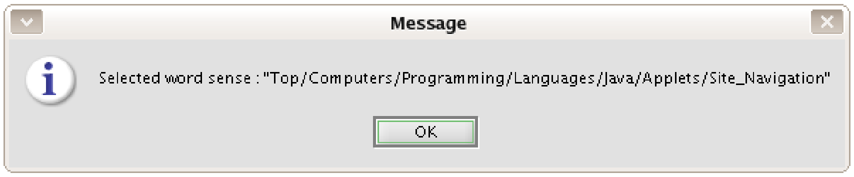
Top/Computers/Emulators/Apple
Top/Computers/Emulators/Apple/Apple_II
Top/Computers/Emulators/Apple/Apple_II/Utilities
Top/Computers/Emulators/Apple/Macintosh
Top/Computers/Emulators/Apple/Macintosh/vMac
Top/Computers/Emulators/Apple/Macintosh/Fusion
Top/Computers/Emulators/Apple/Macintosh/Executor
Top/Computers/Emulators/Apple/Macintosh/SoftMac
Top/Computers/Emulators/Apple/Macintosh/Basilisk_II
Top/Computers/Emulators/Apple/Lisa

Match Exact

Pack Tolerance: 1.5

□ Deep Check

Figure 4.7: Chosing a Word Sense



Selected word sense : "Top/Computers/Programming/Languages/Java/Applets/Site_Navigation"

Figure 4.8: Selected an Expanded Word Sense

4.3.1 Contextualisation of the Path Between Concepts in a Tree

In this section we propose the notion of a context anchor in a topic tree to influence the semantic similarity of concepts. The context anchor was planned to be used with the tree-based similarity measures in order to influence their semantics. The context anchor forces a particular concept to influence the similarity calculation. For example, one could say, "what is the distance between **Coffee** and **Tea**, in the context of **Drink**?" In this case, the contextual distance would be expected to be low, as both coffee and tea will likely share a close least common subsumer node, for example as **Drink**. However, if one asked, "what is the distance between **Coffee** and **Tea**, in the context of **Pens**?", then the lowest common subsumer would be set to **Pens**. This would mean that the distance traversed would most likely be greater than before, meaning that the distance is likely to be greater then the **Drink** context. If one were to think about this, tea and coffee have more in common to beverages than to pens. We illustrate these two context anchors in **Figure 4.9** and **Figure 4.10** .

In order to practically achieve this calculation, two calculations can be used. If we call our concepts **A** and **B** and our context anchor **C**, then we perform two calculations as follows:

A –> C then **C –> B**

4.3.2 Problems Encountered with the Taxonomy Approach

As can be demonstrated through **Figures 4.6** and **4.7** the approach of the Explicit-Disambiguator prototype, the keyword *'Apple'* produces only three stub contexts. Entering the keyword, *'Java'*, however, produces tens of results and illustrates one of the problems when relying on a concept taxonomy tree approach.

By observing the stub results from **Figure 4.6** an issue with using a taxonomy can be seen. The word sense of *'Apple'*, meaning *'Apple Computers'* can be seen to be

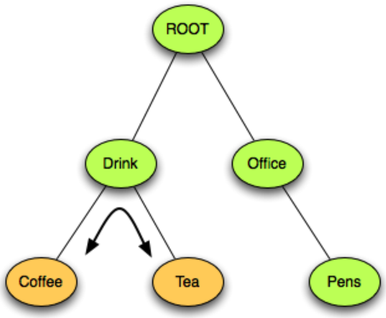


Figure 4.9: Shortest Path between Tea and Coffee

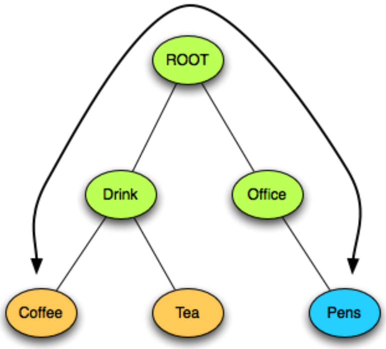


Figure 4.10: Shortest Path between Tea and Coffee with a Context Anchor of Pens

in three places in the taxonomy. This situation is much worse in the *'Java'* example due to the number of positions in the tree where this one word sense is placed.

A further problem is where the word sense is placed in the taxonomy. For example, taking the two word senses of *Camino* and *Firefox*; both of these are web browsers. Camino is a Mozilla-based browser designed for the Apple Macintosh platform, whereas Firefox is a Mozilla-based browser for Linux, Windows, and Macintosh platforms. As a human, intuitively we could see both these word senses being located close to each other in the tree as they both inherit from Mozilla or a web browser concept. This is not the case in DMOZ, however, as Camino is filed under Apple Computers and Firefox is filed under generic WWW Browsers branch.

This case illustrates the problem inherent with a rigid taxonomy. The problem is related to the decision on placement of items within such a static structure as a taxonomy. Depending on authority and ownership of a structure, the view of such information will follow a single perspective on the information, which may or may not be representative of an agreed world opinion. As has been discussed previously within this thesis, there is a tradeoff between using an emergent structure, for example a tag cloud and a controlled engineered structure, for instance a taxonomy.

Chapter 5

Wiki-based Knowledge Processing

Following the discussion of concept similarity and the tradeoff between dynamic evolving social knowledge structures and controlled engineered knowledge structures, this chapter investigates a structure that sits between the two extremes. A structure that we investigate to fit this position is the wiki that we introduced briefly earlier in this thesis as a knowledge-base structure that has become popular in recent years amongst the web community.

This chapter revisits the place of the wiki pattern in the information space of the web, specifically the Wikipedia implementation. This chapter proposes the wiki as a potential bridge between the text-oriented WWW and the data-based Semantic Web as a candidate for a world knowledge-base following a qualification of Wikipedia as a suitable knowledge-base. This knowledge-base will be used as the basis for our proposed contextual concept expansion framework and will be used to populate the proposed context cloud model.

The next part of this chapter is to explore and verify the use of Wikipedia as an evolving community maintained knowledge-base as one that is suitable to base the proposed concept expansion and comparison framework upon for contextual information filtering. From this investigation we discuss how concept associations can be used as a basis for contextual comparisons to aid in filtering. As introduced in

Section 2.4.1, wikis provide a consistent method for web users to collaboratively build up a text-based knowledge-base with relatively low entry requirements and little knowledge as to the underlying technology.

The experiment detailed in this chapter will be used to highlight and reinforce issues related to the information overload problem detailed in **Section 2.1.1** by providing a workflow process to apply the proposed knowledge-base enhanced contextual filtering technique. A scenario is proposed with scope to allow information overload issues to be observed and explored. This scenario is presented based around a real-world case study and vignette which will be expanded upon in **Chapter 6**. We have chosen to use web feeds as an application basis for this scenario. Web feeds and web feed aggregation lend themselves dually as a source domain for an information filtering application and as a means of populating our application model.

5.1 Problem Scenario

One example of information overload, introduced earlier in this thesis, is web feed news syndication. To better explain this problem in the context of web feeds; if, for example, five web news feeds are subscribed to by a user, with each news feed producing ten news items per day (on average), then the user will have to filter through fifty news items in total per day. Depending on the user, this number may (or may not) be manageable. As described by (Liu et al. 2005), *“a user does not have the resources to monitor large number of feeds and hence the user can easily miss information of interest"*. This is a problem that we attempt to work towards a solution to.

It is at this point that the motivations behind searching for items of interest in web feeds should be briefly discussed as web feeds can be used in a variety of domains. On one hand, the user may be a casual reader, who just wants to gain a flavour of the current news of the day and may just scan over the list of items. On the other hand, however, the reader may be a system administrator looking for security vulnerability information updates from software system vendors published over a web feed. For the system administrator, items may have an associated cost of failure attached with missing information contained within them.

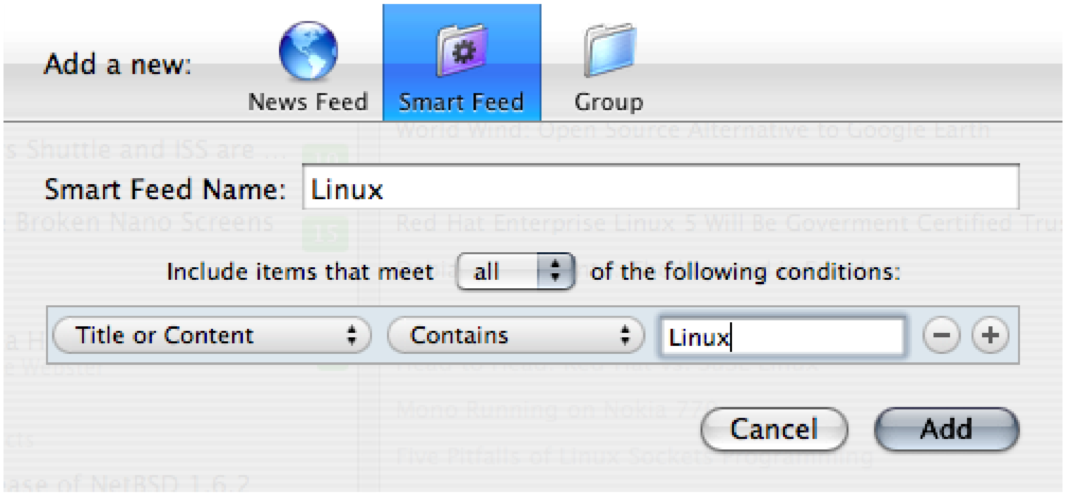
It should be noted that for the first example, if the user only wants to scan through ten news items per day, the web feed aggregator can present to him the ten most recent. This is not a difficult task and we can assume that a loss of information will not pose a problem for the user. However, another (and less trivial) approach would be for the aggregator to present the ten most relevant items for the day. This case is what the thesis' research contribution builds towards.

This scenario relates back to the conceptual difference between 'information retrieval' and 'information filtering' explored in **Section 2.2**. From a user's perspective, he may be searching all items that 'are' relevant to his needs or of interest. Secondly, he may be interested in filtering only the items that are of 'most' interest from the item set.

For all of web feed technology's convenience, the need can be seen for intelligent filtering in web feed aggregation. For our proposed framework, we derive this intelligence from the Wikipedia knowledge-base.

5.2 Solution Motivation

As has been noted in the previous section, the need can be seen for intelligent filtering in web feed news aggregation. Existing web feed aggregation software goes some way towards alleviating information overload by allowing users to perform a number of basic filtering operations in order to aid in filtering feeds based on set criteria or through manual user specification. A typical example of an operation involves a user manually creating categories within the aggregator and adding news feeds to them on a per feed level. Recently news aggregators (for instance *NewsFire* and *NetNewsWire*) have used the idea of 'smart lists', whereby users can create categories based upon criteria and keywords they choose; these are then processed against the content of incoming news items. An example of an application demonstrating this type of filtering can be seen in **Figures 5.1 and 5.2**. At an abstract level, one can easily find parallels between this filtering approach and setting up an email filter or virtual folder. During the course of this thesis, we produced the poster paper at WWW06 as part of this thesis' work (Webster et al. 2006). This poster outlined a high-level framework for web filtering that builds upon this idea of Publish/Subscribe previously detailed in **Section 2.2.3**.

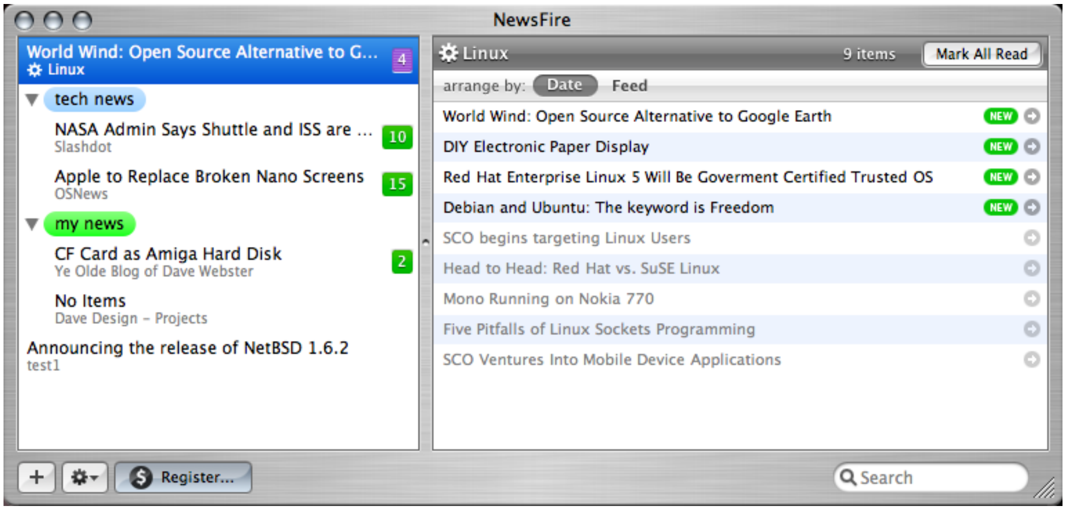


Smart Feed Name: Linux

Include items that meet all of the following conditions:

Title or Content Contains Linux

Figure 5.1: Creating a smart category 'Linux' in NewsFire news aggregator.



World Wind: Open Source Alternative to G...
Linux

tech news

NASA Admin Says Shuttle and ISS are ...
Slashdot

Apple to Replace Broken Nano Screens
OSNews

my news

CF Card as Amiga Hard Disk
Ye Olde Blog of Dave Webster

No Items
Dave Design – Projects

Announcing the release of NetBSD 1.6.2
test1

Linux

World Wind: Open Source Alternative to Google Earth
DIY Electronic Paper Display
Red Hat Enterprise Linux 5 Will Be Goverment Certified Trusted OS
Debian and Ubuntu: The keyword is Freedom
SCO begins targeting Linux Users
Head to Head: Red Hat vs. SuSE Linux
Mono Running on Nokia 770
Five Pitfalls of Linux Sockets Programming
SCO Ventures Into Mobile Device Applications

Figure 5.2: NewsFire news aggregator demonstrating categories and a smart category 'Linux'

Whilst the approach of simple keyword matching will work for manually specified keywords in filtering, such a simple matching system will not match contextually-related topic keywords. Whilst it is possible to perform ad-hoc operations, for example removing a 's' from the end of words to words that are plural, such as 'dogs', to aid in matching, there is still the problem that only those keywords specified by the user will be matched. An example of this problem is one where a user will set up a filter to capture any news items with the title containing the concept keyword 'videogames' because that is their interest. This search, however, will not return relevant news items such as 'Nintendo releases Zelda for the Wii'. In this case, all words 'Nintendo', 'Zelda' and 'Wii' are all brand names related to videogames. The question to ask is: how likely are news items related to a topic/concept going to contain the keyword of the topic?

This method of filtering news items can be described as the *'bag of words'* (BOW) technique, as discussed in **Section 3.2.1**. The BOW technique has the limitation that the keyword from the user interest profile needs to be explicitly mentioned within the document. This technique does not make use of world knowledge (Gabrilovich and Markovitch 2005). As Lang (1995) noted, a personalised news filtering system requires not only the creation of a user profile representing their interests, but also the maintenance of that user profile over time through interaction with the user (M. Marinilli and Sciarrone 1999). Whilst learning-based approaches to information filtering help reduce the user load on maintaining the interest profile, they do not eliminate the maintenance required by the user. A problem with these approaches is that only internal knowledge is used, gained through interactions between users and the information filtering system.

5.2.1 Tagging Content

Community derived world knowledge (for example as collected through social media sites) posses knowledge that can be used to help us to enhance a user's interest profile. This knowledge can come through a number of forms - for instance users annotating content with tags (Lerman 2007). In the case of web feeds and websites that the feeds are sourced from, it is common for news items to be annotated through tags. It is worth noting that during the time of early publication from this thesis (Huang and Webster 2004), topic categorisation was scarcely utilised by web feed content providers, but is now increasingly used by web feed capable blogging software, such as *Wordpress*. More recent web feed formats (for instance, RSS 2.0 and ATOM 1.0) allow the publisher to include multiple category tags for each web feed item and for the web feed itself; this is a feature supported by modern blogging software.

Tagging Stability and Human Judgement

Experimental work by Golder and Huberman (2006) discovered that there were regularities in user activity regarding tagging. One of these regularities was the stability of tag proportions for a given resource against the speculative initial assumption that over time a chaotic pattern of tags would occur. Hassan-Montero and Herrero-Solana (2006) discussed the point that in community tagging a user will be simultaneously tagging a resource for both their own use and for the community.

Halpin, Robu and Shepherd (2007) describes the tag collection and their frequencies rank order for a given resource as a *tag distribution*. A *stable* tag distribution is one where a resource is represented with a group of tags that maintains its frequency despite retagging by new users over time. Experimentation by Helpin, et al. found that collaborative tagging system distributions stabilise according to a power law.

As Golder and Huberman state, in the case of the Del.icio.us bookmark case study *"combined tags of many users' bookmarks give rise to a stable pattern in which the proportions of each tag are nearly fixed"*. After a small number of users tagging resources a *"nascent consensus seems to form, one that is not affected by the addition of further tags"*. One feature of Del.icio.us that contributes towards this stability is that when tagging resources users are presented with existing tags for a resource when tagging it for themselves. One benefit of this approach is that users are discouraged from entering variant or misspelling of tags. The approach goes some way towards validating the assumption that users of tagging systems correctly tag resources, as there is nothing preventing a user from disagreeing with existing tags and tagging a resource with other tags. As Golder and Huberman note, one reason for this is that ideas and characteristics that are represented in tags are stable and minority opinions do not disrupt a stable consensus.

There is a point to note regarding the case study in **Chapter 6** of tagging resources in the OSNews technology news website. Whilst massively collaborative tagging systems have been discussed, the OSNews site is controlled by a group of internal editors. This allows for domain experts to tag/categorise news items based on a defined set of category tags that evolves over time. This is more controlled than ad-hoc tagging and limits the ad-hoc creation of tags.

Human Judgement in Word Pair Similarity

Finkelstein, Gabrilovich, Matias, Rivlin, Solan, Wolfman and Ruppin (2002) produced the WordSimilarity-353 Test Collection dataset of english word pairs with human-assigned similarity judgements. **Figure 5.3**, illustrates the WordSimilarity-353 dataset of mean average word similarity judgements per word pair. For this thesis, the standard deviation from the data set has been plotted to demonstrate the variance in human judgements per word pair, plotted here as a sample. Whilst only briefly analysed here, the plotted standard deviation (illustrated as error bars) demonstrates that there is a consistency across the user participant set in assigning similarities to words and associated concepts.

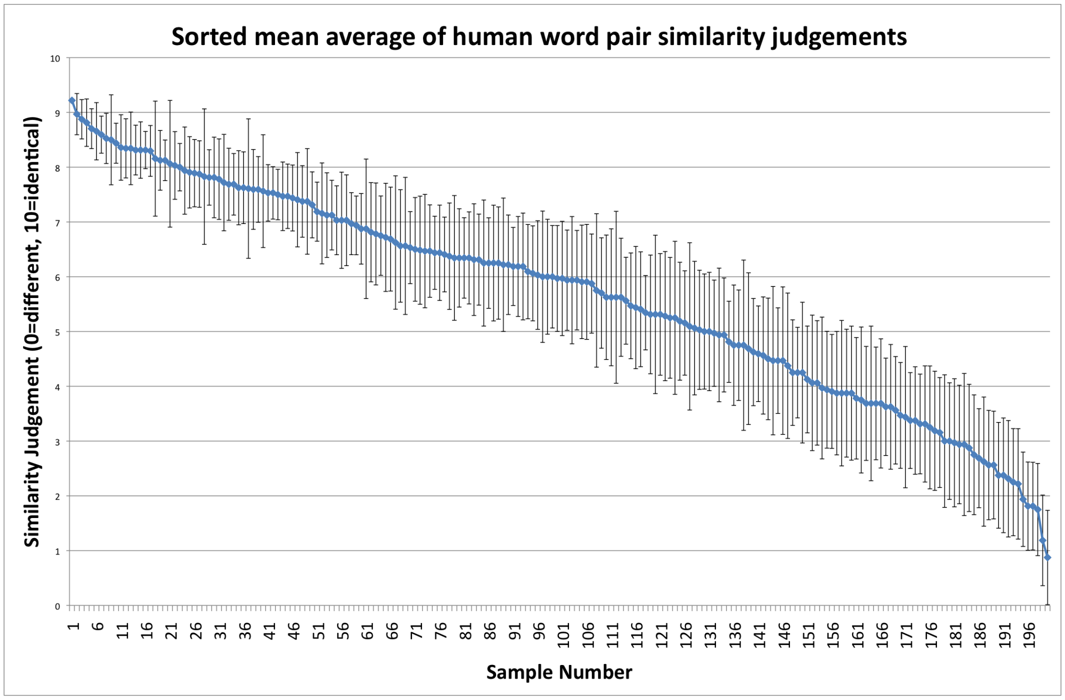


Figure 5.3: Summary of human judgements mean average from WordSimilarity-353 word similarity set with standard deviation plotted

5.2.2 Towards a Solution

In this thesis we propose to combine both explicit knowledge provided by the user in a profile of interests with knowledge gained from a community maintained knowledge-base containing world knowledge. We propose that through the use of a knowledge-base we can implicitly collect related concepts based on a user's given concept. Through a community maintained knowledge-base we can combine explicit knowledge and expand it through implicit knowledge to provide an improved solution to the BOW technique without ongoing maintenance of the user interests profile by the user. As Lang (1995) states, a user will not be willing to spend the time creating the rules required for filtering of the required complexity. Again, we run into the *'barrier to entry'* problem when these approaches present complexity to the user. The proposed framework will minimise this effort.

5.3 Solution Architecture

At this point we present a concise illustrative model as to where this thesis's research contribution will fit into an information filtering process. The overall architecture that enables use to understand the position of our contextual filtering contribution of the proposed framework for this thesis is presented in **Figure 5.4**.

Our framework accepts in a set of terms representing the information item being filtered and then matches them against the contextually related words to filter the items relevant to the user's context. To break this down, our contribution takes in a keyword of interest to the user and expands these keywords based on a knowledge-base. This part of the process is enclosed in a dashed line in **Figure 5.4**.

The expanded keywords are then fed into a simple keyword matcher, whereby the given keywords are matched against textual keyword representations of information items. Out of this matcher are produced the filtered informational items that are contextually relevant to the user.

It is important to note that for this thesis' contribution, the task of extracting keywords is purposely delegated beyond the scope of this investigation, as this is not where the thesis contribution lies. There is much ongoing research in the area of keyword extraction from textual sources, so the framework can potentially be used with any external algorithm that is plugged in. For this experiment we shall simply extract out the words of web feed item titles as a representation of their content. From a human usability perspective, a human scanning a list of items (be they web feeds or emails) will commonly scan the title of the item to determine its relevance.

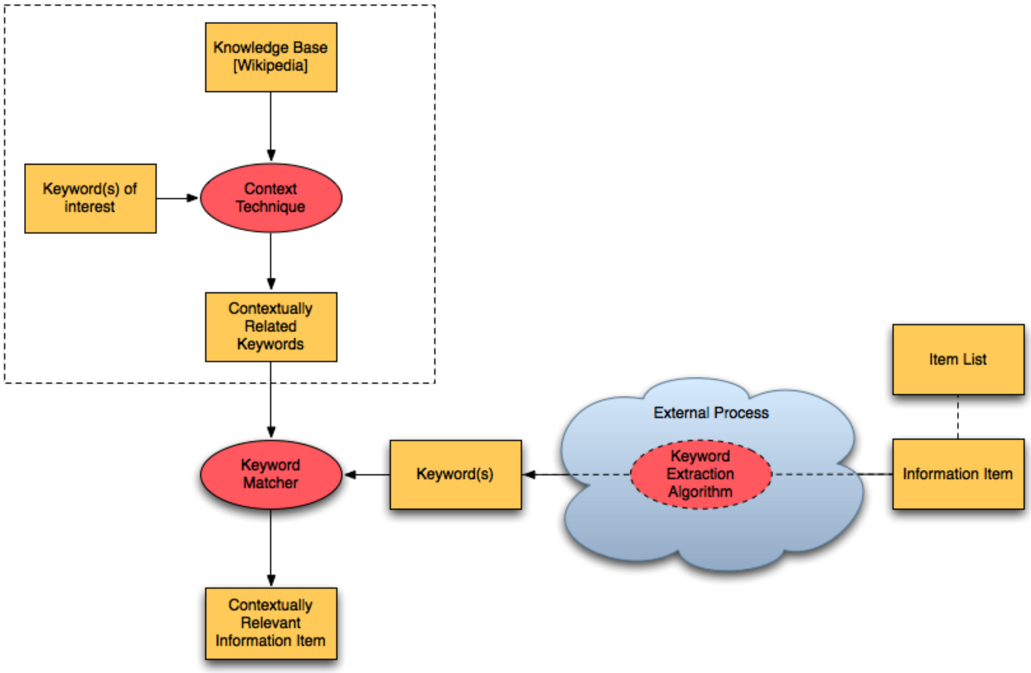

Knowledge Base [Wikipedia]
Keyword(s) of interest
Context Technique
Contextually Related Keywords
Keyword Matcher
Keyword(s)
External Process
Keyword Extraction Algorithm
Item List
Information Item
Contextually Relevant Information Item

Figure 5.4: Overview architecture of the proposed contextual filtering framework

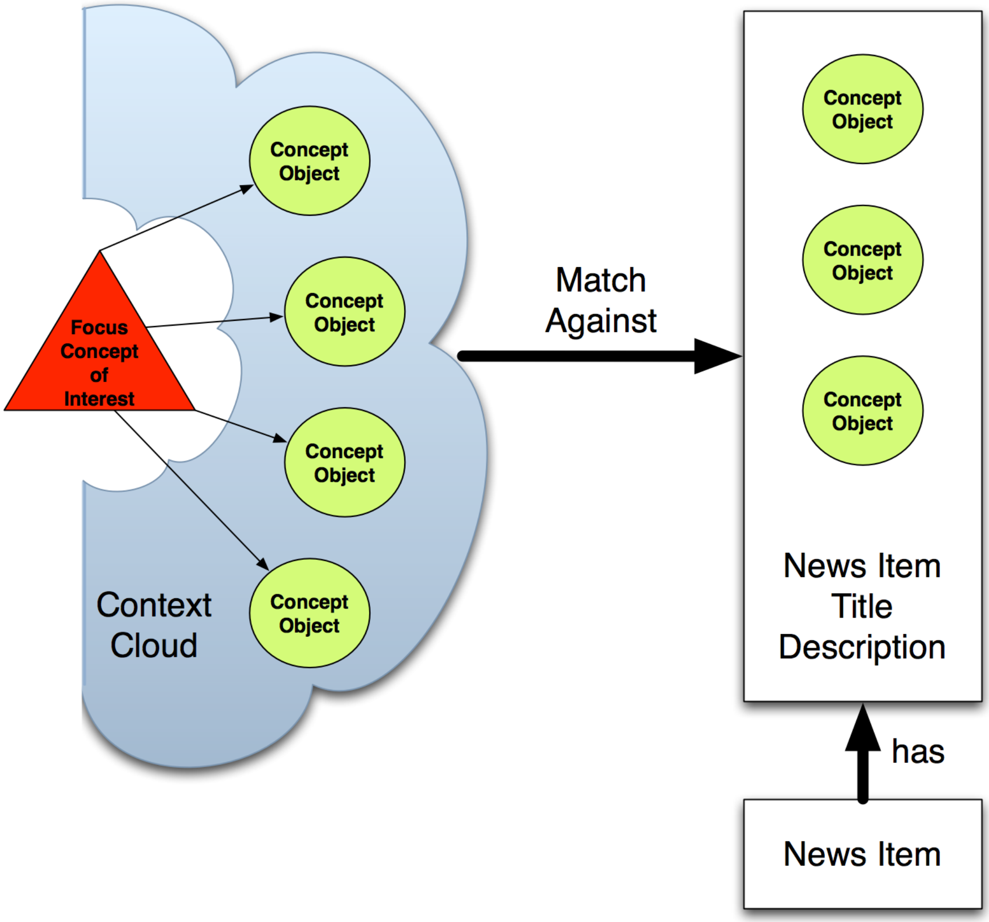


Figure 5.5: Overview of news item title matching against user's interest topic - expanded through the context cloud

Figure 5.5 illustrates the expansion of a user's concept of interest (through the Wikipedia knowledge-base) into a concept cloud. These concepts are matched against concepts derived from a news item's title. Note that for this diagram, only one layer of the context cloud from the F-O contextual concept model is shown to simplify the illustration.

To revisit the Ogden Meaning Triangle, as illustrated in **Figure 5.6**, one can see the relationship between the concepts expanded from the Wikipedia knowledge-base.

The important point of the diagram is to illustrate that a word/symbol cannot fully capture the essence of a concept. Due to this implicit connection to a concept, the concept needs to be explicitly invoked without ambiguity. For the experimentation in this thesis, the assumption is that topics/concepts of interest from the user will be entered as keywords in in a way that will directly map onto a Wikipedia concept. Rather than developing a tool, the user must look up the concept of interest in Wikipedia and use the web-based concept disambiguation features that Wikipedia provides. The development of a semi-automated tool is a possible direction for future work.

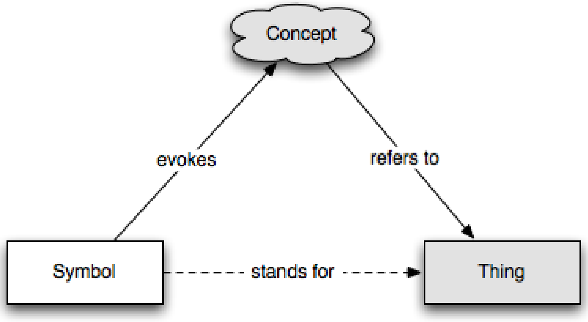


Figure 5.6: Ogden Meaning Triangle, redrawn from illustration from (Fensel et al. 2002)

5.4 Introduction to a Wiki Approach towards Knowledge Extraction and Comparison

As we have discussed earlier within this thesis, there are challenges in encoding human world knowledge. In order for human knowledge to be made accessible, this knowledge would have to be recoded in a format that a machine can understand. Amsler (1987) notes that either there will be a parallelism of effort or that there will be an evolution of publishing methods to accommodate consumption by both humans and computers. As he states; *“researchers in AI and computational* *linguistics therefore have some responsibility to determine how the existing printed knowledge can evolve into usable computational world knowledge".*

Looking at traditional textual documents, texts, such as encyclopedias, go some way towards providing a world knowledge for human readers. From here we look at Wikipedia as a popular example of a web-based encyclopedia that provides a process to engineer the knowledge-base that is human readable, but at the same time is self contained and has a degree of machine understandable semantics.

One of the contributions of this thesis' work involves the usage of a wiki (in this case Wikipedia) as a knowledge-base to achieve a tangible ideal of a global ontology and knowledge-base, whereby all concepts are linked together in a reasonably consistent manner under an overarching control authority. At the same time, the structure and knowledge needs to possess the agility to be able to evolve over time. This instantiation of a knowledge-base through Wikipedia will be used as the basis of world knowledge to base the proposed concept expansion framework upon for use in (news) filtering.

In order to perform contextual comparisons between concepts, the need for a global knowledge-base is great as there are benefits of having a holistic collection of world objects in a structure that can be interrogated. Large scale wikis (for instance, Wikipedia) provide a structure that can be interrogated for contextual information in a manner that cannot easily be achieved through an unstructured WWW or a tightly, but fragmented structured Semantic Web.

As has been detailed previously in this thesis, there are a number of techniques and measures to help determine the semantic distance between concepts in a taxonomy tree and to an extent through RDF graphs and ontologies. Limitations of these approaches, however, is that they rely upon a taxonomy tree structure. As previously discussed, the taxonomy tree approach suffers from decisions on where to place concepts in the structure, leading to forced semantics from a particular authority's view. The original approach of this thesis is based upon a third structure, that of a wiki, along with proposed techniques in order to perform similarity determinations between concepts.

5.5 Wikipedia as a Knowledge-base

Wikipedia, launched in 2001 is an online wiki-based encyclopaedia. Wikipedia allows web users to create pages for concepts along with the ability to edit pages and interlink pages. Wikipedia allows for the community maintained creation and editing of encyclopaedic content by web users.

Wikipedia has been described by and used pragmatically as a knowledge-base by a number of research sources (Nguyen and Cao 2008, Strube and Ponzetto 2006, Gabrilovich and Markovitch 2006) in recent years. However a stronger qualification of Wikipedia as a knowledge-base is required for this thesis, which will be discussed here.

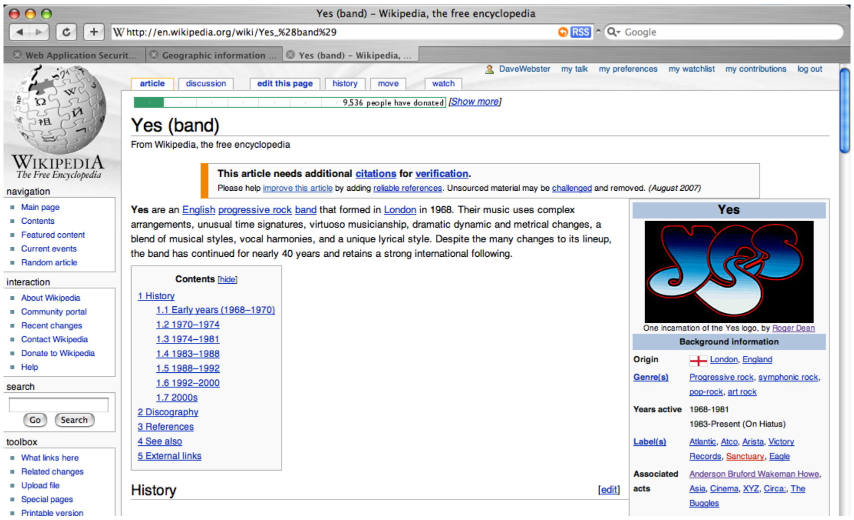
In this section, we will discuss Wikipedia in the context of the characteristics of a knowledge-base introduced in **Section 2.6**. As has been described, a knowledge-base allows for a permanent storage of knowledge and allows inferences to be drawn from it. Existing work (for instance Strube and Ponzetto (2006)) use Wikipedia as if it were a knowledge-base in order to aid in word sense relatedness calculation and make inferences based on a taxonomy of concepts. As Brachman and Levesque (2004) describes, a knowledge-base needs to be told facts in order for it to be useful. Likewise as Fellbaum (1998) describes, common-sense reasoning requires millions of concepts for aiding useful reasoning problems. Wikipedia contains over 3 million items (Wikipedia 2009) and, therefore, qualifies by size as being large enough to use for meaningful reasoning.

5.5.1 Wikipedia as a Knowledge-base by Definition?

The above has described Wikipedia as being used pragmatically as a knowledge-base, but according to the definition in **Section 2.6** does Wikipedia qualify as a knowledge-base by definition? A major distinguishing factor of a knowledge-base from a collection of knowledge is through the definition of knowledge and the controlled and constrained way that knowledge is defined and linked together using propositions, for instance supported through a schema.

Wikipedia takes the form of an online encyclopaedia and represents knowledge in a manner that is easily created and processable by humans. Wikipedia's units of knowledge are concepts that are represented by a textual word. This definition can be correlated with the Ogden Meaning Triangle whereby the textual representation of a Wikipedia concept is the symbol that evokes the concept - pending word-sense disambiguation by the human contributors. This word becomes the title of a textual document whose goal is to describe the concept given certain objective criteria, for instance correctly citing facts and not giving subjective interpretation of information. The definition of a *'document'* by Shapiro, Voiskunskii and Frants (1997) help us with this as she defines a document as, *"a material carrier with information fixed on it"*. The *'goal'* of this textual information is, therefore, to represent knowledge about a concept as expressed by the human author. Propositional logic is, therefore, represented within this textual document through a combination of layout conventions and specific machine processable elements, such as links to key supporting concepts within the document text.

A Wikipedia concept page is typically linked against multiple related concept pages. According to Buriol, Castillo, Donato, Leonardi and Millozzi (2006), the 2006 Wikipedia archive dump contains concept pages with an average number of outlinks numbering 16, following a power law. A typical Wikipedia page can be shown in **Figure 5.7**. This can be abstracted as per **Figure 5.8**. A schema for Wikipedia to link together knowledge is, therefore, present. This takes the form of links between concepts with the textual description of a concept. The propositions represented by the link to other concepts are weak predicates and are unlabelled. However, just because they are not labelled does not diminish their role as predicates; an appropriate label would be *"is related to"*. This simplicity contributes to the popularity of Wikipedia amongst human contributors as the users do not need to spend too much time concerning this overarching concept linking structure. This is because the propositional predicate links between concepts are implicitly created at the point of linking to a concept within the document text.



Yes (band)

From Wikipedia, the free encyclopedia

This article needs additional citations for verification.
Please help improve this article by adding reliable references. Unsourced material may be challenged and removed. *(August 2007)*

Yes are an English progressive rock band that formed in London in 1968. Their music uses complex arrangements, unusual time signatures, virtuoso musicianship, dramatic dynamic and metrical changes, a blend of musical styles, vocal harmonies, and a unique lyrical style. Despite the many changes to its lineup, the band has continued for nearly 40 years and retains a strong international following.

Contents [hide]

History [edit]

Yes

One incarnation of the Yes logo, by Roger Dean

Background information	
Origin	London, England
Genre(s)	Progressive rock, symphonic rock, pop-rock, art rock
Years active	1968-1981 1983-Present (On Hiatus)
Label(s)	Atlantic, Atco, Arista, Victory Records, Sanctuary, Eagle
Associated acts	Anderson Bruford Wakeman Howe, Asia, Cinema, XYZ, Circa:, The Buggles

Figure 5.7: Screenshot of the Wikipedia html page for the rock band Yes.

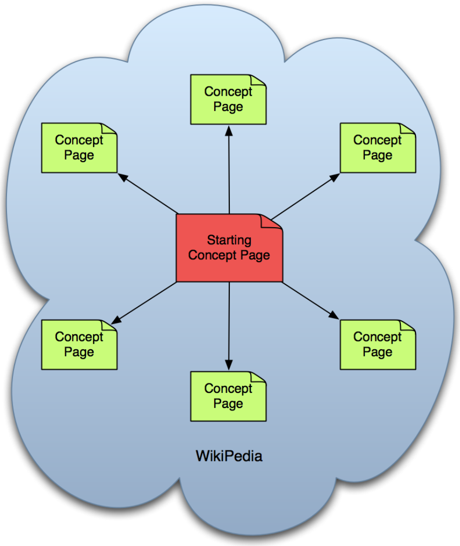


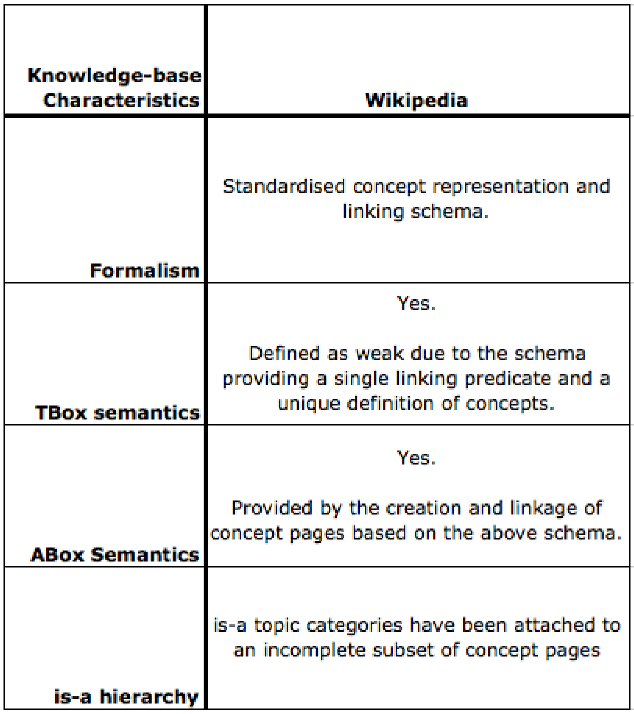
Figure 5.8: Abstraction indicating the relationship between a wiki concept page and its referenced concept pages.

This can be related to the definitions of ABox and TBox features in knowledge-bases respectively. One of Wikipedia's strengths is in that it has a very simple TBox model with relaxed semantics that are inferred through the user's reading context. In addition to this, the rich concept connectivity comes through the ability to link key terms/concepts in the page text of a concept to other concepts in the Wikipedia knowledge-base. This benefits our definition as Brachman and Levesque (2004) and Baader et al. (2003) describe that in their experience a knowledge-base can benefit from rich concept connectivity, but a small set of relation types. The consequence of this is that the semantics of reasoning across this knowledge-base are weak.

Another fundamental characteristic of a knowledge-base is that only one definition for a concept name is allowed. Likewise Wikipedia provides an information space for knowledge documents based on a concept (Gabrilovich and Markovitch 2006), for example a document for telephones called **Telephone**. This pattern can be described as *"one URI per concept"*, whereby concepts can be consistently represented by a URI (Hepp et al. 2007). It is important to note that this pattern bears close resemblance to the pattern present in Semantic Web languages, for example RDF, in relation to the identification of resources. Wikipedia provides us with a method of defining unambiguous terms (Gruber 2005) and allows page authors and editors to create links between these terms through the consistent concept page naming scheme. Wikipedia also provides a consistent access pattern for concept pages. Taking the telephone example, the resource URI for the concept would be
http://en.wikipedia.org/wiki/**Telephone**. In addition to this pattern, Wikipedia provides the ability to create redirects for synonyms or items that have been renamed. Wikipedia, therefore, provides a controlled vocabulary in addition to the one URI per concept. We now have a knowledge-base of concepts that can be accessed in a homogeneous and consistent manner.

In summary, Wikipedia can be described as a knowledge-base by definition, but it should be clarified that Wikipedia is a **weak** knowledge-base compared to a rich and formal knowledge-base as would be created with a logic programming language like Prolog. Wikipedia has weakened this TBox schema in order to aid in contribution by human contributors, but nonetheless does contain a homogeneous schema for defining concepts and linking them.

This can be further summarised in tabulated form against the already defined Knowledge-base characteristics in **Figure 5.9** below.



Knowledge-base Characteristics	Wikipedia
Formalism	Standardised concept representation and linking schema.
TBox semantics	Yes. Defined as weak due to the schema providing a single linking predicate and a unique definition of concepts.
ABox Semantics	Yes. Provided by the creation and linkage of concept pages based on the above schema.
is-a hierarchy	is-a topic categories have been attached to an incomplete subset of concept pages

Figure 5.9: Summary for knowledge-base characteristics applied to Wikipedia

From here we can develop a method to access and probe concept pages based upon a given concept name that maps directly to a Wikipedia concept. It is within the structural content of the wiki upon which our proposed framework and associated methods to determine the semantic relatedness between concepts and expansion thereof can be performed upon.

5.5.2 Summary and Human Judgement in Wikipedia Articles

In looking at the wiki structure and in particular Wikipedia as an example, we now have presented to us a knowledge-base structure that lies between the rigidity of the centrally controlled taxonomy and the weak semantics of free text. Wikipedia also provides us with a method of defining unambiguous terms and allows page authors and editors to create links between these terms through the consistent concept page naming scheme.

Whilst the accuracy of Wikipedia may be disputed, there is at least a social process of creating an accurate (or at least socially challenged) account of the concept being written about in the page (Hepp et al. 2007), although personal bias is likely to play a part in this. This has shown a contradiction to the claim by (Berners-Lee et al. 2006) that *“enforcing consistency checking and trying to outlaw contradiction is a non-starter thanks to the social pressures towards inconsistency on the Web, or indeed other large-scale distributed systems"*. In light of this, Wikipedia can be considered as a good example of a community maintained knowledge-base, which sits between the organically controlled folksonomies and the tightly engineered ontologies.

Work by Stvilia, Twidale, Smith and Gasser (2005) attempts to measure the quality of Wikipedia by defining a number of metrics. In order to perform this analysis the articles and edit history logs are used and the research has been successful in discriminating high quality Wikipedia articles. Hu, Lim, Sun, Lauw and Vuong (2007) similarly developed a number of quality measurement models, this time through collaboration through authors when editing Wikipedia articles. The motivation of this work is that the peer review of Wikipedia articles as they are being built will contribute towards their quality with content surviving the collaborative edits indicating quality. The authors summarise this as the interaction between articles and their contributors. This work concluded that quality is not just judged from interaction data, but also through consideration of the length and maturity of articles - this approach was validated through a comparison with the Wikipedia "best articles" list that undergoes a *“rigorous"* human review process. This work is affirmed by Kittur and Kraut (2008) who notes that *“even explicit coordination via discussion has an implicit structural component: planning is done by a small subset of contributors"*. This coordination is documented as creating a strong association with article quality; quality increases in the early life-cycle of the article, with edits later in the life-cycle distributed to more authors focussing on minor grammatical fixes.

In addition to this discussion Voss (2005)'s measurement of Wikipedia revealed that article sizes are lognormal distributed. Buriol et al. (2006) notes that the average edit size of wikipedia articles is between 300-500 bytes, *"roughly equivalent to a short paragraph of text"*. Buriol et al. (2006) also states that *"50% of the articles have more than 7 different persons involved and about 5% of the articles have more than 50 different editors"*.

(Wilkinson and Huberman 2007)'s experiments observed that the lognormal distribution of edits per article moved from a peak of 3 edits after 120 weeks to 4.5 after 240 weeks. The conclusion from this work was that *"edits correspond on average to an increase in article quality"* and this was reinforced by having a larger number of distinct editors.

Before proposing our approach to determining the contextual relatedness between concepts in subsequent sections, related work is explored for determining semantic similarity of concepts based on a wiki-based knowledge-base.

5.6 Existing Work in Wiki Concept Relatedness

In recent years there has been a small body of work that has utilised Wikipedia as a knowledge-base in order to perform concept similarity measurements. Due to the relatively young nature of research into wiki structures, there is not a large bulk of related work due to the novelty of the domain. Work towards this has originated from the existing work on taxonomy tree based concept similarity determination.

5.6.1 Strube and Ponzetto

A first example of this work comes from Strube and Ponzetto (2006) who use Wikipedia as a knowledge-base to aid in word sense relatedness calculation. Wikipedia has a large coverage of topics/concepts (more-so than WordNet, particularly contemporary concepts, such as brands) and, therefore, is useful for comparison of concepts that are domain and non-domain specific. Strube and Ponzetto (2006) use the category taxonomy provided by Wikipedia, as most pages contain category information attached to the bottom of each page. The authors also noted that Wikipedia categories do not just form to the pattern of a basic taxonomy, but contain elements of a folksonomy. Standard tree-based semantic similarity measures for example, path based (*Leacock and Chodrow*), information content based (*Resnik*) and finally text overlap based (*Lesk*) were used by Strube and Ponzetto (2006).

A limitation of this method is that whilst it takes advantage of the evolutionary nature of Wikipedia, it relies upon the taxonomy structure of the Wikipedia categories. Limitations of this approach include that not all Wikipedia articles contain category tags and, as with the previous taxonomy tree distance work has the limitation of a static tree structure.

5.6.2 Gabrilovich and Markovitch

A second method in wiki-based semantic similarity comes from Gabrilovich and Markovitch (2006). The work is based upon text categorisation by assigning a document to a concept label(s) that is represented by that document. The work is based upon the process that humans go through in human text categorisation and related to their wider world knowledge. The motivation behind Gabrilovich and Markovitch's technique is that they can infer related concepts based upon a concept. The author acknowledges that there are a number of limitations of is-a topic taxonomies and also comments on some of the issues regarding influence by topic branch owners that have been discussed earlier within this thesis. The authors, therefore, acknowledge that Wikipedia articles are non-hierarchical in their structure. A point made through the work of Gabrilovich and Markovitch is that their technique should not need 'pre-catalogued common-sense' or inference rules in order to interpret Wikipedia articles. The technique proposed performs a statistical analysis on the words used within a Wikipedia page and uses an inverted index to map words to the articles in which the concept word appears. Additionally the technique takes in to consideration that the anchor text provided by the hyperlinks between pages can provide a source for additional labels for concepts.

A limitation of this method (as acknowledged by Gabrilovich and Markovitch) is that their work pays limited attention to the linked structure of Wikipedia in that the number of incoming links express a preference for the article. The work proposes the spidering of related articles to the sourced article to enrich the article and further acknowledges that relations can be derived between concepts through leveraging the cross linking between wiki concepts and suggests investigation for further work.

Gabrilovich and Markovitch propose the following hypothesis, however:

> *“Given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from a concept is ill-advised, as this would collect a lot of weakly related material."*

5.7 Development of a Framework for Contextual Comparison and Semantic Concept Expansion for this Thesis

In this section we detail work conducted in the utilisation of Wikipedia to help us infer contextually related concepts given a base concept. As previously explained, the motivation behind this task is to help alleviate the problem of information overload through the use of information filtering. This filtering is assisted through the inferred concepts. The contribution of this thesis is that the inferred concepts are derived from the Wikipedia knowledge-base through a novel link-based technique that we develop upon.

In our framework methods we do not want to just say that inferred concepts *'are'* related to the base concepts in a boolean manner. Instead, we ideally would like to prioritise this relationship and say *'how much'* these concepts are related. Therefore two methods are proposed, first to say that concepts are related and second to attach a weighting to this relationship.

By building upon the foundational knowledge from of Strube and Ponzetto (2006) and Gabrilovich and Markovitch (2006), this thesis proposes an approach to context modelling techniques to aid in information filtering, in part based upon new application of existing similarity determination techniques (Erdös Number, for instance) to the Wikipedia knowledge-base and its concept interlinking structure. The fundamental difference to existing work is that, rather than use the category tags of each page to map to a taxonomy tree or to use the bag-of-words and an inverted vector space model approach, we propose the novel application of techniques based on the inter-page structure of the wiki.

In the following sections we detail the background behind the methods for our contextual concept expansion framework. We then proceed to detail the motivations behind our methods and then explain how we plan to conduct evaluations on them.

5.7.1 Utilisation of Wikipedia Linking Structure

In our proposed framework for determining conceptual similarity and the contextual expansion of concepts, we view the relationship between wiki concept pages as nodes with out-links pointing to nodes that are referenced by that originating node. This can be represented diagrammatically, as shown in **Figure 5.10**.

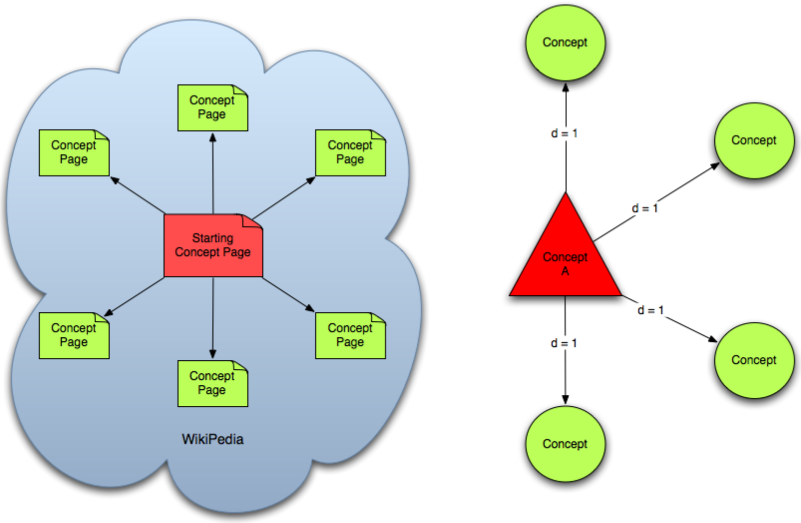
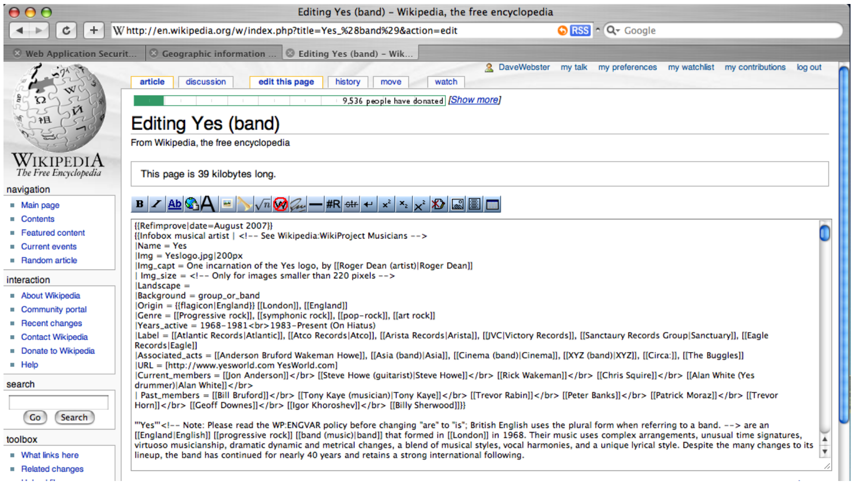


Figure 5.10: Diagram to the left shows the relationship between a wiki concept page and its referenced concept page. The diagram to the right shows how this pattern can be represented as a directed graph, whereby the link depth d can be given a value.

Although Wikipedia (implemented through the MediaWiki software platform) relies upon an HTML web-based output format, Wikipedia uses its own markup for concept pages. The format used to link between intra-wiki concept pages is as follows:

[[PageName]] or **[[PageName | Label]]**

Through this formatting pattern, we can be guaranteed to obtain the concept name for the linked concept page regardless of how the link is labelled and do not have to rely upon complex HTML parsing and natural language processing algorithms to extract these links from the HTML output. As a result of this pattern in the wiki page markup code, we can use a simple regular expression to extract all the intra-wiki out links from a concept page when viewed in editor mode.



Editing Yes (band)

From Wikipedia, the free encyclopedia

This page is 39 kilobytes long.

```
{{Refimprove|date=August 2007}}
{{Infobox musical artist | <!-- See Wikipedia:WikiProject Musicians -->
|Name = Yes
|Img = Yeslogo.jpg|200px
|Img_capt = One incarnation of the Yes logo, by [[Roger Dean (artist)|Roger Dean]]
| Img_size = <!-- Only for images smaller than 220 pixels -->
|Landscape =
|Background = group_or_band
|Origin = {{flagicon|England}} [[London]], [[England]]
|Genre = [[Progressive rock]], [[symphonic rock]], [[pop-rock]], [[art rock]]
|Years_active = 1968-1981<br>1983-Present (On Hiatus)
|Label = [[Atlantic Records|Atlantic]], [[Atco Records|Atco]], [[Arista Records|Arista]], [[JVC|Victory Records]], [[Sanctaury Records Group|Sanctuary]], [[Eagle
Records|Eagle]]
|Associated_acts = [[Anderson Bruford Wakeman Howe]], [[Asia (band)|Asia]], [[Cinema (band)|Cinema]], [[XYZ (band)|XYZ]], [[Circa:]], [[The Buggles]]
|URL = [http://www.yesworld.com YesWorld.com]
|Current_members = [[Jon Anderson]]</br> [[Steve Howe (guitarist)|Steve Howe]]</br> [[Rick Wakeman]]</br> [[Chris Squire]]</br> [[Alan White (Yes
drummer)|Alan White]]</br>
| Past_members = [[Bill Bruford]]</br> [[Tony Kaye (musician)|Tony Kaye]]</br> [[Trevor Rabin]]</br> [[Peter Banks]]</br> [[Patrick Moraz]]</br> [[Trevor
Horn]]</br> [[Geoff Downes]]</br> [[Igor Khoroshev]]</br> [[Billy Sherwood]]}}

'''Yes'''<!-- Note: Please read the WP:ENGVAR policy before changing "are" to "is"; British English uses the plural form when referring to a band. --> are an
[[England|English]] [[progressive rock]] [[band (music)|band]] that formed in [[London]] in 1968. Their music uses complex arrangements, unusual time signatures,
virtuoso musicianship, dramatic dynamic and metrical changes, a blend of musical styles, vocal harmonies, and a unique lyrical style. Despite the many changes to its
lineup, the band has continued for nearly 40 years and retains a strong international following.
```

Figure 5.11: Screenshot of the Wikipedia wikicode for the rock band Yes page.

The Focus-Object context model as discussed in **Chapter 4** can be seen here as applied to the Wikipedia concept linking structure in **Figure 5.12**.

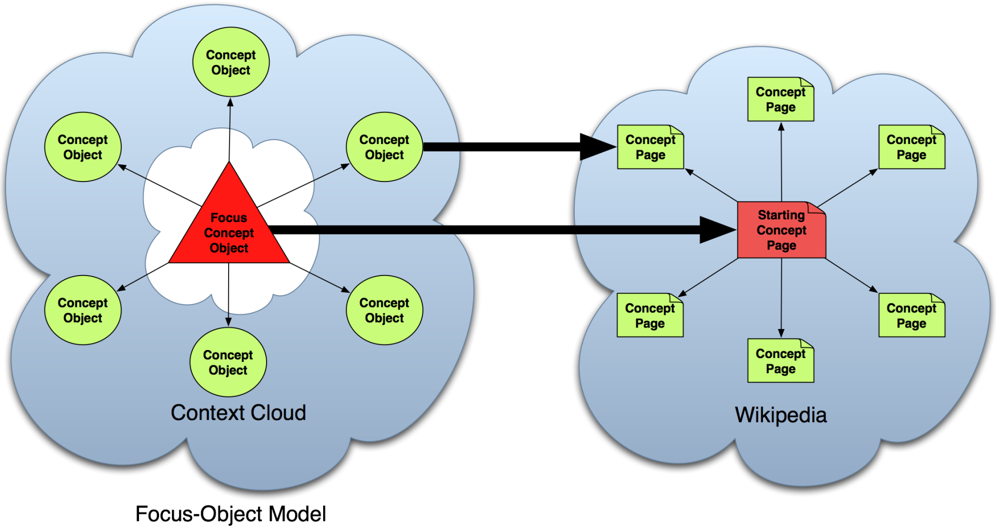


Figure 5.12: Focus-Object model from Context Framework applied to Wikipedia

5.7.2 Contextual Concept Distance

Using a graph-based approach, we can calculate concept distances through the use of an Erdös number to determine the distance between nodes in a linked graph.

The Erdös number originated around the Hungarian mathematician Paul Erdös. Erdös was said to have collaborated with an usually large number of co-authors in the authorship of academic publications. As a result, researchers developed a type of game, whereby the collaboration distance could be measured. The Erdös number is based upon the collaboration distance from Erdös (Grossman 2007). The Erdös number can be be seen illustrated in **Figur 5.13**, with **’D’** representing the collaboration distance. This kind of network is an example of a social network and is related to the theory where any two people in the United States are linked through six degrees of separation (Raghavan 2002).

To give a simple scenario, Paul Erdös collaborates with John Smith. John Smith later collaborates with John Doe. John Doe later collaborates with Anne Onymous. When applying the Erdös distance, Erdös himself has a distance of **0**. Smith has

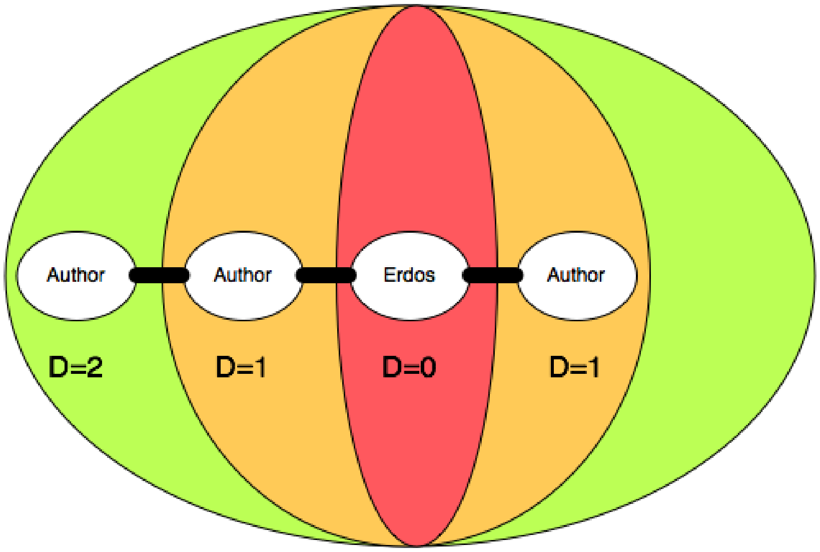


Figure 5.13: Illustration of Erdös Number.

a distance of **1** and Doe has a distance of **2**, with Onymous having a distance of **3**. Variations on this technique have been used by a number of games, for example the Kevin Bacon Game, whereby distance is based on actors co-starring with Kevin Bacon in a film.

In our approach we can use associative reasoning to help us trace a path from one concept object to another, represented by the links between Wikipedia pages in a connected graph. This is based on the *'shortest path heuristic'* (Brachman and Levesque 2004) whereby conclusions are preferred from shorter paths in the network. We use the supposition that if concepts can be connected through a small number of edges, then they are in some way more strongly related than those with a larger (or no) edge path. In this manner, we can say that such concepts are contextually related in a stronger manner.

Following a spidered iteration over the concept graph, by specifying one Focus concept (node), we can aggregate the child concepts of the concepts that the spider has passed through during the spidering process. Supporting the notion of context we can use this approach to extract contextually related (or collateral) concepts. Additionally, we can take into account the distance from the focus concept, therefore attaching semantics as to *how* relevant to the focus concepts the collateral concepts are.

Figure 5.14 demonstrates spidering the graph to aggregate concepts.

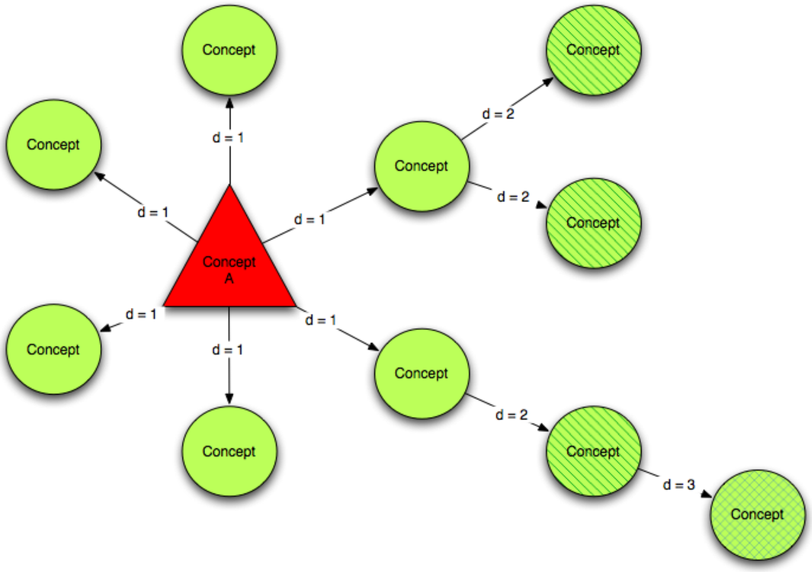


Figure 5.14: Spidering the graph to aggregate concepts

5.8 Details of Methods used within the Proposed Contextual Comparison and Semantic Concept Expansion Framework

This section details the proposed contextual concept expansion and contextual concept distance methods proposed for this thesis. These are known as **Method 1** and **Method 2** respectively. As previously discussed in our framework methods we want to determine that inferred concepts *'are'* related to the base concepts in a boolean manner. Secondly, we ideally would like to prioritise this relationship and say *'how much'* these concepts are related. Therefore two methods are proposed, first to say that concepts are related and second to attach a semantic context based on the path taken through the Wikipedia concept graph.

5.8.1 Method 1: Traversing the Graph for Local Concepts

In this section we detail our first method to determine concept relatedness based on the concept graph. As introduced in the previous section, this method is based upon the notion of spidering a graph for related concepts. As the simplest of our proposed methods, this method collects concepts that are spidered one layer deep from **Focus A**.

The hypothesis of this method is that the set of concepts that are are immediately linked to by a focus concept are more contextually related to that focus concept than concepts further away in the concept graph. We can, therefore, aggregate these concepts and apply them, in addition to the focus concept, to a set of keywords used in information filtering in order to identify related items.

A limitation of this method when viewed in the context of Wikipedia articles is that all linked articles are aggregated without discrimination. This method has the potential of capturing a broad and shallow collection of related concepts, which may prove useful for contextual filtering. A potential problem with this method is that a large number of unwanted concepts have the potential to be aggregated due to the limitation of not contextually partitioning the Wikipedia concept pages. This partitioning is identified primarily as a direction for future work. In addition to future work, it may be possible to assign a weighting to concepts that are spidered more than one layer deep from the root concept.

As discussed by (Gabrilovich and Markovitch 2006):

> *"Given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from a concept is ill-advised, as this would collect a lot of weakly related material. We believe that using similarity analysis and focused crawling techniques can greatly enrich concept representation with that of strongly related concepts, and plan to pursue this direction in our future work."*

For this work, however, we perform one basic partition of the Wikipedia page. Wikipedia pages consisting of more than one section are divided up into an introduction and then followed up by the rest of the page content, divided up by sections. The Wikipedia editing manual (Wikipedia 2008) recommends that this introduction is meant to give an overall context to the concept and should summarise the most important points. We can base the variation around this factor and seek to determine whether prioritising concepts linked from the introduction section of a concept page at the exclusion of all other concepts make a substantial difference to the recall and fallout measures in the experimentation.

Evaluation of Method

In order to evaluate the success of this method as it stands, without further partitioning of the linked concept set, we propose that a collection of items explicitly assigned to a given concept be collected. These items, however, should be described using textual keywords. An important point is that it is not mandated that the keyword for the topic be used in the description. Doing so would create an explicit binding between the focus concept and the information item being compared against. To better explain this, we may want to compare the focus of **"Linux"** against an information item of **"Red Hat Version 9 is Released"**. Red Hat is a linux distribution, so this item would be contextually relevant, but needs not contain the keyword Linux in the textual description.

The use of this collection set is achieved twofold: firstly, the items' textual descriptions are compared against the *Focus A* concept and its immediately expanded concepts. The comparison between the textual names of these expanded concepts and the item descriptions will help us determine a statistical *recall* measure for the information filtering.

Whilst collecting related concepts can be seen as being useful in order to expand a given concept in the context of information filtering, the caveat of this method is that one could easily mine all concepts from Wikipedia and aggregate them to the set of expanded terms. Whilst this would greatly increase the recall measure or items in the filtering process it would also attract a large quantity of unwanted and irrelevant items.

The second part of the evaluation will help us determine the signal-to-noise level when aggregating concepts one layer deep from the concept graph and help us determine the *fallout* of the information filtering based on this method. In order to achieve this evaluation, the information items set will be polluted with items not related to the originating topic.

As previously discussed, for the evaluation of each case, an experimental variation will be applied whereby the whole Wikipedia page and only the introduction section of the Wikipedia page will be tested for both recall and fallout. From this variation we can make a suggestive conclusion and hypothesise as to whether partitioning of the linked concepts based on page location warrants further investigation.

As previously discussed, one of the benefits of using Wikipedia over a semantic lexicon (such as WordNet) is that it contains a large coverage of contemporary concepts and product names and has the ability to quickly integrate new concepts. This will allow a practical implementation of the context framework to be usable with ‘real-world’ user queries.

The Google Zeitgeist search trend tool (Google 2008) listed the fastest rising search terms for 2008 as:

1. sarah palin
2. beijing 2008
3. facebook login
4. tuenti
5. heath ledger
6. obama
7. nasza klasa
8. wer kennt wen
9. euro 2008
10. jonas brothers

This list demonstrates that web users are commonly searching for contemporary concepts, such as product names, events and influential/popular people. One aim for the evaluation scenario for the contextual filtering framework is to be based on real web data; unfortunately, this precludes the comparison with WordNet. The other contradiction with WordNet is in our proposed method of expanding out a concept through the Wikipedia graph with related concepts. **Figure 5.15**, demonstrates that the WordNet term network is limited to an is-a network and does not include the contextually related concepts that our experimentation attempts to validate from Wikipedia.

Additionally, whilst Method 1 is used to expand out concepts based on multiple layers of the concept cloud, there is no obvious mapping on to the taxonomy-based similarity metrics introduced in **Section 2.7**. In order to achieve this, the maximum depth of the concept cloud for Wikipedia would need to be determined; this would allow for the distance to be normalised between 0 and 1. This area of normalisation of Wikipedia distance data is an area envisaged as part of future work through a detailed study, as currently it would require an arbitrary maximum distance to be placed.

In order to evaluate our approach, a baseline will be used, whereby an unexpanded concept will be used for filtering through our real-world case study. The motivation for the experimentation is to determine whether a context-driven approach to information retrieval can yield positive results above a baseline. The root hypothesis behind this motivation is that expanding out a focus concept with contextually related concepts from a knowledge-base will improve title filtering quality over the baseline of just the focus concept on its own. Whilst we want to be able to measure the recall of our method, we also want to measure the fallout of the method. One could describe this as the signal-to-noise ratio. These two measures form a crucial foundation for the experimental design. This is further documented in **Section 6.3**.

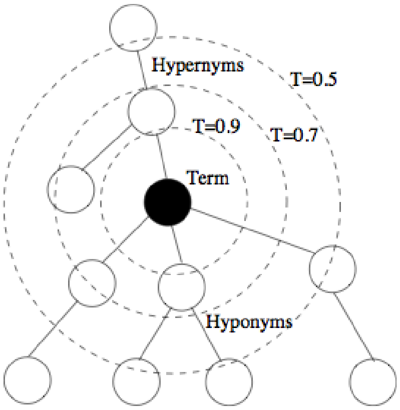


Figure 5.15: WordNet term expansion. Sourced from (Varelas et al. 2005)

.

5.8.2 Method 2: Semantic Context Paths

A second proposed method of concept relatedness is again derived from the Erdös distance applied to semantic distance. In this method we take the approach of identifying two focus concepts; **Focus A** and **Focus B**.

The hypothesis behind this method is that it builds on Method 1 in that concepts closer to **Focus A** in the concept graph are more semantically and contextually related than ones far away. Based upon the tree-based semantic distance we propose finding the shortest path between the focus nodes in the concept graph. Unlike a concept tree, whereby all concepts are linked together by a common ancestor (or root), in a directed graph, a shortest path algorithm will need to be used, for example Djikstra's shortest path algorithm.

To summarise, in this method, we propose calculating the shortest distance between two concepts based on the shortest path between them in the concept graph.

The Context in the Context Paths

Whilst one can view the above method as a contextual approach to determining the semantic distance between concepts, this can also be viewed as just a semantic distance. A further approach can be applied to the paths as we explain.

In an un-weighted cyclic graph, for instance that which is present between concepts in Wikipedia, there can potentially be a number of different candidate paths between nodes with the same (shortest) edge distance. One may call this type of path a chain of nodes and edges. This research recognises this phenomenon in the context of a knowledge-base, whereby different paths between concepts represent different contexts. **Figure 5.16** illustrates the notion of multiple candidate shortest paths between concepts.

Taking this concept further, with the notion of user preferences, it may be more appropriate to take a path longer than the shortest path to avoid a particular node, thereby influencing the context path. In this manner, the problem can be seen as

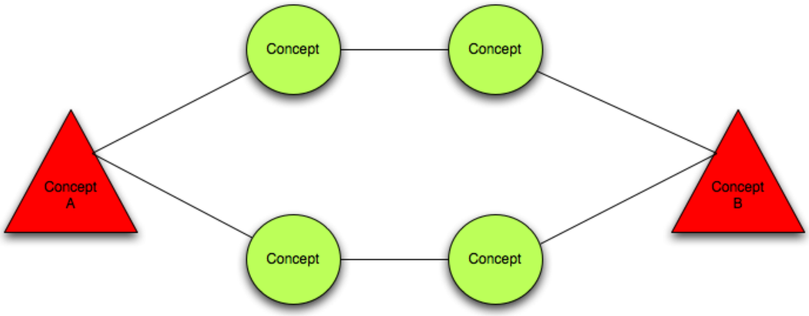


Figure 5.16: An example of two context paths between concepts

similar to that of train routing. This notion relates back to our discussion on context anchors in **Section 4.3.1**.

Traversing the Graph

We acknowledge that the path between **Focus A** to **Focus B** and **Focus B** to **Focus A** may have semantic differences, particularly if the concept graph is treated as being unidirectional in nature. Due to the scope of research, this graph traversal has been decided to be outside the scope, but would make for valid research for future work due to the computational load associated with it. Within the graph, we represent the edges as bidirectional, in order to simplify the problem.

It should be noted, however, that this decision will make paths shorter when tracing the path in a circle, from **Focus A** to **Focus B** and back to **Focus A** again, as it will be possible to reverse the traversal along the path that has already been traversed. To give an example of the shortening of the whole path through a bidirectional approach to links, we note a case that involves the spidering from the return path when spidering to a limited depth. To clarify, it is possible that nodes not spidered in the first direction can be spidered in the second direction. The sub-graph, therefore, may contain a shorter route, when these sub-graphs are combined. Illustrative traversal patterns can be shown in **Figure 5.17** and **Figure 5.18**.

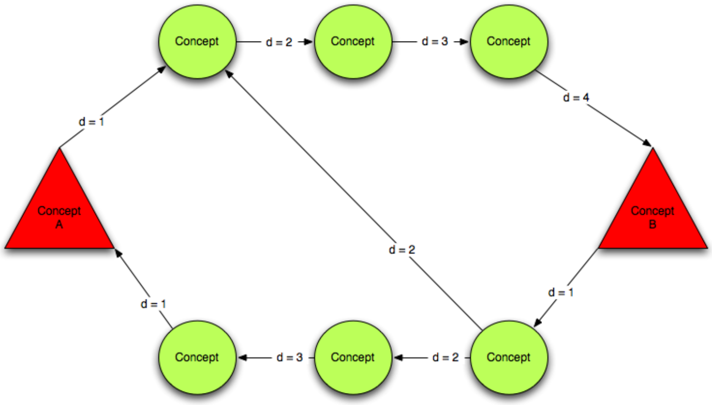


Figure 5.17: Cyclic path following unidirectional links

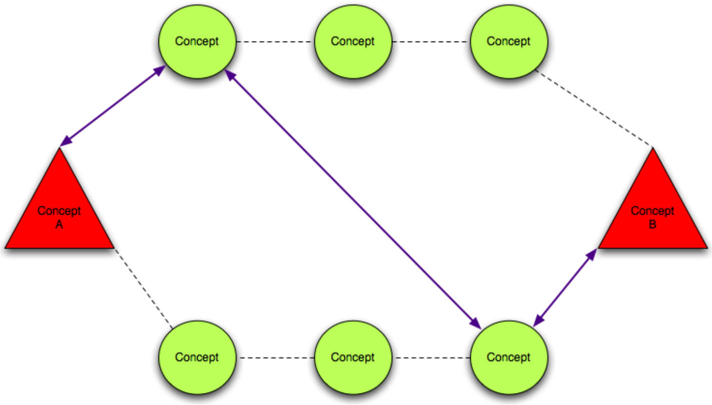


Figure 5.18: Cyclic path following bidirectional links

To continue with this illustrative example, **Figure 5.17** demonstrates that a spider is undertaken with a spider depth of 4, for instance. The spider meets **Focus B** at the 4th depth level. When spidering back from **Focus B** to **Focus A**, the spider proceeds through three nodes that it has not previously encountered. If we add up the path lengths, we have a total path length of **8**. In this case, the first concept that **Focus B** to **Focus A** a node that is passed through links to a node that was passed through in the **Focus A** to **Focus B** traversal. In the case of unidirectional links, this does not matter as none of these nodes link back to **Focus A**. When treating all links as bidirectional, this does matter and we can treat the cross-linking as a shortcut opportunity, giving a total traversal distance of **6**, as can be shown in **Figure 5.18**.

Feasibility Analysis - Discussion of Technical Challenges in Processing Wikipedia

In order to traverse a path between **Focus A** and **Focus B**, we need to adopt a technique for traversal. This technique is to spider the graph from a depth first manner with a cap on the number of depth levels spidered. Due to the size of the Wikipedia graph (the XML dump was 4 GB) there is a computational cost to traversing the graph.

To give a conservative estimate as to the RAM required to store the concept linking graph, Wikipedia contains over 3 million items (Wikipedia 2009) with an average number of out-links numbering 16 (Buriol et al. 2006). If each concept were to contain 16 links with an mean average length of 16.3 characters (standard deviation = 9.6) then the minimum size of the raw graph data in memory would be:

> RAM Required = 3,000,000 pages * 16 links per page * 16 characters per link * 2 Bytes per character (UTF-16)
>
> = 1,536,000,000 Bytes
> = 1536 MB

Note that this figure does not consider memory required for data structures to store this data and the memory required for the operating system (Fedora GNU/Linux 5), Java 1.5 runtime and the Eclipse 3.2 code development environment.

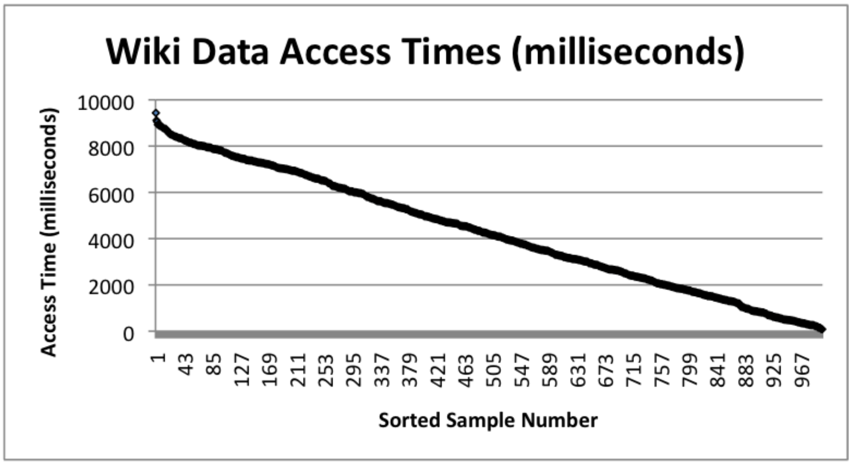


Figure 5.19: Access times for individual Wikipedia page data from the physical disk drive

Figure 5.19 shows a sample of 1000 accesses for Wikipedia data on the physical disk storage device, including both the time to find the page in an index and the lookup and reading of the page data. The data taken from this test shows that the access lookup times came in between at most 9433 milliseconds and least 85 milliseconds. The mean average access time was 4288 milliseconds with a standard deviation of 2492 milliseconds. The calculation of the time to load the entire graph of Wikipedia data into RAM would be:

Total Time = 3,000,000 pages * 4288 milliseconds

= 12,864,000,000 milliseconds
= 3573.33 hours
= 148.89 days

We, therefore, accept that due to the size and complexity of the cyclic graph, traversing the entire graph will not be practically feasible given the resources allocated to this PhD study of a desktop PC with a 2 Ghz Pentium IV processor with 512 MB RAM with USB attached disk storage for the Wikipedia data.

The result of spidering the Wikipedia data dump is to produce a sub-graph that can be held in computer RAM, as the implementation for the thesis has been realised using the Java programming language. Using this sub-graph, a shortest path algorithm can be used on the graph, for instance Djikstra's algorithm, in order to find the shortest path from one focus concept to the other. Whilst from a high-level view this seems sensible, there are a number of subtle problems.

We encountered difficulties in Djikstra's shortest algorithm whereby only one shortest path (length) is found. In our experience, we observed that implementations of the algorithm, for instance Jung (2005) did not contain the ability to provide more than one shortest path, nor did they have the scope to be modified to gain this ability due to the way that the algorithm has been implemented. The conclusion from this observation is that the only option would be able to re-implement the algorithm from scratch, which throws a problem into this method, in that considerable time would be needed in order to thoroughly test this implemented algorithm. We recognise that a sub-standard implementation would have repercussions on the contextual technique that we implement on-top of it and, therefore, assign this to future work outside the scope of this thesis.

Using the Wikipedia link structure, this simple proposition becomes more complex than originally seems, as one needs to consider that a page links out to neighbouring pages. From this we have a directed graph with edges pointing from the originating page to the child neighbours. Semantically there will be milage in recognising that for every directed edge between a concept and its neighbour, there is an implicit edge between the neighbour and the originating concept.

Finally it is difficult to normalise the distance value of this method. Whilst one can easily say that a distance of zero to **Focus A** means that **Focus B** is identical to **Focus A**, it is difficult to assign a maximum value to this distance in order to normalise between 0.0 and 1.0. Taking the six degrees of separation and assigning a maximum value of 6 may promote some milage, but we have deemed this task as outside the scope of this research.

Summary of Method 2

As has been detailed in this chapter, whilst initial planning work was undertaken in the development of Method 2, the practicalities of implementing such a method proved to be a problem as revealed by the feasibility analysis. With a large dataset, for instance as with Wikipedia, the data processing proved to grow beyond the computational resources available. Whilst the research into this area of determining point-to-point distances would have proven to be an interesting and potentially useful line of research, the development of an implementation to implement Method 1 was, therefore, given priority. Experimental work in the latter chapters of this thesis, will consequently focus on Method 1.

5.8.3 Application of Methods to the Semantic Web

In both methods presented in this chapter we produce a sub-graph that can be interrogated for related concepts and of which can be measured for graph distance. These sub-graphs can be represented as an RDF graph, as one can easily see the analogy between the node links in an RDF graph and the node links in the Wikipedia graph. An advantage to this Semantic Web data is that the information can be reused in other applications and can be processed quickly, as the graphs represent only a sub-graph of the entire Wikipedia concept graph.

A limitation with the current work is that the wiki will not provide meaningful predicate information between nodes explicitly. As a 'quick fix', it would be possible, however, to define the predicate of **relatedTo** and apply that to each edge of the graph. An exception to the rule that Wikipedia does not have explicit semantics between links is that of the redirect. In the case of turning the links into bidirectional edges, one would simply apply an inverse predicate between nodes in the opposite direction to the explicit predicate.

For future work in this particular area, it may be possible to imply semantics based upon the position of the page in which the link appears, however as mentioned in this work this has not been explored due to being outside of scope. Initial thoughts suggest that the position of a link, such as inside an information box (which can come under a number of different categories) can form the basis for a semantic predicate to apply to the link.

Chapter 6

Experimentation and Evaluation: Context-Orientated News Feed Filtering Case Study

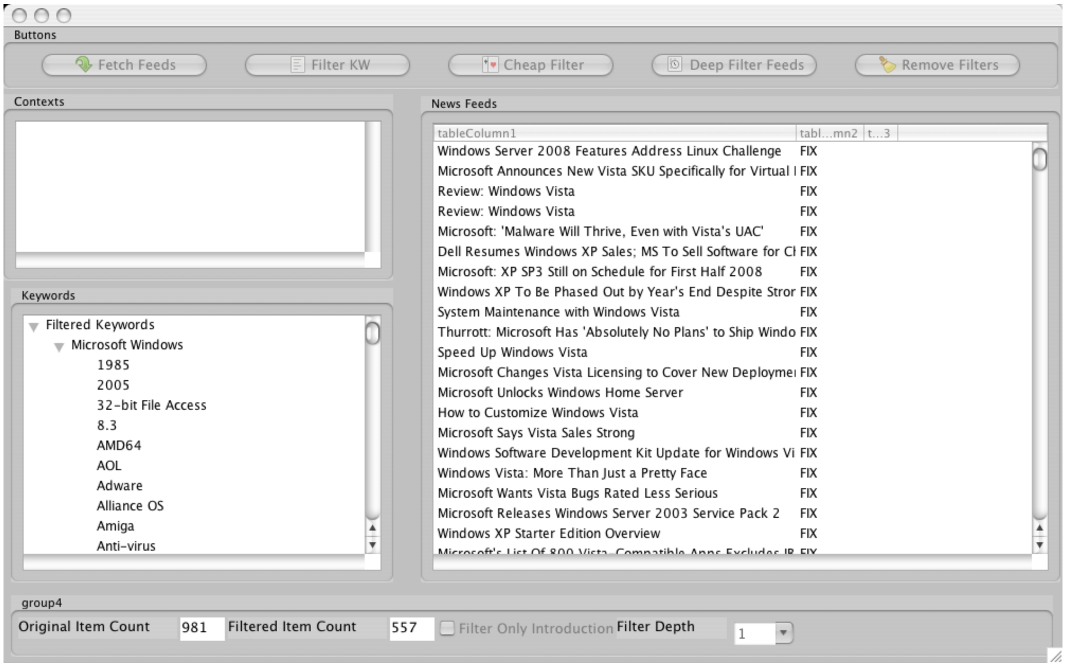
In this chapter the experimental methodology for our knowledge-base concept expansion method is presented and discussed. The practical experimentation will provide a mechanism to demonstrate the validity of the proposed method through the production of empirical data. For the experiment, a scenario is proposed in order to demonstrate the context-orientated solution to the information overload and filtering problem presented throughout this thesis.

6.1 Development of a Prototype Tool

For this experimentation phase, a simple web feed aggregator has been built using the Java programming language that we call **DAV**e's **R**ss **O**rganisation **S**ystem or DAVROS-2. DAVROS-2 is designed to aggregate web feeds from a given URL and store these items internally. The tool then accepts a concept as a focus for filtering. The tool then calls a developed sub-system to perform the concept expansion based upon a cached copy of Wikipedia. Once the expanded list of concept keywords is built, the web feed item titles are compared against the concept keyword list and then boolean filtered based on this concept keyword comparison.

DAVROS-2 was designed within the scope defined in the previous chapter as a testbed environment to demonstrate specific tests used within this experiment and to output data relating to each test. As a result of this scope, the application does not implement any complex text mining algorithms, instead relies upon simple case insensitive keyword matching.

From a technical implementation perspective, DAVROS-2 uses the Informa library for Java in order to import and parse web feeds from given sources. The interface of the application is designed to resemble a simple web feed aggregator, with a layout similar to that present in e-mail and Usenet readers. The interface has been implemented using the SWT library.



- Filtered Keywords
  - Microsoft Windows
    - 1985
    - 2005
    - 32-bit File Access
    - 8.3
    - AMD64
    - AOL
    - Adware
    - Alliance OS
    - Amiga
    - Anti-virus

tableColumn1	tabl...mn2	t...3
Windows Server 2008 Features Address Linux Challenge	FIX	
Microsoft Announces New Vista SKU Specifically for Virtual I	FIX	
Review: Windows Vista	FIX	
Review: Windows Vista	FIX	
Microsoft: 'Malware Will Thrive, Even with Vista's UAC'	FIX	
Dell Resumes Windows XP Sales; MS To Sell Software for Cl	FIX	
Microsoft: XP SP3 Still on Schedule for First Half 2008	FIX	
Windows XP To Be Phased Out by Year's End Despite Stror	FIX	
System Maintenance with Windows Vista	FIX	
Thurrott: Microsoft Has 'Absolutely No Plans' to Ship Windo	FIX	
Speed Up Windows Vista	FIX	
Microsoft Changes Vista Licensing to Cover New Deployme	FIX	
Microsoft Unlocks Windows Home Server	FIX	
How to Customize Windows Vista	FIX	
Microsoft Says Vista Sales Strong	FIX	
Windows Software Development Kit Update for Windows Vi	FIX	
Windows Vista: More Than Just a Pretty Face	FIX	
Microsoft Wants Vista Bugs Rated Less Serious	FIX	
Microsoft Releases Windows Server 2003 Service Pack 2	FIX	
Windows XP Starter Edition Overview	FIX	
Microsoft's List Of 800 Vista-Compatible Apps Excludes IE	FIX	

Original Item Count 981 Filtered Item Count 557 ☐ Filter Only Introduction Filter Depth 1

Figure 6.1: Overview screen of DAVROS-2.

6.2 Vignette: News Filtering based on OSNews Web Feeds

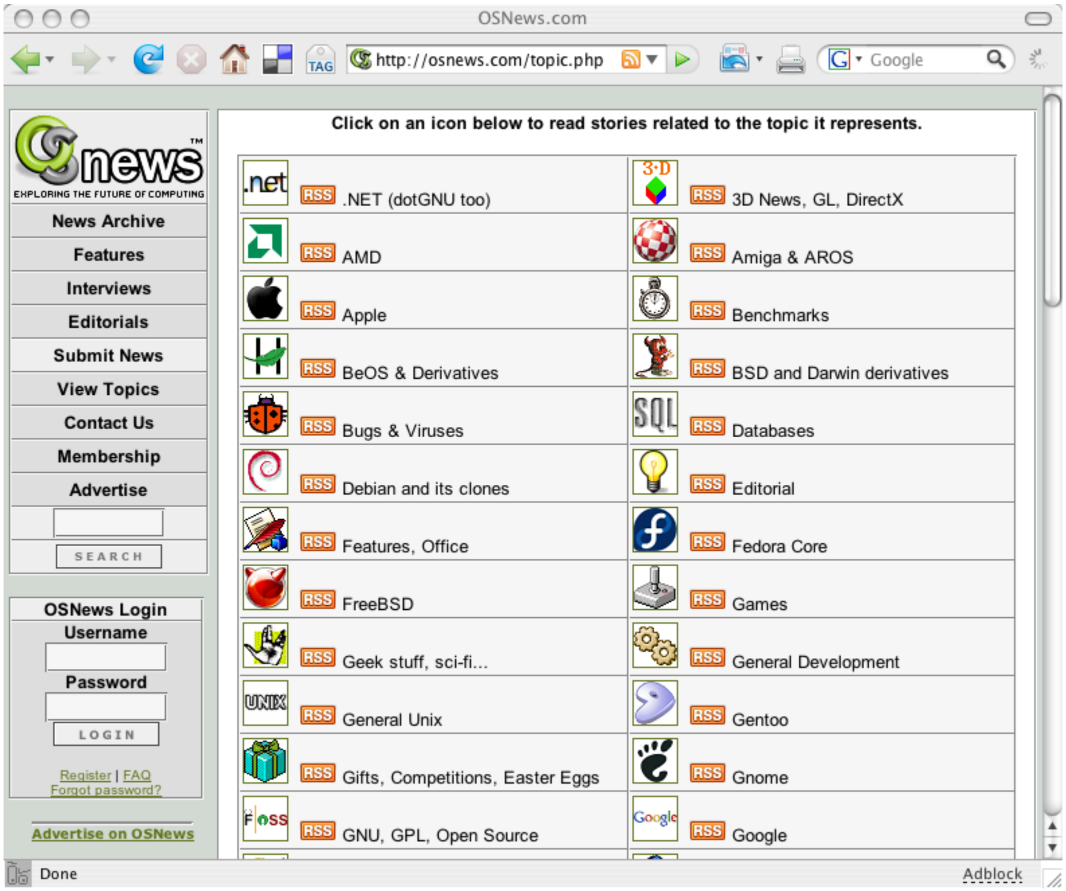
The OSNews website caters as a source of reporting news around the topic of operating systems and related topics. The website allocates each news story a topic sourced from a set of topics that are defined for the website, which are viewable on one of the website's pages. This page allows a reader to click on a topic of his choice and be shown a webpage representing a view of all the news stories under that topic.

6.2.1 Process for Information Sourcing

The OSNews website provides readers with the facility to view and subscribe to a web feed for new news items. A key useful feature of OSNews for this experiment's case study is that OSNews provides web feeds for each of it's topics, meaning that for this experiment we can have relatively easy access to news stories separated by topic. The web site maintainers manually assign a category to news items, so we can make the assumption that some human judgement has been involved in this process. **Figure 6.2** shows the topics page for OSNews.

A problem encountered when we attempted to gather news stories, based on a feed for a given topic, is that only 15 news items are provided in the feed, with these 15 being the most recent. This limitation dates back to the original version of RSS, whereby feeds were limited to 15 items per feed (RSS 2003, RSS n.d.).

A second solution to retrieving the required news stories archive was achieved through the creation of a web scraping tool for the OSNews website, again using Java. OSNews provides an archive for each news topic in the form of a web page. The web scraper was designed to parse the news item titles and dates then 'click' the **next page** button and then repeat the process until all items are retrieved. These parsed entries were converted into an internal model and then exported to an RSS 2.0 web feed using the Informa library and allocated a topic tag in accordance with



Click on an icon below to read stories related to the topic it represents.

RSS .NET (dotGNU too)	RSS 3D News, GL, DirectX
RSS AMD	RSS Amiga & AROS
RSS Apple	RSS Benchmarks
RSS BeOS & Derivatives	RSS BSD and Darwin derivatives
RSS Bugs & Viruses	RSS Databases
RSS Debian and its clones	RSS Editorial
RSS Features, Office	RSS Fedora Core
RSS FreeBSD	RSS Games
RSS Geek stuff, sci-fi...	RSS General Development
RSS General Unix	RSS Gentoo
RSS Gifts, Competitions, Easter Eggs	RSS Gnome
RSS GNU, GPL, Open Source	RSS Google

Figure 6.2: OSNews topics page viewed through a web browser.

the news topic being parsed.

6.2.2 Application and Testing of the Research Contribution

Summarising the experimental proposals within this chapter we outline the following process and divide the experimentation for this vignette up into a number of sub experiments, as summarised below:

A) Simple concept keyword matching against news titles

B) Simple concept keyword expanded to one level against news titles using a shallow spider of Wikipedia

C) Simple concept keyword expanded to one level against news titles using a local copy of Wikipedia

D) Simple concept keyword expanded to two levels against news titles using a local copy of Wikipedia

6.2.3 Choice of Keywords

In this section, we describe the motivation for the choice of OSNews topics that we use as focus concepts for our experiment. For all experiments, eight topics have been chosen from the OSNews archive, namely those of **Linux**, **GNOME**, **KDE**, **GTK**, **QT**, **Games**.

Two criteria were selected in order to simplify the experiment as follows:

- The concept keyword must consist of one word. This simplifies the semantics for experimentation.
- The concept keyword must automatically disambiguate to the correct concept

within Wikipedia. This mitigates the issue of word sense disambiguation and helps reduce experimental side-effects.

The topics of GNOME, KDE, GTK and QT present us with an interesting semantic map. GNOME and KDE are both graphical desktop environments for Linux. GTK and QT are graphical widget toolkits for GNOME and KDE respectively.

6.3 Design of Experimental Verification

The motivation for the following experiment is to determine whether a context-driven approach to information retrieval can yield positive results above a baseline. The root hypothesis behind this motivation is that expanding out a focus concept with contextually related concepts from a knowledge-base will improve title filtering quality over the baseline of just the focus concept on its own.

Whilst we want to be able to measure the recall of our method, we also want to measure the fallout of the method. One could describe this as the signal-to-noise ratio. These two measures form a crucial foundation for the experimental design.

The original contribution to our research revolves around the use of Wikipedia, as a common knowledge-base, whereby related keywords are retrieved at run-time and are not pre-defined in the application (Kiyavitskaya, Zeni, Cordy, Mich and Mylopoulos 2005). It can be seen, therefore, that domains are created on the fly depending on the user's search context. As Wikipedia is a constantly updating knowledge-base, these knowledge domains can be updated to align with new keywords present in the news titles. As stated before, if a news title is retrieved by the web feed aggregator with the title of *'Nintendo releases Zelda for the Wii'*, then particular words, such as brands, may only be a month old.

In the proceeding sections we detail the methodology for each experiment and provide an output of the results gained from the conduction of the experiments.

6.4 Experiment A

The following section will describe the motivation for this experiment and will set up the control and baseline measures in order for subsequent experiments to be compared against.

6.4.1 Methodology: The Control Test

In order to set up a control test, a single keyword has been chosen to be tested against each news title for a given topic. For the experiment the news items were pre-collected and stored in a set before any filtering was conducted.

In order to provide a baseline for the recall and fallout measures we first collected a homogeneous set of news items for each focus topic/concept. As previously discussed, we rely upon the human judgment of the OSNews administrators to perform the manual allocation of news items to categories.

To assist in creating a control measure for the baseline, we created a filter of the keyword in order to discard news titles that do not contain the focus concept keyword. The number of news titles returned will form a baseline that subsequent experiments are compared against.

As a result of a feed being used from a specific topic, over a mixed topic feed, the assumption can be made that all the news items in that feed are relevant to the feeds's topic.

6.5 Experiment B

The second experiment of the testing involves parsing links from the Wikipedia page for a given topic. For this experiment, for a given topic, all the links contained within the Wikipedia page will be captured from the wiki page text and converted into keywords based on their labels. Note, that for this experiment, the pages are not spidered and re-directs ignored. Due to the low resource usage of this experiment, we refer to this as the **Shallow Filter**.

There are a number of experimental factors to take into consideration here:

1) The spidering of the page is shallow, meaning that page link redirection and disambiguation page linking does not occur. Consequently meaning that some labels will be out of date.

2) Since only the pages are spidered at a shallow level, this method is limited to one layer deep.

For this experiment, all retrieved keywords are stored and treated as a flat list, whereby all keywords are treated as equal and unweighted.

Experimental Variation: Page Partitioning

A variation of this experiment involves the extraction of concepts (and in turn keywords) from only the introduction lead section of the Wikipedia page. As previously discussed, the Wikipedia editing manual (Wikipedia 2008) recommends that this introduction is meant to give an overall context to the concept and should summarise the most important points. We can base the variation around this factor and seek to determine whether prioritising concepts linked from the introduction section of a concept page at the exclusion of all other concepts make a substantial difference to the recall and fallout measures in the experimentation.

6.6 Experiment C

In this experiment we use the experimental template from Experiment B but instead of performing a light text based extraction of concepts for Wikipedia concepts, we attempt to resolve the concepts and extract the title from the resolved pages. For each link, the target page will be read and (if needed) redirects will be dereferenced so that the page name can be retrieved. This experiment will expose differences with retrieving titles of resolved and unresolved concepts.

For this third experiment, a local copy of the Wikipedia XML dump is used in order to avoid 'hammering' the Wikipedia web server with http requests, in accordance with the Wikipedia terms of use. For this experiment, for a given topic, all the links contained within the Wikipedia page will be spidered. It is worth noting that due to the number of page requests required to spider each child page of a parent page, an offline XML dump of Wikipedia has been used, so that all page requests are private on a local disk. The XML dump is dated June 2006, so a factor that may influence the experiment is that there is a one year gap between this knowledge-base information and the web feeds being used.

The advantage of this method of spidering is that spidering can be undertaken iteratively, spidering pages that are many layers deep. A major problem with this method, as opposed to the shallow scanning method used in experiment B, is that all concepts need to be resolved, which adds an extra level of spidering overhead on-top of the spider.

As with experiment B, the variation of only extracting links from the introduction section of the page will be used in addition to extracting links from the whole page.

6.6.1 Computational Issues

Upon downloading the Wikipedia XML dump, the file size amounted to around 4GB. In order to speed up XML parsing and to avoid the use of a memory intensive DOM parse of the data, a simple Java tool was developed to perform a lightweight

SAX parse through the 4GB XML file. Upon inspection of the XML document tree, it could be seen that pages were represented using the element **<page>**. The Java application in this case extracted out all pages and created an XML file for each. Each file name was given an incremental index number beginning with 0. An index file was created mapping all page titles with the index number of the XML file. The index file, however came in at around 100MB. When scanning for pages in the index through the DAVROS-2 backend library, the access speed for lookup came in at between one and eight seconds. This access speed presents a problem for experimentation, as a Wikipedia page may contain tens of page links. This problem increases if the links are spidered to two or more layers deep. Given the time and resources to develop an distributed index, this index lookup may be achieved in less time than with the current available computational resources of a desktop personal computer. It is, therefore, not feasible to spider more than one layer deep with the whole page, but, rather the scanning will be limited to the introduction section of the page only. This effect will mainly come into play during Experiment D.

Wikipedia presents the researcher with a vast knowledge-base. Whilst the obvious suggestion may be to attempt to extract out a small section of Wikipedia, the highly interlinked nature of the page link structure means that this is not a reasonable possibility.

6.7 Experiment D

This experiment explores the effect of spidering to two levels deep on the retrieved concepts. Specifically we want to examine the effect on the recall and fallout measures based on the addition of this extra level of concept aggregation. A hypothesis that we present is that the extra layer will reduce the fallout of the filtering process due to the concepts being less contextually relevant based on their distance from the focus concept. It should be noted that no weighting is applied to this second layer of concepts and they are aggregated into the expanded concept set in addition to the first layer concepts.

As described in the previous experiment, and for the given reasons, the spidering for this experimentation will be limited to only two layers deep.

As noted in Experiment C, due to computation resources it is not feasible to spider more than one layer deep with the whole page, but, rather the concept scanning will be limited to the introduction section of the page only in order to restrict the size of this overhead spidering layer.

Chapter 7

Analysis and Evaluation: Context-Orientated News Feed Filtering Case Study

In this chapter we record and analyse the results of the experimentation detailed in **Chapter 6**. We analyse the results of the experimentation in order to assess the effectiveness of the proposed information filtering method proposed by this thesis to provide contextual concept expansion. The analysis will not only compare the variations of the filtering methods against a baseline, but, in addition, will compare variation methods against each other. This analysis will aid in the evaluation of the concept expansion method proposed and developed within this thesis and will contribute to a demonstration of how successful the filtering methods are in alleviating the information overload problem.

7.0.1 Overview of Analysis Measures

As previously introduced in **Section 2.2.1**, we focus on traditional measures for information retrieval, and specifically focus on the following measures for this experimentation, as were both described by Van Rijsbergen (1979):

Recall:

> "*the recall of the system, that is, the proportion of relevant material actually retrieved in answer to a search request*"

and **Fallout**:

> "*an estimate of the conditional probability that an item will be retrieved given that it is non-relevant*"

We can calculate recall as: *number of relevant hits in results / number of relevant document in the collection.*

We can calculate recall as: *number of non-relevant hits in results / number of non-relevant document in the collection.*

The motivations behind using these measures are twofold. Firstly, we want to be able to evaluate the value of our method in terms of how many of the items that we want to retrieve are retrieved. Secondly, we want to determine, out of those items retrieved, how many of them are actually useful. Both of these can be achieved through the use of a given set of items that we know are relevant.

One can illustrate this in a simple example. If we were to retrieve all items possible, then we would have returned all relevant items, but the result set would also contain many more unwanted items and, therefore, has the potential to damage the value of the retrieval filter to the end user.

7.1 Output Data from Experimental Methods across Topics

In **Appendix Section A** we present the results data from each experiment method performed across the OSNews topics. For the result graphs we plot the calculated recall and fallout values corresponding to each test. We present each topic as a profile consisting of the results from each experimental method. The specific tests reported are:

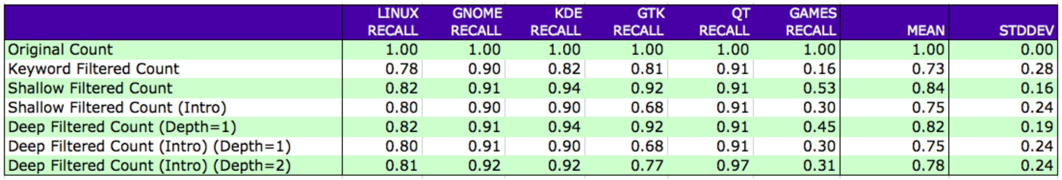
1. All values retrieved from each topic item set.
2. Keyword filter method applied to each topic item set.
3. Shallow filter method applied to each topic item set. Full page.
4. Shallow filter method applied to each topic item set. Introduction section only.
5. Deep resolved filter method applied to each topic item set. 1 layer deep. Full page.
6. Deep resolved filter method applied to each topic item set. 1 layer deep. Introduction section only.
7. Deep resolved filter method applied to each topic item set. 2 layers deep. Introduction section only.

7.2 Analysis of Test Results across Experimental Filtering Methods

In this section we present the summarised results of each profile in a single view; each presented against the baseline method. To achieve this view we simply take the mean average of each method's test across the topics. In the table and graph presented in **Figure 7.1** and **Figure 7.2** respectively we present these mean averages along with associated standard deviations to one standard deviation of the mean.

7.2.1 Analysis of Recall Measure

In this subsection we present a summary of the recall measures for each filtering method across all topics.



	LINUX RECALL	GNOME RECALL	KDE RECALL	GTK RECALL	QT RECALL	GAMES RECALL	MEAN	STDDEV
Original Count	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Keyword Filtered Count	0.78	0.90	0.82	0.81	0.91	0.16	0.73	0.28
Shallow Filtered Count	0.82	0.91	0.94	0.92	0.91	0.53	0.84	0.16
Shallow Filtered Count (Intro)	0.80	0.90	0.90	0.68	0.91	0.30	0.75	0.24
Deep Filtered Count (Depth=1)	0.82	0.91	0.94	0.92	0.91	0.45	0.82	0.19
Deep Filtered Count (Intro) (Depth=1)	0.80	0.91	0.90	0.68	0.91	0.30	0.75	0.24
Deep Filtered Count (Intro) (Depth=2)	0.81	0.92	0.92	0.77	0.97	0.31	0.78	0.24

Figure 7.1: Recall Mean Average of Experimental Methods.

Baseline

Figure 7.2 demonstrates that the baseline keyword filter performed lowest out of all filter methods with respect to the recall value. This result is expected, as the baseline keyword filtering method can only match news items that explicitly contain the topic keyword, whereas the other methods are designed to find related news items based on expanded context concepts. We can conject from this result that the OSNews website editors have in practice allocated news items containing a topic name to that topic correctly.

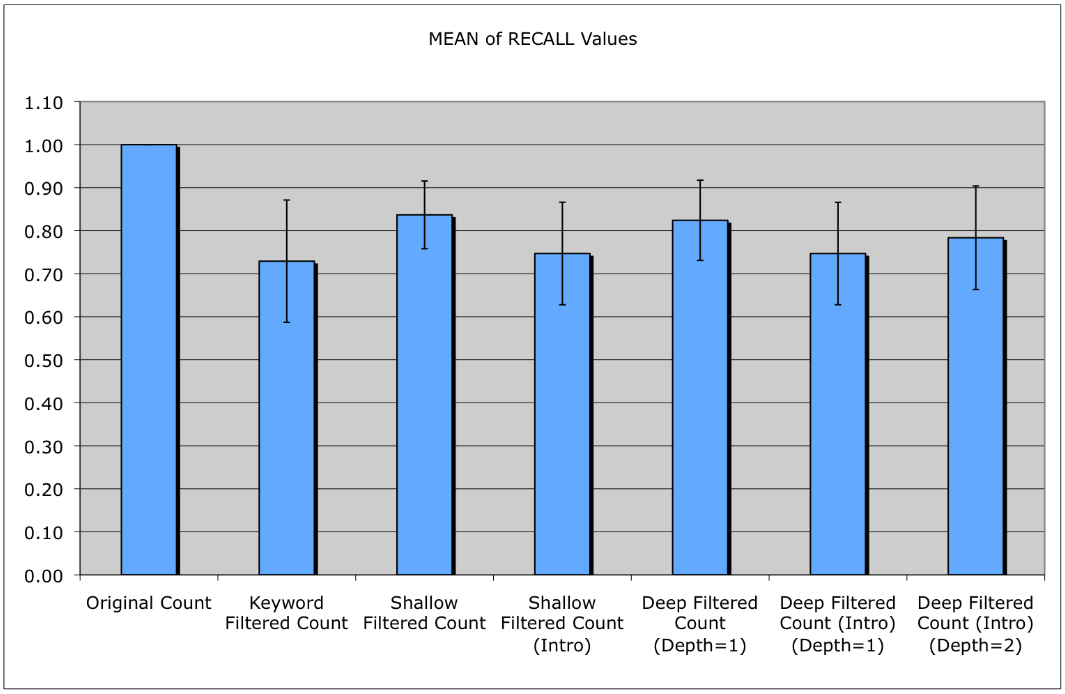


Figure 7.2: Recall Mean Average Graph of Experimental Methods. Bars represent 1 Standard Deviation.

We do note and accept, however, that the standard deviation on the mean of the baseline keyword filtering recall is 28% and is larger than all other filtering methods. We acknowledge that this large standard deviation will diminish the reliability of the keyword filter as a stable baseline to compare filtering methods against.

Filtering Methods Performance

We found that the shallow filter and the deep filter both performed the most optimal with the highest recall mean averages. The similarity between the recalls of these two filtering methods was expected due to the only difference between them being the resolution of Wikipedia concept names to associated pages. If any difference, we expected the deep (resolved) filter to perform slightly lower in terms of recall than the shallow filter. This is due to the occasions where concept links that did not resolve with the deep filter were blindly treated as valid concept pages by the shallow filter.

Likewise, the shallow and deep filters with limitation to the introduction section of Wikipedia pages performed slightly lower than the full page equivalents. This result helps us to invalidate the hypothesis that the introduction section for concept pages fully captures the context of the concept as far as related concepts go, but we do acknowledge the improvement over the baseline filter method. We can, therefore, say that the introduction partition of a Wikipedia page may not be the most optimal solution in order to expand the context of the page concept. Note that at this point, we can only make this conclusion for the recall measure.

Spider Depth Performance

The deep filter with a spider depth of 2 (limited to the introduction section) performed better for recall than both the introduction section limitations of the shallow filter and the single deep filter. We can attribute this to the additional concepts that were aggregated from the second layer. We note that this recall value was not as high as the full page filter methods. This can be viewed as a confirmation that concepts that are further away from the focus concept are not as contextually relevant as those closer to the focus concept and, therefore, do not best match against as many relevant news items.

It would have been interesting to process the deep filter to two or more spider levels deep using the complete page. However, as we previously commented, this method would require computational resources beyond the restrictions of a desktop PC. A method to work towards this in the future would be to utilise the computation resources of a GRID cluster and redevelop the software tools to utilise this type of platform. However, we acknowledge, that would take significant research effort to develop a methodology to maintain such a large and interlinked graph over a distributed system.

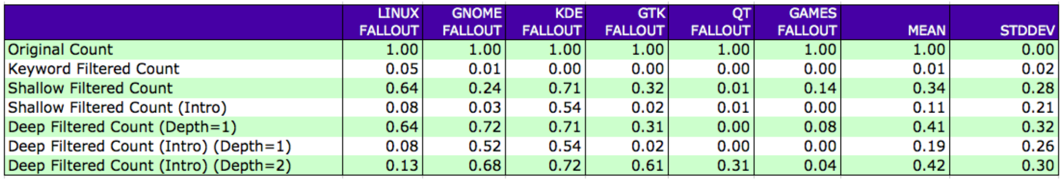
Final Comments for Section

Finally, the original count recall is 1.0. The original count represents a case where all items are captured, and, therefore, the recall value is 1.0 because all relevant items are captured for each topic.

In summary, we have demonstrated that the contextual filtering techniques provide an effective solution for news filtering in the OSNews web site environment in respect to the recall results. In addition, we have demonstrated that concepts further away from the focus concept in the Wikipedia graph are not as successful at matching contextually related news topics to the focus concept as those concepts that are closer to the focus concept.

7.2.2 Analysis of Fallout Measure

In this section we present a summary of the fallout measures for each filtering method across all topics.



	LINUX	GNOME	KDE	GTK	QT	GAMES		
	FALLOUT	FALLOUT	FALLOUT	FALLOUT	FALLOUT	FALLOUT	MEAN	STDDEV
Original Count	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Keyword Filtered Count	0.05	0.01	0.00	0.00	0.00	0.00	0.01	0.02
Shallow Filtered Count	0.64	0.24	0.71	0.32	0.01	0.14	0.34	0.28
Shallow Filtered Count (Intro)	0.08	0.03	0.54	0.02	0.01	0.00	0.11	0.21
Deep Filtered Count (Depth=1)	0.64	0.72	0.71	0.31	0.00	0.08	0.41	0.32
Deep Filtered Count (Intro) (Depth=1)	0.08	0.52	0.54	0.02	0.00	0.00	0.19	0.26
Deep Filtered Count (Intro) (Depth=2)	0.13	0.68	0.72	0.61	0.31	0.04	0.42	0.30

Figure 7.3: Fallout Mean Average of Experimental Methods.

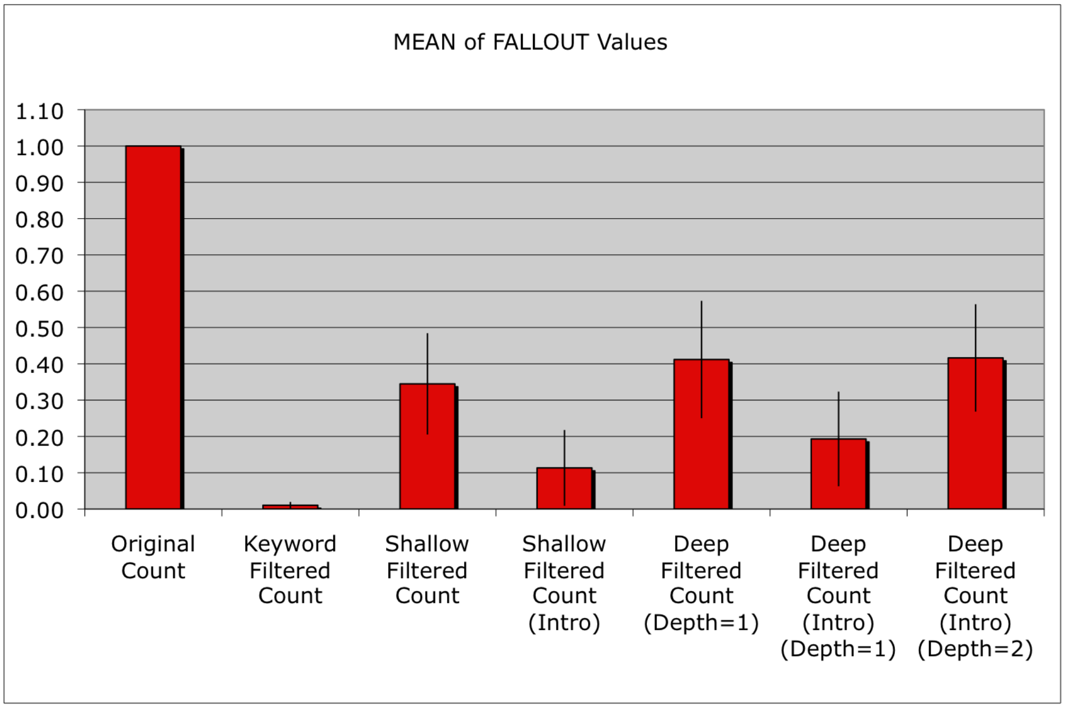


Figure 7.4: Fallout Mean Average Graph of Experimental Methods. Bars represent 1 Standard Deviation.

Baseline

Figure 7.4 demonstrates that the keyword filter has the lowest average value with respect to fallout measure. We can conject from this result that the OSNews web site editors have, in practice, consistently allocated news items containing a site topic keyword to that topic. For example, topics containing the keyword *Linux* would be allocated to the Linux topic category.

Introduction Section Partitioning

We can observe that for both single depth filters (Shallow Filtered and Deep Filtered with Depth=1) the fallout is lower for the introduction section partitioned variations than the full page equivalents. This result contributes to our demonstration that discriminate partitioning of the Wikipedia page is required in order to focus the context of concept expansion.

We can confirm from our results that Gabrilovich and Markovitch (2006)'s hypothesis holds true. To recap Gabrilovich and Markovitch (2006) proposed that:

> *"Given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from a concept is ill-advised, as this would collect a lot of weakly related material."*

Spider Depth Performance

Spidering the Wikipedia page introduction section out-links to 2 levels deep produced an improvement in recall over the 1 level equivalents. This recall value, however, was not as high as full page out-link expansion method. We did observe, in addition, that this method produced a higher fallout average value than all other methods tested; this was expected due to our hypothesis that items spidered to larger depths are less contextually relevant to the focus concept than concepts at smaller depths.

Final Comments for Section

Finally, the original count fallout is 1.0. The original count represents a case where all items are captured, and, therefore, the fallout value is 1.0 because all irrelevant items are captured.

7.2.3 Summary of Recall and Fallout across Experimental Filtering Methods

Upon analysis of the Recall and Fallout mean averages across filtering methods, we observed the following features of the filtering methods:

1. All methods tested improved the recall for the filtering of news topics over the baseline, however, full page variations of the filters performed better than introduction section variations.

2. Spidering the Wikipedia page introduction section out-links to 2 levels deep produced an improvement in recall over the 1 level equivalents. This recall value, however, was not as high as full page out-link expansion method. We did observe, in addition, that this method produced a higher fallout average value than all other methods tested; this was expected due to our hypothesis that items spidered to larger depths are less contextually relevant to the focus concept than concepts at smaller depths.

3. The partitioning of the Wikipedia page improved the fallout of the filter over the full page equivalent filter. However, this partitioning produced a recall value that only marginally improved over the baseline keyword filter and was below that of their full page equivalents. We, therefore, have a trade-off, between recall and fallout depending on how we partition the Wikipedia page. This opens up an avenue for future work on a partitioning strategy for Wikipedia pages.

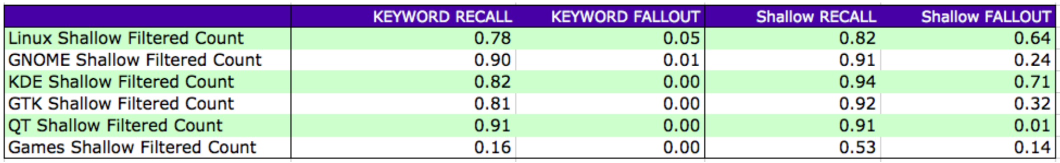
As can be seen, the properties that we present in this section have been demonstrated for the mean recalls and fallouts for the tests conducted. As topics were chosen based on their semantic properties, we analyse the results; focussing on one filtering method across each topic in the following section.

7.3 Summary of Shallow Filter Method

In this section we focus on a partial analysis of our context expansion method with focus specifically on the shallow method. This method, aside from the baseline keyword filter method, was the most resource efficient due to the requirement of software implementation. This efficiency is due to the software only needing to scan a Wikipedia concept page once for the focus topic, rather than having to resolve each link.

This method, therefore, is the method that has scope to scale to cover more topics in future work and implementation, due to the deep filtering methods requiring many lookups and resolving of each Wikipedia concept linked to from the focus concept page. This method also produced a higher mean average recall than all other filter methods.

The tabular data for this summary can be shown in **Figure 7.5** and in graph form in **Figure 7.6**.



	KEYWORD RECALL	KEYWORD FALLOUT	Shallow RECALL	Shallow FALLOUT
Linux Shallow Filtered Count	0.78	0.05	0.82	0.64
GNOME Shallow Filtered Count	0.90	0.01	0.91	0.24
KDE Shallow Filtered Count	0.82	0.00	0.94	0.71
GTK Shallow Filtered Count	0.81	0.00	0.92	0.32
QT Shallow Filtered Count	0.91	0.00	0.91	0.01
Games Shallow Filtered Count	0.16	0.00	0.53	0.14

Figure 7.5: Summary of Shallow Filter Method.

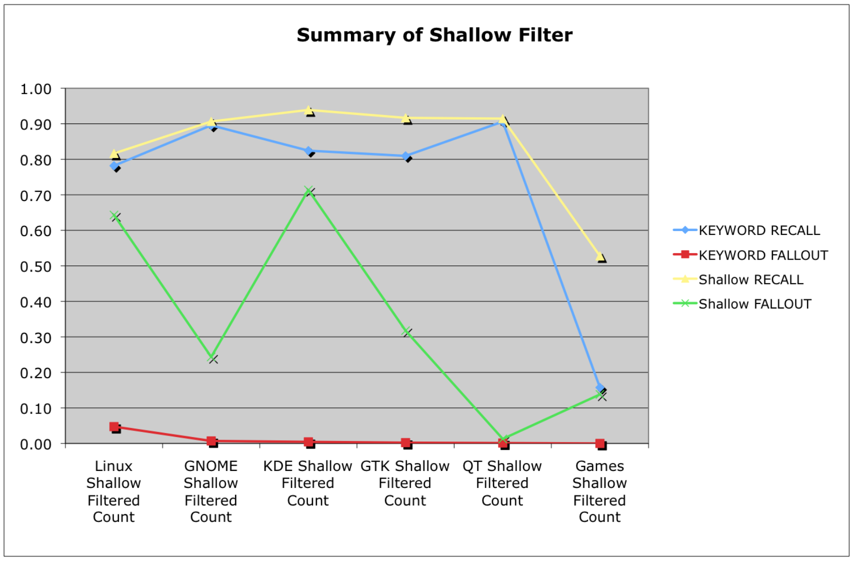


Figure 7.6: Summary of Shallow Filter Method Graph.

Recall

The graph of the recalls and fallouts (**Figure 7.6**) demonstrates that there is a differing success in the recall of various topics for the shallow filter method. There is little variance in the GNOME, KDE, GTK and QT topics. Linux and Games, however, produced a lower recall value with Games producing a recall of almost half of the other topics. Linux produced a lower recall value than GNOME, etc., but this value is not as drastic as Games. Looking at the data, we can see that there is a correlation between the plots of Keyword Filter Recall and Shallow Filter Recall.

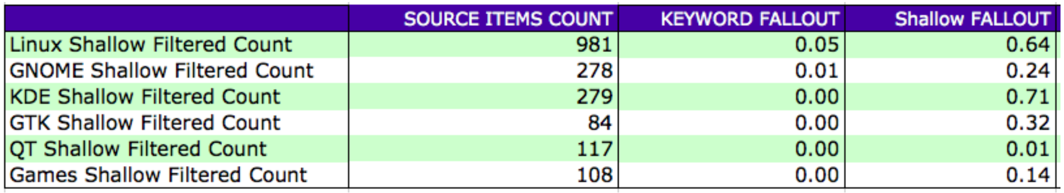
One hypothesis that we can propose is that Games is a generic topic, with particular news stories referring to a title of a game that is of interest, whereas GNOME, KDE, GTK and QT will be used explicitly in news titles due to the fact that they are project/product names. From this hypothesis we can conject that due to the generic nature of the games topic, the context expander will need to not only match the topic Games, but also the related game title or game publisher names in order to match the news titles. Linux is a product name, but due to it referring to an operating system may encompass a number of sub or related projects that can be described using differing keywords.

Fallout

For the fallout measures produced from from the shallow filter method, there does not appear to be any consistency of value across the different topics, nor does there appear to be any correlation between the keyword filter and the shallow filter for fallout. Due to this lack of perceived consistency across topics in regards to the fallout value, we pay further attention to this measure and perform further analysis in the following section.

Fallout of Tests against Source Items Count

In this section we present a view across the topics tested for the source item count against fallout. We obtain this data from the shallow filter method. The tabular data for this summary can be shown in **Figure 7.7**. The motivation behind this analysis is to better understand why the Games topic fallout value was so low in the previous test. One view that we will look at is the relationship between the total number of items per topic category and the fallout for the shallow filter method.



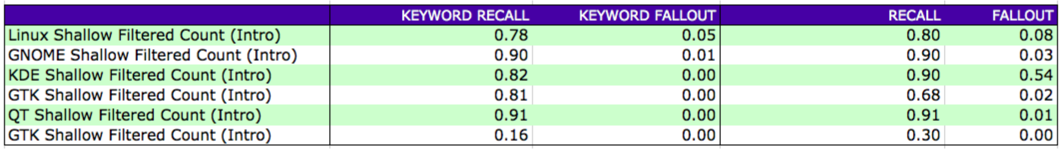
	SOURCE ITEMS COUNT	KEYWORD FALLOUT	Shallow FALLOUT
Linux Shallow Filtered Count	981	0.05	0.64
GNOME Shallow Filtered Count	278	0.01	0.24
KDE Shallow Filtered Count	279	0.00	0.71
GTK Shallow Filtered Count	84	0.00	0.32
QT Shallow Filtered Count	117	0.00	0.01
Games Shallow Filtered Count	108	0.00	0.14

Figure 7.7: Table of Fallout with Source Item Count.

Looking at the graph (**Figure 7.6**) it is difficult to tell if there is a trend between source items and fallout for the shallow filter method. This difficulty can be seen illustrated through the fluctuating fallout (shallow fallout) values across all topics. A reason for this issue is that there is a clustering of the topics chosen between 100 and 150 topics with a few outlying counts. In order to better analyse this data we would need to choose a wider range of topics.

7.4 Summary of Shallow Filter (Introduction Section) Method

In this test, we repeat the analysis of the shallow filter method, but this time we use data from the shallow filter method limited to just the introduction section of each Wikipedia page.



	KEYWORD RECALL	KEYWORD FALLOUT	RECALL	FALLOUT
Linux Shallow Filtered Count (Intro)	0.78	0.05	0.80	0.08
GNOME Shallow Filtered Count (Intro)	0.90	0.01	0.90	0.03
KDE Shallow Filtered Count (Intro)	0.82	0.00	0.90	0.54
GTK Shallow Filtered Count (Intro)	0.81	0.00	0.68	0.02
QT Shallow Filtered Count (Intro)	0.91	0.00	0.91	0.01
GTK Shallow Filtered Count (Intro)	0.16	0.00	0.30	0.00

Figure 7.8: Summary of Shallow Filter (Introduction Section) Method.

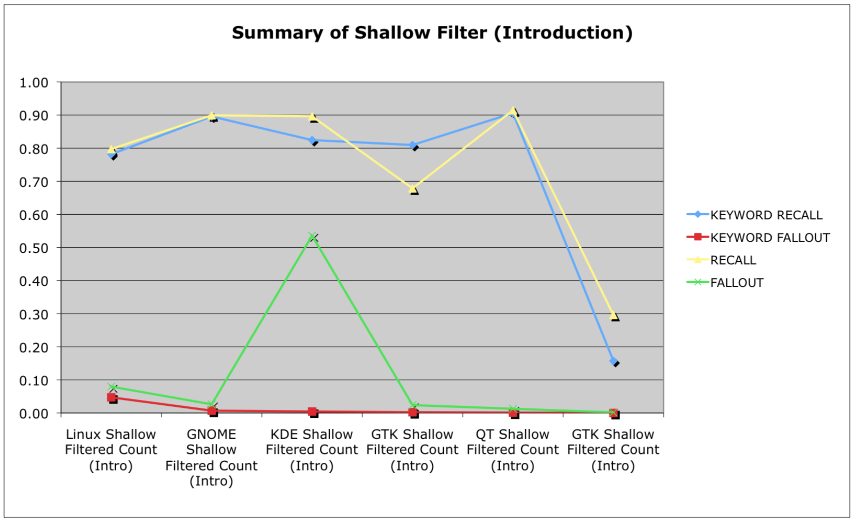


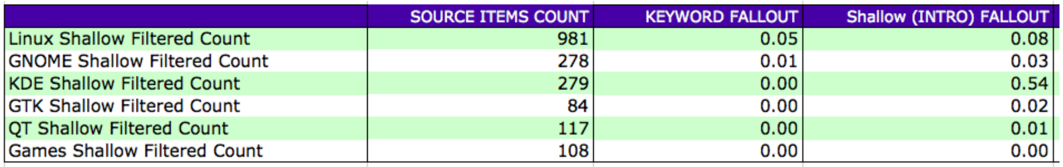
Figure 7.9: Summary of Shallow Filter (Introduction Section) Method Graph.

As with the shallow filter method, we can see that there is a correlation between the graphs of Keyword Filter Recall and Shallow Filter Recall.

For the fallout measures produced from from the shallow filter (introduction) method, there appears to be a correlation between the keyword filter and the shallow filter for fallout, save for one anomalous value for KDE. At this point we are unable to explain this anomalous value without further investigation.

Fallout of Tests against Source Items Count

In this section we present a view across the topics tested for the source item count against fallout. We obtain this data from the shallow filter (introduction) method. The tabular data for this summary can be shown in **Figure 7.10**.



	SOURCE ITEMS COUNT	KEYWORD FALLOUT	Shallow (INTRO) FALLOUT
Linux Shallow Filtered Count	981	0.05	0.08
GNOME Shallow Filtered Count	278	0.01	0.03
KDE Shallow Filtered Count	279	0.00	0.54
GTK Shallow Filtered Count	84	0.00	0.02
QT Shallow Filtered Count	117	0.00	0.01
Games Shallow Filtered Count	108	0.00	0.00

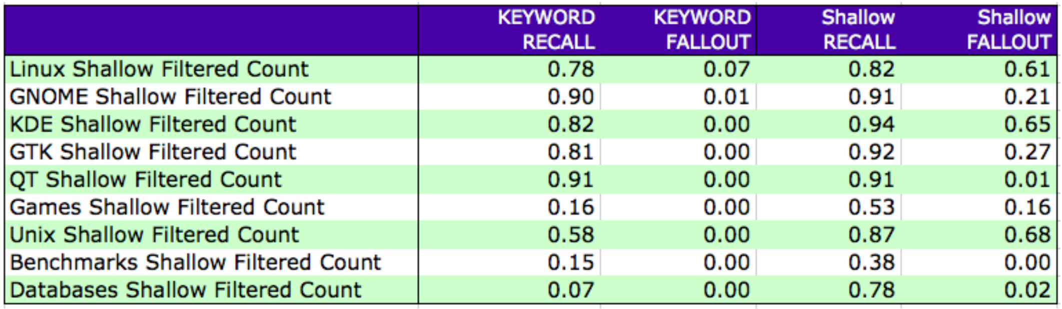
Figure 7.10: Table of Fallout with Source Item Count.

7.5 Expanded Tests for Shallow Method

Following the previous analysis, we decided to repeat the testing for the shallow methods, but this time we added more topics, as the shallow methods take a relatively short time to process. These tests are aimed at better informing out evaluation twofold:

1. We hope to better determine if there is a reason why generic topics produce lower recalls than specific topics.

2. We hope to provide a greater sample to plot source items count against fallout values.



	KEYWORD RECALL	KEYWORD FALLOUT	Shallow RECALL	Shallow FALLOUT
Linux Shallow Filtered Count	0.78	0.07	0.82	0.61
GNOME Shallow Filtered Count	0.90	0.01	0.91	0.21
KDE Shallow Filtered Count	0.82	0.00	0.94	0.65
GTK Shallow Filtered Count	0.81	0.00	0.92	0.27
QT Shallow Filtered Count	0.91	0.00	0.91	0.01
Games Shallow Filtered Count	0.16	0.00	0.53	0.16
Unix Shallow Filtered Count	0.58	0.00	0.87	0.68
Benchmarks Shallow Filtered Count	0.15	0.00	0.38	0.00
Databases Shallow Filtered Count	0.07	0.00	0.78	0.02

Figure 7.11: Summary of Shallow Filter Method with Expanded Topics.

From the graph in **Figure 7.12** the recall and fallout values for the previous topics followed a similar pattern to the previous tests, despite the extra irrelevant news items introduced through the introduction of the new topics.

We do notice, from the graph (**Figure 7.12**), that UNIX performed similarly to Linux for both recall and fallout. It should be noted here that UNIX and Linux are both Operating Systems and are semantically similar. We noticed that the generic topics, benchmarks, performed similarly to Games.

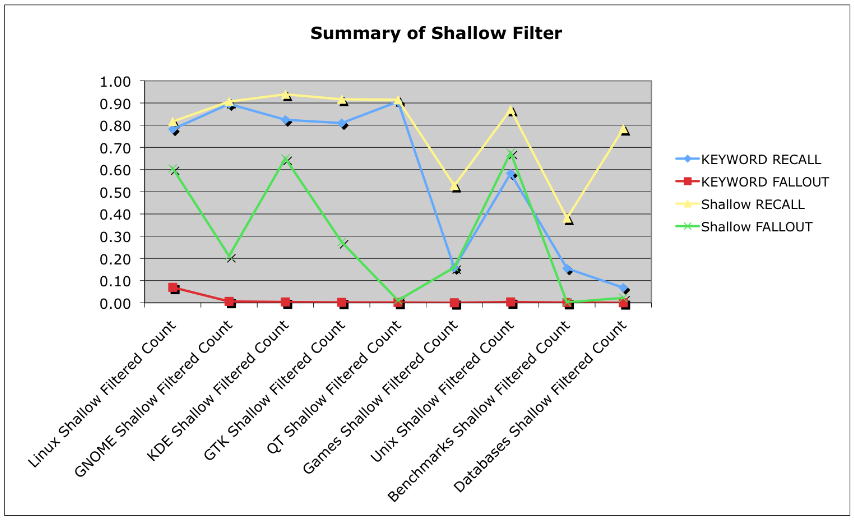
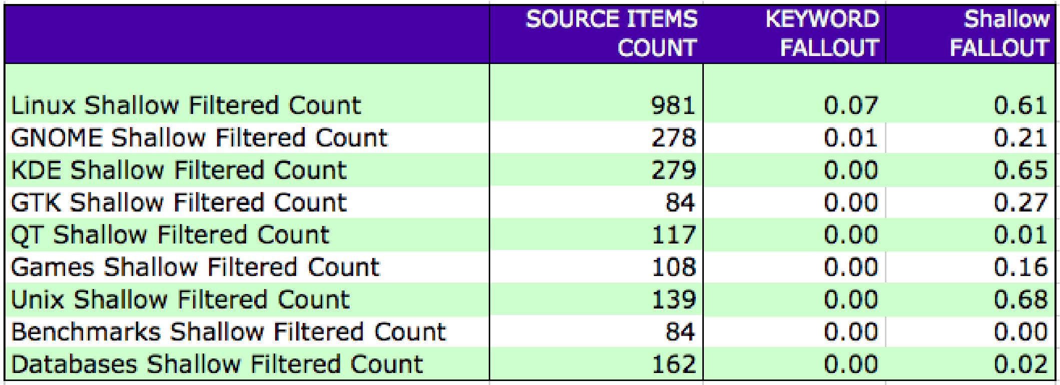


Figure 7.12: Summary of Shallow Filter Method Graph with Expanded Topics.

Fallout of Tests against Source Items Count

In this section we present a view across the topics tested for the source item count against fallout. We obtain this data from the shallow filter (introduction) method. The tabular data for this summary can be shown in **Figure 7.13**.



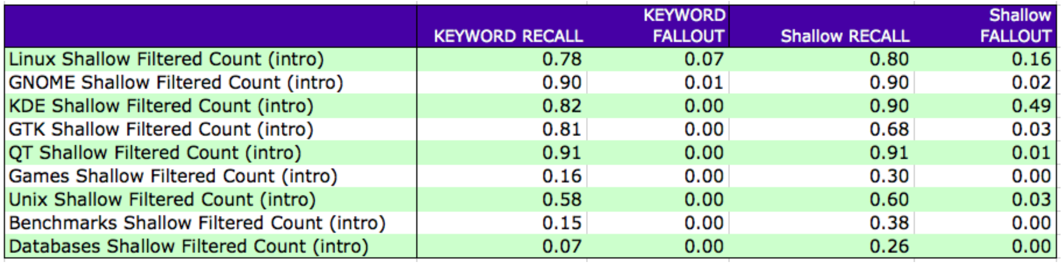
	SOURCE ITEMS COUNT	KEYWORD FALLOUT	Shallow FALLOUT
Linux Shallow Filtered Count	981	0.07	0.61
GNOME Shallow Filtered Count	278	0.01	0.21
KDE Shallow Filtered Count	279	0.00	0.65
GTK Shallow Filtered Count	84	0.00	0.27
QT Shallow Filtered Count	117	0.00	0.01
Games Shallow Filtered Count	108	0.00	0.16
Unix Shallow Filtered Count	139	0.00	0.68
Benchmarks Shallow Filtered Count	84	0.00	0.00
Databases Shallow Filtered Count	162	0.00	0.02

Figure 7.13: Table of Fallout with Source Item Count.

Despite the addition of extra plot points in the graph for source items count against fallout, there is still no discernible pattern between both axes.

7.6 Expanded Tests for Shallow (Introduction Section) Method

In this section we present a view across the expanded topics tested for the source item count against fallout. This time, we use data from the shallow filter method limited to just the introduction section of each Wikipedia page. We obtain this data from the shallow filter (introduction) method. The tabular data for this summary can be shown in **Figure 7.14** and in graph form in **Figure 7.15**.



	KEYWORD RECALL	KEYWORD FALLOUT	Shallow RECALL	Shallow FALLOUT
Linux Shallow Filtered Count (intro)	0.78	0.07	0.80	0.16
GNOME Shallow Filtered Count (intro)	0.90	0.01	0.90	0.02
KDE Shallow Filtered Count (intro)	0.82	0.00	0.90	0.49
GTK Shallow Filtered Count (intro)	0.81	0.00	0.68	0.03
QT Shallow Filtered Count (intro)	0.91	0.00	0.91	0.01
Games Shallow Filtered Count (intro)	0.16	0.00	0.30	0.00
Unix Shallow Filtered Count (intro)	0.58	0.00	0.60	0.03
Benchmarks Shallow Filtered Count (intro)	0.15	0.00	0.38	0.00
Databases Shallow Filtered Count (intro)	0.07	0.00	0.26	0.00

Figure 7.14: Summary of Shallow Filter (Introduction Section) Method with Expanded Topics.

As with the expanded full page test, the introduction test performed similarly for the already existing topic s in respect to the recall values. For the additional topics, the fallout values closely followed the keyword baseline, with the only exception being the aforementioned KDE value.

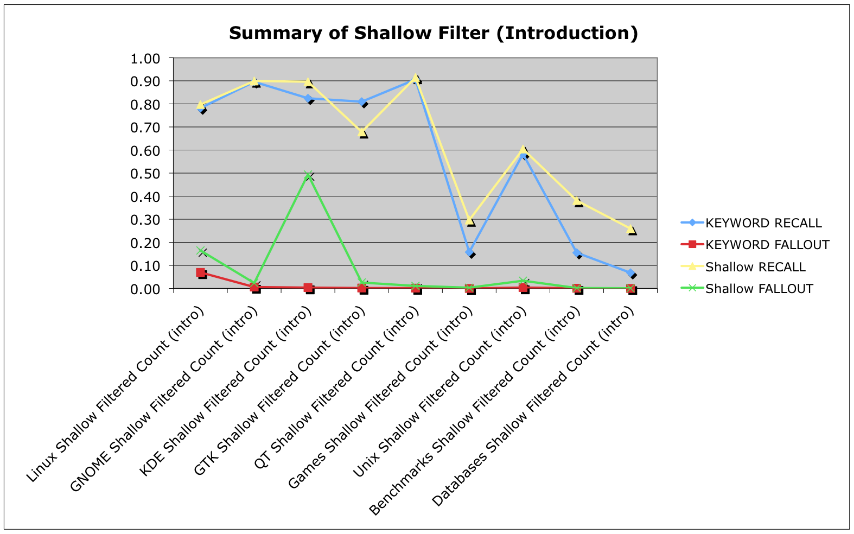
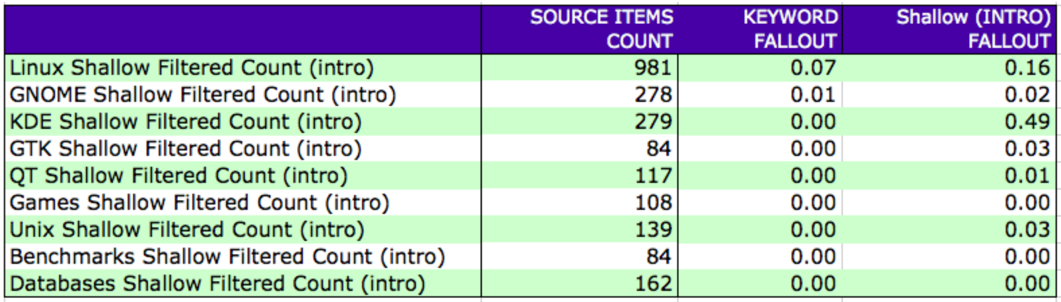


Figure 7.15: Summary of Shallow Filter (Introduction Section) Method Graph with Expanded Topics.

Fallout of Tests against Source Items Count

In this section, we repeat the tabulation for source items against fallout for the extended test results, but this time show the values obtained from the introduction section variant of the shallow filter in **Figure 7.16**.



	SOURCE ITEMS COUNT	KEYWORD FALLOUT	Shallow (INTRO) FALLOUT
Linux Shallow Filtered Count (intro)	981	0.07	0.16
GNOME Shallow Filtered Count (intro)	278	0.01	0.02
KDE Shallow Filtered Count (intro)	279	0.00	0.49
GTK Shallow Filtered Count (intro)	84	0.00	0.03
QT Shallow Filtered Count (intro)	117	0.00	0.01
Games Shallow Filtered Count (intro)	108	0.00	0.00
Unix Shallow Filtered Count (intro)	139	0.00	0.03
Benchmarks Shallow Filtered Count (intro)	84	0.00	0.00
Databases Shallow Filtered Count (intro)	162	0.00	0.00

Figure 7.16: Table of Fallout with Source Item Count.

7.7 Overall Assessment of Experimentation

In regards to the experiments conducted in the course of this thesis, we can conclude that all methods have proven successful in evaluating the tradeoffs of using the proposed contextual concept expansion method. Overall, from our analysis of the experiments performed as part of this thesis, we summarise to the points below:

- *All filter methods improved over the baseline.* All methods tested improved the recall for the filtering of news topics over the baseline, however, full page variations of the filters performed better than introduction section variations.

- *The shallow method was the most effective in terms of recall.* Despite taking longer to process, the deep filtered version of the full page filter did not perform significantly more effectively than the shallow method.

- *Spidering to 2 levels deep proved effective at aggregating weakly related links.* Spidering the Wikipedia page introduction section out-links to 2 levels deep produced an improvement in recall over the 1 level equivalents. This recall value, however, was not as high as full page out-link expansion method. We did observe, in addition, that this method produced a higher fallout average value than all other methods tested; this was expected due to our hypothesis that items spidered to larger depths are less contextually relevant to the focus concept than concepts at smaller depths.

- *Partitioning the Wikipedia page using the introduction was partially effective in prioritising contextually related concepts.* The partitioning of the Wikipedia page improved the fallout of the filter over the full page equivalent filter. However, this partitioning produced a recall value that marginally improved over the baseline keyword filter and was below that of their full page equivalents. We, therefore, have a tradeoff, between recall and fallout depending on how we partition the Wikipedia page. This opens up an avenue for future work on a partitioning strategy for Wikipedia pages.

- *We found that there was little relationship between the fallout value and the number of source items.*

7.8 Limitation of the Experimentation

Whilst the filtering methods demonstrated an improvement over the baseline filtering method, there was a large standard deviation for recall and fallout for the mean averages of both. In order to improve the accuracy of the tests, a larger topic and news item set would have prompted a more conclusive result.

Whilst we were able to make conclusions for the filtering of news items for OSNews, we would like to see in the future if we receive comparable performance results across different datasets. We would like to test the methods against a richer range of topics across domains, rather than the limited OSNews domain.

It would have been interesting to process the deep filter to two or more spider levels deep using the complete page. However, as we previously commented, this method would require computational resources beyond the restrictions of a desktop PC. A method to work towards this in the future would be to utilise the computation resources of a GRID cluster and redevelop the software tools to utilise this type of platform. However, we acknowledge, that would take significant research effort to develop a methodology to maintain such a large and interlinked graph over a distributed system.

Chapter 8

Conclusion and Reflection

This thesis has presented the development of a framework in order to contextually expand concepts through the use of the Wikipedia knowledge-base. As has been verified Wikipedia is a weak knowledge-base due to its simple TBox and implicit predicates. This framework has been developed in response to the information overload problem and consequently, has been evaluated through the use of a case-study consisting of topic-based web feed filtering.

Chapter 2 begins with an overview of the area of information consumption within service based environments. In this section, we present the notion of information information overload and how this is a problem that we are increasingly encountering within information environments. We go on to present the WWW and highlight the motivations behind its development; we then present some of the limitations of the WWW regarding information processing.

Information Retrieval is a focus for this chapter in order for us to better understand how current information filtering and retrieval is achieved. We presented a number of commonly used and well understood measures for information retrieval. Regarding the WWW and Semantic Web, we presented an overview of the differences between statistical and logical processing of the information present on them.

The development of the Semantic Web is presented along with the motivations behind its development. Whilst we initially focus on a technical overview of the Semantic Web stack and associated technologies, we then proceed to discuss a number of cited issues that are present in the implementation of the Semantic Web vision. Web 2.0 was presented with associated developments as a bottom-up social alternative to the formal Semantic Web. We discussed a number of benefits that Web 2.0 presents over the Semantic Web and the limitations of Web 2.0

Finally, we explored the way that concepts can be represented in information systems and looked at how they are commonly compared for similarity. In this chapter, we cited the Ogden Meaning Triangle to enable us to better understand how concepts relate to words and the world. We then preceded to explore current work in knowledge-bases and present a number of relevant internet-based options. For each of these knowledge-bases we explored the area of concept similarity and present current research pertaining to each type of knowledge-base.

In Chapter 3, we introduced the notion of context in information retrieval. We began with an understanding of the notion of context. Existing definitions of context were explored, which allowed us to understand not only dictionary definitions of context, but also semantically what context means regarding related objects. We presented a definition of context for this thesis before giving an overview as to how context helps us to understand the relationship between an object and its world.

Focus was then passed on to what context means for information search. We cited a number of properties of contextual information search including personalised filtering and historical context.

In Chapter 4, the development of the initial context model developed for this thesis was discussed. We presented our F-O model and explored how this helps us to represent contextual clouds in a lightweight manner through the relaxing of predicate semantics. The latter half of Chapter 4 focussed on our experience with using a taxonomy to determine contextual similarity between concepts and introduced our method of contextualisation of similarity calculation for concepts . Through the use of prototyping, we uncovered a number of problems inherent with a rigid taxonomy and provide motivation that would ultimately lead us to favour the wiki structure over the taxonomy structure.

Wiki-based Knowledge Processing was presented in Chapter 5. We explored existing methods of determining semantic similarity within Wikipedia. We reviewed these methods and proposed a new novel method to allow the semantic similarity between and expansion of concepts to be achieved. We presented our methods and explained why the concept expansion method was the one chosen for experimental validation within this thesis.

In Chapter 6, we presented the experimental method for our proposed concept expansion method. We introduced a news feed filtering case study utilising source news items from the OSNews technology news website. From this case study we were able to design a number of experiments along with an associated baseline experiment to allow us to exploit our overarching concept expansion method through the development of a number of news item filters.

Experimental analysis was presented in Chapter 7, where the effectiveness of the proposed experimental filtering methods was evaluated across our selected OSNews topics. For each filter method, we were able to take a mean average of the recall and fallout measures across all topics and then make conclusions based on these views. We found that all filter methods improved recall over the baseline, and that full-page filters performed more effectively than introduction section filters for recall. For fallout, we found that the introduction section filters were more effective, due to the contextual prioritisation of topics in the introduction lead section of the Wikipedia page.

8.1 Satisfaction of Research Objectives

For this thesis, a number of research aims and objectives have been achieved. We summarise the contributions towards these below:

- *Critique current literature in the field to understand past and current practice in the areas of the conceptualisation of a single object and understanding the* *context under which that object exists (or can exist). Furthermore this critique will focus on the topics of information filtering and concept comparison based on a knowledge-base.* - This objective has been satisfied in **Chapter 2** and **Chapter 3** through the exploration of current literature in different paradigms concept representation; for instance, Semantic Web, Web 2.0 tagging and taxonomies. To give an example, the Ogden Meaning Triangle has been used to enable understanding of the term *'concept'* for the work within this thesis. We understand that the concept that a word refers to may only exist as an abstract concept - this concept buffer between a symbol and a thing is fundamental as we are not necessarily talking about a physical instance of a real-world object, but rather to an abstract definition of a concept, which can be equated to a mental model.

- *Establish the current limitations of the field in areas described in the previous point including identifying shortfalls of existing knowledge-bases for concept comparison.* - This objective has been achieved in a number of sections within this thesis. 1) **Section 2.6** and **Section 2.3.2** explore the limitations of ontology-based knowledge-bases with complex TBox models have been discussed, focussing on reasoning across large scale knowledge-bases. 2) Limitations of taxonomy and semantic lexicon (WordNet) approaches to similarity have been discussed in **Section 2.7** and **Section 5.8.1**, with the Wiki weak knowledge-base structure justified as a more appropriate structure to use for mining contextually related concepts due to the ease of collaborative knowledge creation that it enables.

- *Develop a solution that provides a clear conceptual identification of an object including implicit situational information and detail of surrounding objects.* - This objective has been satisfied in **Chapter 4**. In this chapter the F-O context model has been derived based on a study of existing related work in context modelling; this has influenced the conceptual and structural development of the final F-O model which provides a contextual cloud model based on relaxed predicate semantics.

- *Verify if a weakly formalised knowledge-base (by implementation, Wikipedia) is a suitable a knowledge-base to base the above concept expansion and comparison framework upon.* - This objective has been satisfied in **Chapter 5**. This chapter has explored human judgement in both the Wikipedia knowledge-base and also to the Web 2.0 style tagging of resources based on existing peer reviewed work. Wikipedia has been validated as a weekly formalised knowledge-base based on the characteristics of a knowledge-base introduced in **Section 2.6**, specifically Wikipedia has been classified as a weak knowledge-base due to its simple TBox schema and implicit predicates.

- *Utilise the weakly formalised Wikipedia knowledge-base for the basis of a framework to expand out a given concept using contextual information provided by the knowledge-base. In order to achieve this, the derivation of contextual concept information is based on the topological structure of the wiki knowledge-base.* - This objective has been satisfied in **Chapter 5**. In this chapter a proposed method to expand out a given concept based on neighbouring concepts in the Wikipedia knowledge-base has been developed. This has been based on a study of existing work on concept relatedness and has applied a shortest path heuristic to the Wikipedia concept linking structure. The developed F-O context model has been used here as a foundation for machine processing of this contextual concept data.

- *Develop a web feed aggregator to utilise the developed concept expansion framework. This will allow the experimental testing of the developed concept expansion technique and web feed filters using it.* - This objective has been satisfied in **Chapter 6**. A simple web feed aggregator has been built using the Java programming language that we call **DAV**e's **R**ss **O**rganisation **S**ystem or DAVROS-2. DAVROS-2 is designed to aggregate web feeds from a given URL and store these items internally. The tool then accepts a concept as a focus for filtering. The tool then calls a developed sub-system to perform the concept expansion based upon a cached copy of Wikipedia. Once the expanded list of concept keywords is built, the web feed item titles are compared against the concept keyword list and then boolean filtered based on this concept keyword comparison.

- *Develop a testbed environment to demonstrate specific tests used within the experimental aspects of this investigation. Additionally determine measures to evaluate the effectiveness of the web feed filtering solution.* - This objective has been satisfied in **Chapter 6**. The practical experimentation has provided a mechanism to demonstrate the validity of the proposed experimental filtering methods through the production of empirical data. For the experiment, a scenario has been proposed in order to demonstrate the context-orientated solution to the information overload and filtering problem presented throughout this thesis.

8.2 Reflection on Hypotheses

In this thesis we proposed the following research hypotheses in order to assist research and validation into the areas explored in this thesis:

> 1: *In order to perform contextual comparisons between concepts in a non-domain independent manner, there is the need to utilise a knowledge-base of community maintained world knowledge. The second part of this hypothesis is that a weak knowledge-base is a satisfactory knowledge-base for the task.*

In **Chapter 2** and **Chapter 3** current literature in different paradigms concept representation; for instance, Semantic Web, Web 2.0 tagging and taxonomies has been explored. Limitations of ontology-based knowledge-bases with complex TBox models have been discussed in the context of reasoning across large scale knowledge-bases with the conclusion that a knowledge-base with a simpler TBox model lends itself better to reasoning over a large knowledge-base of world common-sense knowledge. We have explored human judgement in both the Wikipedia knowledge-base and also to the Web 2.0 style tagging of resources based on existing peer reviewed work. Wikipedia has been validated as a knowledge-base based on the characteristics of a knowledge-base introduced in **Section 2.6**.

In summary, Wikipedia has been validated as a world knowledge-base and existing work on the quality of human judgements in Wikipedia has been reviewed. This knowledge-base has been demonstrated through the population of our F-O context model and applied to the problem of information overload through a news filtering scenario.

> 2: *Large scale wikis (for instance Wikipedia) provide a structure that can be interrogated for contextual information in a manner that cannot easily be achieved through: an unstructured WWW; traditional thesauri; or through the Semantic Web that is richly structured but in a fragmented manner.*

> 3: Gabrilovich and Markovitch (2006) propose the following unverified hypothesis in the context of a wiki structure, that this thesis aims to verify: *“given a concept, we would like to use related articles to enrich its text. However, indiscriminately taking all articles pointed from a concept is ill-advised, as this would collect a lot of weakly related material.*"

In this thesis we have demonstrated the effectiveness of our concept expansion method in respect to the above two hypotheses. In order to achieve this we conducted experimental tests in respect to the proposed case study. We designed tests in order to evaluate both recall and fallout measures. Whilst we demonstrated that the recall value for the filtering methods, particularly the full page method, provided an improvement over the baseline, the fallout of the methods, highlighted that there is scope for improvement. We envisage this improvement to come from the the optimisation of a Wikipedia page partitioning strategy.

We view this point as a confirmation of the criteria for success involving the development of a method and framework to expand concepts with contextually related concepts. We also demonstrated the success of the software tool developed during this thesis (DAVROS-2) in implementing our proposed and developed concept expansion methods and, in addition, the enabling of empirical testing of the methods through the tool.

In this thesis we have validated this hypothesis through the use of the framework and associated context expansion method. Through this method and our experimentation we have been able to evaluate the success in collecting related concepts based on the user’s focus concept.

Regarding the comparison between concepts, we only achieved a partial solution due to the analysed complexities and processing requirements of processing a graph with the size that Wikipedia encompasses. Unfortunately, for this thesis we were unable to experimentally evaluate the envisaged methods of creating multiple contextual chains through the Wikipedia graph. However, we did achieve some success with the evaluated concept expansion method of determining if concepts further away from the focus concept are less contextually relevant than nearer concepts. This has been assisted through the use of our proposed F-O model and context cloud.

8.3 Future Work

In this work, we have demonstrated the novelty of our contextual concept expansion method. Regarding future work we divide this commentary into a number of sub areas to aid in discussion:

- *Expansion of page partitioning strategy.* - In this thesis, we have demonstrated the limited success of partitioning of a Wikipedia page. For this thesis we partitioned the page using the introduction section as a partition. In future work we would like to invest resources into the semantic analysis of Wikipedia pages in order to better focus the contextual expansion of concepts using out linking concept links from the Wikipedia page partition.

- *Implementation of concept expansion method using full page deep spidering.* - Whilst we were able to exploit the depth of spidering Wikipedia pages to more than one level deep, we were limited to just the introduction partition for 2 levels deep due to computational processing requirements. In future work we would like to spider to a larger number of depth levels with the usage of the full page of resolved concept links.

- *Development of the proposed concept comparison method.* - Finally, we would like to invest resources in the development of the proposed concept comparison method proposed in this thesis. This method, above all, would require a significant computational resource and a robust and distributed graph building and pathfinding algorithm to uncover routes between nodes in the concept graph. We would like to not only determine the shortest path between noted, but also examine the semantic implications of taking different and influenced paths through the graph.

- *Evaluation of contextual concept expansion technique on a strongly formalised knowledge-base.* - A limitation of the proposed approach within this thesis is due to the chain of decisions that determined the use of Wikipedia as a weakly formalised knowledge-base. This is for the requirement of a knowledge-base that contains a comprehensive collection of world concepts, for instance events, influential people and products. The knock-on effect of this is that the contextual comparison and evaluation framework will be based on a weakly formalised knowledge-base. The limitation, therefore, is that the contextual concept expansion technique was only evaluated against a weakly formalised knowledge-base. In the future, if a sufficiently comprehensive strongly formalised knowledge-base of world concepts is developed, then future work will take the opportunity to assess this technique against that class of knowledge-base.

Bibliography

Adams, D.: 1979, *The Hitchhiker's Guide to the Galaxy*, Pan Books.

Akerkar, R. and Sajja, P.: 2009, *Knowledge-Based Systems*, Jones and Bartlett Publishers International.

Allen, D.: 2003, *Getting Things Done: the Art of Stress-Free Productivity*, Penguin Books, New York.

Ambrosini, L., Cirillo, V. and Micarelli, A.: 1997, A hybrid architecture for user-adapted information filtering on the world wide web, *in* A. Jameson, C. Paris and C. Tasso (eds), *Proceedings of the Sixth International Conference on User Modeling (UM97)*, Springer, Berlin, pp. 59–61.

Amsler, R. A.: 1987, Words and worlds, *Proceedings of the 1987 workshop on Theoretical issues in natural language processing*, Association for Computational Linguistics, Morristown, NJ, USA, pp. 16–19.

Ankolekar, A., Krötzsch, M., Tran, T. and Vrandecic, D.: 2007, The two cultures: mashing up web 2.0 and the semantic web, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 825–834.

Antoniou, G. and Harmelen, F. v.: 2008, *A Semantic Web Primer, 2nd Edition (Cooperative Information Systems)*, The MIT Press.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D. and Patel-Schneider, P. F. (eds): 2003, *The Description Logic Handbook: Theory, Implementation and Applications*, Cambridge University Press. Second Edition, 2007.

Baader, F., Ganter, B., Sertkaya, B. and Sattler, U.: 2007, Completing description logic knowledge bases using formal concept analysis, *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 230–235.

Becker, C. and Nicklas, D.: 2004, Where do spatial context-models end and where do ontologies start? a proposal of a combined approach, *First International Workshop on Advanced Context Modelling, Reasoning And Management*, UbiComp.

Belkin, N. J., Muresan, G. and Zhang, X. M.: 2004, Using user's context for ir personalization, *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, ACM Press, pp. 23–25.

Benatallah, B., Hacid, M.-S., young Paik, H., Rey, C. and Toumani, F.: 2006, Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities, *Inf. Syst.* **31**(4), 266–294.

Berghel, H.: 1997, Cyberspace 2000: dealing with information overload, *Comm. ACM* **40**(2), 19–24.

Bergmann, R.: 2002, *Experience Management: Foundations, Development Methodology, and Internet-Based Applications*, Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Berners-Lee, T.: 1999, *Weaving The Web*, Orion Business.

Berners-Lee, T. and Miller, E.: 2002, The semantic web lifts off, *ECRIM News* (51), 9–11.

Berners-Lee, T., Hall, W., Hendler, J. A., O'Hara, K., Shadbolt, N. and Weitzner, D. J.: 2006, A framework for web science, *Foundations and Trends in Web Science*.

Berners-Lee, T., Hendler, J. and Lassila, O.: 2001, The Semantic Web, *Scientific American*.

Bourbaki, N.: 1992, Worse is better is worse, Retrieved August 18th, 2009, from `http://www.dreamsongs.com/Files/worse-is-worse.pdf`.

Brachman, R. and Levesque, H.: 2004, *Knowledge Representation and Reasoning (The Morgan Kaufmann Series in Artificial Intelligence)*, Morgan Kaufmann.

Buitelaar, P., Cimiano, P., Frank, A., Hartung, M. and Racioppa, S.: 2008, Ontology-based information extraction and integration from heterogeneous data sources, *Int. J. Hum.-Comput. Stud.* **66**(11), 759–788.

Buriol, L. S., Castillo, C., Donato, D., Leonardi, S. and Millozzi, S.: 2006, Temporal analysis of the wikigraph, *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society, Washington, DC, USA, pp. 45–51.

Bussler, C.: 2008, Is semantic web technology taking the wrong turn?, *IEEE Internet Computing* **12**(1), 75–79.

Chakrabarti, S., Punera, K. and Subramanyam, M.: 2002, Accelerated focused crawling through online relevance feedback, *Proceedings of the eleventh international conference on World Wide Web*, ACM Press, pp. 148–159.

Cho, J. and Tomkins, A.: 2007, Guest editors' introduction: Social media and search, *IEEE Internet Computing* **11**(6), 13–15.

Christensen, W.: 1992, Collection of memories of writing and running the first bbs by ward christensen (circa 1992), Retrieved August 18th, 2009, from `http://www.bbsdocumentary.com/software/AAA/AAA/CBBS/memories.txt`.

Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A.: 2003, Methodologies, tools and languages for building ontologies: where is their meeting point?, *Data Knowl. Eng.* **46**(1), 41–64.

Davies, J., Fensel, D. and van Harmelen, F.: 2003, *Towards the Semantic Web*, Wiley.

DCMI: 2008, Dublin core metadata element set, version 1.1: Reference descriptionk, Retrieved August 18th, 2009, from `http://dublincore.org/documents/dces/`.

De Bruijn, J., Eiter, T., Polleres, A. and Tompits, H.: 2007, Embedding non-ground logic programs into autoepistemic logic for knowledge-base combination, *IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 304–309.

Deng, X., Haarslev, V. and Shiri, N.: 2007, Measuring inconsistencies in ontologies, *ESWC '07: Proceedings of the 4th European conference on The Semantic Web*, Springer-Verlag, Berlin, Heidelberg, pp. 326–340.

Dey, A. K.: 2001, Understanding and using context, *Personal Ubiquitous Computing* **5**(1), 4–7.

Doan, A., Madhavan, J., Dhamankar, R., Domingos, P. and Halevy, A.: 2003, Learning to match ontologies on the semantic web, *The VLDB Journal* **12**(4), 303–319.

Echarte, F., Astrain, J. J., Cordoba, A. and Villadangos, J. E.: 2007, Ontology of folksonomy: A new modelling method., *in* S. Handschuh, N. Collier, T. Groza, R. Dieng, M. Sintek and A. de Waard (eds), *SAAKM*, Vol. 289 of *CEUR Workshop Proceedings*, CEUR-WS.org.

Egghe, L.: 2008, The measures precision, recall, fallout and miss as a function of the number of retrieved documents and their mutual interrelations, *Inf. Process. Manage.* **44**(2), 856–876.

Embley, D. W., Xu, L. and Ding, Y.: 2004, Automatic direct and indirect schema mapping: experiences and lessons learned, *SIGMOD Rec.* **33**(4), 14–19.

Fellbaum, C. (ed.): 1998, *WordNet: An Electronic Lexical Database (ISBN: 0-262-06197-X)*, first edn, MIT Press.

Fensel, D., Wahlster, W. and Lieberman, H. (eds): 2002, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, MIT Press, Cambridge, MA, USA.

Fileto, R., Liu, L., Pu, C., Assad, E. D. and Medeiros, C. B.: 2003, Poesia: An ontological workflow approach for composing web services in agriculture, *The VLDB Journal* **12**(4), 352–367.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E.: 2002, Placing search in context: the concept revisited, *ACM Trans. Inf. Syst.* **20**(1), 116–131.

Flake, G. W., Pennock, D. M. and Fain, D. C.: 2003, The self-organized web: The yin to the semantic webÃŢs yang, *IEEE Intelligent Systems* **18**(4), 75–77.

Gabariel, R. P.: 1989, Lisp: Good news, bad news, how to win big, Retrieved August 18th, 2009, from `http://www.dreamsongs.com/Files/LispGoodNewsBadNews.pdf`.

Gabrilovich, E. and Markovitch, S.: 2005, Feature generation for text categorization using world knowledge, *Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence*, Edinburgh, Scotland, pp. 1048–1053.

Gabrilovich, E. and Markovitch, S.: 2006, Overcoming the brittleness bottleneck using wikipedia: enhancing text categorization with encyclopedic knowledge, *Twenty-First AAAI Conference on Artificial Intelligence*.

Giunchiglia, F. and Shvaiko, P.: 2003, Semantic matching, *Knowl. Eng. Rev.* **18**(3), 265–280.

Goble, C.: 2003, Guest editorial: the semantic web: an evolution for a revolution, *Comput. Networks* **42**(5), 551–556.

Golder, S. A. and Huberman, B. A.: 2006, Usage patterns of collaborative tagging systems, *J. Inf. Sci.* **32**(2), 198–208.

Gong, L.: 2005, Contextual modeling and applications, *Systems, Man and Cybernetics, 2005 IEEE International Conference on* **1**, 381–386 Vol. 1.

Google: 2007, Google maps, Retrieved August 18th, 2009, from `http://maps.google.com/`.

Google: 2008, Google zeitgiest 2008, Retrieved August 18th, 2009, from `http://www.google.com/intl/en/press/zeitgeist2008/index.html`.

Greaves, M. and Mika, P.: 2008, Semantic web and web 2.0, *Web Semantics: Science, Services and Agents on the World Wide Web* **6**(1), 1 – 3. Semantic Web and Web 2.0.

Grossman, J. W.: 2007, Information about the erdos number project, Retrieved January 21st, 2007, from `http://www.oakland.edu/enp/`.

Gruber, T.: 2005, Ontology of folksonomy: A mash-up of apples and oranges, Retrieved August 18th, 2009, from `http://tomgruber.org/writing/ontology-of-folksonomy.htm`.

Gruhl, D., Meredith, D. N., Pieper, J. H., Cozzi, A. and Dill, S.: 2006, The web beyond popularity: a really simple system for web scale rss, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 183–192.

Guarino, N.: 1995, Formal ontology, conceptual analysis and knowledge representation, *Int. J. Hum.-Comput. Stud.* **43**(5-6), 625–640.

Guo, Y., Qasem, A., Pan, Z. and Heflin, J.: 2007, A requirements driven framework for benchmarking semantic web knowledge base systems, *IEEE Trans. on Knowl. and Data Eng.* **19**(2), 297–309.

Haarslev, V. and Möller, R.: 2001, Racer system description, *IJCAR '01: Proceedings of the First International Joint Conference on Automated Reasoning*, Springer-Verlag, London, UK, pp. 701–706.

Halpin, H., Robu, V. and Shepherd, H.: 2007, The complex dynamics of collaborative tagging, *WWW '07: Proceedings of the 16th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 211–220.

Hassan-Montero, Y. and Herrero-Solana, V.: 2006, Improving Tag-Clouds as Visual Information Retrieval Interfaces, *Proceedings of Multidisciplinary Information Sciences and Technologies, InSciT2006*, Merida, Spain.

Hepp, M., Siorpaes, K. and Bachlechner, D.: 2007, Harvesting wiki consensus: Using wikipedia entries as vocabulary for knowledge management, *IEEE Internet Computing* **11**(5), 54–65.

Hirst, G. and St-Onge, D.: 1997, Lexical chains as representation of context for the detection and correction malapropisms, *WordNet: An electronic lexical database and some of its applications. Cambrige, MA: The MIT Press.*

Honle, N., Kappeler, U.-P., Nicklas, D., Schwarz, T. and Grossmann, M.: 2005, Benefits of integrating meta data into a context model, *PERCOMW '05: Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications Workshops*, IEEE Computer Society, Washington, DC, USA, pp. 25–29.

Hu, M., Lim, E.-P., Sun, A., Lauw, H. W. and Vuong, B.-Q.: 2007, Measuring article quality in wikipedia: models and evaluation, *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, New York, NY, USA, pp. 243–252.

Huang, W. and Webster, D.: 2004, Enabling context-aware agents to understand semantic resources on the www and the semantic web., *Web Intelligence*, pp. 138–144.

Ingwersen, P. and Jarvelin, K.: 2004, Information retrieval in contexts, *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, ACM Press, pp. 6–9.

iri: 2004, Information retrieval in context workshop (irix), Retrieved August 18th, 2009, from `http://ir.dcs.gla.ac.uk/context/`.

Jiang, J. J. and Conrath, D. W.: 1997, Semantic similarity based on corpus statistics and lexical taxonomy, *eprint arXiv:cmp-lg/9709008*, pp. 9008+.

Jones, G. J. F. and Brown, P. J.: 2004, The role of context in information retrieval, *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, ACM Press, pp. 20–22.

Jones, K. S.: 2004, What's new about the semantic web?: some questions, *SIGIR Forum* **38**(2), 18–23.

Jung: 2005, Jung: Java universal network/graph framework, Retrieved August 18th, 2009, from `http://jung.sourceforge.net/index.html`.

Katsuno, H. and Mendelzon, A. O.: 1991, Propositional knowledge base revision and minimal change, *Artif. Intell.* **52**(3), 263–294.

Kittur, A. and Kraut, R. E.: 2008, Harnessing the wisdom of crowds in wikipedia: quality through coordination, *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, ACM, New York, NY, USA, pp. 37–46.

Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L. and Mylopoulos, J.: 2005, Semiautomatic semantic annotations for web documents, *Proceedings of the 2nd Italian Semantic Web Workshop University of Trento, Trento, Italy, 14-15-16*.

Knight, W.: 2005, 'info-mania' dents iq more than marijuana, Retrieved August 18th, 2009, from `http://www.newscientist.com/article.ns?id=dn7298`.

Kraft, R., Chang, C. C., Maghoul, F. and Kumar, R.: 2006, Searching with context, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 477–486.

Kraft, R., Maghoul, F. and Chang, C. C.: 2005, Y!q: contextual search at the point of inspiration, *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, ACM Press, New York, NY, USA, pp. 816–823.

Lacy, L. W.: 2005, *Owl: Representing Information Using the Web Ontology Language*, Trafford Publishing.

Lang, K.: 1995, Newsweeder: learning to filter netnews, *Proceedings of the 12th International Conference on Machine Learning*, Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, pp. 331–339.

Lawrence, S.: 2000, Context in web search, *IEEE Data Engineering Bulletin* **23**(3), 25–32.

Leacock, C. and Chodorow, M.: 1998, Combining local context with wordnet similarity for word sense identification, *WordNet: A Lexical Reference System and its Application*.

Lehmann, D.: 1989, What does a conditional knowledge base entail?, *Proceedings of the first international conference on Principles of knowledge representation and reasoning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 212–222.

Lerman, K.: 2007, Social information processing in news aggregation, *IEEE Internet Computing* **11**(6), 16–28.

Lieberman, H.: 1995, Letizia: An agent that assists web browsing, *in* C. S. Mellish (ed.), *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*, Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, Montreal, Quebec, Canada, pp. 924–929.

Lin, D.: 1997, Using syntactic dependency as local context to resolve word sense ambiguity, *Meeting of the Association for Computational Linguistics*, pp. 64–71.

Liu, H., Petrovic, M. and Jacobsen, H.-A.: 2005, Efficient and scalable filtering of graph-based metadata, *J. Web Sem.* **3**(4), 294–310.

M. Marinilli, A. M. and Sciarrone, F.: 1999, A case-based approach to adaptive information filtering for the www, *Proc. of the 2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web, Sixth International Conference on User Modeling UM-99*, Canada.

Maes, P. and Sheth, B.: 1993, Evolving agents for personalized information filtering, *Artificial Intelligence for Applications, 1993. Proceedings., Ninth Conference on*, IEEE, pp. 345–352.

Mann, M.: 2007, Inbox zero, Retrieved August 18th, 2009, from `http://video.google.com/videoplay?docid=973149761529535925`.

Melnik, S., Garcia-Molina, H. and Rahm, E.: 2002, Similarity flooding: A versatile graph matching algorithm and its application to schema matching, *ICDE '02: Proceedings of the 18th International Conference on Data Engineering*, IEEE Computer Society, Washington, DC, USA, p. 117.

moz: 2007, Mozilla thunderbird, Retrieved August 18th, 2009, from `http://www.mozilla.com/en-US/thunderbird/`.

Murthy, K. and Keerthi, S.: 1999, Context filters for document-based information filtering, *Fifth International Conference on Document Analysis and Recognition. ICDAR '99, 20-22 September, Bangalore,India*.

Mylopoulos, J., Borgida, A., Jarke, M. and Koubarakis, M.: 1990, Telos: representing knowledge about information systems, *ACM Trans. Inf. Syst.* **8**(4), 325–362.

Mylopoulos, J., Chaudhri, V., Plexousakis, D., Shrufi, A. and Topologlou, T.: 1996, Building knowledge base management systems, *The VLDB Journal* **5**(4), 238–263.

Nakai, K. and Kanehisa, M.: 1992, A knowledge base for predicting protein localization sites in eukaryotic cells., *Genomics* **14**(4), 897–911.

Nardi, D. and Brachman, R. J.: 2003, An introduction to description logics, pp. 1–40.

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T. and Swartout, W. R.: 1991, Enabling technology for knowledge sharing, *AI Mag.* **12**(3), 36–56.

Nguyen, H. and Cao, T.: 2008, Named entity disambiguation on an ontology enriched by wikipedia, *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*, pp. 247–254.

Nonaka, I.: 1994, A dynamic theory of organizational knowledge creation, *Organization Science* **5**(1), 14–37.

OED: 2002, *Concise Oxford English Dictionary - Thumb Index Edition*, tenth edn, Oxford University Press.

Ogden, C. K. and Richards, I. A.: 1923, *The Meaning of Meaning: A Study of the Influence of Language Upon Thought and of the Science of Symbolism*, Routledge & Kegan Paul, London.

O'Reilly, T.: 2006, O'reilly radar > web 2.0 compact definition: Trying again, Retrieved August 18th, 2009, from `http://radar.oreilly.com/archives/2006/12/web_20_compact.html`.

owl: 2004, Owl web ontology language overview, Retrieved August 18th, 2009, from `http://www.w3.org/TR/owl-features/`.

Pan, J. Z.: 2007, A flexible ontology reasoning architecture for the semantic web, *IEEE Transactions on Knowledge and Data Engineering* **19**(2), 246–260.

Paolucci, M., Kawamura, T., Payne, T. R. and Sycara, K.: 2002, Semantic matching of web services capabilities, *International Semantic Web Conference (ISWC)*.

Passin, T. B.: 2004, *Explorer's Guide to the Semantic Web*, Manning Publications Co., Greenwich, CT, USA.

Patwardhan, S., Banerjee, S. and Pedersen, T.: 2003, Using measures of semantic relatedness for word sense disambiguation, *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S. and Chute, C. G.: 2007, Measures of semantic similarity and relatedness in the biomedical domain, *J. of Biomedical Informatics* **40**(3), 288–299.

Pitkow, J., Schutze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E. and Breuel, T.: 2002, Personalized search, *Communications of the ACM* **45**(9), 50–55.

Powers, S.: 2003, *Practical RDF*, O'Reilly & Associates, Inc., Sebastopol, CA, USA.

Pressman, R. S.: 2001, *Software Engineering: A Practitioner's Approach*, fifth edn, McGrap-Hill.

Prud'hommeaux, E. and Seaborne, A.: 2007, SPARQL query language for RDF, *Proposed recommentation*, W3C.

Qu, Y., Hu, W. and Cheng, G.: 2006, Constructing virtual documents for ontology matching, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM, New York, NY, USA, pp. 23–31.

Raghavan, P.: 2002, Social networks: From the web to the enterprise, *IEEE Internet Computing* **06**(1), 91–94.

Resnik, P.: 1995, Using information content to evaluate semantic similarity in a taxonomy, *IJCAI*, pp. 448–453.

Robbins, A. and Beebe, N. H. F.: 2005, *Classic Shell Scripting*, O'Reilly.

RSS: 2003, Rss 0.92 specification, Retrieved August 18th, 2009, from `http://backend.userland.com/rss092`.

RSS: n.d., Rss 2.0 specification, Retrieved August 18th, 2009, from `http://blogs.law.harvard.edu/tech/rss`.

Russomanno, D. J.: 2006, A plausible inference prototype for the semantic web, *J. Intell. Inf. Syst.* **26**(3), 227–246.

Ruthven, I.: 2004, and this set of words represents the user's context..., *ACM SIGIR 2004 Workshop on Information Retrieval in Context*, ACM Press, p. 10.

Sadofsky, J. S.: 2005, Bbs: The documentary, Retrieved August 18th, 2009, from `http://www.bbsdocumentary.com/`.

Schmidt, A., Beigl, M. and Gellersen, H.-W.: 1999, There is more to context than location, *Computers and Graphics* **23**(6), 893–901.

Sebastiani, F.: 2002, Machine learning in automated text categorization, *ACM Comput. Surv.* **34**(1), 1–47.

Shadbolt, N.: 2006, The semantic web revisited, *IEEE Intelligent Systems*.

Shapiro, J., Voiskunskii, V. G. and Frants, V. I.: 1997, *Automated information retrieval: theory and methods*, Academic Press Professional, Inc., San Diego, CA, USA.

Shepherd, M., Watters, C. and Marath, A.: 2002, Adaptive user modeling for filtering electronic news, *HICSS '02: Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4*, IEEE Computer Society, Washington, DC, USA, p. 102.2.

Shipman, III, F. M. and McCall, R.: 1994, Supporting knowledge-base evolution with incremental formalization, *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, NY, USA, pp. 285–291.

Sirin, E., Parsia, B. and Hendler, J.: 2004, Composition-driven filtering and selection of semantic web services, *In AAAI Spring Symposium on Semantic Web Services*, p. 2004.

Strube, M. and Ponzetto, S. P.: 2006, Wikirelate! computing semantic relatedness using wikipedia, *AAAI' 06*.

Stvilia, B., Twidale, M. B., Smith, L. C. and Gasser, L.: 2005, Assessing information quality of a community-based encyclopedia, *Proceedings of the International Conference on Information Quality - ICIQ 2005*, pp. 442–454.

Swartout, B., Ramesh, P., Knight, K. and Russ, T.: 1997, Toward distributed use of large-scale ontologies, *AAAI Symposium on Ontological Engineering*.

swi: n.d., Inference engines for the semantic web, Retrieved October 29, 2005, from `http://semanticweb.org/inference.html`.

Sycara, K., Paolucci, M., Ankolekar, A. and Srinivasan, N.: 2003, Automated discovery, interaction and composition of semantic web services, *Web Semantics: Science, Services and Agents on the World Wide Web* **1**(1), 27 – 46.

Tazari, M., Grimm, M. and Finke, M.: 2003, Modelling user context, *10th International Conference on Human-Computer Interaction*, Crete (Greece).

Van Rijsbergen, C. J.: 1979, *Information Retrieval, 2nd edition*, Dept. of Computer Science, University of Glasgow.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G. and Milios, E. E.: 2005, Semantic similarity methods in wordnet and their application to information retrieval on the web, *WIDM '05: Proceedings of the 7th annual ACM international workshop on Web information and data management*, ACM, New York, NY, USA, pp. 10–16.

Voss, J.: 2005, Measuring wikipedia, *Proceedings International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden.

w3c: n.d.a, W3 seminar: Protocols, Retrieved August 18th, 2009, from `http://www.w3.org/Talks/General/Protocols.html`.

W3C: n.d.b, World wide web consortium, Retrieved August 18th, 2009, from `http://w3.org/`.

Webster, D., Huang, W., Mundy, D. and Warren, P.: 2006, Context-orientated news filtering for web 2.0 and beyond, *WWW '06: Proceedings of the 15th international conference on World Wide Web*, ACM Press, New York, NY, USA, pp. 1001–1002.

Wikipedia: 2008, Wikipedia:lead section, Retrieved August 18th, 2009, from `http://en.wikipedia.org/wiki/Wikipedia:Lead_section`.

Wikipedia: 2009, Wikipedia statistics, Retrieved August 18th, 2009, from `http://en.wikipedia.org/wiki/Special:Statistics`.

Wilkinson, D. M. and Huberman, B. A.: 2007, Cooperation and quality in wikipedia, *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, ACM, New York, NY, USA, pp. 157–164.

Winer, D.: 2006, Taking theory into practice, Retrieved August 18th, 2009, from `http://blog.broadbandmechanics.com/2006/01/14/taking-theory-into-practice/`.

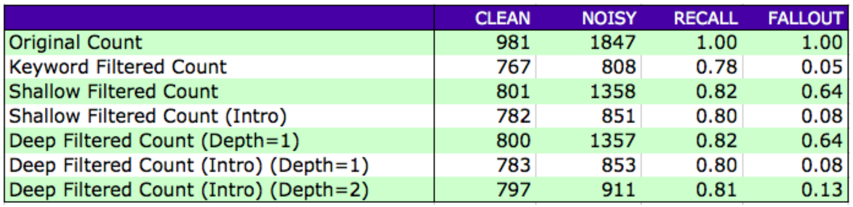
Yahoo!: 2007, Yahoo pipes, Retrieved August 18th, 2009, from `http://pipes.yahoo.com/pipes/`.

Zakon, R. H.: 2005, Hobbes' internet timeline, Retrieved August 18th, 2009, from `http://zakon.org/robert/internet/timeline/`.

Appendix A

Extended Results Data

A.1 Linux Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	981	1847	1.00	1.00
Keyword Filtered Count	767	808	0.78	0.05
Shallow Filtered Count	801	1358	0.82	0.64
Shallow Filtered Count (Intro)	782	851	0.80	0.08
Deep Filtered Count (Depth=1)	800	1357	0.82	0.64
Deep Filtered Count (Intro) (Depth=1)	783	853	0.80	0.08
Deep Filtered Count (Intro) (Depth=2)	797	911	0.81	0.13

Figure A.1: Linux Profile.

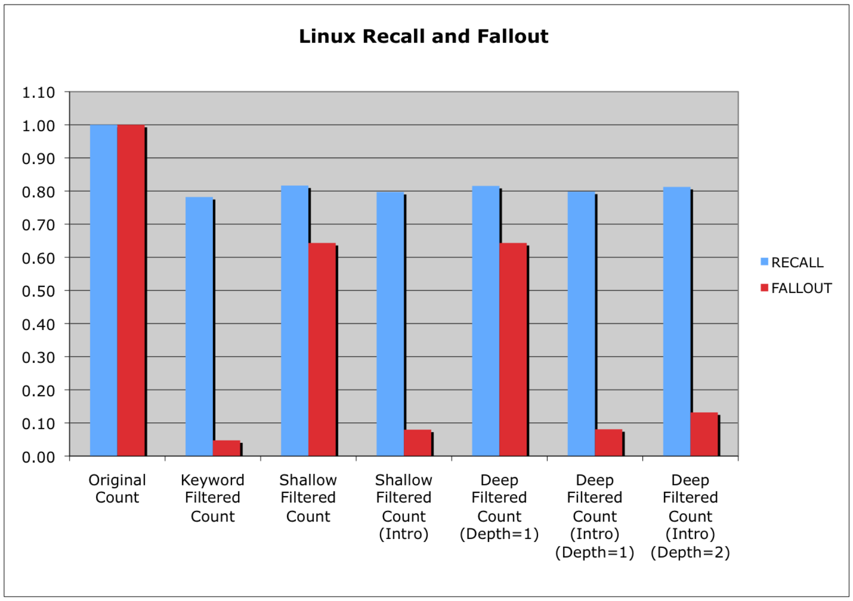
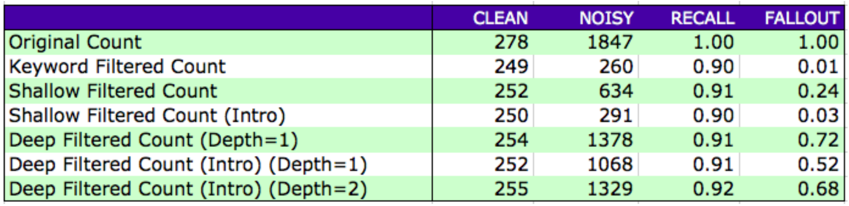


Figure A.2: Linux Profile Graph.

A.2 GNOME Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	278	1847	1.00	1.00
Keyword Filtered Count	249	260	0.90	0.01
Shallow Filtered Count	252	634	0.91	0.24
Shallow Filtered Count (Intro)	250	291	0.90	0.03
Deep Filtered Count (Depth=1)	254	1378	0.91	0.72
Deep Filtered Count (Intro) (Depth=1)	252	1068	0.91	0.52
Deep Filtered Count (Intro) (Depth=2)	255	1329	0.92	0.68

Figure A.3: GNOME Profile.

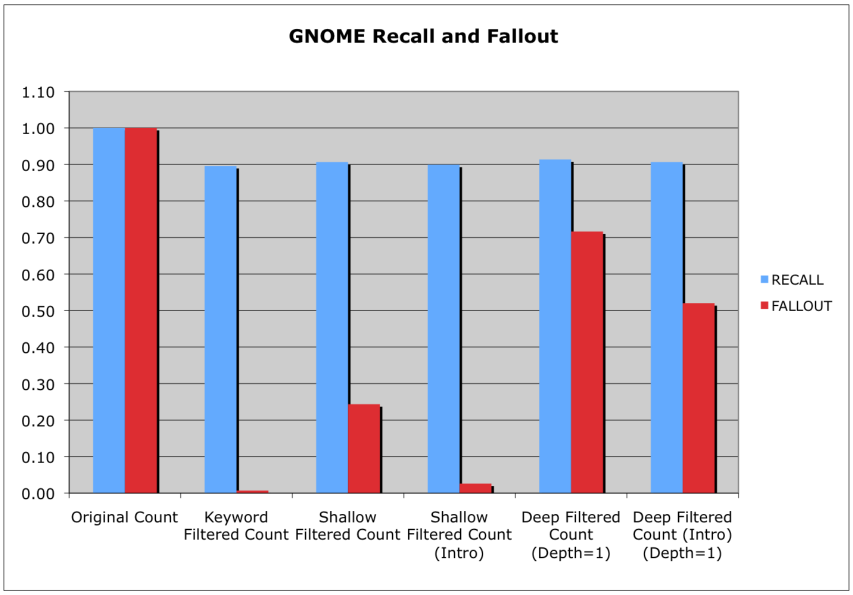
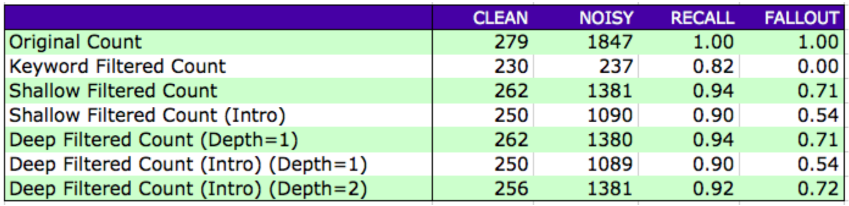


Figure A.4: GNOME Profile Graph.

A.3 KDE Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	279	1847	1.00	1.00
Keyword Filtered Count	230	237	0.82	0.00
Shallow Filtered Count	262	1381	0.94	0.71
Shallow Filtered Count (Intro)	250	1090	0.90	0.54
Deep Filtered Count (Depth=1)	262	1380	0.94	0.71
Deep Filtered Count (Intro) (Depth=1)	250	1089	0.90	0.54
Deep Filtered Count (Intro) (Depth=2)	256	1381	0.92	0.72

Figure A.5: KDE Profile.

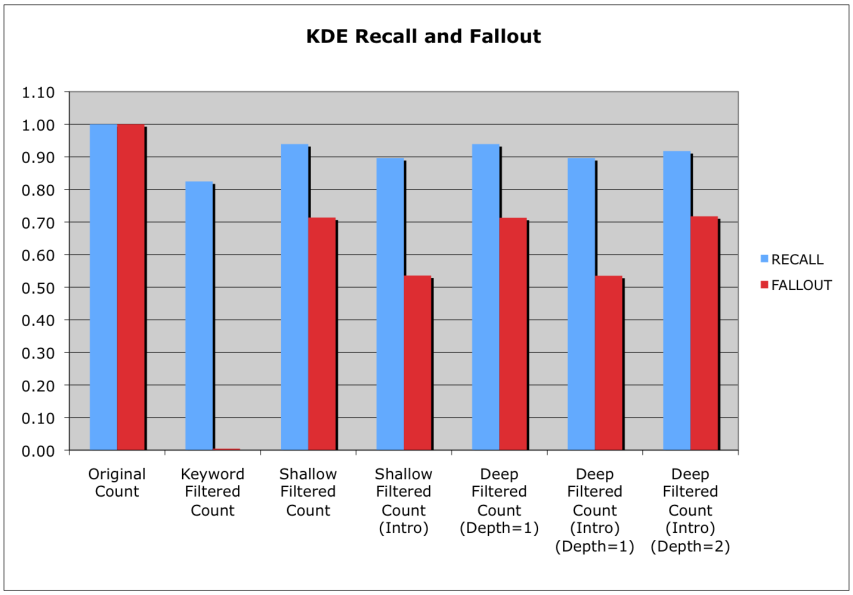
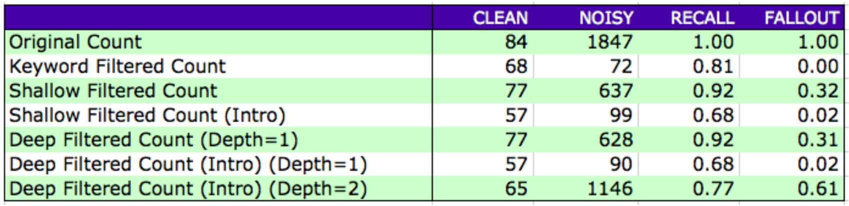


Figure A.6: KDE Profile Graph.

A.4 GTK Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	84	1847	1.00	1.00
Keyword Filtered Count	68	72	0.81	0.00
Shallow Filtered Count	77	637	0.92	0.32
Shallow Filtered Count (Intro)	57	99	0.68	0.02
Deep Filtered Count (Depth=1)	77	628	0.92	0.31
Deep Filtered Count (Intro) (Depth=1)	57	90	0.68	0.02
Deep Filtered Count (Intro) (Depth=2)	65	1146	0.77	0.61

Figure A.7: GTK Profile.

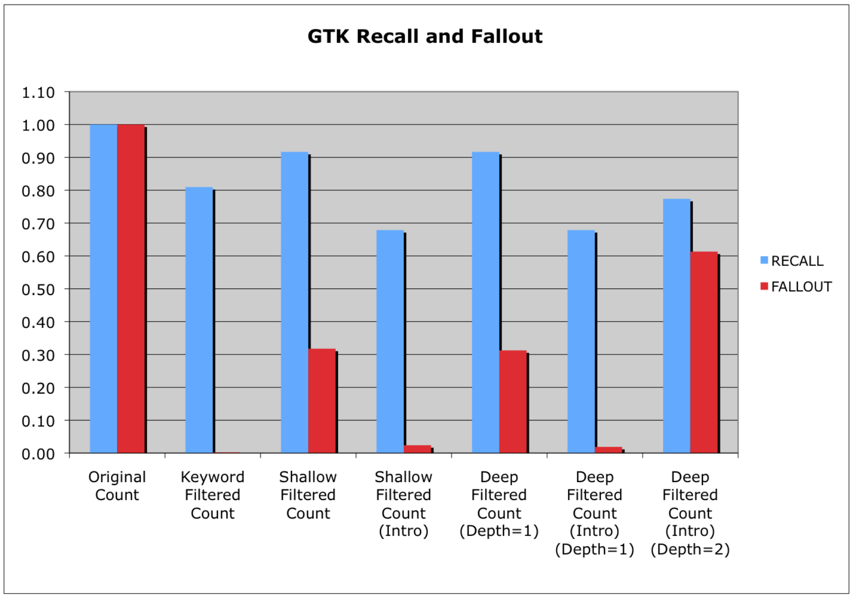
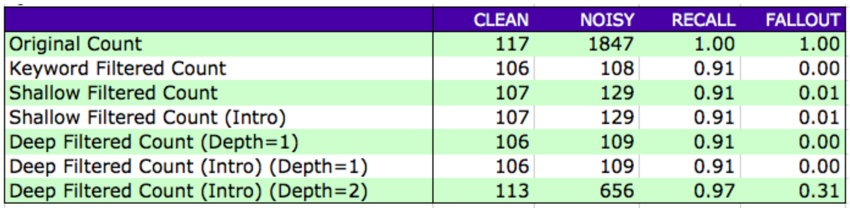


Figure A.8: GTK Profile Graph.

A.5 QT Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	117	1847	1.00	1.00
Keyword Filtered Count	106	108	0.91	0.00
Shallow Filtered Count	107	129	0.91	0.01
Shallow Filtered Count (Intro)	107	129	0.91	0.01
Deep Filtered Count (Depth=1)	106	109	0.91	0.00
Deep Filtered Count (Intro) (Depth=1)	106	109	0.91	0.00
Deep Filtered Count (Intro) (Depth=2)	113	656	0.97	0.31

Figure A.9: QT Profile.

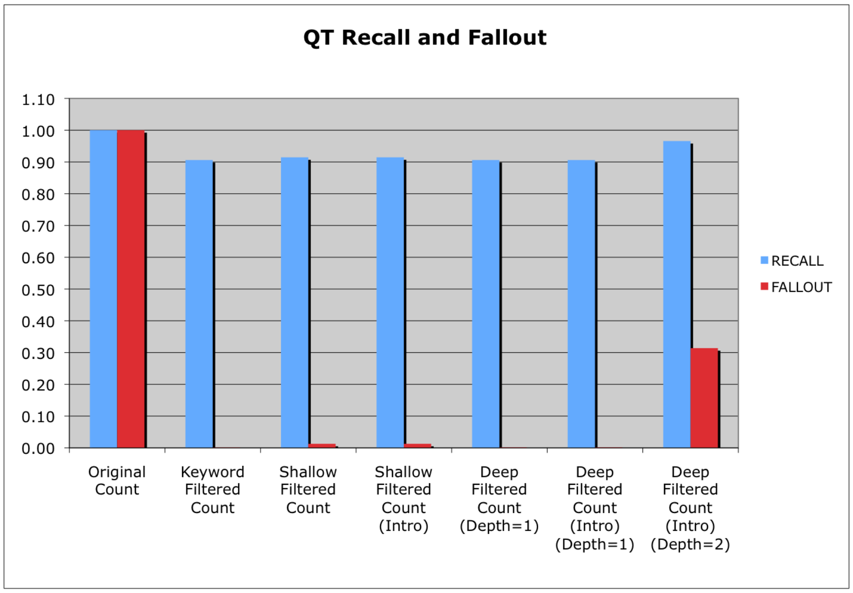
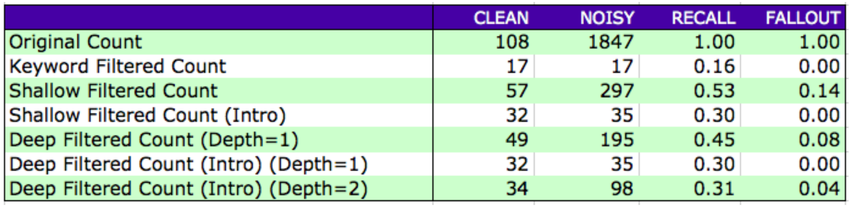


Figure A.10: QT Profile Graph.

A.6 Games Profile



	CLEAN	NOISY	RECALL	FALLOUT
Original Count	108	1847	1.00	1.00
Keyword Filtered Count	17	17	0.16	0.00
Shallow Filtered Count	57	297	0.53	0.14
Shallow Filtered Count (Intro)	32	35	0.30	0.00
Deep Filtered Count (Depth=1)	49	195	0.45	0.08
Deep Filtered Count (Intro) (Depth=1)	32	35	0.30	0.00
Deep Filtered Count (Intro) (Depth=2)	34	98	0.31	0.04

Figure A.11: Games Profile.

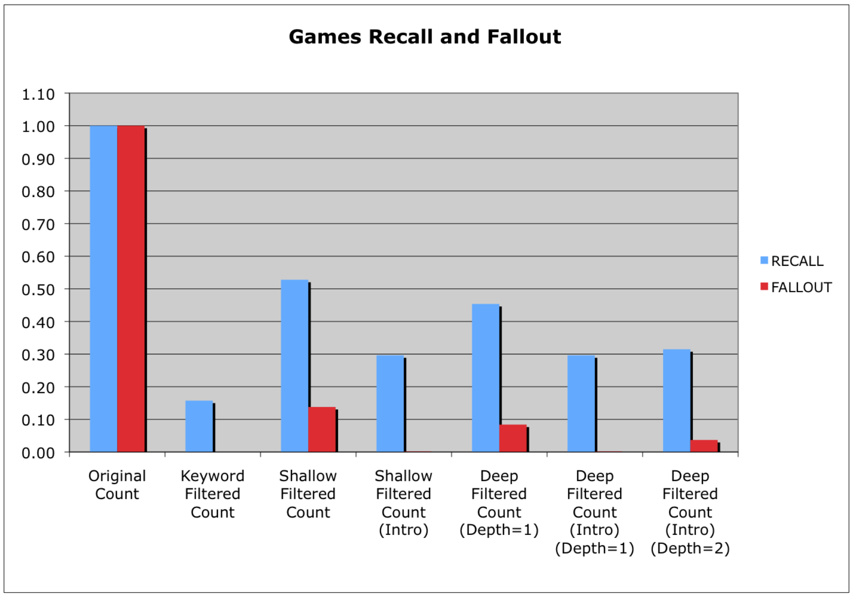


Figure A.12: Games Profile Graph.