Department of Computer Science

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Hamilton, New Zealand

Applying Wikipedia to Interactive Information Retrieval

by

David Milne

This thesis is submitted in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy in Computer Science
at The University of Waikato

September 2010

Abstract

There are many opportunities to improve the interactivity of information retrieval systems beyond the ubiquitous search box. One idea is to use knowledge bases—e.g. controlled vocabularies, classification schemes, thesauri and ontologies—to organize, describe and navigate the information space. These resources are popular in libraries and specialist collections, but have proven too expensive and narrow to be applied to everyday web-scale search.

Wikipedia has the potential to bring structured knowledge into more widespread use. This online, collaboratively generated encyclopaedia is one of the largest and most consulted reference works in existence. It is broader, deeper and more agile than the knowledge bases put forward to assist retrieval in the past. Rendering this resource machine-readable is a challenging task that has captured the interest of many researchers. Many see it as a key step required to break the knowledge acquisition bottleneck that crippled previous efforts.

This thesis claims that the roadblock can be sidestepped: Wikipedia can be applied effectively to open-domain information retrieval with minimal natural language processing or information extraction. The key is to focus on gathering and applying human-readable rather than machine-readable knowledge.

To demonstrate this claim, the thesis tackles three separate problems: extracting knowledge from Wikipedia; connecting it to textual documents; and applying it to the retrieval process. First, we demonstrate that a large thesaurus-like structure can be obtained directly from Wikipedia, and that accurate measures of semantic relatedness can be efficiently mined from it. Second, we show that Wikipedia provides the necessary features and training data for existing data mining techniques to accurately detect and disambiguate topics when they are mentioned in plain text. Third, we provide two systems and user studies that demonstrate the utility of the Wikipedia-derived knowledge base for interactive information retrieval.

Acknowledgements

I would like to thank my supervisor, Ian Witten. The acknowledgements of my Masters Thesis, written four years ago, start with exactly the same line. The gist then was "thank you for investing so much time in me, and for teaching me how to be a researcher." If you were to keep reading that thesis (please don't) you would find that I still had much to learn. Ian has continued to be patient, encouraging and hands-on. The lessons have not slowed, and with them have come sailing trips, cocktails, star-lit clarinet recitals and even the occasional house. Thank you Ian, for being such a great mentor.

Thanks also to my backup supervisors, Dave Nichols and Sally Jo Cunningham. Dave in particular has put up with near constant intrusions (his office is inconveniently just across the hall), and has been a great source of advice.

I am very proud to have belonged to Ian's gaggle of PhD students: the stars of the NZCSRSC and the scourge of the Build-IT publicity awards (will you ever let anyone else have one?). Our well-earned escapes to Waiheke Island, powered by rocket-fuel gin martinis and multicultural feasts, have been great fun.

The standouts of this group—for me at least—are Aly and Veronica. Aly—short for Olena Medelyan, a name oft cited in the pages that follow—has worked closely with me for much of this investigation, and has had her toes stepped on repeatedly as a consequence. Thank you for being such a patient collaborator and friend. Veronica has been chief design consultant for all the pretty diagrams. Sorry for all of the times I've made you jump two feet up from your seat for the most inane questions.

Thank you to the others I have collaborated with, particularly Cathy Legg and Vivi Nastase. The same goes to those I've played with, especially my "surfing for sanity" support group, Andreas and Doris. Anu, your thesis is done now. Sorry it took so long.

When I was a young teenager, my ever-enterprising father made a split decision to start a local newspaper for our home on Great Barrier Island. Within weeks I was lead designer and resident tech support, and so began my career in computers and my apprenticeship in visual and written communication. Mum and Dad, thank you for this initial push and all of the faith and support that has come since. Hopefully I will soon reward you with another cheesy graduation photo.

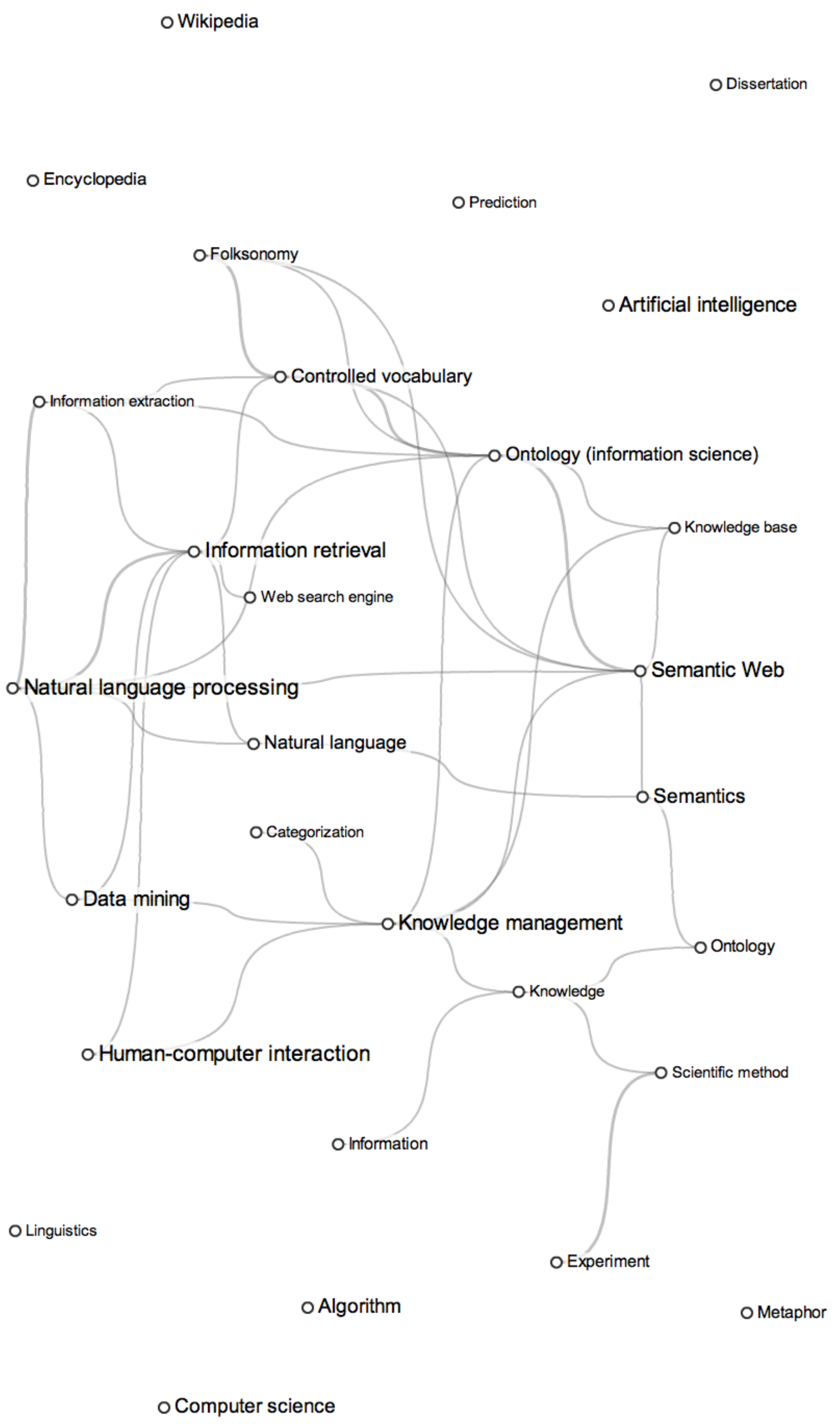
I have received funding from an embarrassing array of sources: the University of Waikato, the New Zealand Tertiary Education Commission, the New Zealand Digital Library Group, and Build-IT. Thank you all for the generous support.

Finally, thank you Wikipedians! I fear this will instigate a flurry of anti-plagiarism measures, but I must admit: I wouldn’t have much of a thesis without you.

Table of Contents

List of Tables

List of Figures



This visualization of topics discussed in Chapter 1 was created automatically (see Section 6.5). The topics were gathered using our techniques for detection and disambiguation (Chapter 6). The relations between them were measured using the WLM algorithm (Chapter 5). The layout is based on Hōpara (Chapter 7).

1. Introduction

> *And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of your enquiry? And if you find out what you want, how will you ever know that this is the thing that you did not know?*
>
> *Plato's Meno, 380 BC*

Whenever we seek out new knowledge—whenever we turn to the ubiquitous search engines—we must grapple with a fundamental paradox: how can one describe the unknown? A search engine query is not a question or a statement of intent. Instead it is an excerpt, a few words or phrases, from within a relevant document. To form an effective query, the searcher must predict not only what information this relevant document contains, but also the terms by which it express it. In short, they must already know a great deal of what is being sought, in order to find it.

Of course, the situation cannot be entirely circular. Search engines are constantly used to discover new information. We can leave the philosophical implications to Socrates, Plato and Meno, but it is important to understand how this occurs. An obvious answer is that information needs do not simply pop into existence; something must occur for a searcher to recognize a gap in their knowledge and be prompted to fill it. Along with the prompt comes the first clues of how to resolve the gap; the first sketchy queries to issue. Even if these do not immediately provide the complete answer, they should generate some clues as to what to try next, and put the user on a path by which they can work iteratively towards a solution.

The point of this philosophical aside is to demonstrate that information retrieval is more than just matching queries to documents. It is more than just search. There is a broader context in which searcher and retrieval system enter into a dialog and work together to gradually resolve an information need. The study of this broader context is known as interactive information retrieval.

This thesis investigates how access to Wikipedia can make interactive information retrieval more effective and efficient. This online, collaboratively generated encyclopaedia has already proven itself as a valued source of knowledge; it is one of the largest and most consulted reference works in existence. Here we investigate how it can be used to augment other information sources—any collection of textual documents—to make them easier to search and navigate.

This idea of having search engines consult external sources of knowledge—ontologies, taxonomies, thesauri, glossaries and gazetteers—in addition to the documents being searched is by no means original. The idea is old, obvious, and compelling, but results thus far have been singularly unimpressive, at least for large-scale, open-domain document retrieval. The best performing and most widely used search systems are still those that deal in lexical character patterns without using structured knowledge to understand them.

Wikipedia has the potential to change all that. This open, constantly evolving encyclopaedia represents a vast pool of topics and semantic relations. It easily dwarfs the controlled vocabularies, taxonomies and thesauri that have been put forward to support information seekers in the past. Perhaps, at last, we have a manually-constructed knowledge base that is sufficiently broad, deep, and timely to be applicable to open-domain information retrieval.

If so, then the resource does not come without its own challenges. While the knowledge contained within Wikipedia is undeniably useful to its many readers, it is produced and organized by anonymous contributors with unknown qualifications and motivations. This use of crowd-sourced labour has garnered many sceptics and critics. The result is a resource that is controversially generated, somewhat haphazardly organized, and only partially machine-readable. This thesis investigates how best to apply it to organizing and retrieving information, given these shortcomings.

1.1 The thesis

This thesis makes the following central claim:

> *Wikipedia can be applied effectively to open-domain information retrieval without deep natural language processing.*

For any given document collection—regardless of domain—it is likely that Wikipedia knows something about the subject material discussed within, and that this knowledge could make the material easier to understand and navigate. However, applying Wikipedia as a knowledge base for information retrieval appears, at first glance, to be extremely challenging. The resource has been built with human readers in mind. How can structured knowledge be extracted from it? How could this knowledge be cross-referenced and applied automatically to other document collections? How can it be applied usefully to the retrieval process?

One way to answer these three questions would be to apply natural language processing and information extraction techniques to Wikipedia. Many researchers are doing exactly that. However, the ambitiousness of this approach and the limitations of NLP mean that progress is limited. Researchers are almost entirely focussed on the first question; on extracting knowledge from Wikipedia. It will be some time before they can move on to the latter questions, and apply the knowledge they are currently struggling to extract.

This thesis aims to sidestep the difficulties inherent in rendering Wikipedia machine readable, by capitalizing on its existing structural features. As a result, it is able to provide contributions and partial answers for each of the three questions posed above.

1.1.1 Extracting knowledge from Wikipedia

> *A significant amount of semantic knowledge is defined explicitly in Wikipedia's structure, and can be captured without sophisticated extraction techniques.*

Essentially, we claim that Wikipedia's structure can be used directly—or with minimal automatic processing—as a knowledge base to support information retrieval. "Knowledge base" is a deliberately vague term: there are many different kinds, ranging from minimal to complex, from unsophisticated to highly expressive. Most interest currently resides at the ends of the spectrum; in the simple folksonomies that have emerged in recent years and the highly complex ontologies that form the cornerstone of Berners-Lee's vision of the Semantic Web.

This thesis revisits the middle of the spectrum, and explains how Wikipedia's raw structure bears remarkable resemblance to moderately expressive knowledge bases: controlled vocabularies, glossaries, taxonomies and thesauri. With little computational effort, it can yield exactly the same structural elements that were previously encoded exhaustively by hand, but on a greater scale. Many others have recognized Wikipedia's strengths for providing machine-readable knowledge, but most are ambitiously applying sophisticated knowledge extraction techniques to render it into a formal ontology. In contrast, this work is unapologetically oriented towards short-term gains, and focuses on making use of what Wikipedia already provides or makes easily accessible. In short, it demonstrates that there is a great deal of low-hanging fruit that can be harvested and immediately put to widespread use.

1.1.2 Connecting Wikipedia to textual documents

> *Wikipedia provides sufficient training data to allow existing data mining techniques to accurately detect and disambiguate Wikipedia topics when they are mentioned in plain text.*

One limitation of our shallow approach to knowledge extraction is that the resulting structure has little chance of resolving information needs directly. It does not have the expressivity or formalism required to provide answers on its own, as an ontology might. It is better suited to assisting retrieval within other resources such as web pages and document collections. Consequently, connecting the two resources—the derived knowledge base and the documents being searched—is a key challenge.

The current popularity of folksonomies suggests that a certain degree of manual effort is reasonable; readers will tag documents if they are given the tools to do so, and it not would be impractical to have them select tags from a Wikipedia-derived vocabulary—assuming it were broad and detailed enough—rather than typing them directly. However, reliance on manual labour immediately places limits: on how many documents are supported, how quickly the collection is updated, and to what level of detail the documents are described. It would be a great advantage if the connections between documents and the knowledge base could be made automatically, or at least semi-automatically.

One of the key observations of this thesis is that every single Wikipedia article is an example of how to connect a textual document to the appropriate Wikipedia topics. Articles are peppered with hundreds of millions of links, which explain the topics being discussed and provide an environment where serendipitous encounters with information are commonplace. We demonstrate how these links can be used as training data for detecting and disambiguating Wikipedia topics when they are mentioned in other documents. The techniques described here can provide structured knowledge about any unstructured fragment of text, and are therefore applicable to a wide variety of tasks.

1.1.3 Applying Wikipedia to the retrieval process

> *Wikipedia-derived knowledge, having been extracted and connected appropriately to documents, can be applied to the information seeking process in ways that are helpful and intuitive to users.*

The result of the previous two claims is a large-scale, domain-independent knowledge base that can be easily connected to any textual document collection. All that is needed to confirm the central thesis claim is to demonstrate the utility of this structure for information retrieval over the documents to which it is connected. This can be measured indirectly by looking at the properties of the structure and its intersection with documents—e.g. its breadth and depth, the accuracy of its relations, etc.—but more concrete findings require evaluating algorithms and systems that directly apply it to information retrieval.

Information retrieval is not simply a matter of matching queries to documents. Users do not always know what they are looking for in advance, and can have far reaching information needs that cannot be satisfied by a single document. They often need assistance in exploring the information space and gathering what they need. Intuitively, it is this broader, more interactive setting in which knowledge bases—essentially road maps of the available terminology—are likely to provide the greatest value.

This is especially true for the knowledge base studied here: it is based on Wikipedia's skeleton structure, which was built manually to assist people in navigating and retrieving information from the resource. That structure was crafted with human users in mind, and its utility for automatic inference or intelligent retrieval systems is limited. To continue the map metaphor, it is more like directions scrawled on a napkin than waypoints encoded into a GPS. Consequently, our efforts to apply it to the retrieval process are oriented towards human-computer interaction rather than natural language processing, information extraction or artificial intelligence. The chapters that follow describe several interactive retrieval systems, and several user studies to investigate their usefulness.

1.2 Contributions

The contributions made during this investigation are as follows:

Obtaining knowledge from Wikipedia:

- A comprehensive survey of extracting knowledge from Wikipedia.
- A comparison between Wikipedia and a traditional knowledge base (the domain-specific thesaurus *Agrovoc*).
- An efficient yet accurate method for measuring semantic relatedness between Wikipedia concepts, which makes textual explanations of relations easily accessible.
- A test set for evaluating semantic relatedness measures that has been manually disambiguated against Wikipedia (Appendix B).

Connecting Wikipedia to textual resources:

- A comprehensive survey of detecting Wikipedia topics when they are mentioned in text.
- An efficient yet accurate method for resolving ambiguous terms and phrases, to identify the Wikipedia articles they refer to.
- An efficient yet accurate method for identifying Wikipedia topics when they are mentioned in free text, and predicting to what extent they are likely to be of interest to the reader.

- A new dataset, with multiple human judgments, for evaluating techniques for Wikipedia topic detection and disambiguation (Appendix B).

Applying Wikipedia to the retrieval process:

- A user study that investigates the advantages of resolving search terms to the appropriate Wikipedia topics.
- An approach for providing categorized recommendations in response to single- and multi-topic queries.
- An approach for visualizing and interacting with the recommendations described above.
- A user study that investigates the effectiveness of the recommendations and visualization described above.

In addition, the following contributions have been made that do not contribute directly to the thesis claims:

- Several publications, including an article in the *International Journal of Human-Computer Studies*, and the best paper of the *2008 International Conference on Information and Knowledge Management*. A full list, including succinct summaries, appears in Appendix A.
- An open-source toolkit that provides efficient programmatic access to Wikipedia, and the relatedness measures and topic detection/disambiguation algorithms described above. This is described in Appendix C.
- A suite of publicly available web services, which provide immediate human- and machine-readable access to the main features of the toolkit described above. These services are also described in Appendix C.

1.3 Thesis structure

The remainder of the thesis is structured as follows. Chapter 2 provides background for how knowledge bases can be applied to information retrieval. It first describes several models that capture the highly interactive nature of the retrieval process. It then surveys existing knowledge bases, and explains related topics such as the knowledge acquisition bottleneck and the Semantic Web. It also surveys a range of strategies for making search engines more interactive and proactive, and how knowledge bases are applied to them. The chapter concludes with some general insights into what is needed from a knowledge base for supporting interactive information retrieval.

Chapter 3 introduces Wikipedia. We document the resource's origins as a speculative side-project with limited ambitions, its growing pains, and its rapid rise to become one of the largest, most popular reference works in existence. Its inherent strengths and weaknesses are surveyed and related back to its eventful beginnings. Each structural feature is described, and the extensive body of work that seeks to extract knowledge from them is surveyed. A comparison is made between Wikipedia and a traditional knowledge base: the domain-specific thesaurus *Agrovoc*. Finally, the conclusions of the previous chapter are revisited, and compared to the specific brand of knowledge Wikipedia provides.

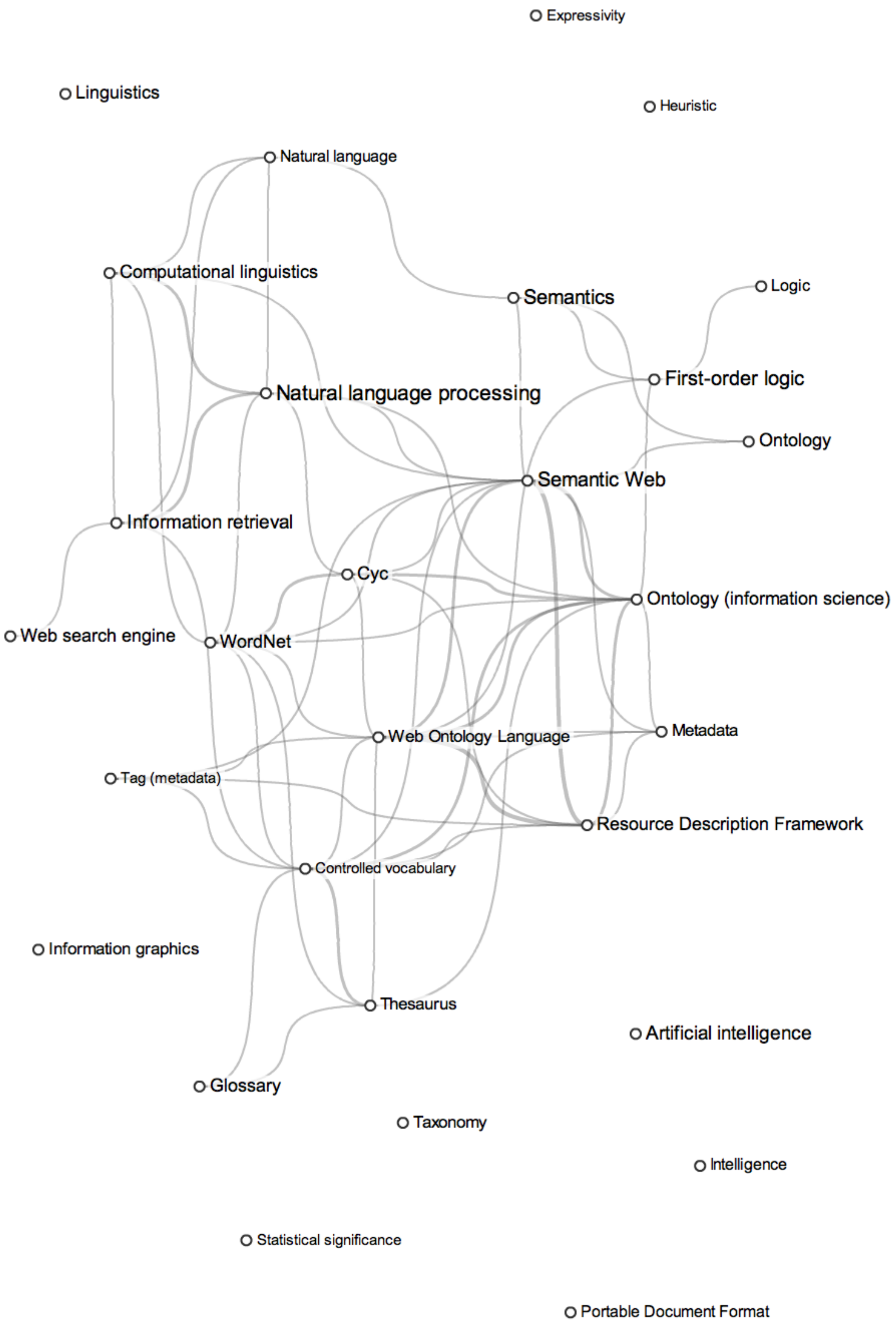
Chapter 4 makes an early attempt to address the entire thesis claim. It presents a prototype retrieval system that uses Wikipedia to assist retrieval within a collection of newswire stories. To do so it presents ad hoc techniques for extracting the knowledge base, connecting it to documents, and applying it to the retrieval process. The system is evaluated in a detailed user study that compares it against traditional keyword search. The results of this experiment guide the remainder of the thesis.

Chapter 5 revisits the first sub-claim; the extraction of knowledge from Wikipedia. In the previous two chapters it becomes clear that deriving semantic relations from Wikipedia is non-trivial, despite its abundance of structural features. Although its articles cross-reference each other extensively, the links they make do not necessarily indicate relatedness. In this chapter the relation extraction task is isolated and slightly refocused: we aim to quantify the strengths of semantic relations rather than merely identify their existence. This allows an extensive body of related work to be drawn on for comparison and evaluation. The result is a new, highly efficient yet accurate measure of semantic relatedness between Wikipedia concepts, which has a wide range of applications. The chapters that follow rely heavily on this measure.

Chapter 6 focuses on the second sub-claim; connecting the Wikipedia-derived knowledge base to textual documents. This is attempted briefly in Chapter 4, where an ad hoc algorithm is presented without justification, evaluation, or reference to related work. Chapter 6 isolates the task and investigates it systematically. It provides an extensive survey, and explains how approaches can be evaluated elegantly at a large scale using Wikipedia itself as ground truth. It also explains how cross-referencing documents can be broken down into two subtasks: detecting terms that are suitable link anchors, and disambiguating terms that would otherwise link to multiple topics. In both cases new approaches are presented and evaluated, with significant gains over the previous state of the art. The chapter concludes with a discussion of the implications, which are broad: the techniques described here can provide structured knowledge about any unstructured fragment of text, and are therefore applicable to a wide variety of tasks.

The third sub-claim—applying Wikipedia to the retrieval process—is isolated in Chapter 7. The task is separated from the cross-referencing problem by supporting retrieval within a document collection that has been manually connected to Wikipedia's structure: namely, Wikipedia itself. The new system concentrates on interactive query expansion, a feature that was poorly addressed in the previous study (from Chapter 4). To do so it utilizes the new semantic relatedness measures described in Chapter 5 to suggest topics in response to queries, Wikipedia's raw category structure and lightweight sentence extraction to organize and explain them, and visualization to convey the relations between them. A user study demonstrates the advantages of this new system over the incumbent Wikipedia interface.

The thesis concludes with a critical look at how successfully each sub-claim has been isolated and addressed, and directions for future work.


Expressivity
Linguistics
Heuristic
Natural language
Computational linguistics
Semantics
Logic
First-order logic
Natural language processing
Ontology
Semantic Web
Information retrieval
Cyc
Ontology (information science)
Web search engine
WordNet
Metadata
Web Ontology Language
Tag (metadata)
Resource Description Framework
Controlled vocabulary
Information graphics
Thesaurus
Artificial intelligence
Glossary
Taxonomy
Intelligence
Statistical significance
Portable Document Format

2. Interactivity, knowledge and intelligence in information retrieval

The key idea of this thesis is to allow search engines to tap into knowledge contained in Wikipedia. Equipping information retrieval systems with background knowledge and the ability to apply it is by no means a new idea. Smith (1976), for example, draws connections between Vannevar Bush's vision of the Memex (Bush 1945) and Turing's test for artificial intelligence (Turing 1950). She claims that all information retrieval can be viewed as "question answering", and that IR systems are judged on their ability to answer questions competently; the same test Turing proposes for intelligent machines. Spärck Jones (1991) predicted that retrieval systems would use prior knowledge to infer intentions and needs from fuzzy, ill-specified requests. Maes (1994) envisioned intelligent retrieval agents, capable of acting autonomously to gather information with minimal instruction or direction.

These early predictions and efforts viewed the role of intelligence in information retrieval as shifting workload and responsibilities from users to machines; as saving people from expressing their needs carefully or expending too much effort delving into the information space. They bring to mind the image of Star Trek's Captain Kirk, casually conversing with the ship's computer as if it were just another (extremely knowledgeable) member of the crew.

This classic NLP/AI view of intelligent information retrieval has been attacked repeatedly—e.g. by Brooks (1987), Bates (1990) and Belkin (1996)—but received a resurgence of interest with the Semantic Web initiative and the bold ambitions that have come along with it.

> *I have a dream for the Web [in which computers] become capable of analysing all the data on the Web – the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines.* ***.*** *The 'intelligent agents' people have touted for ages will finally materialize.*

(Berners-Lee and Fischetti 1999)

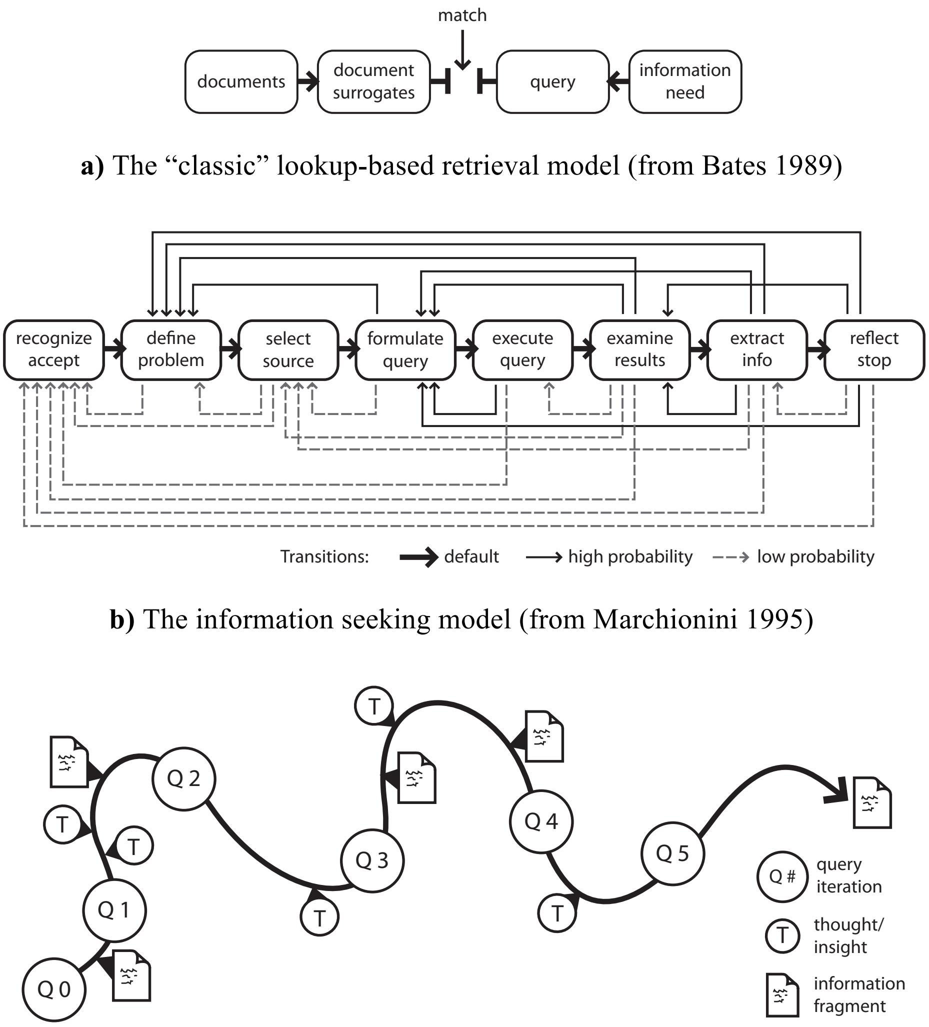
Berners-Lee's vision is still very much alive more than a decade later. The Semantic Web movement continues to make structured knowledge more accessible and deep automated reasoning of information spaces more plausible. Nevertheless, the points raised by Brooks, Bates, Belkin and others still hold. Information retrieval—particularly when placed in its broader context of learning, investigation and sense making—is profoundly interactive. This thesis advocates using knowledge to augment interaction with retrieval systems rather than reducing the need for it; allowing human intelligence to be more easily applied rather than seeking to imitate or replace it. This perspective of the roles of interactivity, knowledge and intelligence in information retrieval is outlined by Belkin (1996):

> *In such a scenario, the user plays a central role, guiding the system, making evaluative judgments, deciding about what to do and when to stop. The other processes contribute by understanding something about what is likely to help the user in supporting the interactions in which that person is engaged, in knowing something about what the likely course of the interaction as a whole might be, and in using their knowledge about the resources at their disposal to inform the user about the system and its contents so that the user can interact effectively.*
>
> *Thus, using this model, intelligent IR turns out to be IR in which intelligence is explicitly distributed throughout the system, all of the actors contributing according to their specific roles and knowledge to support the user's effective interaction with information.*

This chapter is centred on one question: what sort of knowledge bases are most likely to bring about the kind of intelligent information retrieval described above? Further support for viewing information retrieval as a highly interactive process is given in Section 2.1. Section 2.2 describes the different forms knowledge bases can take, and relevant topics such as the Semantic Web and the knowledge acquisition bottleneck. Section 2.3 surveys how knowledge bases are currently being applied to make retrieval systems more interactive. The chapter closes with a first attempt to answer the question posed above. It will be revisited in Chapter 3, which examines the unique form of structured knowledge provided by Wikipedia.

2.1 The information retrieval process

Much has been done to observe information seekers as they interact with retrieval systems, and to model the difficulties they encounter and the strategies they employ. We will not delve deeply into this work, as it has already been competently and exhaustively surveyed by many others—Marchionini (1995), White and Roth (2009) and Hearst



a) The “classic” lookup-based retrieval model (from Bates 1989)

b) The information seeking model (from Marchionini 1995)

c) The berrypicking model (from Bates 1989)

Figure 2.1: Models of the information retrieval process.

(2009) to name a few. This section instead gives a shallow overview, to provide context for the more detailed discussions that follow.

Figure 2.1a illustrates the “classic” lookup-based model of the retrieval process—taken from Bates (1989)—in which a user approaches the retrieval system with some need in mind and translates it into a query or command suited to the system. Internally, the system has preprocessed a collection of documents so that they can be automatically and efficiently matched against relevant queries. This simple model is directly reflected in how retrieval systems like Google are built, used and evaluated.

The classic model contains many gaps. What about vague, broad, or multifaceted information needs, which cannot be satisfied with a single query or result set? What about users who are exploring new domains, where unfamiliarity cripples their ability to construct effective queries? Should users always have to issue queries explicitly, or can

Task	Feelings (affective)	Thoughts (cognitive)		Actions (physical)
Initiation	Uncertainty			
Selection	Optimism	Vague		Seeking relevant info, exploring
Exploration	Confusion, frustration, doubt			
Formulation	Clarity	↓	↓	↓
Collection	Sense of direction/confidence	Focused	Increased Interest	Seeking pertinent info, documenting
Presentation	Satisfaction or disappointment			

Table 2.1: The information search process (from Kuhlthau 1991)

retrieval systems be more proactive in bringing material to their attention? Are there no connections from one document to the next; one query to the next; or one information need to the next?

Many more informative models of the information retrieval process have been advanced. A common theme is iteration, where interactions are not considered in isolation, but allowed to build upon each other. Query formulation and results evaluation, for example, can form a tight loop as users struggle to express their information needs effectively. Figure 2.1b shows Marchionini's (1995) model of information seeking behaviour within electronic environments, which breaks the search process into eight steps. There is a default linear sequence, but users often loop back to previous states as they encounter difficulties or gain new insights. Some transitions are more likely than others: e.g., users commonly extract information from a result set and immediately use it to formulate a new query, but rarely turn to an entirely new information source without first mulling over and reflecting on the results they have found. Several models—e.g., Shneiderman et al. (1997), Sutcliffe and Ennis (1998)—provide similar descriptions of states, transitions and loops, but differ on the semantics. Bates' (1989) berry-picking model, illustrated in Figure 2.1c, glosses over the specific stages of interaction, but allows items encountered during the search to inspire entirely new information needs (not just refinements of the original need), and for users to be satisfied not by individual documents or result sets but by scattered fragments of knowledge collected along the way. The information foraging theory compares information seeking to the mechanisms by which organisms forage for food (Pirolli and Card 1999).

Information retrieval can range from lightweight fact finding to deep, extended investigations. Kuhlthau (1991) provides a longitudinal study of the latter by tracking the behaviour and emotional state of 385 library users—primarily students with lengthy research projects—for several months. This reveals common patterns in both actions and emotions, which are summarized in Table 2.1. Roughly speaking, the initial stages of *initiation* and *selection*—identifying the general problem to work on—are characterized by optimism. This gives way to confusion and doubt during *exploration*, when information is gathered with only vague goals and boundaries. The information encountered during this stage allows users to *formulate* their problem more carefully, focus on *collecting* only the information that is most pertinent, and finally *present* the result. These later stages are characterized by growing clarity, confidence, focus and interest.[1]

The above study alludes to broader activities surrounding the retrieval of information, where users collate, learn from and act upon the knowledge they gather. This broader context is out of scope for this investigation but worth mentioning in passing. Two of the more well-known activities in this area are *sensemaking,* which studies how users form conceptual representations of large information spaces (Klein et al. 2006), and *exploratory search*, which synthesizes a large toolkit of search activities and strategies, and organizes them under three key headings: *lookup, learning* and *investigation* (Marchionini 2006, White and Roth 2009).

2.2 Knowledge bases

The purpose of a knowledge base is to describe meaning in a structured, unambiguous fashion. Given this goal, it is somewhat ironic that there is a wide variety of names—ontologies, taxonomies, thesauri, etc.—by which these structures are known. Sometimes the terms denote subtly different kinds of resources, but they are often used inconsistently and swapped interchangeably (Gilchrist 2003). This section clarifies how these terms can be interpreted consistently throughout this thesis (but not necessarily within the surrounding literature).

Figure 2.2 illustrates a spectrum on which knowledge bases can be organized, which has been adapted from McGuinness (2005). The same spectrum is mirrored in Figure 2.3, which provides examples for each step along the scale. On the left of Figure 2.2 and the top left of Figure 2.3 are the simplest and least expressive structures, called *controlled vocabularies*. These seek only to define a consistent terminology, so that any relevant concept or topic is known unambiguously by one—and only one—name. These are often

[1] The author can confirm the same patterns from personal experience, as (we speculate) can anyone who has surveyed related work for a thesis!

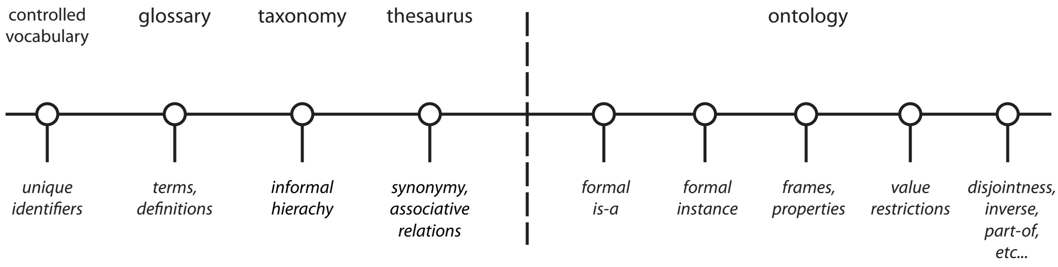


Figure 2.2: The knowledge base spectrum

used to achieve consistency when assigning identifiers, keywords, tags or subject headings to documents. It may be useful to provide succinct, human-readable definitions of terms in the vocabulary, so that indexers and searchers are clear about what they refer to. At this point the structure becomes a *glossary.*

Information is commonly organized hierarchically, so that one can easily navigate from general topics to more specific ones. Such hierarchies, which we refer to as *taxonomies*, can define whatever organizational scheme its creators deem fit; they do not have to correspond to any formal relationships. For example, the taxonomy in Figure 2.3 does not distinguish between strict inheritance (e.g. *hiking* is a subclass of *outdoor activity*) and looser associations (e.g., *hiking equipment* is more specific but only thematically associated with *hiking*).

The one-to-one mapping enforced by controlled vocabularies is extremely restrictive, and makes both indexing and searching difficult. There is no guarantee that the author of a document, the librarian who indexed it, and the searcher who needs it will all converge on the same terminology. Consequently, an important next step is to encode synonymy, so that alternative labels can be connected to the preferred term for each concept. The resulting structure is often referred to as a *thesaurus*, but this thesis reserves the term for a slightly more complex structure. We instead take our definition from the *Guidelines for the establishment and development of monolingual thesauri* (ISO 2788), which describes three types of relations that thesauri are expected to express:

- The equivalence relation, which connects one or more terms (non-descriptors) to a single preferred term (descriptor), if they are synonymous.
- The hierarchical relation, which occurs between general and specific terms.
- The associative relation, which stands of any other kind of semantic relation.

In other words, a thesaurus is a controlled vocabulary that defines synonymy and a hierarchical organization scheme, and allows pairs of concepts to be associated with each other if they are, to quote the ISO standard, “mentally associated to such an extent that the link between them should be made explicit.” This last type of relation is difficult to

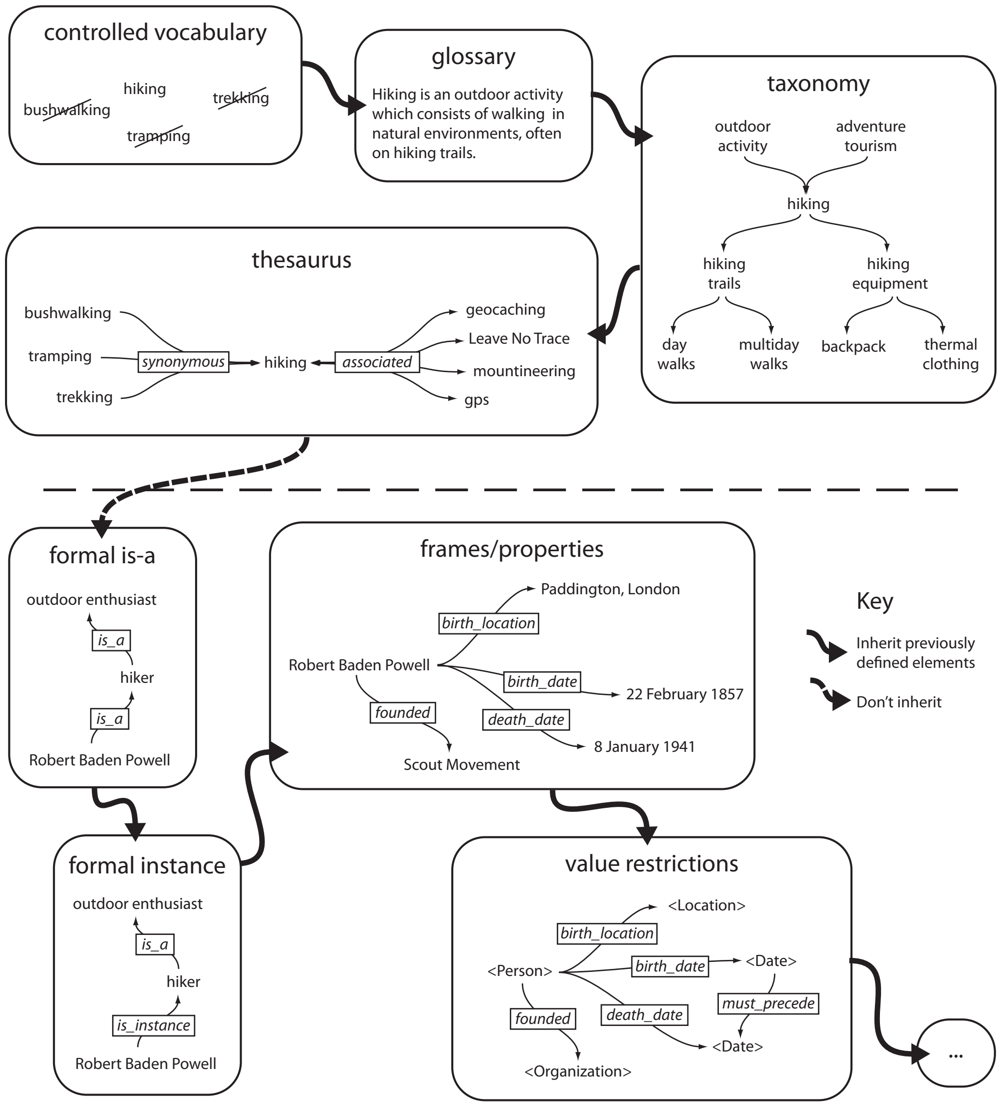


Figure 2.3: A running example of the knowledge base spectrum

define, but its use is recommended when it would be difficult to define or understand one concept without knowledge of another (e.g. *aesthetics* and *beauty*, or *aircraft* and *flight*).

Beyond this point the spectrum crosses a divide. The above structures are intended for human consumption, while those that follow can support artificial intelligence and automated inference. All of the remaining structures are referred to as *ontologies* within this thesis, although they continue to vary in complexity, expressiveness and ambition, as the spectrum indicates.

The first step towards ontologies, according to McGuinness (2005), is to tidy up hierarchical relations so that knowledge can be cleanly inherited. Here the loose categorization schemes described previously must be pruned to enforce strict subclass hierarchies: if A is a subclass of B, and B is a subclass of C, then it necessarily follows that A is a subclass of C. The chain *outdoor enthusiast* → *hiker* → *Robert Baden Powell* is valid, but *adventure tourism* → *hiking* → *hiking trails* is not. A related step is the separation of instances and the classes they instantiate. *Hiker* is a class that can be instantiated many times, but there is only one *Robert Baden Powell.*

The next step allows ontologies to describe an arbitrarily large vocabulary of relations and properties; that *Robert Baden Powell* has a *birth_location* and *birth_date*, for example. Such properties are more useful when they are specified at a general class level and inherited consistently by subclasses and instances: e.g., all subclasses and instances of *person* should have *birth_locations*. The next step places value restrictions on properties: e.g., that the *founded_by* relation must be between a *person* and an *organization*, and a *birth_date* must precede a *death_date*. The spectrum continues into yet more expressive structures, which may have the ability to describe parthood (the *Tongariro Alpine Crossing*—a famous hiking trail in New Zealand—is part of the *Tongariro Northern Circuit*), disjointness (an *organization* cannot also be a *person*), and a host of other features.

The following two subsections survey some of the more significant knowledge bases, and organize them by the breadth they aspire to: domain independent resources that attempt to describe everything (Section 2.2.1), and domain restricted resources that focus on a particular field (Section 2.2.2). Table 2.2 provides an overview. Section 2.2.3 briefly surveys approaches for gathering knowledge. Section 2.2.4 describes the Semantic Web.

2.2.1 Domain-independent knowledge bases

Domain independent knowledge bases are designed to be generally applicable to any subject area. One of the most well known and widely used examples is WordNet,[2] the lexical database produced by Princeton University (Miller 1995). This resource has a unique structure that was inspired by theories of human lexical memory. Within it, terms are organized into 120K "synsets", or groups of synonyms. The groups are connected to each other via a wide vocabulary of relationship types, including hyponomy, meronymy, antonymy, etc. It is generally considered to be an ontology, given that it describes a formal *is-a* and *is-instance-of* hierarchy in which knowledge can be cleanly inherited.

WordNet is a lexical database, and as such is primarily concerned with describing general concepts and lexical relations, rather than entities and factual information. For example, it contains 16 different nouns and 41 different verb definitions for the term *run*, but does not include a single competitive runner. It has seen extensive use in natural language processing, including part of speech tagging (Segond et al. 1997), information extraction

[2] WordNet can be downloaded or browsed online at *http://wordnet.princeton.edu*

Knowledge base	expressiveness	concepts
Domain Independent		
WordNet	Ontology	120K
Roget	Thesaurus	1K
Library of Congress Subject Headings	Taxonomy+Synonyms	370K
Open Directory Project	Taxonomy	590K
ResearchCyc	Ontology	250K
Freebase	Ontology	50M
Domain restricted		
AGROVOC	Thesaurus	17K
Medical Subject Headings	Taxonomy+Synonyms	24K
MusicBrains	Ontology	10M

Table 2.2: A selection of knowledge bases

(Chai and Biermann 1997, Califf and Mooney 1999) and most extensively in word sense disambiguation (Voorhees 1993, Resnik 1995a, Banerjee and Pedersen 2002). Mandala et al. (1998) provide a somewhat negative survey of various attempts to apply WordNet to information retrieval, and conclude that improvements are only possible when the resource is automatically expanded, using techniques similar to those described in Section 2.2.3. As Table 2.2 shows, without such efforts the resource has a limited vocabulary of 120K topics.

One of the oldest linguistic resources is Roget's Thesaurus,[3] which was first published in 1852. It is intended as a tool for writers—for those who are "painfully groping their way and struggling with the difficulties of composition" (Roget 1852). It contains limited structured knowledge and describes only a shallow hierarchy and groups of related terms. Although it has seen some use in natural language processing—Jarmasz (2003) provides a survey—it has significantly less use than WordNet: it is less comprehensive, less structured (it cannot be considered an ontology), and more difficult to obtain (the only open-source digital version we could locate is based on an edition released in 1911).

Another venerable resource is the Library of Congress subject headings (LCSH).[4] This thesaurus has been actively maintained by the Library of Congress since 1898 and currently contains 370K headings (or descriptors). The LCSH is used for indexing documents (a task described in Section 2.3.2) by many libraries in the United States and other countries. Unfortunately its owners have until recently been protective of the

[3] A (dated) open source version of Roget is available at *http://rogets.site.uottawa.ca*

[4] The LCSH can be downloaded or browsed online at *http://id.loc.gov/authorities/search*

resource (Summers 2008), and it has consequently garnered little interest from researchers. Paynter (2005) provides one exception: he attempts to automatically index documents against the resource.

The open directory project (ODP)[5] is a looser, web-based equivalent of the LCSH. This large-scale taxonomy recruits the public to organize the web hierarchically. It begins with 16 top-level nodes, including *arts*, *business*, *health* and *news*. These descend into more specific categories (e.g. *arts* → *visual arts* → *calligraphy* → *Japanese*) and individual web pages, such as *Shodo's Room* (the gallery of a calligrapher at Shinshu University). Each web page is summarized by a short description. The structure allows documents and categories to be accessed via multiple paths: e.g., *Wikipedia* is categorized as *reference* → *encyclopaedias* → *open content* and *computers* → *open source* → *open content*.

Given that the purpose of the ODP is to organize, group, and summarize web pages, it has naturally been put to use for related tasks such as text categorization (Davidov et al. 2004), clustering (Osiriski and Weiss 2004) and summarization (Berger and Mittal 2000). For most applications, it provides evaluation or training data, rather than background knowledge; unlike the previous resources, most efforts seem to be focused on automatically improving or expanding the directory rather than applying it to new problems. One exception is Gabrilovich and Markovitch (2005), who use the ODP and the pages it organizes to provide new features for organizing texts against arbitrary categorization schemes (not merely the one expressed by the ODP). This work is described in Section 2.3.2.

The Cyc ontology[6] is an ambitious project on which ontologists and philosophers have toiled for more than 20 years (Lenat et al. 1990). It was created to allow artificial intelligence systems to perform common-sense reasoning, by capturing the facts one would expect any normal adult to know. Its modest size (Table 2.2) is due to the sophistication and care with which the knowledge is captured: it is rigorously defined and wrapped by an inference engine that supports first-order logic.

Cyc was never intended to directly resolve people's information needs; after all, its aim is to cover facts that the average person should already know and wouldn't need to search for. For retrieval purposes, it is instead expected to assist in gathering information from *other* resources. It could, for example, sanity check facts (e.g. to reject *1980* as a person's age) or extrapolate and fill in missing assertions (e.g. by calculating someone's age given

[5] The ODP can be browsed online at *http://www.dmoz.org* and downloaded from *http://rdf.dmoz.org*

[6] Academic licenses for ResearchCyc can be obtained from *http://research.cyc.com*

their birth date). Cyc's founders see it as a central component of the Semantic Web (Section 2.2.4), which inevitably consists of a scattered array of divergent and contradictory ontologies that requires extensive common-sense reasoning to unify (Reed and Lenat 2002). These visions are compelling, but it is disappointing that Cyc has not yet found practical applications, and has failed to reach any natural endpoint—an article from New Scientist (Anonymous 2006), its designers estimate that it contains only 2% of what it needs to emulate human intelligence. Table 2.2 shows the size of ResearchCyc, a version that is available for academic use. There is also a substantially smaller open source version called OpenCyc.

Freebase[7] is a highly structured open-domain ontology that is mined from a wide variety of sources, including Wikipedia, MusicBrainz and the Notable Names Database. It is another commercial rather than academic venture, so again little is known about the algorithms used to mine these sources. One significant difference between it and the above resources is the inclusion of extensive mechanisms by which the public can contribute new knowledge via an online wiki-like interface and programmatic APIs. Depending upon the extent to which these mechanisms are relied upon—something we could not quantify—this puts Freebase into the realm of crowd-sourced knowledge and raises many interesting implications. The pros and cons of crowd sourcing are surveyed, with respect to Wikipedia, in Section 3.2, but here they are complicated by an additional expectation for laypeople to become amateur ontologists. Their ability to take on this role is unknown, and to our knowledge the resulting structure has never been evaluated. Its sheer scale is certainly encouraging (see Table 2.2).

2.2.2 Domain-specific knowledge bases

Many knowledge bases sacrifice breadth for depth by focusing on a particular domain or subject area. The multilingual thesaurus Agrovoc[8] is particularly pertinent to this thesis. Developed and maintained by the UN Food and Agriculture Organization to organize and facilitate access to its extensive repository of reports on agriculture, agriculture, forestry, fisheries, food and related domains, it contains all of the elements of thesauri and the less expressive structures. The English version defines 17K descriptors (many with glossaries/scope notes), 11K non-descriptors (i.e. alternative labels for descriptors), 16K hierarchical and 27K associative relations. Agrovoc is revisited in Section 3.5, where it is directly compared to Wikipedia.

[7] Freebase can be browsed online, downloaded, or queried programmatically from *http://www.freebase.com*

[8] Agrovoc can be browsed online at *http://aims.fao.org/website/Search-AGROVOC/sub*. An English version can be download from *http://www.nzdl.org/Kea/Download*

The Medical Subject Headings (MeSH)[9] is developed by the U.S. National Library of Medicine for indexing the PubMed repository. It is essentially a domain specific version of the LCSH, and contains 24,000 descriptors and 117,000 non-descriptors, organized into a taxonomy of 32,000 hierarchical relations.

When discussing domain-specific ontologies, it is useful to distinguish between factual and problem-solving resources. The former provide machine-readable information about concepts of interest, and are essentially more formalized and expressive versions of the thesauri described above. A well-known example is MusicBrainz,[10] a crowd-sourced ontology for describing commercial music in a highly structured fashion (Swartz 2002). As of March 2010 it includes 540K artists, 790K releases, and 9M tracks. Problem solving ontologies, in contrast, focus on describing strategies and algorithms for achieving goals. Some examples are the Music Ontology, which specifies how to capture and describe music-related information consistently (Raimond et al. 2007), and the COMUS ontology for mood-based music recommendation (Rho et al. 2009). This thesis is concerned only with factual knowledge bases.

2.2.3 The knowledge acquisition bottleneck

The cost of obtaining knowledge is well known to be prohibitive. The resources described above are generally produced and maintained by small panels of professional domain experts, indexers and ontologists, and are expensive and time-consuming to construct and maintain. As a result they are often protected through copyright, have limited ability to cover broad or swiftly evolving domains, and are prone to inconsistency and bias. Table 2.3 provides an example of the latter hazard, in which thesaurus.com—an online version of Roget's Thesaurus II—describes *conservative* and *liberal* in decidedly different tones. These words admittedly have many interpretations, but the inclusion of *Tory, right* and *left* demonstrates that the political senses are encompassed here.

One option for overcoming these problems is to extract knowledge bases automatically from text. The resulting resources would be economical to construct and update, and more closely tailored to individual corpora than even the most carefully matched domain specific structures. Unfortunately, the inherent complexity and ambiguity of natural language makes deriving knowledge bases from text extremely challenging. It encompasses entire fields like computational linguistics, natural language processing and

[9] MeSH can be used online at *http://www.nlm.nih.gov/mesh/MBrowser.html* and downloaded from *http://thesauri.cs.vu.nl/eswc06*

[10] MusicBrainz can be browsed online, downloaded and queried programmatically at *http://musicbrainz.org*

Conservative	*http://thesaurus.com/browse/conservative*
Tory, bourgeois, fearful, firm, fogyish, fuddy-duddy, guarded, in a rut, inflexible, middle-of-the-road, obstinate, orthodox, reactionary, redneck, *right*, sober, stable, timid, traditional, uncreative, unimaginative, white bread.	
Liberal	*http://thesaurus.com/browse/liberal*
Advanced, avant-garde, broad-minded, catholic, enlightened, high-minded, humanitarian, indulgent, intelligent, *left*, lenient, magnanimous, rational, receptive, reformist, tolerant, unbiased, unbigoted, unprejudiced.	

Table 2.3: Roget's associations for *conservative* and *liberal*

information extraction, and is out of scope for this investigation. Those who are interested in it are directed to Widdows (2004) and Grefenstette (1994): the first book provides a gentle, non-technical introduction, while the second delves deeper into generating thesauri and lesser structures automatically. Section 3.4.4 of this thesis provides a survey of deriving ontologies from Wikipedia's structure and textual content.

This thesis pursues a different solution to the knowledge acquisition bottleneck: utilising large communities of human volunteers. Table 2.2 demonstrates the advantage of "crowd sourcing" over the traditional expert-driven approach: two of the three largest resources (Freebase and MusicBrainz) are crowd-sourced. The pros and cons of relying on volunteers are examined more carefully in Section 3.2. For now, we roughly characterize the types of knowledge bases crowd sourcing and automatic generation are likely to yield.

Formal ontologies are problematic for both crowd sourcing and automatic generation. The bar that separates these structures from lesser ones in Figure 2.2 is not a gradual, continuous transition from one type of knowledge base to a slightly more expressive one. It is instead a leap in expectations and ambitions. The key transition is from human-readable to machine-readable knowledge, and from loose associations to hard facts. Another strong—though not universal—shift is from metadata to data. The human-readable structures are not typically used on their own, but rather support other information sources with index terms, query suggestions, term definitions, and additional semantics to enhance document representations. Towards the right of the spectrum, there appears to be a growing ambition to produce self-sufficient resources that people and machines can turn to directly to resolve their information needs (e.g., Freebase and Wolfram|Alpha). This is an entirely different proposition, and requires a far greater volume of knowledge.

Putting aside the metadata vs. data issue, both scenarios—automatic generation and crowd sourcing—are better suited to the left side of the spectrum than the right. Extracting thesauri and lesser structures automatically is challenging enough, but generating ontologies from raw text is much more so. Whereas informal relations can be obtained via grammatical patterns and broad statistical trends, formal relations cannot be obtained without deep natural language processing. As for crowd sourcing, it is naturally easier and more forgiving to educate and motivate human volunteers to contribute loose, human-readable knowledge. Although machine-readable knowledge can be obtained via the same mechanism—e.g. the Freebase ontology has extensive mechanisms to support crowd sourcing—it is not clear that this could achieve widespread, unbiased enthusiasm from volunteer contributors. Even systems that allow users to contribute freely in their own words suffer bias because of the kind of people they attract and retain (Section 3.2.4).

2.2.4 The Semantic Web

The Semantic Web is a detailed plan to render the web—or at least great chunks of it—machine-readable: "to take us, step by step, from the Web of today to a Web in which machine reasoning will be ubiquitous and devastatingly powerful (Berners-Lee 1998a)." This section provides a brief overview of the vision, if only to explain why it is out of scope.

Currently, the web keeps a great deal of information—concert times, weather forecasts, camera specifications, television guides—frustratingly out of reach for machines. Even data that is stored neatly in relational databases and the like becomes a jumbled mess (from the perspective of a machine) when published on the Internet. As a result, gathering and reusing information is much more arduous and error-prone than it needs to be. The Semantic Web is an emerging initiative to capture and publish machine-readable knowledge in a carefully planned and thoroughly documented fashion that encourages reuse.

The above proposal sounds reasonable and pragmatic, but is the start of a slippery slope that becomes extremely ambitious, as indicated by Berners-Lee's dream of having machines handle our daily lives (see the start of this chapter). Deep open-domain reasoning over the entire web is a far cry from encouraging data reuse. Berners-Lee's grand vision has received widespread criticism (Shirky 2003, McCool 2005, Schoop et al. 2006, Legg 2007). It opens up enormous, arguably insurmountable challenges: how could the entirety of the web—gigantic, messy, ever evolving and often contradictory—be exhaustively rendered machine-readable and stored within a consistent schema that could be efficiently traversed and applied automatically to arbitrary problems?

Even the less ambitious vision—where data is rendered machine-readable by hand; where scale, domain coverage, expressivity and interconnectivity are all determined organically; and where the services that use and combine data sources are handcrafted to do so—involves both technical and social challenges. On the technical side, how can a single description language balance the flexibility required to describe anything against the consistency needed for reuse and the simplicity required for voluntary adoption? On the social side, how can individuals and organizations be motivated and educated to adopt the same description language and apply it properly?

Detailed plans have emerged for resolving the technical issues (Antoniou and Van Harmelen 2004). At the lowest level is the Resource Description Framework (RDF), which simply allows relations and properties to be described in subject-predicate-object triples (e.g. *Resource Description Framework—has acronym—RDF)*. Built on top of this is an ever-growing array of problem-solving ontologies (see Section 2.2.2) to specify how various domains should be described using the framework. Some notable examples are the *Friend Of A Friend (FOAF)* ontology for describing social networks; the *Music Ontology* (introduced in Section 2.2.2) for music-related information; the *Simple Knowledge Organization System (SKOS)* for describing thesauri, taxonomies, and lesser structures; and the *Web Ontology Language* family *(OWL Lite, OWL DL, OWL Full)* for describing more complex knowledge bases—in other words, a set of ontologies for specifying ontologies. Surrounding the specifications is an ever-expanding toolkit of parsers, validators, storage solutions, search and inference engines, and so on. The social issues are less thoroughly addressed: Alani et al. (2008) provide one of the few discussions about motivating and educating individuals and businesses.

The Semantic Web has limited relevance to this investigation. It would be desirable to feed any structured knowledge we obtain from Wikipedia into the Semantic Web, but we do not have to pursue this intersection directly. The whole point of building description vocabularies on top of the RDF framework is to allow any form of structured knowledge to be incorporated. Thus any knowledge base we extract from Wikipedia—taxonomy, thesaurus, ontology or something else entirely—will be translatable and applicable.

	scale	accuracy	expressiveness				
Section 2.3.1							
automatic query expansion	high	low	○	✓	✓	✓	○
interactive query expansion	high	medium	○	✓	✓	✓	○
relevance feedback	high	low	○	✓	○	○	○
Section 2.3.2							
tagging and indexing	high	medium	✓	✓	✓	✓	○
categorization	medium	high	✓	✓	✓	✓	○
Section 2.3.3							
clustering	high	low	○	✓	○	○	○
Section 2.3.4							
faceted browsing	medium	high	○	○	○	○	✓
Section 2.3.5							
personalization	high	low	○	✓	✓	○	○
adaptive hypermedia	high	low	○	✓	✓	○	○
Section 2.3.6							
visualization*	N/A	N/A	○	○	○	○	○
			↑ controlled vocabulary	↑ glossary	↑ taxonomy	↑ thesaurus	↑ ontology

* No knowledge base stands out as more or less applicable in this area.

Figure 2.4: Applying knowledge bases to interactive information retrieval

2.3 *Applying knowledge bases to interactive information retrieval*

This section surveys a range of strategies for facilitating interactive information retrieval, and discusses the relevance and applicability of knowledge bases to them. Figure 2.4

provides an overview, listing each strategy and charting the priorities it places on background knowledge. *Scale* refers to the depth, breadth and adaptability of the knowledge base: its ability to exhaustively describe the topics and concepts mentioned within documents. *Accuracy* refers to how carefully the knowledge base must be constructed; from rough automatic processing, through crowd sourcing, and finally the use of professional indexers and domain experts. *Expressiveness* refers to the spectrum of knowledge bases—from simple controlled vocabularies to heavyweight ontologies—that was described in Section 2.2. The values shown in this figure are explained and justified in the sections that follow. They are intended only for rough comparison and are open to debate.

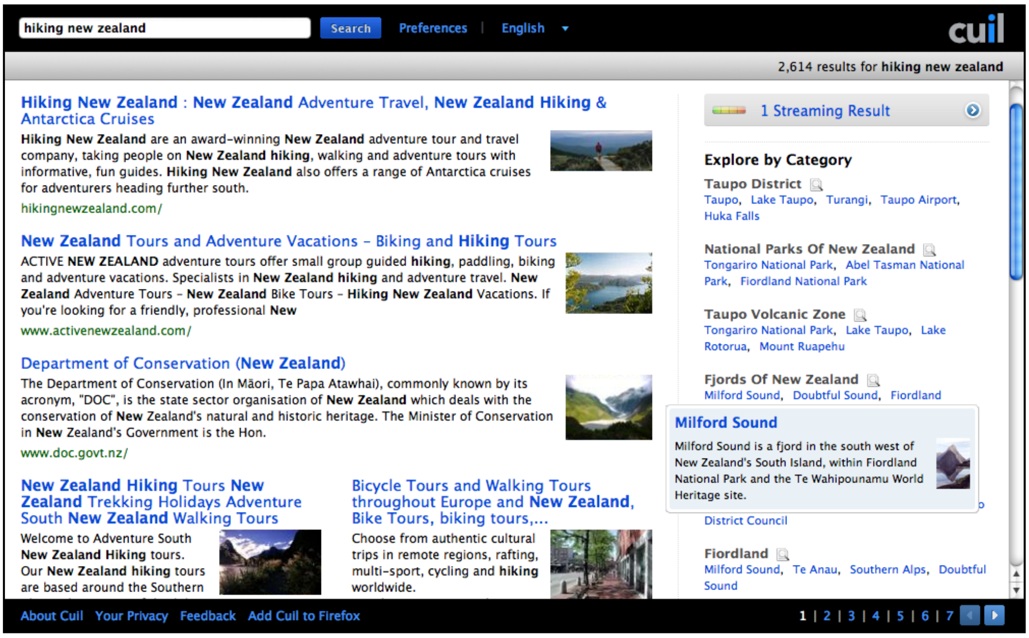
2.3.1 Query expansion and refinement

Queries are the primary—in many cases the only—means of interaction with current retrieval systems. There exists a great deal of work in helping users describe their information needs more effectively, or allowing them to easily transition to new queries as their information needs evolve.

Query expansion and refinement can be either automatic or interactive. Automatic question answering assumes searchers know what they are looking for, but are not using the most effective terminology to locate it. It is widely performed without any form of knowledge base. For example, almost all search engines expand queries through stemming (Porter 1980). Google and other well-established search engines use analysis of query behaviour and click-through data to augment queries. At first glance the synonyms and narrower terms available in manually defined thesauri would seem immediately applicable, but in practice their performance has been disappointing. Better results are obtained with automatically generated corpus-based thesauri, which are much less accurate and expressive but many times larger and more closely connected to the documents being searched (Schütze and Pedersen 1997). Mandala et al. (1999) demonstrate that the two sources—manually-defined and automatically-derived thesauri—complement each other well and can be combined for effective query expansion. Specifically, they augment WordNet (Section 2.2.1) with thesauri gathered through co-occurrence and grammatical analysis of corpora. The key limitations they identify in WordNet are its separation of nouns and verbs even when they are conceptually related, and its inability to describe domain-specific topics (particularly proper names) and relations (which can be attributed to its modest size). Large, manually defined knowledge bases that overcome these limitations without sacrificing quality would be valuable. Accuracy and expressivity are not strong priorities compared to scale, however, given that automatically derived loose association thesauri currently outperform manually defined resources.

Information needs evolve (Section 2.1). As searchers continue to interact with a retrieval system, they gain a clearer idea of the information space and refine their requirements accordingly. Interactive query expansion can accelerate the process by constantly suggesting options for where to go next. Manually defined knowledge is more useful in this interactive setting, where terms and relations are directly exposed to the user and not hidden behind the search box.

Suitable knowledge bases for interactive query expansion must be comprehensive (otherwise they have little chance of providing useful paths) and accurate (the relations



2,614 results for **hiking new zealand**

Hiking New Zealand : **New Zealand** Adventure Travel, **New Zealand Hiking** & Antarctica Cruises

Hiking New Zealand are an award-winning **New Zealand** adventure tour and travel company, taking people on **New Zealand hiking**, walking and adventure tours with informative, fun guides. **Hiking New Zealand** also offers a range of Antarctica cruises for adventurers heading further south.

hikingnewzealand.com/

New Zealand Tours and Adventure Vacations - Biking and **Hiking** Tours

ACTIVE **NEW ZEALAND** adventure tours offer small group guided **hiking**, paddling, biking and adventure vacations. Specialists in **New Zealand hiking** and adventure travel. **New Zealand** Adventure Tours - **New Zealand** Bike Tours - **Hiking New Zealand** Vacations. If you're looking for a friendly, professional **New**

www.activenewzealand.com/

Department of Conservation (**New Zealand**)

The Department of Conservation (In Māori, Te Papa Atawhai), commonly known by its acronym, "DOC", is the state sector organisation of **New Zealand** which deals with the conservation of **New** Zealand's natural and historic heritage. The Minister of Conservation in **New** Zealand's Government is the Hon.

www.doc.govt.nz/

New Zealand Hiking Tours **New Zealand** Trekking Holidays Adventure South **New Zealand** Walking Tours

Welcome to Adventure South **New Zealand Hiking** tours. Our **New Zealand hiking** tours are based around the Southern

Bicycle Tours and Walking Tours throughout Europe and **New Zealand**, Bike Tours, biking tours,...

Choose from authentic cultural trips in remote regions, rafting, multi-sport, cycling and **hiking** worldwide.

1 Streaming Result

Explore by Category

Taupo District
Taupo, Lake Taupo, Turangi, Taupo Airport, Huka Falls

National Parks Of New Zealand
Tongariro National Park, Abel Tasman National Park, Fiordland National Park

Taupo Volcanic Zone
Tongariro National Park, Lake Taupo, Lake Rotorua, Mount Ruapehu

Fjords Of New Zealand
Milford Sound, Doubtful Sound, Fiordland

Milford Sound
Milford Sound is a fjord in the south west of New Zealand's South Island, within Fiordland National Park and the Te Wahipounamu World Heritage site.

District Council

Fiordland
Milford Sound, Te Anau, Southern Alps, Doubtful Sound

Figure 2.5: Cuil, with categorized query suggestions for *hiking new zealand*

will be used directly) but not particularly expressive. Synonymy and associative relations are the minimal requirement, but many systems do not bother to distinguish between them: e.g., Lee et al. (2001). Glossary descriptions are helpful for explaining the suggestions, and hierarchical relations can be used to organize them. Both of these features are demonstrated by the search engine Cuil, pictured in Figure 2.5: here suggestions for *hiking new zealand* are organized into categories, and explained through tooltips. Hierarchical relations can also be used to distinguish between query refinement and generalization—making overly general queries more specific and vice versa. A user study by Shiri and Revie (2006) suggests that separating hierarchical, associative and synonymy relations is desirable, because they are used differently by users with different levels of experience. The utility of more formal knowledge bases for query expansion is debatable. García and Sicilia (2003) provide an exhaustive list of opportunities that ontologies provide. However, the features they propose can be facilitated by lesser structures, and the system built to demonstrate their ideas—OntoIR—has never been tested against other search systems.

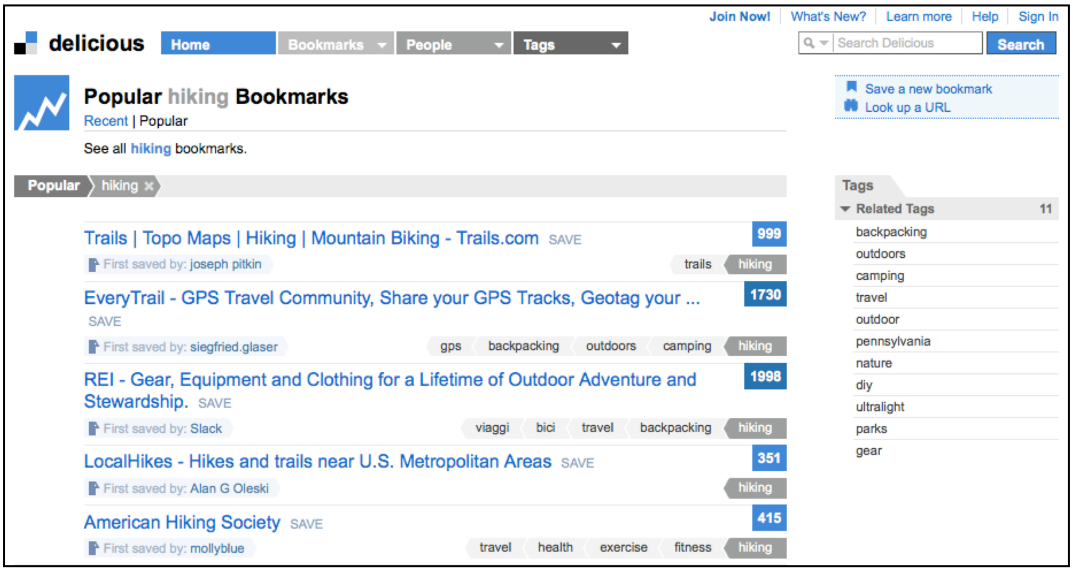
Real-time query completion is a popular form of query expansion, where suggestions are made as the user types their query. The list of suggestions is refined with each character entered. This assists users when they are unsure of the terminology, and provide shortcuts when they know exactly what to type. Unfortunately it can also lead to query drift, where a user is distracted from the original intent before they have had a chance to satisfy it (White and Marchionini 2007). Real-time query suggestions are typically generated without consultation with knowledge bases, by mining the available documents or queries that have been issued over them. A knowledge base could potentially be used to compare suggestions against the wider context of who the person is and what they are looking for (more on this in Section 2.3.5). This would help to reduce the disorientation and query drift that White and Marchionini warn against, but to our knowledge has never been attempted. If it were, the required knowledge base would have to be extremely comprehensive, but not particularly accurate (it would not be presented directly to the user) or expressive (only loose associations would be required).

A common and extensively explored strategy for query expansion is relevance feedback, where the documents returned are mined for significant terms, which are fed back into the query (Ruthven and Lalmas 2003). In pseudo-relevance feedback, the documents are gathered automatically and the user is unaware of the query expansion being performed. True relevance feedback is more interactive: here the user specifies the documents that are used to augment the query, either by explicitly marking them in some way or implicitly expressing interest via click-through data or other means.

To our knowledge, the only successful application of knowledge bases to relevance feedback is Egozi et al. (2008), which uses Wikipedia to provide a concept-based representation—as opposed the to the usual bag-of-words—of the documents. This representation was used to augment the four top performing systems from the TREC-8 competition and provided 4%–15% improvement in mean average precision, depending on the system being augmented. These gains may sound modest, but are impressive given that the performance of relevance feedback systems plateaued a decade ago (Armstrong et al. 2009). Egozi et al. argue that the size and accuracy of the knowledge base are critical factors: neither small, carefully crafted resources nor large automatically generated ones are likely to be of any use. However, the expressivity of the structure does not seem to be a factor: in this work Wikipedia is essentially treated as a glossary in which only article titles and textual content are used.

2.3.2 Tagging, indexing and categorization

Users of sites such as Delicious and Flickr invest a great deal of effort to tag documents with descriptive terms (Golder and Huberman 2006, Marlow et al. 2006). Figure 2.6 demonstrates some of the rich opportunities for exploration that these tags offer. They can succinctly summarize search results, group related documents together, provide related tags to explore, and describe trends in the collection as a whole (e.g. recency, popularity). The tagging activity also generates a community in which users can locate others who share similar interests. This social aspect is explored in Section 2.3.5.



Popular hiking Bookmarks

Recent | Popular

See all hiking bookmarks.

Trails | Topo Maps | Hiking | Mountain Biking - Trails.com SAVE 999
First saved by: joseph pitkin — trails hiking

EveryTrail - GPS Travel Community, Share your GPS Tracks, Geotag your ... SAVE 1730
First saved by: siegfried.glaser — gps backpacking outdoors camping hiking

REI - Gear, Equipment and Clothing for a Lifetime of Outdoor Adventure and Stewardship. SAVE 1998
First saved by: Slack — viaggi bici travel backpacking hiking

LocalHikes - Hikes and trails near U.S. Metropolitan Areas SAVE 351
First saved by: Alan G Oleski — hiking

American Hiking Society SAVE 415
First saved by: mollyblue — travel health exercise fitness hiking

Tags

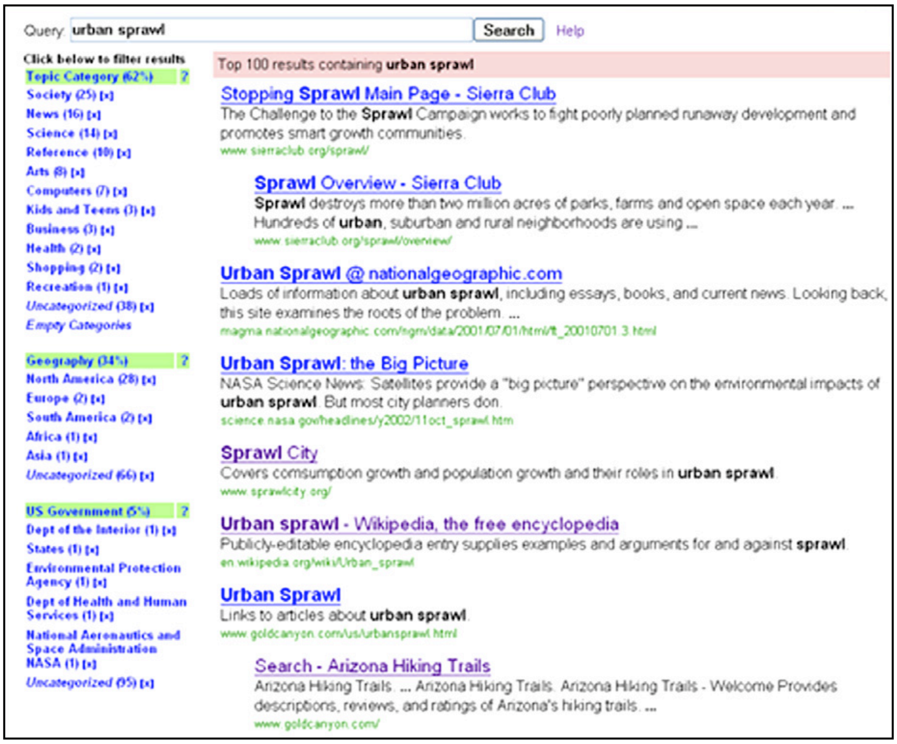
Related Tags 11

backpacking
outdoors
camping
travel
outdoor
pennsylvania
nature
diy
ultralight
parks
gear

Figure 2.6: Browsing popular *hiking* bookmarks with Delicious

Tagging is typically done without background knowledge: users are free to type any term or phrase as a tag. The resulting vocabularies—known as folksonomies—are easy to construct and maintain, but significantly less accurate and consistent than more traditionally derived structures (Noruzi 2007). Delicious, for example, contains the tags *hiking*, *hike*, *tramping*, *hikers*, *walking* and *bushwalking*, with no explicitly defined relations to connect them. None of these synonymous tags are made available in Figure 2.6. This results in brittle searching, where resources are easily missed because the anticipated tag was not used (Morrison 2008).

Knowledge bases offer direct advantages to tagging with folksonomies: a controlled vocabulary immediately enforces consistency; a glossary explains what the tags mean; a taxonomy allows resources tagged with *hiking* to be implicitly tagged with broader terms like *outdoor recreation* and *leisure activities*; synonymy relations make searching less brittle; and associative relations provide better suggestions for new tags to browse (Section 2.3.1). Having some ability to reason about what tags are and how they are connected greatly increases the accuracy of automated and semi-automated tagging systems (Medelyan 2009). Despite these clear advantages, tagging against structured knowledge—a task known as controlled indexing—is rarely seen outside libraries and specialist collections. This mismatch is a strong indication of how difficult it is to obtain knowledge bases at the scale required for open-domain information retrieval: if suitable resources were available, social tagging is one of the first scenarios in which they would find widespread use.



Query: urban sprawl

Top 100 results containing **urban sprawl**

Stopping Sprawl Main Page - Sierra Club
The Challenge to the **Sprawl** Campaign works to fight poorly planned runaway development and promotes smart growth communities.
www.sierraclub.org/sprawl/

Sprawl Overview - Sierra Club
Sprawl destroys more than two million acres of parks, farms and open space each year. ... Hundreds of **urban**, suburban and rural neighborhoods are using ...
www.sierraclub.org/sprawl/overview/

Urban Sprawl @ nationalgeographic.com
Loads of information about **urban sprawl**, including essays, books, and current news. Looking back, this site examines the roots of the problem. ...
magma.nationalgeographic.com/ngm/data/2001/07/01/html/ft_20010701.3.html

Urban Sprawl: the Big Picture
NASA Science News: Satellites provide a "big picture" perspective on the environmental impacts of **urban sprawl**. But most city planners don.
science.nasa.gov/headlines/y2002/11oct_sprawl.htm

Sprawl City
Covers consumption growth and population growth and their roles in **urban sprawl**.
www.sprawlcity.org/

Urban sprawl - Wikipedia, the free encyclopedia
Publicly-editable encyclopedia entry supplies examples and arguments for and against **sprawl**.
en.wikipedia.org/wiki/Urban_sprawl

Urban Sprawl
Links to articles about **urban sprawl**.
www.goldcanyon.com/us/urbansprawl.html

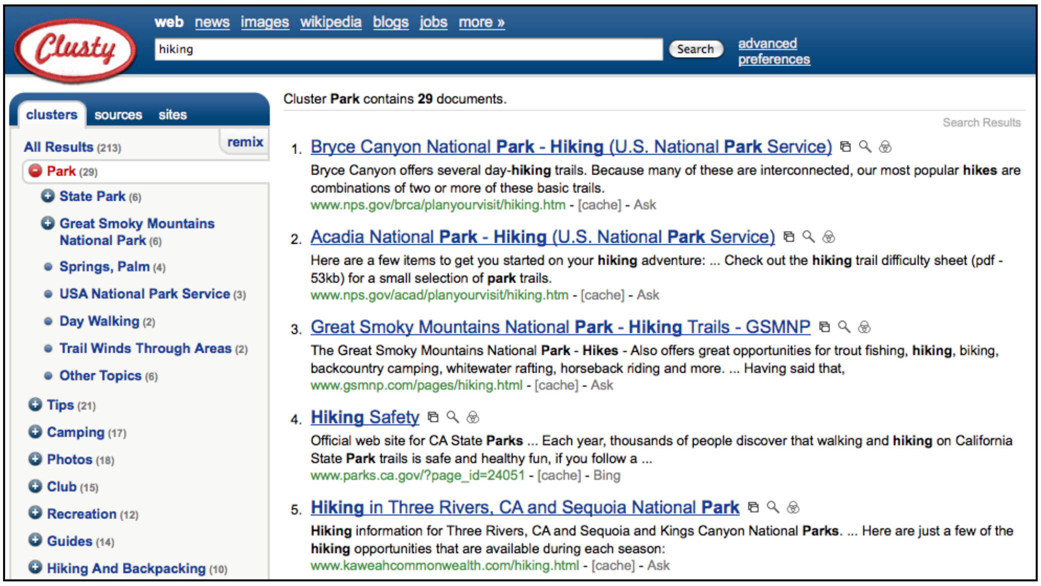
Search - Arizona Hiking Trails
Arizona Hiking Trails. ... Arizona Hiking Trails. Arizona Hiking Trails - Welcome Provides descriptions, reviews, and ratings of Arizona's hiking trails. ...
www.goldcanyon.com/

Figure 2.7: Using categories to filter search results about *urban sprawl* (from Kules and Shneiderman 2008)

Scale is the paramount concern for applying knowledge bases to tagging and indexing. Accuracy is only moderately important: relations are directly placed in front of the user, but popular systems like Delicious and Flickr make do with automatic approximations. Expressivity is only important up to a point: the opportunities offered by controlled vocabularies, glossaries, taxonomies and thesauri were described above, but the utility of more formalized knowledge is not obvious. Angeletou et al. (2008) augment folksonomies with knowledge obtained from ontologies, but do not explain why this is better than using simpler structures.

Text categorization (or classification) involves labelling documents with categories, such business, sport, or entertainment. This is a similar but subtly different problem to tagging and indexing, the difference being that the pool of labels is much smaller and carefully controlled, and the activity is always performed with at least a controlled vocabulary (the category labels themselves). Categories can be browsed directly, or used to interactively filter search results as shown in Figure 2.7. The latter strategy encourages users to explore results more deeply, without increasing the perceived complexity of the system (Kules and Shneiderman 2008).

When performing automatic categorization, it is helpful to have background knowledge about what the labels mean. For example, Davidov et al. (2004) automatically categorize



hiking

- All Results (213)
- Park (29)
  - State Park (6)
  - Great Smoky Mountains National Park (6)
  - Springs, Palm (4)
  - USA National Park Service (3)
  - Day Walking (2)
  - Trail Winds Through Areas (2)
  - Other Topics (6)
- Tips (21)
- Camping (17)
- Photos (18)
- Club (15)
- Recreation (12)
- Guides (14)
- Hiking And Backpacking (10)

Cluster **Park** contains **29** documents.

1. Bryce Canyon National **Park** - **Hiking** (U.S. National **Park** Service)
   Bryce Canyon offers several day-**hiking** trails. Because many of these are interconnected, our most popular **hikes** are combinations of two or more of these basic trails.
   www.nps.gov/brca/planyourvisit/hiking.htm - [cache] - Ask
2. Acadia National **Park** - **Hiking** (U.S. National **Park** Service)
   Here are a few items to get you started on your **hiking** adventure: ... Check out the **hiking** trail difficulty sheet (pdf - 53kb) for a small selection of **park** trails.
   www.nps.gov/acad/planyourvisit/hiking.htm - [cache] - Ask
3. Great Smoky Mountains National **Park** - **Hiking** Trails - GSMNP
   The Great Smoky Mountains National **Park** - **Hikes** - Also offers great opportunities for trout fishing, **hiking**, biking, backcountry camping, whitewater rafting, horseback riding and more. ... Having said that,
   www.gsmnp.com/pages/hiking.html - [cache] - Ask
4. **Hiking** Safety
   Official web site for CA State **Parks** ... Each year, thousands of people discover that walking and **hiking** on California State **Park** trails is safe and healthy fun, if you follow a ...
   www.parks.ca.gov/?page_id=24051 - [cache] - Bing
5. **Hiking** in Three Rivers, CA and Sequoia National **Park**
   Hiking information for Three Rivers, CA and Sequoia and Kings Canyon National **Parks**. ... Here are just a few of the **hiking** opportunities that are available during each season:
   www.kaweahcommonwealth.com/hiking.html - [cache] - Ask

Figure 2.8: Exploring *hiking*—and the *park* cluster specifically—with Clusty

documents against the open directory project (described in Section 2.2.1) by consulting the text of web pages already organized by the taxonomy. Here the knowledge base is a direct extension of the category structure that documents are to be assigned to, but this need not be the case. Gabrilovich and Markovitch (2005), for example, use the open directory project to inform categorization against arbitrary controlled vocabularies, and in later work do the same with Wikipedia (Gabrilovich and Markovitch 2006). None of these systems require knowledge bases more expressive than a glossary, but size and accuracy are significant: Wikipedia outperforms the open directory project because it is a larger, cleaner source of concepts.

2.3.3 Clustering

Document clustering—automatically identifying and grouping documents that are conceptually related—is an extensively researched tool for information retrieval. It was originally proposed for improving searches automatically by emphasizing documents that are similar to those that were previously identified as relevant (Willett 1988). Relevance feedback (Section 2.3.1) has proven to be a cheaper and more effective technique, and clustering is now rarely used to improve query results automatically.

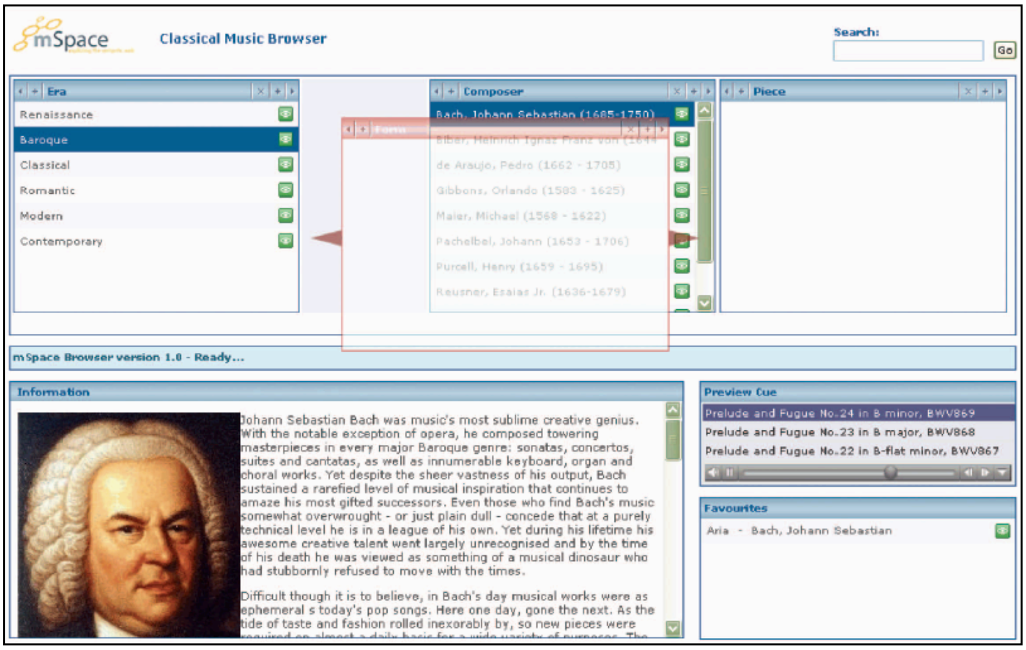
Cutting et al. (1992) were the first to recognize clustering's potential to facilitate interactive browsing. They proposed the famous *Scatter/Gather* technique, in which users are presented with clusters to represent the main themes of a document collection. They select one or more clusters of interest, and re-cluster the relevant documents. The process continues until the search space is sufficiently narrow for linear browsing. Hearst and Pedersen (1996) found this technique to be labour-intensive for the user and inferior to keyphrase search, but also found that the two techniques complemented each other well: a search could be issued as a first step, and clustering could be used to organize and explore the results. Figure 2.8 demonstrates this idea with a screen-shot from the Clusty[11] search engine, in which the user has searched for *hiking* and chosen to focus on the *park* cluster.

Clustering has much the same requirements for structured knowledge as relevance feedback (Section 2.3.1) and categorization (Section 2.3.2). It is almost always performed without any form of background knowledge, but can stand to benefit from reducing bag-of-words document representations to—or augmenting them with—key concepts (depending on whether efficiency or accuracy is a priority), and consulting the relations between concepts to reduce brittleness (Huang et al. 2009a). As before, expressivity is not of concern, and the utility of ontologies is not obvious. Hotho et al. (2001) is a widely cited paper on applying ontologies to the task, but on close inspection the knowledge base they use contains only hierarchical and synonymy relations. Pratt et al. (1999) advocate using both small, handcrafted ontologies and large domain-specific taxonomies when clustering query results. Factual knowledge is provided exclusively by the latter, and the use of ontologies is restricted to problem-solving knowledge on how to identify query types.

Of course, a clustered document collection is only a rough approximation of what that collection would be like if indexed or categorized against a well-crafted taxonomy. Figure 2.8 provides several examples of the limitations of clustering: Why is *day walking* organized under *parks*? Where is its obvious counterpart, *multi-day walking*? What kind of a cluster is *trail winds through areas*? Why is the fourth search result assigned to the *park* cluster, when it is clearly about safety and therefore belongs under *tips*?

Document categorization and clustering provide the same opportunities for exploration, as a quick comparison of Figure 2.7 and Figure 2.8 will demonstrate. The former has proven to be more useable—more consistent, coherent and predictable—in head-to-head comparisons (Hearst 2006). The only advantage of the latter is that it doesn't require a manually crafted set of categories, or annotation against them. Thus, it makes little sense to insist on clustering if a suitable taxonomy is available. Like tagging (Section 2.3.2), the prevalence of clustering is a symptom of how difficult it is to obtain knowledge bases at the scale required, and annotate documents against them.

[11] http://clusty.com/


mSpace
Classical Music Browser
Search:
Go
Era
Renaissance
Baroque
Classical
Romantic
Modern
Contemporary
Form
Composer
Bach, Johann Sebastian (1685-1750)
Biber, Heinrich Ignaz Franz von (1644
de Araujo, Pedro (1662 - 1705)
Gibbons, Orlando (1583 - 1625)
Maier, Michael (1568 - 1622)
Pachelbel, Johann (1653 - 1706)
Purcell, Henry (1659 - 1695)
Reusner, Esaias Jr. (1636-1679)
Piece
mSpace Browser version 1.0 - Ready...
Information
Preview Cue
Prelude and Fugue No.24 in B minor, BWV869
Prelude and Fugue No.23 in B major, BWV868
Prelude and Fugue No.22 in B-flat minor, BWV867
Favourites
Aria - Bach, Johann Sebastian

Johann Sebastian Bach was music's most sublime creative genius. With the notable exception of opera, he composed towering masterpieces in every major Baroque genre: sonatas, concertos, suites and cantatas, as well as innumerable keyboard, organ and choral works. Yet despite the sheer vastness of his output, Bach sustained a rarefied level of musical inspiration that continues to amaze his most gifted successors. Even those who find Bach's music somewhat overwrought - or just plain dull - concede that at a purely technical level he is in a league of his own. Yet during his lifetime his awesome creative talent went largely unrecognised and by the time of his death he was viewed as something of a musical dinosaur who had stubbornly refused to move with the times.

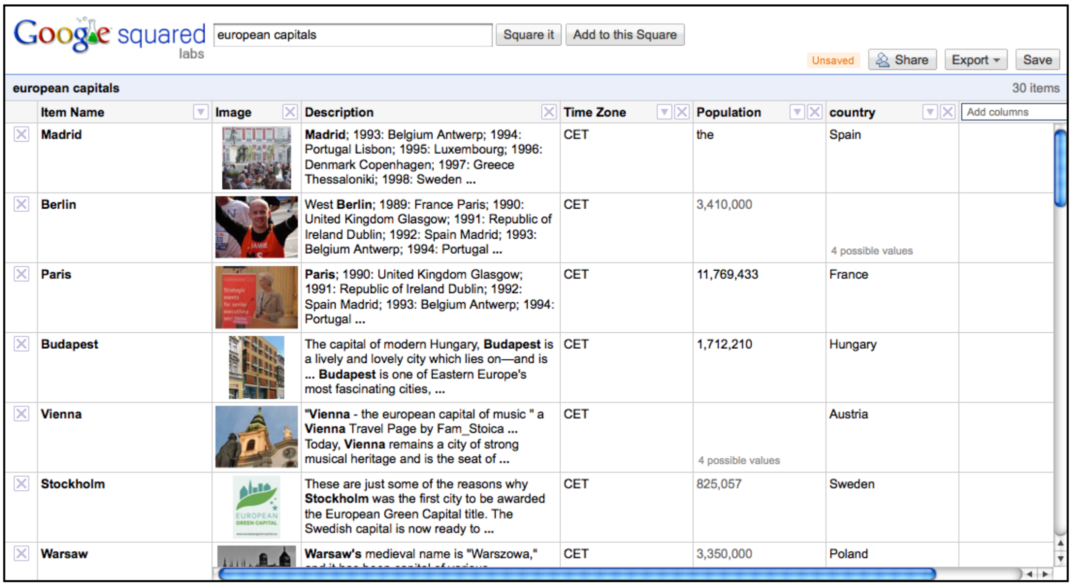
Difficult though it is to believe, in Bach's day musical works were as ephemeral s today's pop songs. Here one day, gone the next. As the tide of taste and fashion rolled inexorably by, so new pieces were

Figure 2.9: Exploring music from the *Baroque* era composed by *Bach* using mSpace (from Schraefel et al. 2006)

2.3.4 Faceted browsing

Faceted browsing uses structured background knowledge to support interactive, iterative exploration of large information spaces (Yee et al. 2003). A facet is a dimension by which a document collection—or whatever is being searched—could be organized. Reasonable facets for organizing a CD collection, for example, are title, artist, album, genre, mood, release date, etc. With these facets in place and appropriate values assigned for each item in the collection, a large information space can be systematically narrowed by applying constraints. In Figure 2.9, for example, the searcher has specified their interest in music from the *Baroque* era, composed by *Bach*. Facets facilitate exploration of unknown information spaces, since the system provides opportunities to focus the search at each step, without requiring users to define explicit queries. Furthermore, search can complement browsing, either by searching individual facets for certain values, or restricting search results to the specified subset of the information space.

Faceted browsing requires highly expressive and accurate knowledge bases. The background structure—a faceted classification—is not generally considered to be an ontology, but there are strong similarities. Choi (2008) explains the subtle differences between them. Scale is not such an issue, because the technique is only applicable within homogeneous information spaces—music collections, recipe libraries, e-commerce sites—where facets can be applied consistently. It is unclear how the technique could be



european capitals

european capitals — 30 items

	Item Name	Image	Description	Time Zone	Population	country
☒	**Madrid**		**Madrid**; 1993: Belgium Antwerp; 1994: Portugal Lisbon; 1995: Luxembourg; 1996: Denmark Copenhagen; 1997: Greece Thessaloniki; 1998: Sweden ...	CET	the	Spain
☒	**Berlin**		West **Berlin**; 1989: France Paris; 1990: United Kingdom Glasgow; 1991: Republic of Ireland Dublin; 1992: Spain Madrid; 1993: Belgium Antwerp; 1994: Portugal ...	CET	3,410,000	4 possible values
☒	**Paris**		**Paris**; 1990: United Kingdom Glasgow; 1991: Republic of Ireland Dublin; 1992: Spain Madrid; 1993: Belgium Antwerp; 1994: Portugal ...	CET	11,769,433	France
☒	**Budapest**		The capital of modern Hungary, **Budapest** is a lively and lovely city which lies on—and is ... **Budapest** is one of Eastern Europe's most fascinating cities, ...	CET	1,712,210	Hungary
☒	**Vienna**		"**Vienna** - the european capital of music " a **Vienna** Travel Page by Fam_Stoica ... Today, **Vienna** remains a city of strong musical heritage and is the seat of ...	CET	4 possible values	Austria
☒	**Stockholm**		These are just some of the reasons why **Stockholm** was the first city to be awarded the European Green Capital title. The Swedish capital is now ready to ...	CET	825,057	Sweden
☒	**Warsaw**		**Warsaw's** medieval name is "Warszowa,"	CET	3,350,000	Poland

Figure 2.10: A Google Squared grid of *European capitals*

adapted to open-domain search, where there is one dominant facet (what the documents are about), and all others (format, size, author, etc.) are of much less importance.

Google Squared (Figure 2.10) provides an interesting example of interaction with structured knowledge, where users specify some group of entities they want to have aggregated; in this case *European capitals*. The system then gathers the entities and builds a grid associating them with whatever properties it assumes are most appropriate—for the capital cities these are initially *image, description*, *time zone*, and *population*. The user is able to manually add or remove properties—in Figure 2.10 the property *country* has been added—and view sources, confidence values and possible alternatives for each cell or fact that is presented. This provides an intuitive interface for fact-finding and building comparisons, as well as a tool for document retrieval (it provides immediate access to the pages from which facts are mined) and crowd sourcing knowledge (erroneous facts can be corrected by the public).

Google Squared's simple interface belies an ambitious attempt to mine structured knowledge from the web. Its performance is somewhat hit-and-miss. Figure 2.10, for example, did not initially include the most obvious discriminating property (*country*) until this was manually added. It contains many unhelpful descriptions, and does not state what country *Berlin* belongs to. The population counts are extremely spotty: many are missing, and Paris' population is gigantic because it includes the wider metropolitan area, while the other statistics do not. It does not distinguish between capitals of countries and capitals of individual regions and provinces (e.g. the grid goes on to list *Florence*, the capital of the Italian province *Tuscany*). This system demonstrates the immense challenges involved in dealing with properties and facts rather than informal relations at the scale of open-domain retrieval.

2.3.5 Personalization and adaptive hypermedia

Relevance is relative. The usefulness of a particular search result depends on the person who entered the query: on their interests, levels of expertise, age, geographical location, and a host of other factors. It is useful for search engines to take this broader context into account and customize themselves to individuals rather than behaving in the same way for everybody. Google, for example, re-ranks search results by tracking individual query histories, bookmarks, and click-through data (Sullivan 2007). This activity is known as *search engine personalization* (Pitkow et al. 2002).

As retrieval systems become more informed about users and their intentions, they can be much more proactive. They can recommend documents and other resources without being explicitly asked for them (Resnick and Varian 1997). They can augment not just searching but also browsing, by emphasizing or hiding existing navigational paths, or constructing new connections on the fly. They can even alter how the documents themselves are presented, by hiding, dimming or scaling down irrelevant content, adding explanatory definitions and images when unfamiliar concepts are introduced, and so on. Brusilovsky (2001) provides an excellent overview of these diverse efforts, which fall under the broad heading of *adaptive hypermedia.*

A very common strategy in search engine personalization is collaborative filtering, which involves clustering users into groups of people with similar interests and needs, and using the behaviour of these groups to filter search results and recommendations for each individual (Hofmann 2004). Adaptive hypermedia focuses on user modelling: following a user during their information seeking journey, tracking their behaviour, and building a model of their short- and long-term goals, interests, levels of expertise, and so on (Chi et al. 2001). Both collaborative filtering and user modelling depend strongly on similarity metrics to group related queries, documents and users together, and separate unrelated ones. This suggests that knowledge bases could be applied to personalization and adaptive hypermedia in a similar fashion to text clustering (Section 2.3.3), which in turn suggests that all three place roughly same requirements on knowledge bases: i.e. that size is more important than accuracy or expressivity.

The utility of highly formalized factual knowledge is not immediately obvious in either search engine personalization or adaptive hypermedia. There are several approaches that describe the use of ontologies for personalization and reccomendation—e.g., Gauch et al. (2003), Trajkova and Gauch (2004), Ziegler et al. (2005)—but on closer inspection use only subject hierarchies—the Yahoo Directory and the Open Directory Project, among others—and their associated Web pages. These resources amount to taxonomies with extended glossaries.

Collaborative filtering is much easier in the social tagging systems described in Section 2.3.2. Here users are cleanly identified and the effort involved in tagging provides much stronger indications of interest and expertise than tracking search behaviour. These systems provide a social dimension in which users can explicitly seek out people who share their interests, rather than the search engine doing so behind the scenes. The introduction of human-readable knowledge bases to tagging—as was advocated above—would not endanger the social context, as long as tagging remained a manual or semi-automated activity. Annotation with more formalized knowledge is generally more ambitious. Berners-Lee's vision for the Semantic Web (Section 2.2.4) would involve extending annotation down to the level of individual entities mentioned within documents and facts asserted about them. This is too tedious to be performed manually at a large scale (Cimiano et al. 2004)—but automatic systems would not provide social metadata.

2.3.6 Visualization

Over the course of human evolution, our perceptions have been gradually honed and attuned to deal with the spatial-visual arrangement of objects. We have specialized in using vision to navigate the world. It follows that retrieval systems should allow us to apply our visual skills to navigate virtual spaces.

Unfortunately, attempts to apply visualization to information retrieval have met with limited success. Chen and Yu (2000) conducted a meta-analysis of 35 information visualization studies published between 1991 and 2000. At the time they were only able to locate six studies that satisfied the criteria for analysis and direct comparison—e.g., task type, use of spatial-visual interfaces, comparison to control. Broadly, their conclusions based on the six studies were:

- The cognitive abilities of participants had a greater effect than choice of interface (visual vs. control).
- Simple, lightweight visualizations were more effective than complex ones.
- The combined effect of visualization was not statistically significant (due in part to the small sample size).

Admittedly, this meta-analysis is somewhat small and dated. A recent book by Hearst (2009) provides an updated and broadened survey—an informal discussion rather than analysis—that does not contradict Chen and Yu's conclusions. This does not imply that

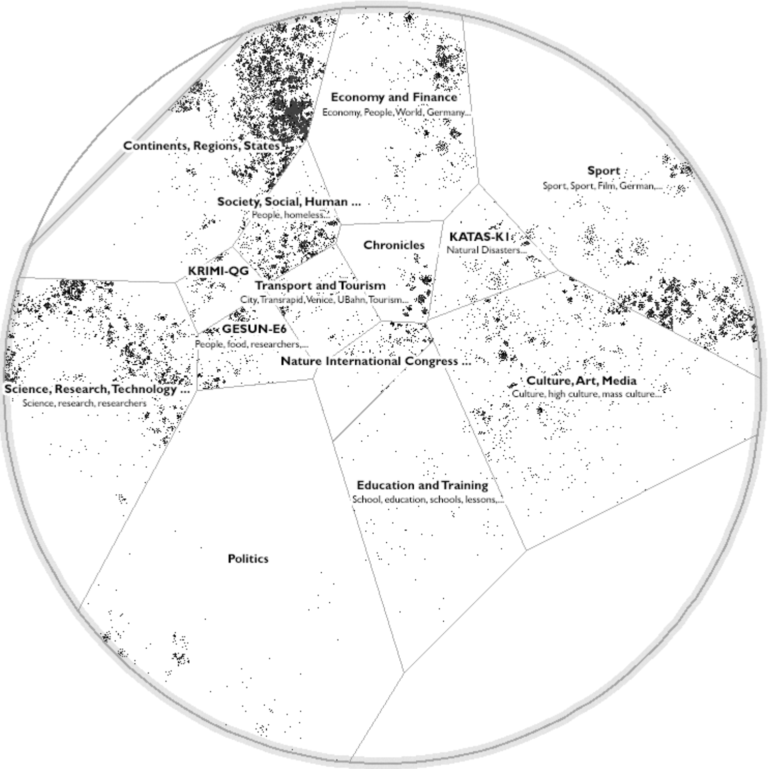

Economy and Finance
Economy, People, World, Germany...
Continents, Regions, States
Sport
Sport, Sport, Film, German,...
Society, Social, Human ...
People, homeless...
Chronicles
KATAS-K1
Natural Disasters...
KRIMI-QG
Transport and Tourism
City, Transrapid, Venice, UBahn, Tourism...
GESUN-E6
People, food, researchers,...
Nature International Congress ...
Science, Research, Technology ...
Science, research, researchers
Culture, Art, Media
Culture, high culture, mass culture...
Education and Training
School, education, schools, lessons,...
Politics

Figure 2.11: An InfoSky visualization (from Granitzer et al. 2004)

information visualization is not useful. It has been applied successfully to information *analysis* (as opposed to *retrieval),* and can effectively expose and communicate general trends, patterns and outliers in data (Fayyad et al. 2002). This is extending into sense making and the broader context surrounding retrieval, which is beyond the scope of this investigation (see Section 2.1).

Why is visualization so difficult to apply to search? This setting is dominated by nominal data: text, category labels, names, etc. Nominal data, by definition, cannot be meaningfully graphed in isolation. It does not lend itself well to visualization unless it has some quantitative property—significance, popularity, similarity—that can be conveyed spatially, or is coupled with something else—dates, quantities, faces and places, etc. As Sections 2.3.2 and 2.3.3 pointed out, connections from text to more structured information are hard to come by.

Shneiderman (1996) provides an early set of set of heuristics for effectively applying visualization to information seeking. His thoughts are succinctly put in his *visual information seeking mantra*: "Overview first, zoom and filter, then details-on-demand." In other words, visualize the entire information space and the relations between items, provide tools for homing in on interesting items or filtering out uninteresting ones, and allow details to be easily obtained for individual items or significant groups.

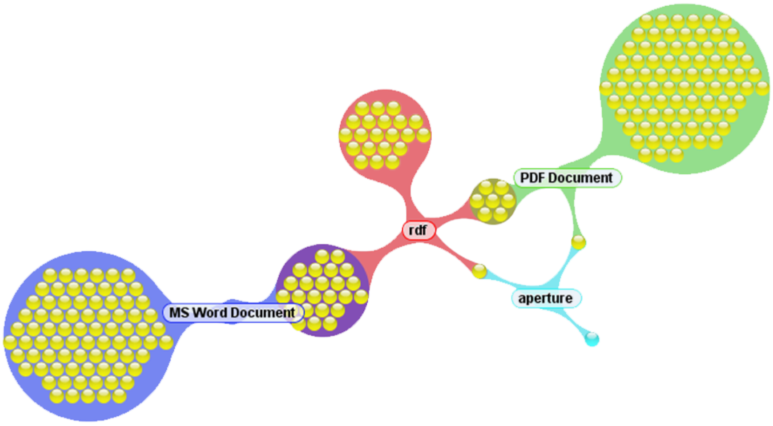
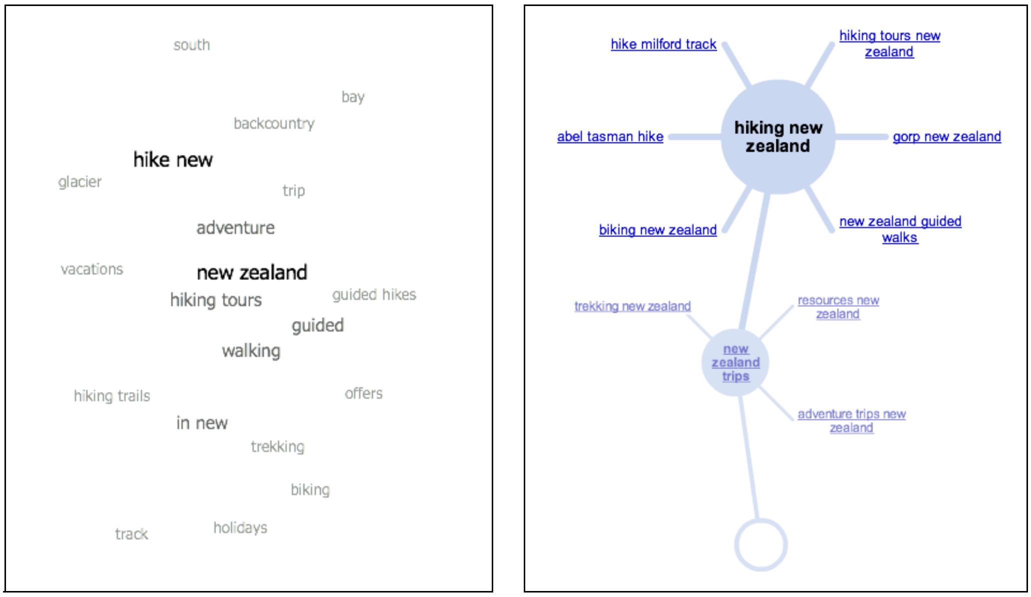


Figure 2.12: A cluster map, with *pdf* and *word* documents containing *rdf* or *aperture*

Providing informative visual overviews of large corpora—particularly open domain web search—is extremely challenging. Categorization (Section 2.3.2) and faceted classification (Section 2.3.4) provide obvious mechanisms for navigating and subdividing large information spaces, and there are several techniques for navigating these structures visually. For example, Yang et al. (2003) visualize category hierarchies, and Ahlberg and Shneiderman's (1994) oft-cited starfield displays seem well suited to faceted browsing. If documents are not organized against background knowledge bases, then visual overviews invariably fall back to automatically generated document clusters. Plotting these on 2D or 3D displays has been investigated many times but evaluations have so far been discouraging. For example, when the InfoSky starfield shown in Figure 2.11 was compared against a standard tree browser, the latter was found to be consistently simpler and faster to use (Granitzer et al. 2004). Difficulties in combining clustering and visualization are to be expected: as Section 2.3.3 explained, clustering has many drawbacks, which are only exacerbated in large overviews, where document representations are typically reduced to uninformative pixels.

It may be more fruitful to wait until the information space has been narrowed with a query before applying visualization. Figure 2.12 illustrates a cluster map from Fluit et al. (2003), where the user has searched for *pdf* or *word* documents containing the terms *rdf aperture*. It separates the various pairings of these constraints—something that many searchers are unable to do via Boolean operators (Dinet et al. 2004)—and communicates a great deal of information about the results: e.g. there are few documents that mention *aperture,* only one that also mentions *rdf*, and this is neither a *word* or *pdf* document. The same visualization can convey the distribution of tags, index terms and category headings across search results.



a) Quintura **b)** Wonder Wheel

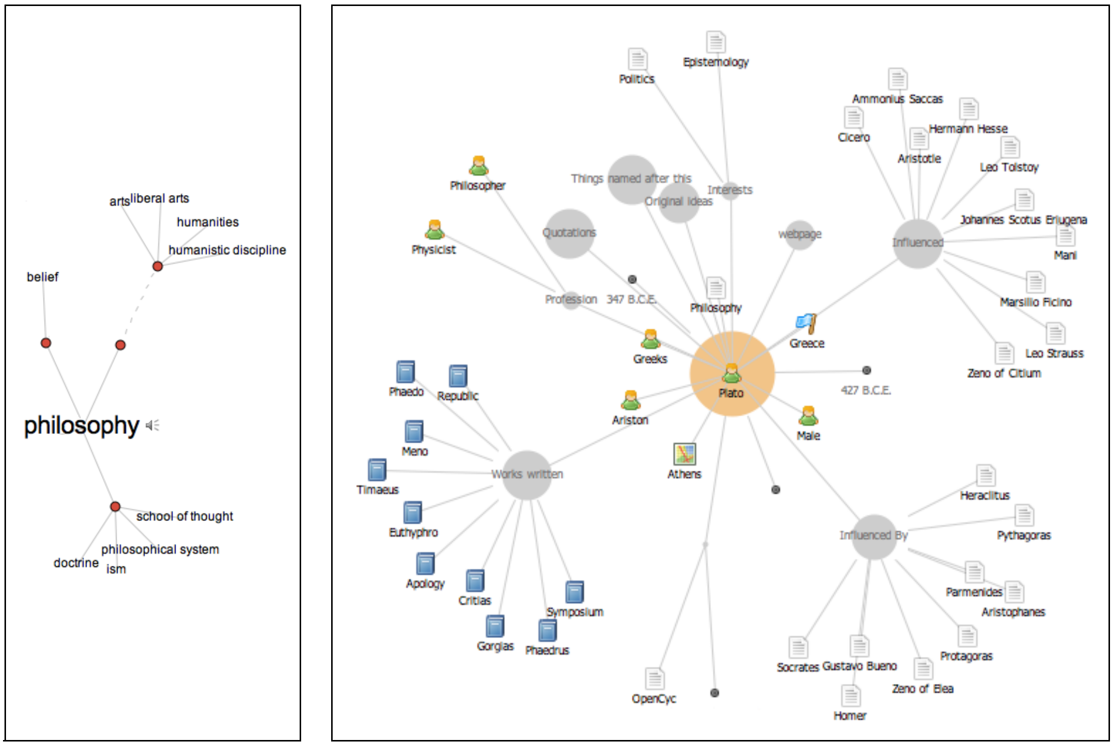
Figure 2.13: Visualizing query suggestions for *hiking new zealand*

Query suggestions are a popular feature to visualize. Figure 2.13 provides two examples: the first from Quintura,[12] and the second from Google's wonder wheel.[13] Quintura's interface resembles the widely used tag cloud, in which the most relevant or important suggestions are emphasized through scale and contrast. It attempts to arrange suggestions spatially, so that related items are found close together. Kaser and Lemire (2007) describe a range of algorithms for achieving such a layout, but focus their evaluations on algorithm efficiency rather than utility to users. One wonders how useful these vague spatial hints really are, particularly in comparison to the explicitly categorized suggestions provided by Cuil (see Figure 2.5). Google's wonder wheel focuses on navigation paths rather than significance and relatedness. Visual cues are used to emphasize and group suggestions that were presented and explored most recently. In Figure 2.13b they communicate the user's path from some initial query, through the suggestions to *new zealand trips*, and finally to *new zealand hiking.*

Knowledge bases can be visualized directly, commonly as graphs where nodes are concepts and edges between them indicate relations. As with visualizing textual information spaces, providing informative high-level overviews is problematic. Using graphs to visualizing the entirety—or large portions—of a knowledge base is a common yet largely unproductive strategy (Karger and Schraefel 2006). Graphs have been more successfully used to visualize small portions of the knowledge base, centred on the user's

[12] http://www.quintura.com/

[13] http://www.google.com/search?q=stuff&tbs=ww:1



a) Visual Thesaurus **b)** Thinkbase

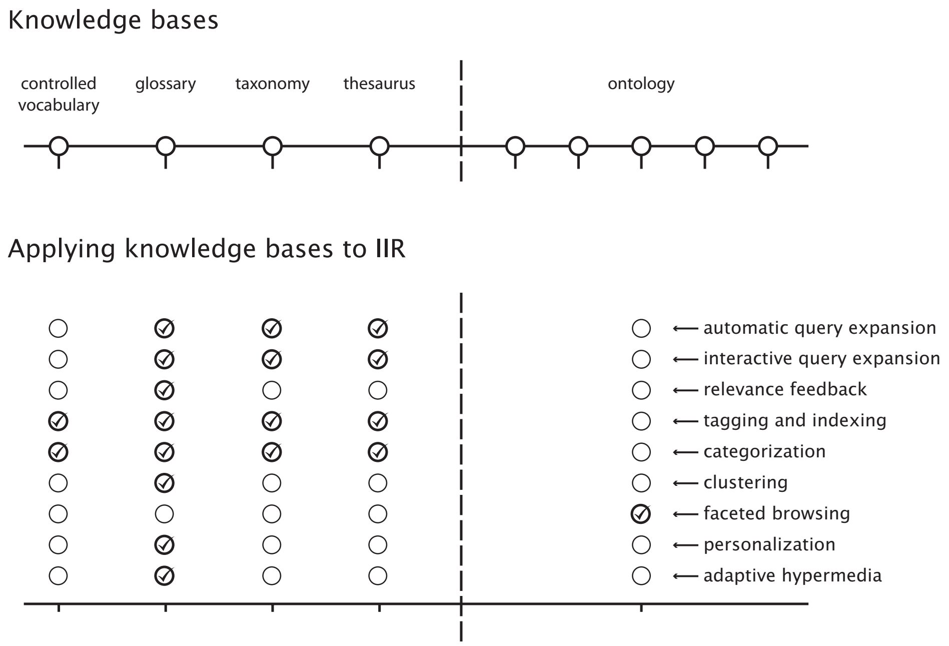
Figure 2.14: Visualizing knowledge bases directly

current location. Figure 2.14 shows two examples: the Visual Thesaurus[14] interface for WordNet, and the Thinkbase[15] interface for Freebase. Both systems have received praise for cleanly illustrating interconnectivity and encouraging exploration—e.g. Mangis (2005) for the former and Catone (2008) for the latter—but have never been formally evaluated against non-visual baselines. The utility of such interfaces is revisited in Section 7.1, which describes Thinkpedia, a visualization of Wikipedia that is almost identical to Thinkbase.

The introduction of visualization to interactive information retrieval has not altered any of the claims made in previous sections. The difficulties in combining clustering and visualization (described above), mean that categorization—particularly faceted classification—is still more effective for providing overviews of the information space or query results, whether that overview is communicated visually or not. Manually defined relations still provide better paths to navigate from one query to the next—particularly when those paths can be classified and grouped meaningfully. No one type of knowledge base stands out as being more amenable to visualization.

[14] http://www.visualthesaurus.com

[15] http://thinkbase.cs.auckland.ac.nz



	controlled vocabulary	glossary	taxonomy	thesaurus	ontology
automatic query expansion	○	✓	✓	✓	○
interactive query expansion	○	✓	✓	✓	○
relevance feedback	○	✓	○	○	○
tagging and indexing	✓	✓	✓	✓	○
categorization	✓	✓	✓	✓	○
clustering	○	✓	○	○	○
faceted browsing	○	○	○	○	✓
personalization	○	✓	○	○	○
adaptive hypermedia	○	✓	○	○	○

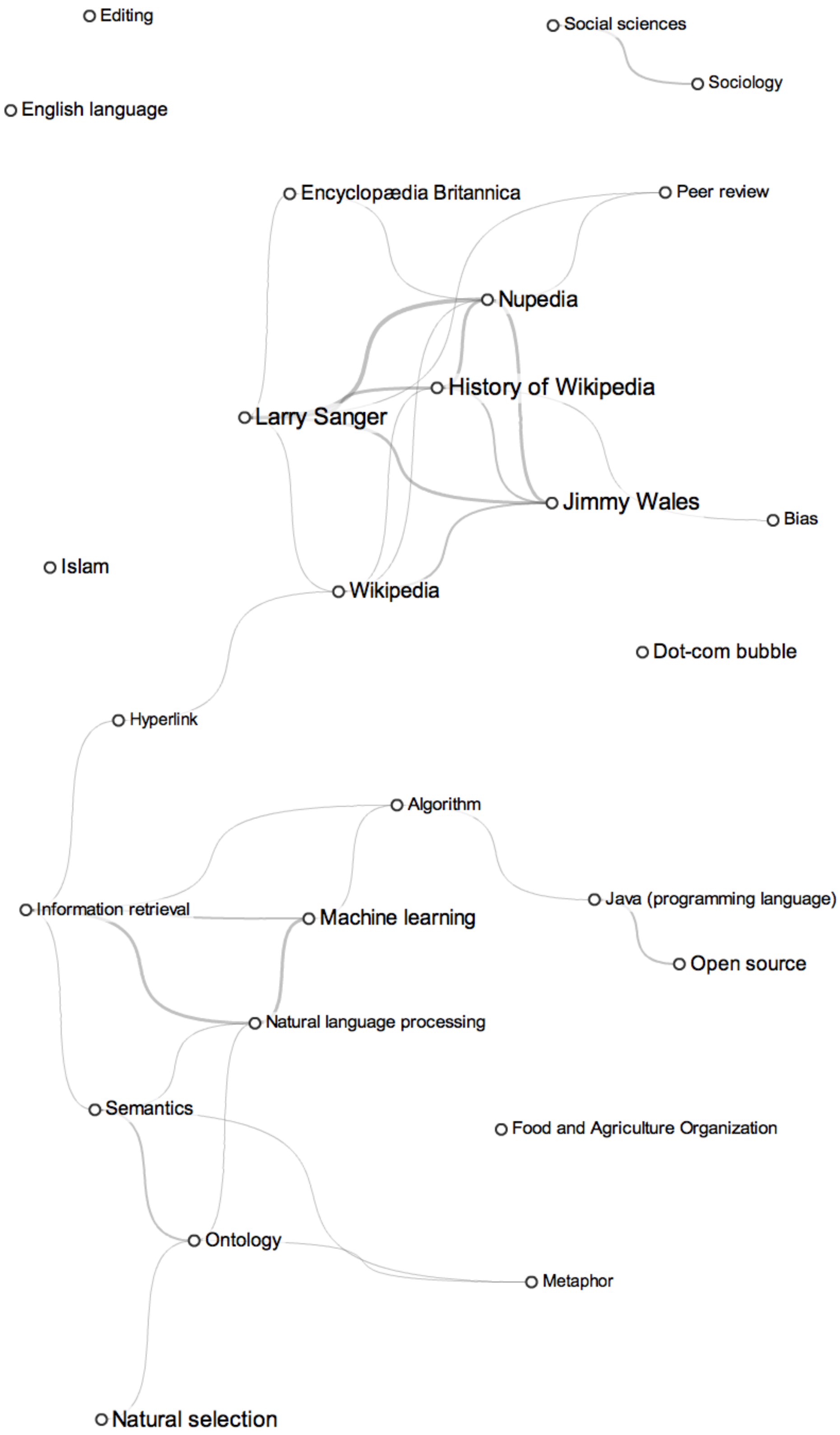
Figure 2.15: Applying knowledge bases to Interactive Information Retrieval

2.4 Discussion

This chapter has described several ways of looking at the information process, plotted the spectrum of knowledge bases from simple controlled vocabularies to highly expressive formal ontologies, and surveyed how these resources can support interactive information retrieval. Before moving on to the next chapter, which is specific to extracting knowledge from Wikipedia, it is worth reflecting on what has been learned.

Figure 2.15 revisits the knowledge base spectrum from McGuinness (2005). Each step along this scale represents an increase in complexity and expressivity. As Section 2.2.3 explained, the steps are not free. Each requires an increased investment to construct the resource, and should ideally be justified with a direct payoff for the end user.

A great deal of effort has been expended to build formal, machine-readable knowledge bases. While this effort has undoubtedly taught us much about knowledge representation, extraction and inference, it has not had a great impact in how we interact with information. In our survey of interactive information retrieval (Section 2.3), there is only one strategy for which formalized knowledge has been demonstrated as directly applicable: the faceted browsing technique described in Section 2.3.4. In contrast, lesser structures have seen extensive use. Human-readable thesauri and taxonomies have proven effective in query expansion, indexing and categorization. Simple glossaries, which cleanly separate concepts from each other but represent them only with human readable text, are useful in any situation where some metric of similarity is required between documents. In other words, machine-interpretable is not a prerequisite for machine-usable: much can be done with fuzzy, human-oriented knowledge, particularly in highly interactive systems.


Editing
Social sciences
Sociology
English language
Encyclopædia Britannica
Peer review
Nupedia
History of Wikipedia
Larry Sanger
Jimmy Wales
Bias
Islam
Wikipedia
Dot-com bubble
Hyperlink
Algorithm
Java (programming language)
Information retrieval
Machine learning
Open source
Natural language processing
Semantics
Food and Agriculture Organization
Ontology
Metaphor
Natural selection

3. Wikipedia as a knowledge base

This chapter investigates a new source of structured knowledge: the collaboratively built encyclopaedia Wikipedia. This resource offers terms and relations defined by human intelligence (as opposed to statistical or lexical approximations), at a larger scale than the expert driven efforts described Section 2.2. Its unique editing mechanism promises continual maintenance, coverage of rapidly evolving domains, and incorporation of contemporary language and interests. Unfortunately, the same mechanism raises concerns about accuracy, bias, and susceptibility to vandalism and exploitation.

Section 3.1 provides a brief history of the Wikipedia project. These roots contribute directly to the inherent strengths and limitations of the resource, which are described in Section 3.2. Section 3.3 outlines the structural features of Wikipedia, and Section 3.4 surveys how they are mined for the types of knowledge bases—controlled vocabularies, taxonomies, thesauri and ontologies—that were described in the previous chapter. Section 3.5 compares Wikipedia to the domain specific thesaurus Agrovoc. The chapter concludes by revisiting the discussion begun in Section 2.4. Here the specific features of Wikipedia are taken into account, to make clear recommendations about how it can be applied to interactive information retrieval.

3.1 Brief history of Wikipedia

Wikipedia was first proposed in January 2001 with modest ambitions (Sanger 2005). It began life as a mere side-project of the now defunct Nupedia, a project that was founded in 2000 by Jimmy Wales with the now familiar vision of a collaboratively produced, freely licensed encyclopaedia constructed under the same model as open-source software. The goal was the same, but the details were different: the original project was largely run by editor-in-chief Larry Sanger, an academic who felt that careful oversight would be essential to the project's success. Sanger felt that the resource would only be credible if closely managed by experts; that the inclusion of the public would necessitate an unusually rigorous review process, even by academic standards. Although he welcomed public contribution, participants were largely recruited from academia with the expectation that authors and reviewers could be readily identified and held accountable within their field. Articles would pass through many hands—a seven-step program of review—before finally being offered up to the reader. The downside of this cautious model became apparent after 18 months and $250,000USD, at which point it had produced only 20 fully approved articles.

To address Nupedia's shortcomings, Sanger suggested using the then-new wiki technology to allow the public to more easily propose and contribute to articles before they underwent review. He also began simplifying and relaxing the review process to deal with the influx of new content. Under his vision, the content of Wikipedia—as the new resource was coined—would have been carefully vetted by a board of academics and experts, and then fed into the more authoritative Nupedia. At this time the dot-com bubble burst, and funds dried up for both projects. Nupedia and its expert oversight were dropped, Sanger moved on, and Wikipedia was left to manage itself in a unique cauldron of new technology, scarce funding, a leadership vacuum, and rapidly increasing public exposure.

Astonishingly, Wikipedia did not merely survive under these circumstances—it thrived. It is a testament to the culture and ideology behind these two projects, and the willingness of people to share what they know, that Wikipedia rapidly grew to become one of the largest and most visited reference work in the world.

3.2 Pros and cons of crowd-sourced knowledge

The goal of this section is to clarify Wikipedia's strengths and limitations, where they came from, and what they mean for anyone seeking to use Wikipedia in particular—and crowd sourcing in general—to provide knowledge bases for information retrieval.

3.2.1 Scale

The most obvious advantage offered by a crowd-sourced knowledge base is scale; it will be more capable of covering the breath of information than expert-driven resources. Figure 3.1 plots the growth of the English language Wikipedia since its release in January 2001. If one were to drill down at the origin of this graph, Wikipedia would be seen outstripping its predecessor Neupedia within days of release. The number of edits, articles and registered contributors—known as Wikipedians—all grew super-linearly until approximately January 2006. Since then the metrics have held steadily at (roughly) 1M edits, 50K new articles, and 10K new Wikipedians per month.

It is difficult to imagine generating an encyclopaedia of this size without adopting Wikipedia's open editing policy. Wikipedia's nearest competitor, Encyclopaedia Britannica, has been in constant development since the 1700s but in 2002 consisted of only 65,000 articles. Wikipedia reached a comparable size sometime between July and October 2002, at less than two years old. A reliable figure could not be found of

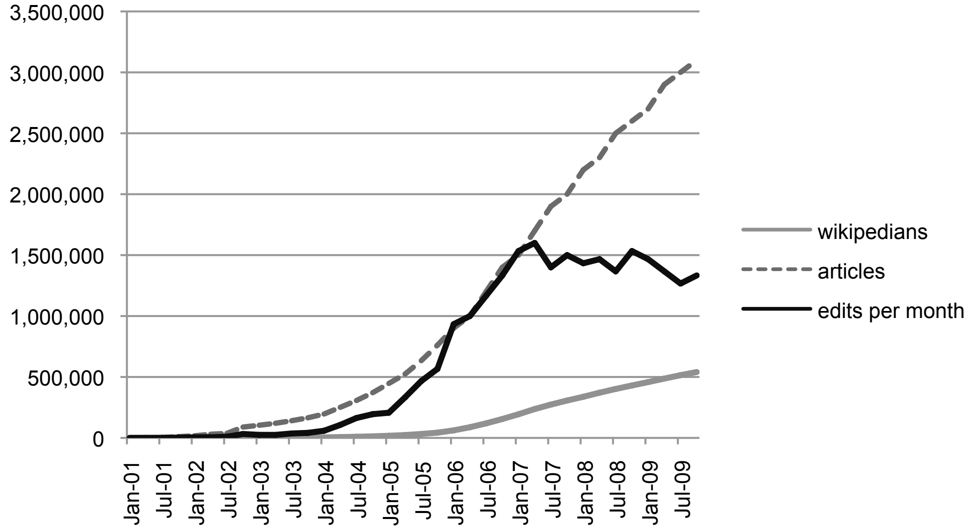


Figure 3.1: Monthly growth of Wikipedia

Britannica's current size, but its online edition[16] quantifies itself as "hundreds of thousands of articles, biographies, videos, images, and web sites"—at least an order of magnitude smaller than Wikipedia in terms of number of articles.

Wikipedia's growth is not entirely unchecked. Approximately a quarter of all articles are deleted, typically soon after their creation (Lam and Riedl 2009). Most deletions are due to a lack of notability; there is an expectation that any new content must have received significant coverage in independent sources.[17]

Knowledge bases that are extracted from Wikipedia, or are crowd-sourced in a similar fashion, will likely gain the same advantages in scale compared to their traditionally constructed counterparts. For example, Freebase (Section 2.2.1) is derived partially from Wikipedia, partially from other resources (MusicBrains, the CIA World Fact Book, etc), and partially crowd-sourced. It was only released to the public in 2007, yet it already dwarfs Cyc (Section 2.2.1), an expert-built ontology that has been under construction for twenty years.

3.2.2 Adaptability

Another closely related advantage of crowd sourcing is the ability to keep up with the deluge of new topics and developments that appear every day. The traditional expert-driven model involves a lot of inertia, because existing content is carefully polished and

[16] http://www.britannica.com

[17] http://en.wikipedia.org/wiki/Wikipedia:Notability

people trusted and qualified to change it are a scarce resource. Maintenance is expensive, so most resources release new editions annually or more intermittently. This is only sufficient for covering narrow domains that remain relatively static; thus encyclopaedias and traditional knowledge bases focus on historical subjects, biology, geography, etc, rather than technology, politics, or current affairs.

In contrast, the crowd sourcing model is able to adapt quickly to new events, and forms a close relationship with the press. Lih (2004) identifies many instances of current events precipitating flurries of increased activity in the corresponding Wikipedia articles, and describes Wikipedia as a "working draft of history" that fills the wide gap between newspaper and book. It may even encroach on the former: a sister project called WikiNews[18] was launched in 2004 to provide a constantly updated news source built on the same crowd-sourcing model. Acceptance of this resource has been somewhat shaky, although concerns have not been about speed or timeliness, but reliability (more on this below) and whether a separate resource is needed—Wikipedia arguably fulfils the role already (Dee 2007). Few seem to doubt that the crowd sourcing mechanism behind Wikipedia and WikiNews will have any trouble keeping up with current events.

The main implication of this adaptability is that deriving a knowledge base from Wikipedia should not be a one-step process: new entries and edits should be able to flow through immediately, or at least be fed through with regular maintenance. The process will need to be automatic or crowd-supervised. Constructing algorithms or crowd-based infrastructures that are accurate and robust enough will be a challenge, but the payoff will be immense: it would yield up-to-the-minute structured knowledge for even the most turbulent topics and domains.

Wikipedia's ability to adapt rapidly also means that it has the potential to be guided. Wikipedia's workforce is quick to address missing or unsatisfactory content if they are informed about it (Viégas et al. 2004), and there are many facilities (e.g. templates and article stubs described in Section 3.3.8) by which they could be notified and guided by some automatic process. Any knowledge-based system that uses Wikipedia could potentially form a symbiotic relationship with it. An IR system, for example, could use Wikipedia to provide additional knowledge about the items being searched for, and in return inform Wikipedia about things that people are looking for that it does not yet cover. In this case, however, the external system would have to submit to Wikipedia's guidelines of notability.

[18] http://www.wikinews.org

Not only can Wikipedia be guided in terms of the knowledge that it does and does not cover, it can also be guided in the ways it describes that knowledge. Wikipedia's contributors have high computer literacy, and the markup they use has been under constant development. Völkel et al. (2006) describe several ways in which it could be subtly altered to create a "Semantic Wikipedia." If their initiatives—e.g. allowing inter-article connections to be typed—gained widespread adoption, the amount of structured knowledge that is explicitly encoded by the resource would be vastly increased. Section 3.4.4 revisits this plan.

3.2.3 Accuracy

The crowd sourcing model raises concerns about accuracy and reliability. Traditionally, authors are expected to prove their qualifications, and are checked for accuracy by peer review and experienced editors. Wikipedia provides no guarantees. Articles are only kept in check by a form of natural selection, where modifications and contributions are like genetic mutations: if they are not rejected they build upon each other and the article gradually evolves. The metaphor has its limits: whereas mutations in nature are random and only built upon if they improve the odds of survival, modifications to Wikipedia are guided by human hand and accepted by default (but easily reversed). Clearly, the reliability of Wikipedia hinges on human nature: on the proportion of modifications that are well intentioned and well informed; whether qualified people are willing to contribute; and whether the dedication of volunteers outweighs the troublemakers.

The issue of Wikipedia's accuracy has captured the interest of mainstream media since 2004. Initially opinion was strongly divided: some saw the project as a bold step that could soon render traditional encyclopaedias obsolete (Bray 2004); others as dangerously naïve and fatally flawed (McHenry 2004); most as a fascinating phenomenon and a promising resource, but one that should be used with caution (Waldman 2004). The most basic concern of Wikipedia's detractors—that the resource leaves itself wide open to abuse—is addressed by Viégas et al. (2004), who find that Wikipedia is robust against blatant vandalism due to the ease with which new modifications can be monitored and reversed. More subtle errors—hoaxes, unverified facts, and errors of omission—have been known to persist, however (Siegenthaler 2005, Finkelstein 2006).

It is easy to cite specific flaws in Wikipedia, but one cannot expect complete infallibility: Britannica does not live up to that either (Einbinder 1964). The important question is how widespread the flaws are. This question generated a flurry of non-peer-reviewed investigations by *The Guardian* (2005), *Nature* (Giles 2005), *PC Pro* (Andrews 2007) and several other sources. The general consensus of these articles—which typically compare a handful of Wikipedia articles against corresponding entries in more traditional resources—is that the gap between Wikipedia and its counterparts is small. Wikipedia articles are occasionally poorly written or miss important details, but are generally factually accurate.

The most widely cited study regarding the accuracy of Wikipedia is Giles (2005), in which academics were recruited for a single-blind comparison between scientific Wikipedia articles and their equivalent entries in Encyclopaedia Britannica. In 41 article pairs there were 162 mistakes in Wikipedia versus 123 for Britannica. Both sources were equally prone to significant errors, such as misinterpretation of important concepts. More subtle errors, however, such as omissions or misleading statements, were more common in Wikipedia. The conclusion of the study is that Wikipedia approaches the accuracy of Britannica; a finding that has attracted heated debate. Britannica (2006) attacked Giles' study as "fatally flawed" and demanded a retraction. Nature (2006) defended itself and declined to retract.

To our knowledge the only peer-reviewed study of Wikipedia's accuracy is Clauson et al. (2008), who compare it against the Medscape Drug Reference. They found no factual errors, but did find many errors of omission: Wikipedia was only able to provide answers to 40% of questions, compared to 83% for the more specialized resource. This finding is revisited in the next section, which describes Wikipedia's coverage—or lack thereof.

As Fallis (2008) explains, even if the comparisons between Wikipedia and more authoritative resources are inconclusive, it compares favourably against the sources that people would likely turn to if Wikipedia did not exist (i.e. blogs and other freely accessible websites), and that Wikipedia has many other strengths (such as those highlighted in the surrounding sections) that arguably outweigh the deficiencies.

3.2.4 Bias

Wikipedia relies on volunteers, who inevitably focus on topics that pique their interest. It has consequently been accused of systemic bias, in both the topics it covers, and content and opinions it expressed.

Who are Wikipedia's contributors? Some obvious implications of the crowd sourcing mechanism are that most will be technologically savvy people with uncensored internet access and plenty of free time; in other words, young to middle aged, middle- to upper-class Westerners. It is also reasonable to assume that most—or at least those whose contributions are allowed to persist—will be educated and articulate. Wikipedia does not collect personal information of the people who contribute to it, so more specific details are difficult to obtain. Most studies of Wikipedia focus not on who these people are, but on how they go about their job (more on this in Section 3.2.5). Nov (2007) provides one exception with a partially self-selected survey of Wikipedians. He randomly selected 370 registered users and received 151 responses. 93% of respondents were male, with a median age of 31. Their primary motivations were fun and ideology. The latter incentive has raised red flags with people labelling the project as a cult of opinionated leftists (Correa et al. 2006), but this concern does not appear to be justified. Nov (2007) found that "talk is cheap"—that those motivated by ideology contribute significantly less than those with other reasons. Unfortunately the study did not touch on the qualifications of contributors.

The above survey covers only the core of dedicated, registered Wikipedians, who are surrounded by a much larger pool of casual, anonymous contributors that is even more difficult to profile. Priedhorsky et al. (2007) showed that these causal users have little impact, however. They ranked people by the number of edits, and inspected the number of views their contributions received before being reverted. They found that the top 10% of editors were responsible for 86% of Wikipedia's viewable (i.e., surviving) content, and that 0.01% (approximately 4400 editors) were responsible for 46%. Anyone *can* edit the resource, but only those with lasting motivations to do so have any significant impact. These power-editors tend to establish themselves early with an immediate burst of activity—which slows little during their careers—rather than growing gradually into the role (Panciera et al. 2009).

It is difficult to quantify any bias of opinion that Wikipedia might have. Philosophically, the resource aims for a *neutral point of view*. Its manual of style dictates:

> *"All Wikipedia articles and other encyclopaedic content must be written from a neutral point of view, representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources. This is non-negotiable and expected of all articles and all editors."*

This policy is one of the fundamental principles behind Wikipedia, and dates back to its inception (Sanger 2005). Obviously not all contributors will be familiar with the rule, and others may deliberately subvert it. As the worst possible example, Jimmy Wales has been caught doctoring his own Wikipedia biography (Schiff 2007). Many other high-profile abuses came to light with the invention of WikiScanner,[19] a tool that allows even "anonymous" edits to be traced (Hafner 2007). However, it is not known how long such

[19] *http://wikiscanner.virgil.gr*

abuses are allowed to persist—particularly now that WikiScanner has been added to the Wikipedia watchdog toolbox—and it is encouraging that the resource is firmly in control of a core of elite editors (Priedhorsky et al. 2007) who are thoroughly committed to neutrality.

Bias of coverage is somewhat easier to identify. Anecdotally, it appears that Wikipedia has a penchant for modern trivia. For example, there are 600 different articles dedicated to *The Simpsons*, but only half as many pages about the namesake of the cartoon's main character, the Greek poet *Homer*, and all the literary works he created and inspired. Lih (2004) shows that Wikipedia's bias is driven to a large extent by the press. Section 3.5.2 identifies a bias towards concepts that are general or introductory, and therefore more relevant to "everyman." This, in conjunction with Clauson et al. (2008), suggests that only a certain level of depth can be expected from Wikipedia within any given domain. However, Lam and Riedl (2009) show that that newer articles tend to be on more obscure, specific topics and thus Wikipedia is likely growing deeper rather than wider. Additionally, there exist many sub versions of Wikipedia (hosted by Wikia[20]) which each describe a particular domain—some in exhausting detail. Unfortunately, these wikis are strongly skewed towards popular culture—three of the largest are about song lyrics, the online role-playing game World of Warcraft, and the Star Wars universe—so their appeal as knowledge bases is limited.

For open-domain knowledge, Wikipedia and the crowd-sourcing model is likely to dwarf the breadth and depth of any other resource. Any concerns about bias of coverage are moot; as Section 2.2.3 explained, it is likely that any traditionally crafted alternative would also suffer bias, which would be exacerbated by their inherently smaller size.

The concerns become more acute for specific domains. In this case Wikipedia would only compete with traditional resources in domains that match the interests of its contributors: one is far more likely to encounter detailed knowledge about *communication technology* or *U.S. politics* than *medieval architecture* or *marine biology*. Fortunately the two sources complement each other: traditional resources excel in well-established technical or historical domains for which Wikipedians have limited interest or expertise, while Wikipedia excels in covering contemporary, swiftly evolving domains that traditional resources struggle with.

[20] ***http://www.wikia.com***

3.2.5 Transparency

Some of Wikipedia's most vocal detractors have, for obvious reasons, been those invested in more traditional resources. For example, consider the following quote from Encyclopaedia Britannica's former editor, Robert McHenry (2004):

> *"The user who visits Wikipedia to learn about some subject, to confirm some matter of fact, is rather in the position of a visitor to a public restroom. It may be obviously dirty, so that he knows to exercise great care, or it may seem fairly clean, so that he may be lulled into a false sense of security. What he certainly does not know is who has used the facilities before him."*

Ironically, this unflattering metaphor highlights one of the strengths of Wikipedia and crowd sourcing: each visitor can see the fingerprints (or worse) of those who came before. While this communal and changeable nature raises many concerns regarding accuracy (discussed in Section 3.2.3), it also offers many advantages.

For one thing, it allows Wikipedia to abandon any pretence of perfection. Traditional resources depend on their reputations for survival, and must preserve an air of impeccability. Wikipedia has no such concerns. It describes itself and its often-embarrassing growing pains with impressive objectivity. Consequently, few are "lulled into false sense of security." Wikipedia's limitations are clearly advertised both inside and outside the resource.

To push McHenry's metaphor further, Wikipedia is a restroom that is never properly cleaned. That is better than it sounds, because it allows those with a forensic bent to trace the entire history of the resource. The sequence of edits behind every article is preserved, and can be mined to discover patterns of collaboration (Viégas et al. 2004), predict the quality of an article (Thomas and Sheth 2007, Wilkinson and Huberman 2008), and presumably document the rise and fall of a topic's popularity and controversy. This feature is also drawn on by the WikiScanner system described in the section above, to expose the motivations behind individual edits.

Wikipedia is built collaboratively, and effective cooperation requires clear communication. One supposes that debates and arguments are inevitable. A Wikipedia entry is "like a sausage: you might like the taste of it, but you don't necessarily want to see how it's made," as Jimmy Wales has colourfully put it (Waldman 2004). Nevertheless, the ingredients are a matter of public record: every Wikipedia article is paired with a freely available talk page where contributors and critics discuss its faults and how it might be improved or extended. Viegas et al. (2007a) looked inside the discussions and found that they were largely productive and civil, and predominantly supported strategic planning of edits (users would discuss their intentions before actually making changes) and enforcement of standard guidelines and conventions. These debates are freely available, whereas one can only guess at the doubts, opinions and compromises that go on behind traditional, pristinely presented resources.

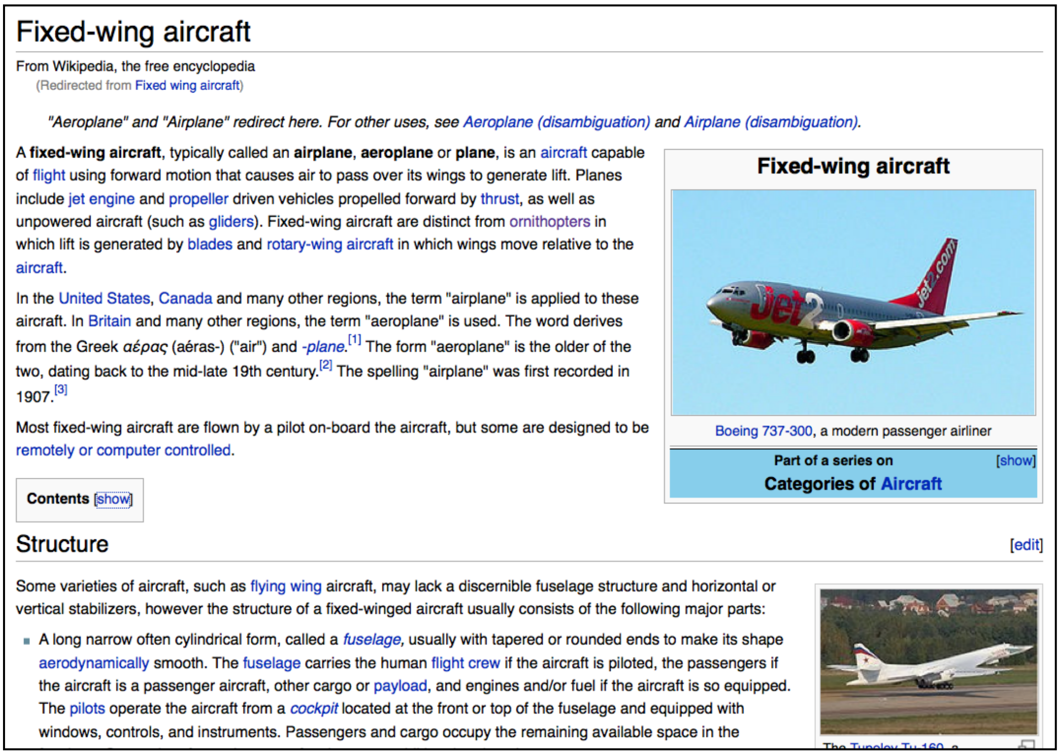
There is an extensive, publicly available library of articles dedicated to explaining Wikipedia's policies and what is expected from its contributors. The guidelines are dense and rapidly growing (Viégas et al. 2007a). This is both a positive and a negative: it represents the level of thought and strategic planning behind the project, but also the height of the hurdle that bars newcomers from contributing.

The discussion pages have limited use for the causal reader or automatic algorithm; they are simply too lengthy and chaotic. Fortunately Wikipedia provides many lightweight, structured tools for flagging items in Wikipedia. There are templates (described in Section 3.3.8) to mark statements that are contentious or require citations, articles that are poorly written or of dubious significance, and many others. Viégas et al. (2007b) provides an exhaustive review of just one of these mechanisms—the process by which pages attain *feature article* status. The template flags provide readers with the necessary warnings and reassurances at a glance, and can be easily rendered machine-readable.

3.2.6 Fecundity

Wikipedia employs a copy-left policy, which allows content to be freely copied, distributed and modified, as long as the results preserve the same freedoms. This is probably a matter of necessity as much as generosity, given Wikipedia's complete reliance on volunteer labour. Many of those who become committed Wikipedians rather than casual contributors are motivated by the ideology that "information should be free" (Nov 2007). These registered users are almost exclusively responsible for the content of the resource (Priedhorsky et al. 2007). Thus it is unlikely that Wikipedia would have been successful if its ideology had been compromised. Any knowledge bases derived from Wikipedia—and probably the crowd sourcing mechanism in general—inherit the same obligations.

Wikipedia's openness is a key reason behind the interest it has received from computer scientists. Wikipedia is pre-dated by several other electronic encyclopaedias—e.g. Encarta, Britannica Online—that were left almost entirely untapped for AI and IR research because of concerns about copyright and ownership. Wikipedia provides a compelling example of what can happen when information is free: it can evolve and be applied in ways far beyond what its creators originally envisage.



Fixed-wing aircraft

From Wikipedia, the free encyclopedia
(Redirected from Fixed wing aircraft)

"Aeroplane" and "Airplane" redirect here. For other uses, see Aeroplane (disambiguation) and Airplane (disambiguation).

A **fixed-wing aircraft**, typically called an **airplane**, **aeroplane** or **plane**, is an aircraft capable of flight using forward motion that causes air to pass over its wings to generate lift. Planes include jet engine and propeller driven vehicles propelled forward by thrust, as well as unpowered aircraft (such as gliders). Fixed-wing aircraft are distinct from ornithopters in which lift is generated by blades and rotary-wing aircraft in which wings move relative to the aircraft.

In the United States, Canada and many other regions, the term "airplane" is applied to these aircraft. In Britain and many other regions, the term "aeroplane" is used. The word derives from the Greek *αέρας* (aéras-) ("air") and *-plane*.[1] The form "aeroplane" is the older of the two, dating back to the mid-late 19th century.[2] The spelling "airplane" was first recorded in 1907.[3]

Most fixed-wing aircraft are flown by a pilot on-board the aircraft, but some are designed to be remotely or computer controlled.

Fixed-wing aircraft

Boeing 737-300, a modern passenger airliner

Part of a series on [show]
Categories of Aircraft

Contents [show]

Structure [edit]

Some varieties of aircraft, such as flying wing aircraft, may lack a discernible fuselage structure and horizontal or vertical stabilizers, however the structure of a fixed-winged aircraft usually consists of the following major parts:

- A long narrow often cylindrical form, called a *fuselage*, usually with tapered or rounded ends to make its shape aerodynamically smooth. The fuselage carries the human flight crew if the aircraft is piloted, the passengers if the aircraft is a passenger aircraft, other cargo or payload, and engines and/or fuel if the aircraft is so equipped. The pilots operate the aircraft from a *cockpit* located at the front or top of the fuselage and equipped with windows, controls, and instruments. Passengers and cargo occupy the remaining available space in the

The Tupolev Tu-160, a

Figure 3.2: A Wikipedia article about *Fixed-wing aircraft*

3.3 Structural elements

We have explained why one would want to use Wikipedia as a knowledge base, and the general properties one could expect from the result. This section gives details of how Wikipedia can be adapted to provide structured knowledge, and the specific elements and features that have been exploited for this purpose. Unless otherwise stated, all statistics here are derived from a version of Wikipedia released on January 30, 2010.

3.3.1 Articles

The basic building block of Wikipedia is the article: a page that is dedicated to describing a particular topic or concept. These represent the conceptual units of a knowledge base. Their titles are unique identifiers and can be used as URIs in ontologies and descriptors in thesauri. Ambiguous titles are qualified by appending parenthetical expressions. For example, the title *Ontology* is reserved for the philosophical study of what is and how it can be categorized. The structures used to represent knowledge are described under the article entitled *Ontology (information science)*. Capitalization in titles is often significant. For example, Wikipedia distinguishes between *Optic nerve*—the connection between eye and brain—and *Optic Nerve*—a comic book series set in Northern California.

Articles are written in free text, but follow a strict and comprehensive set of guidelines. Take the *Fixed-wing aircraft* article shown in Figure 3.2. The items in bold indicate terms by which *fixed-wing aircraft* can be referred to. The first sentence is always a concise description of the topic—the equivalent of a gloss in more traditional resources. The first paragraph and the remainder of the first section provide extended glosses. If length justifies, the remaining content is organized into hierarchical sections, such as *structure, propulsion, history* and *safety*. All content will be relevant to the concept, and can be mined to inform systems about what *fixed-wing aircraft* are.

Internationally, Wikipedia contains 15M articles in its 270 different languages. The English version—which this thesis concentrates on exclusively—contains 3.1M articles (not counting redirects, disambiguation pages, or lists, which are discussed below). Many are placeholders—"stubs" in the Wikipedia lingo—but 2.4M are bona fide articles containing at least 50 words of descriptive text.

3.3.2 Redirects

There are often many terms that denote a given concept; many different queries that a user could issue to search for it, or in Wikipedia's case many titles that could be assigned to the same article. Wikipedia supports this linguistic phenomenon with redirects: pseudo articles that serve only to provide an alternative title for an article. There are about a dozen for *Fixed wing aircraft* and more than 4.0M in the entire English Wikipedia. They encode synonyms (*aeroplane, airplane),* plurals (*aeroplanes, airplanes*), technical or scientific terms, common misspellings, and other variants. Section 3.5.3 demonstrates that redirects accurately match the non-descriptors (use—use-for relations) found in thesauri, and can be used in exactly the same way. They are also easy to gather, as there is no natural language involved.

3.3.3 Disambiguation pages

Just as concepts can be referred to by many terms, terms can be used to refer to multiple concepts. *Plane*, for example, could refer to the aircraft, to a two dimensional surface, or to a woodworking tool. In a traditional resource this ambiguity is encoded using homonym relations. Wikipedia provides disambiguation pages that list each sense, along with a link to the appropriate article and brief scope notes to explain and differentiate between them. These special pages are identified (and separated from normal articles or list pages) by invoking certain templates (discussed in Section 2.2.8) or assigning them to certain categories (Section 2.2.6). The English Wikipedia contains 140K disambiguation pages. Unfortunately they are difficult to mine automatically, because they are written in free text and often refer to items that are only peripherally relevant. Inter-article links provide a cleaner alternative.

3.3.4 Inter-article links

Wikipedia is a densely interconnected structure, in which pages reference each other extensively. There are currently 68M directional links between articles, so each article references an average of 22 others. According to the manual of style,[21] these connections are used to:

- Group articles with related subject matter (for example, the article in Figure 3.2 contains links to *Aircraft* and *Flight*).
- Provide details of certain components or sub-topics of an article (for example, the article in Figure 3.2 contains a brief section about the *aerodynamics* of fixed wing flight, and a link to a dedicated article with more specific details).
- Explain jargon, technical terms or geographic locations that the reader may not be familiar with (e.g. *ornithopters*—flying craft that mimic the motion of birds).

Authors are expected to explicitly disambiguate links using a special syntax known as piping. For example, the markup [[*plane*]] produces a link to a disambiguation page and is unlikely to be helpful to readers. It should be "piped" as [[*Fixed-wing aircraft*|*plane*]] or [[*Plane (surface)*|*plane*]], depending on the desired destination. This is nothing more or less than word sense annotation and provides an alternative to disambiguation pages. To discover the different senses of *plane*, for example, one can simply gather the different destinations to which the anchor is piped across the resource. Link anchors are easier to mine than disambiguation pages. They are also easier and more lightweight to construct and thus more widespread. For example, there is no disambiguation page for *ontology*, but the link anchors specify that the term can refer to the broad philosophical discipline, or a specific type of knowledge base. Anchors provide statistics about the prior probability of each sense (e.g. 56% of *plane* links refer to 2D surfaces, 16% to aircraft, and 4% to the tool).

Piping allows links to encode synonymy and other surface form variants. Authors can, for example, freely choose between *plane*, *airplane*, and *fixed-wing aircraft* to suit their writing style. Thus links are an additional source of synonyms that may not have been captured by redirects, and again provide statistics of use. For example, authors are four times more likely to use *airplane* than *aeroplane*.

[21] Wikipedia's guidelines for inter-article links can be found at *http://en.wikipedia.org/wiki/Wikipedia:Linking*

A link from one article to another is a manually defined, manually disambiguated directional indication of relatedness between source concept and target. Thus Wikipedia already provides an extremely large graph of explicitly defined concepts and relations, without requiring natural language processing or information extraction. Unfortunately the graph is rather haphazard and bears little resemblance to the clean relations found in ontologies and thesauri. There is no explicitly defined information to identify the type of relation (e.g., a glider is a *kind of* aircraft*,* while a propeller is a *part of* one). Many links are of dubious utility to Wikipedia's readers and to those extracting knowledge bases from it (e.g. The article about *Dogs* contains a link to *Chocolate*, even though the connection is extremely tenuous, and the concept requires little explanation). Section 3.5.3 measures how inter-article links match up to relations in traditional thesauri, and Chapter 5 explores how to distinguish weak, irrelevant links from strong ones.

Inter-article links are particularly pertinent for this thesis, and are relied on extensively in the work that follows. One point of controversy is weither they are manually or automatically constructed. We assume that they are almost entirely hand-crafted by Wikipedians, but Huang et al. (2009b) claim that many are not. Admittedly, Wikipedia does provide a framework for building "bots" to carry out repetitive and mundane tasks. However, all bots are listed publicly,[22] and to our knowledge any that manipulate the link graph do so at a narrow scope for specific purposes (e.g. to bypass redirects that point to other redirects). Bots must undergo a strict approval process[23] that requires them to be "harmless" and adhere to the relevant policies and guidelines. It is exceedingly unlikely that widespread unsupervised link creation would be approved, given the complexity of Wikipedia's guidelines regarding linking.[21]

3.3.5 Categories

Wikipedians are encouraged to assign categories to their articles. For example, the article *Fixed-wing aircraft* falls under the category *Aircraft configurations*. Authors are also encouraged to assign the categories themselves to other more general categories: *Aircraft configurations* belongs to *Aircraft*, which in turn belongs to *Vehicles*. The resulting hierarchical structure extends upwards for approximately 16 levels, capped by a *Fundamental* category that contains the broad areas *Concepts, Life, Nature* and *Society.*

[22] A complete list of bots running on Wikipedia is provided at *http://en.wikipedia.org/wiki/Wikipedia:Bots/Status*

[23] Wikipedia's guidelines for bots can be found at *http://en.wikipedia.org/wiki/Wikipedia:Bot_policy*

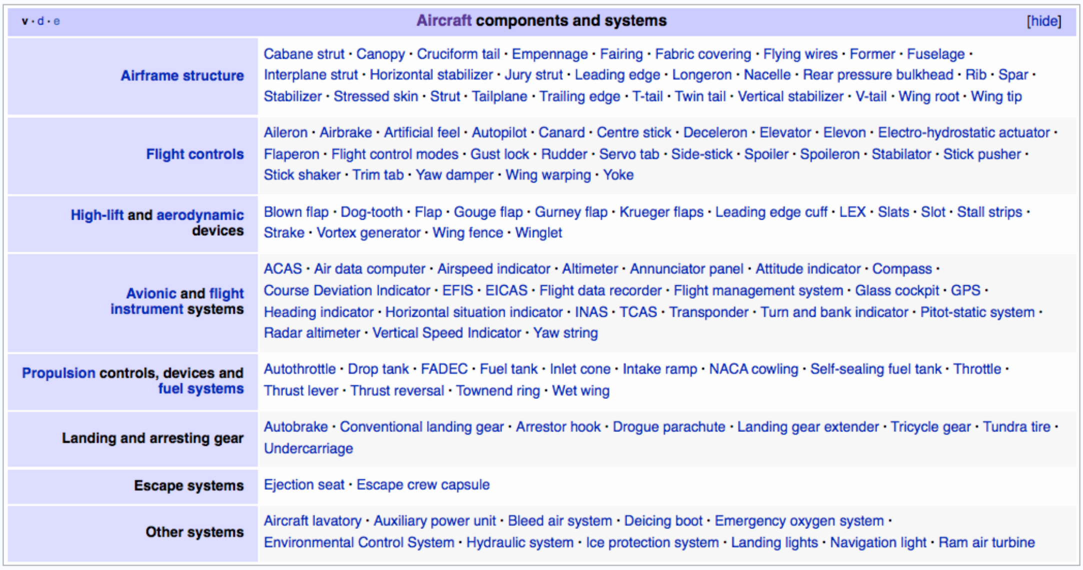
Categories are not themselves articles. They exist only to organize the articles they contain, with a minimum of explanatory text. Often (in about a third of cases), categories correspond to a concept that requires further description. In these cases they are paired with an article: the category *Aircraft* is paired with the article of the same name, and the category *Billionaires* with the article *Billionaire*. Other categories, such as *Aircraft by era*, have no corresponding articles and serve only to organize the content.

The goal of the category structure is to represent information hierarchy. It is not a simple tree-structured taxonomy, but a graph in which multiple parents are permitted. Both articles and categories can belong to more than one category. *Aircraft*, for example, belongs to two: *Vehicles* and *Aviation*. The overall structure approximates an acyclic directed graph: all relations are directional, and although cycles sometimes occur, they are rare. According to Wikipedia's own guidelines, cycles are generally discouraged but may be acceptable in rare cases. For example, *Education* is a field within *Social Sciences*, which is an *Academic discipline*, which belongs under *Education.* In other words, you can educate people about how to educate.

Like hyperlinks, the connections between categories and articles—or categories and their parents—represent a wide variety of types and strengths of relationships, which are not explicitly stated. Many links represent class membership, while others describe physical part-hood, geographical location and many other merely thematic associations between entities—as well as meta-categories used for editorial purposes, such as *Disambiguation*. Converting this informal structure into a strict taxonomy, or extracting typed relations from it, is far from trivial.

3.3.6 List pages

List pages do not provide a great deal of textual content on their own, but instead organize lists of links to other pages. They are similar in function to categories, but tend to be flat rather than hierarchical, and allow links to be explicitly ordered, grouped, and explained with free text. *List of Aircraft*, for example, organizes aircraft alphabetically and by type (*civil* or *military*, with the latter further subdivided by nation, class and the conflicts they were used in). Links have some potential to be mined for knowledge bases. Links collocated within a list page share some kind of relation, and the title of the list indicates the relation type. To our knowledge, list pages have never been mined in this fashion.



v · d · e	Aircraft components and systems [hide]
Airframe structure	Cabane strut · Canopy · Cruciform tail · Empennage · Fairing · Fabric covering · Flying wires · Former · Fuselage · Interplane strut · Horizontal stabilizer · Jury strut · Leading edge · Longeron · Nacelle · Rear pressure bulkhead · Rib · Spar · Stabilizer · Stressed skin · Strut · Tailplane · Trailing edge · T-tail · Twin tail · Vertical stabilizer · V-tail · Wing root · Wing tip
Flight controls	Aileron · Airbrake · Artificial feel · Autopilot · Canard · Centre stick · Deceleron · Elevator · Elevon · Electro-hydrostatic actuator · Flaperon · Flight control modes · Gust lock · Rudder · Servo tab · Side-stick · Spoiler · Spoileron · Stabilator · Stick pusher · Stick shaker · Trim tab · Yaw damper · Wing warping · Yoke
High-lift and aerodynamic devices	Blown flap · Dog-tooth · Flap · Gouge flap · Gurney flap · Krueger flaps · Leading edge cuff · LEX · Slats · Slot · Stall strips · Strake · Vortex generator · Wing fence · Winglet
Avionic and flight instrument systems	ACAS · Air data computer · Airspeed indicator · Altimeter · Annunciator panel · Attitude indicator · Compass · Course Deviation Indicator · EFIS · EICAS · Flight data recorder · Flight management system · Glass cockpit · GPS · Heading indicator · Horizontal situation indicator · INAS · TCAS · Transponder · Turn and bank indicator · Pitot-static system · Radar altimeter · Vertical Speed Indicator · Yaw string
Propulsion controls, devices and fuel systems	Autothrottle · Drop tank · FADEC · Fuel tank · Inlet cone · Intake ramp · NACA cowling · Self-sealing fuel tank · Throttle · Thrust lever · Thrust reversal · Townend ring · Wet wing
Landing and arresting gear	Autobrake · Conventional landing gear · Arrestor hook · Drogue parachute · Landing gear extender · Tricycle gear · Tundra tire · Undercarriage
Escape systems	Ejection seat · Escape crew capsule
Other systems	Aircraft lavatory · Auxiliary power unit · Bleed air system · Deicing boot · Emergency oxygen system · Environmental Control System · Hydraulic system · Ice protection system · Landing lights · Navigation light · Ram air turbine

Figure 3.3: A navbox for navigating between *aircraft components*

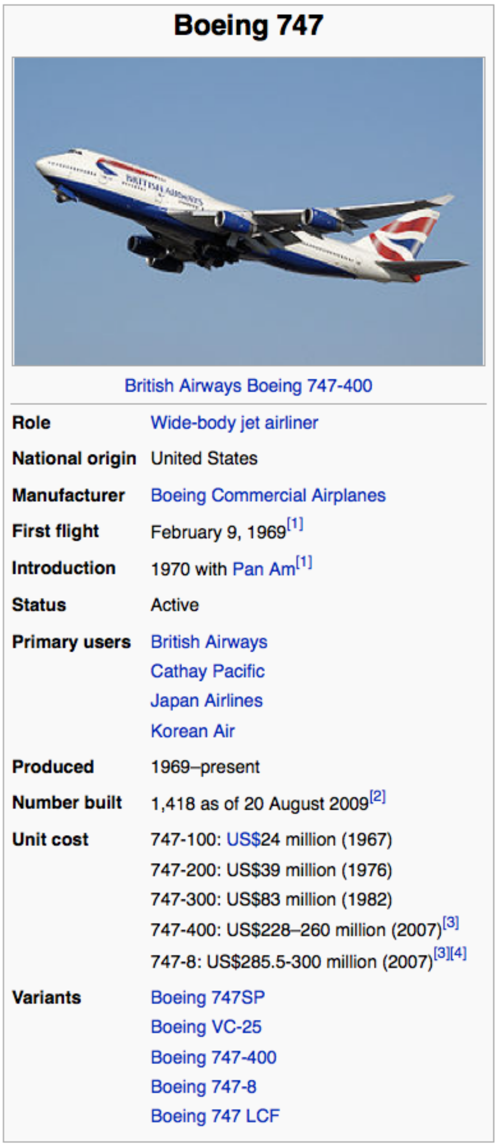
3.3.7 Portals

Portals are intended to provide a hub or home page for a particular field. The *Aviation* portal, for example, showcases a couple of high-quality articles including a selected aircraft, selected biography, and a constantly updated *Today in Aviation* list of anniversaries. It also organizes links to some of the most important articles into groups like *History* and *Aircraft components*. It also provides connections to projects outside Wikipedia (e.g., related events from WikiNews, or books from WikiBooks) and information for potential contributors (e.g., a list of articles requiring attention). For knowledge bases, these elements might help to group related topics together, or provide a starting point for gathering topics that are relevant to a particular domain. Like list pages, however, these have never been tapped.

3.3.8 Templates

Templates are pages that are not viewed in isolation, but are instead invoked to add information or additional formatting to other pages in a reusable fashion. They are designed to allow content to be duplicated or consistently formatted across more than one page. Wikipedia contains 210K different templates, which have been invoked 29M times. They are commonly used to identify articles that require attention due to bias, poor writing, a lack of citations, etc. They can also define pages of different types, such as disambiguation pages or featured (high quality) articles.

A navbox is a specific type of template that provides reusable lists or tables of links for navigating between groups of related pages. Typically, a navbox is included in every article it links to, making the same navigational hub consistently available on each. The *Aircraft components* navbox shown in Figure 3.3, for example, provides an attractively formatted table that organizes components into groups like *airframe structure* and *flight controls.* Every article that is referred to here contains a copy of the navbox, so it is easy to browse from one topic to another. These boxes provide the same opportunities for knowledge extraction as list pages, and to our knowledge have also not yet been tapped.



Boeing 747

British Airways Boeing 747-400

Role	Wide-body jet airliner
National origin	United States
Manufacturer	Boeing Commercial Airplanes
First flight	February 9, 1969[1]
Introduction	1970 with Pan Am[1]
Status	Active
Primary users	British Airways Cathay Pacific Japan Airlines Korean Air
Produced	1969–present
Number built	1,418 as of 20 August 2009[2]
Unit cost	747-100: US$24 million (1967) 747-200: US$39 million (1976) 747-300: US$83 million (1982) 747-400: US$228–260 million (2007)[3] 747-8: US$285.5-300 million (2007)[3][4]
Variants	Boeing 747SP Boeing VC-25 Boeing 747-400 Boeing 747-8 Boeing 747 LCF

Figure 3.4: An infobox for *aircraft*

Figure 3.4 demonstrates another type of template, called an infobox. These are designed not for navigation, but for summarization: to provide factsheets of individual topics that can be taken in at a glance. There is an *Aircraft* infobox, for example, which specifies the attributes (*Role*, *Manufacturer, Unit cost*) that one expects an aircraft to posses. This template is invoked in Figure 3.4 with the specific properties of the Boeing 747 (*Wide-body jet liner, Boeing Commercial Airlines*, etc). Infoboxes have received a great deal of attention from the research community because of their ability to express machine readable, typed relations. They have also been used as training data for extracting typed relations from Wikipedia's text (Section 3.4.4).

3.4 *Extracting knowledge*

This section revisits the types of knowledge bases that were described in Section 2.2, and explains how they can be extracted from Wikipedia.

3.4.1 Controlled vocabularies and glossaries

Controlled vocabularies provide one-to-one mappings between concepts and the terms (or descriptors) by which they should be referred. Wikipedia provides controlled vocabularies trivially, because each article can be considered a concept, and has only one title. Extending from controlled vocabulary to glossary is also trivial; the first sentence, paragraph and section of an article almost always provide succinct definitions that can be used for this purpose.

A one-to-one mapping between terms and concepts is required for consistency and clarity, but makes both indexing and searching difficult. Sections 3.3.2 and 3.3.4 outlined two options for mining synonymy relations from Wikipedia: redirects and link anchors. There are 4.0M redirects, which—combined with article titles—yields 7.1M distinct terms and phrases by which the 3.1M Wikipedia concepts can be referred to. Link anchors are slightly more widespread, with a total vocabulary size (when combined with article titles) of 8.8M distinct terms and phrases. However, since both items are trivial to extract, it seems reasonable to pool them. Doing so yields a vocabulary of 9.5M distinct terms and phrases.

There are also two options for identifying ambiguity: disambiguation pages (Section 3.3.3) and link anchors (Section 3.3.4). Assuming that every link from a disambiguation page indicates a sense of the term for which the page was created, these pages yield 140K ambiguous terms and 1.2M senses. That assumption is not always true, however, so the figure is artificially high. Grouping link anchors together and retaining those that link to multiple destinations yields 470K ambiguous terms and 1.8M senses. Link anchors are clearly a more useful source: they are more widespread, require no natural language processing, and provide useful prior-probability statistics (Chapters 5 and 6 capitalize on these).

3.4.2 Taxonomies

Taxonomies apply some kind of hierarchical categorization scheme, systematically subdividing domains into smaller, more specific chunks. Wikipedia's extensive network of categories would seem directly applicable, but has received a great deal of criticism and is generally regarded as requiring extensive clean up (Chernov et al. 2006, Muchnik et al. 2007, Ponzetto and Strube 2007). Its critics, on close inspection, all require something more formal than a human-readable categorization scheme, however. They instead aim towards a strict hyponymy/meronomy tree, such as McGuinness (2005) describes as the first step towards ontologies. Their efforts are described in Section 3.4.4. We argue that Wikipedia's category structure should be considered ground truth for how to organize Wikipedia concepts hierarchically. It was, after all, manually constructed for exactly this purpose.

3.4.3 Thesauri

The international standard for monolingual thesauri for information retrieval (ISO 2788) defines three types of relations that they are expected to express: synonymy relations to group equivalent terms; hierarchical relations to organize general and specific concepts; and associative relations to group topics horizontally. Extracting equivalence and hierarchical relations from Wikipedia has already been described in Sections 3.4.1 and 3.4.2. All that is left is to extract associative relations, which are defined only in vague terms. Roughly speaking, an associative relation should be made between two topics if, and only if, there is a reasonable chance that anyone interested in one topic would also be interested in the other (the relation is symmetric). The internal hyperlinks between Wikipedia articles were created to facilitate browsing from one article to the next, which suggests a simplistic approach for extraction: assume that all hyperlinks represent associative relations unless they are already expressed as hierarchical or equivalence relations. Unfortunately hyperlinks are also created for other purposes (e.g., to explain a technical but unrelated term that would otherwise confuse the reader) and often express relations that are tenuous at best (e.g., Figure 3.2 contains a link from *Fixed-wing aircraft* to *Britain*). Many irrelevant connections can be filtered out by insisting on mutual links—e.g. insisting that *Britain* must also link to *Fixed-wing aircraft* (which it does not). This matches the ISO standard's expectation for symmetry, but may be too strict a requirement: it reduces the size of Wikipedia's link graph from 68M connections to just 5.6M. These options are investigated in Section 3.5.3.

Given that associative relations are subjective and vaguely defined, it may not be sensible to restrict algorithms to binary yes/no decisions. It may be better to quantify relations and place them on a spectrum from unrelated through to weakly, moderately, and strongly associative. In this way, an algorithm can be more forgiving in the choices it makes, without diluting the most important relations. There are several techniques for gathering semantic relatedness measures from Wikipedia; they are surveyed in Chapter 5.

3.4.4 Ontologies

Revisiting the spectrum illustrated in Figure 2.2, the first step towards an ontology is to provide a formal hierarchy in which knowledge can be inherited vertically. The most important relations involved are hyponymy/*is-a* (e.g., *Boeing 747* is a *Commercial Aircraft*, *Aircraft* is a *Vehicle*) and meronomy/*is-a-part-of* (e.g., *Wing* is a part of

Aircraft). Wikipedia provides rigid hierarchies for certain types of topics. Biological organisms, for example, are organized using *taxoboxes* (a kind of infobox) that specify *class*, *order*, *genus*, *species,* etc. Cities and towns are organized using the *settlement* infobox, which specifies the surrounding region and country. However, templates also specify other kinds of relations (e.g. taxoboxes specify conservation status, settlement infoboxes specify population and size statistics). To our knowledge there has never been an attempt to systematically separate out the hierarchical relations and construct a clean taxonomy. Additionally, templates are somewhat difficult to author, so their coverage is limited; they don't exist for all types of articles, and aren't invoked on all instances of those types.

Wikipedia's category structure is much more widespread, but too loose and haphazard to be used as a formal hierarchy directly. Ponzetto and Strube (2007) aim to classify article–category and subcategory–category associations as either *is-a* or *not-is-a* using a range of heuristics. They assume an *is-a* relation, for example, if parent and child titles share a lexical head (e.g., *British computer scientist* and *Computer scientist*), and the opposite if they share a modifier (e.g., *Islamic mysticism* and *Islam*). If both titles are found as noun phrases in a sentence, the intervening text may provide clues: "such as" indicates an *is-a* relation (e.g., *fruit such as apples and pears*), while "are used in" does not (e.g., *fruit are used in cooking*). These clues, when combined with co-occurrence statistics, produced 100,000 *is-a* relations from Wikipedia with an f-measure of 88% when evaluating against ResearchCyc (an ontology described in Section 2.2.1).

The next step is *formal instance*, which involves separating *is-a* relations into *is-a-subclass* (e.g. *fruit* is a subclass of *food*) and *is-an-instance* (e.g. *New Zealand* is an instance of *country*). Yago and DBPedia—two Wikipedia derived ontologies that are described in the next section—assume that all categories are classes and all articles are instances. Unfortunately the assumption is often false: almost any sufficiently popular or complex topic will have a category to organize the related articles, and many of these—*New Zealand, Barack Obama* and the *Java Programming Language* to name a few—are instances rather than classes. Zirn et al. (2008) explore several techniques for separating classes and instances on a case-by-case basis. They achieve 86% accuracy when evaluating against 8000 categories defined in ResearchCyc.

The next step in extracting ontologies from Wikipedia involves specifying properties by expanding the vocabulary of relations (e.g., <book> *written-by* <author>, <book> *has-genre* <genre>) and introducing value types (<book> has-page-count <*number*>, <book> first-published <*time-stamp*>). The challenge of automatically gathering subject-predicate-object triples from Wikipedia has captured the interest of many researchers.

Wikipedia's infobox templates are, at first glance, an promising resource for gathering triples. The infobox in Figure 3.4 states in machine-readable form that every *aircraft* has a *role* and *manufacturer*, and that a *Boeing 747's* role is *Wide-body jet airliner*. Unfortunately, extracting knowledge from infoboxes is not as trivial as one might hope. They are designed case-by-case to provide fact sheets for certain types of articles, not for defining resource-wide schema. Consequently they contain many inconsistencies. There are, for example, four different templates for describing *US Counties*, and four different attributes for stating which year a population statistic was obtained in (Wu and Weld 2006). Infoboxes often contain natural language text that is difficult to parse (e.g., Figure 3.4 contains four different date formats, and the *unit price* property would be difficult to process). Krötzsch et al. (2006) and Völkel et al. (2006) provide detailed plans for improving Wikipedia's framework for describing structured knowledge (by specifying attribute types, for example), but implementing their vision would require a significant investment of manual labour and has so-far gained little traction with the Wikipedia community.

Given the lack of momentum and enthusiasm this semantic Wikipedia intuitive has gained so far, it seems—for now, at least—that infoboxes are the furthest Wikipedians can be expected to go towards explicitly defining structured knowledge. Even these are not widespread; they are not set up for all classes, and not invoked for all instances. They are only present in 40% of articles.

There have been many efforts to extract structured knowledge from elsewhere in Wikipedia. Nastase and Strube (2008), for example, work with Wikipedia's categories. Where previous work (Ponzetto and Strube 2007) seeks only to classify existing relations between parent and child, Nastase and Strube identify new relations by parsing category and article titles for certain manually-defined noun and verb phrase patterns. As Table 3.1 demonstrates, each association between parent and child can potentially produce several triples. When run over the entire category hierarchy, this approach identifies a total of 3.4 million *is-a* and 3.2 million *spatial* relations, along with 43,000 *member-of* relations and 44,000 other relations such as *caused-by* and *written-by*. Precision ranges from 84% to 98%, depending on relation type.

The bulk of research that has been invested in Wikipedia is focused on extracting ontological relations from its textual content. After all, this is where most of its information resides. Wikipedia's encyclopaedic nature offers some advantages over similar attempts to automatically extract relations from other corpora. Each article is focused on describing a particular concept, which will be the *subject* for most subject-predicate-object triples that are extracted from the text. The *objects* are all entities mentioned in the text. These are often helpfully marked up and disambiguated via piped links. Once subject and object have been identified, all that remains is to decide what type of relation exists between them.

Ruiz-Casado et al. (2005b) focus on expanding the vocabulary of relations that WordNet (Section 2.2.1) describes between its entities. In previous work they describe an algorithm for disambiguating articles in the Simple English Wikipedia to the corresponding entries in WordNet (Ruiz-Casado et al. 2005a). This allows them to detect when WordNet nouns (e.g., *Lisbon, Portugal)* are located within a sentence in the Simple English Wikipedia ([[*Lisbon*]] *is the capital city of* [[*Portugal*]]). If the nouns are related according to WordNet (in this case, via *is-part-of*), the intervening text is stored as a pattern for the given relation, after generalizing via comparison with similar extracted texts (e.g., [[*Auckland*]] *is a city in* [[*New Zealand*]]). These patterns are then used to draw new relations between WordNet nouns by locating new sentences that mention them. The results are modest, however: only 1200 new semantic relations are identified with 61–69% precision, depending on relation type.

In later work, Ruiz-Casado et al. (2006) focus on seven relations that WordNet does not describe: *birth-year*, *death-year*, *birth-place*, *actor-film*, *country-chief_of_state*, *writer-book*, *country-capital* and *player-team*. The technique is much the same, except that training is done with seed lists of term pairs that demonstrate each relation, and the vocabulary of topics is defined not by intersecting Wikipedia with WordNet but by crawling out from certain Wikipedia pages such as *List of authors* and *List of national capitals.* The results vary wildly (8% accuracy for *player-team*, 90% for *death-year*) unless the crawler is manually seeded for each relation—accuracy for the *player-team* relation increases to 93% if the crawler starts from *list of football (soccer) players*. Recall

Category name	Pattern	Relation
Queen (band) members	X members Members of X	Freddy Mercury *member_of* Queen (band) Brian May *member_of* Queen (band)
Movies directed by Woody Allen	X [VBN IN] Y	Annie Hall *directed_by* Woody Allen Annie Hall *is_a* Movie
Villages in Brandenburg	X [IN] Y	Siethen *located_in* Brandenburg Siethen *is_a* Villiage
Mixed Martial Arts Television Programs	X Y	Mixed martial arts *unknown_relation* Television programs Tapout (TV series) *unknown_relation* Mixed martial arts Tapout (TV series) *is_a* Television Program
Albums by Artist ↓ Miles Davis Albums	X by Y	Artist *attribute_of* Album Miles Davis *is_a* Artist Big Fun *is_a* Album Big Fun *artist* Miles Davis

Table 3.1: Decoding Wikipedia categories (from Nastase and Strube 2008)

is not assessed, but the approach gathered 16,000 *birth-years*, 6000 *death-years* and 2000 other relations.

The previous approaches generalize patterns lexically through stemming, stopword removal, and discarding words that differ across otherwise similar patterns. They still rely on sentences following a consistent pattern where the relation is located between subject and object, with a minimum of intervening text. They break down on sentences that are complex or break the pattern—e.g., *the capital city of Portugal is Lisbon.* More success has been obtained by generalising syntactically, through dependency parsing.

There are earlier examples that apply dependency parsing to relation extraction—Pantel and Lin (2001), for example—but to our knowledge Herbelot and Copestake (2006) were the first to apply it to the Wikipedia corpus. They use a dependency parser to identify subject, object and relationship in a sentence, irrespective of word order, and focus on extracting *is_a* relations from biological articles. By manually annotating 100 articles, they identify patterns that indicate relations of interest (such as A *is a* B, A *is a species of* B, A *is a* B *species*, etc). The parser is used to simplify the sentence, to extract *opah is_a fish*, for example, from *Opah (also known colloquially as moonfish, sunfish, kingfish, and Jerusalem haddock) are large, colourful, deep-bodied pelagic Lampriform fish comprising the small family Lampridae (also spelt Lamprididae).* This gives high precision (92%) but low recall (14%). Herbelot and Copestake also attempt to automatically identify patterns, which improves recall (37%) at the expense of precision (65%).

Suchanek et al. (2006) generalize to different subject matter—one set of Wikipedia articles about musical composers, another about geographical locations, and a third (randomly selected) corpus—and different relations—*birth_date*, *synonomy,* and *instance_of.* They also introduce machine learning. Like Ruiz-Casado et al. (2005a), they begin with seed lists of term pairs that demonstrate each relation. They then locate linkages—grammatical patterns detected by the Link Grammar Parser (Sleator and Temperly 1993)—between the example pairs, to produce positive examples for a classifier. The algorithm runs through the sentences again and finds all linkages that match a positive pattern but produce an incorrect relation (a potential *birth_date* that does not match the correct birth date, or a potential *synonym* or *instance_of* that is not defined as such in WordNet). These become negative examples. The approach works better for detecting *birth_dates* (74% f-measure) and *synonyms* (68% f-measure) than *instances* (48% f-measure), and better when restricted to the appropriate corpus (48% for *instance_of* over the composer corpus vs. 33% over the general corpus).

Wu and Weld (2006) map out a compelling vision in which Wikipedia becomes a semi-supervised relation-generating machine. They see relation extraction as a by-product of refining and improving Wikipedia's infoboxes. These templates provide training data and inform the system what classes and attributes are of interest, and the system automatically adds infoboxes to articles that do not have them, or fills in missing attributes. In this way Wikipedia and the relation extraction system—called Kylyn—become symbiotic: Wikipedia gains better infoboxes, and Kylyn gains ongoing supervision from human editors.

The Kylyn system involves three main steps. First, it identifies articles that belong to a given class and therefore should be augmented with the appropriate infobox. Wu and Weld explain that this step would ideally draw on highly successful machine-learned document classification techniques, but instead present an ad-hoc baseline that inspects the categories of pages to which the infobox has already been applied. On the four classes they tested—*Country*, *Airline*, *Actor* and *University*—this achieved an average of 98.5% precision and 68% recall. Second, Kylyn learns to predict which attribute values—if any—are contained in a sentence text. Training data is obtained from articles with existing infoboxes, to which a series of heuristics are applied to automatically match sentences with the corresponding attributes. Features are sentence tokens and part-of-speech tags. The performance of this sentence classifier is not reported. The last step is to extract the attributes from sentences, for which conditional random fields are individually trained for each attribute. These achieve an average of 87% precision and 72% recall, and even outperformed human annotators for the *County* class (where attributes are typically numeric and therefore easier to extract). It should be noted, however, that the evaluation was restricted to attributes for which the training data was abundant and could be easily processed.

Now that the main developments in gathering ontologies from Wikipedia have been charted—i.e. lexical to syntactic generalization, the introduction of machine learning, the use of Wikipedia's infoboxes for training, and the possibility of ongoing manual supervision—the remaining work will be covered in less depth. As Section 3.6 explains, it has only passing relevance to this investigation.

Many experiments draw on related natural language processing research. Nguyen et al. (2007b) and Wang et al. (2007b) focus on increasing precision by typing entities—a well known problem—using first sentence definitions, parts of speech and parent categories. This produces a relation extraction system that knows, for example, that *birth_date* should be between a person and a date, and *founded_by* between an organization and a person. Nguyen et al. (2007b) and Nakayama (2008) use anaphora and co-reference resolution to increase coverage (so that *Barak Obama* can be detected as *Obama*, *the President*, *he*, etc.). Wu and Weld (2006) use an ad hoc technique for adding new links to Wikipedia articles (again, to assist with detection of entities); this problem is addressed more systematically in Chapter 6. Zhang et al. (2008) focus on time-dependent relations (such as Obama's presidency). Adar et al. (2009) align infoboxes across different language versions of Wikipedia, to improve the coverage of all.

Other experiments focus on obtaining training data or stretching it as far as possible. The best source of training data appears to be Wikipedia's infoboxes, but both these and their attributes follow a long-tailed distribution where most are invoked very rarely. Wang et al. (2007a) deal with the scarcity of training data through bootstrapping and automatically gathering negative examples. Blohm and Cimiano (2007) and Wu et al. (2008) tackle the same problem, but use the web to gather additional training examples. Yan et al. (2009) avoid training altogether—their algorithm for unsupervised relation extraction is not informed what types of relations to extract or given examples to learn from.

The extraction of structured knowledge from Wikipedia has gained great momentum, but there is a need for consolidation. There is little overlap between the corpora being used or the relations being extracted, and few comparisons are drawn directly between the systems. Many of the methods seem complementary, and could likely be combined, but it is difficult to assess their individual contributions or predict their performance if they operated together. Another limitation is that—with the exception of Yan et al. (2009)—systems start with known relation types (*is_a, birth_date*, etc*)* and then locate entity or entity-attribute pairs for which the relation should hold. They will only provide relations and connections that are explicitly anticipated, or for which training data is provided by existing infoboxes. Such limited coverage may be sufficient for focused question answering, but has limited use for exploratory search, in which useful connections cannot easily be anticipated in advance. Researchers could plausibly start from the other direction: first detect entities, and then ask how they are related. This problem is known as open relation extraction (Banko et al. 2008). Unfortunately, Wikipedia is only just beginning to be exploited in this fashion.

3.4.5 Stand-alone knowledge bases

Table 3.2 lists the larger and more significant resources that have been derived from Wikipedia. The first of these—the Wikipedia Thesaurus[24]—is, despite the name, not a thesaurus as defined in Section 2.2. It contains 4M entitles and 244M associative relations

[24] *http://dev.wikipedia-lab.org/WikipediaThesaurusV3*

Knowledge base	Entities	Relations or facts
Wikipedia Thesaurus	3.8M	240M
YAGO	2M	20M
DBpedia	2.9M	190M
Freebase	50M	430M

Table 3.2: Knowledge bases derived from Wikipedia

between them, but does not describe synonymy or hierarchy. The sheer number of concepts suggests they were gathered directly from Wikipedia articles without cleanup. The associative relations provided between concepts are automatically derived via analysis of inter-article links. The algorithm behind this—from Hara et al. (2008)—is described in Section 5.3.

YAGO[25] is the result of Suchanek et al.'s (2006, 2007) efforts—described in Section 3.4.4—to augment WordNet with topics and relations obtained from Wikipedia. It contains 2M entities, which are either WordNet synsets or Wikipedia articles that could be cleanly integrated into WordNet's hierarchy. Articles are integrated if their parent category matches an existing synset, and can be automatically classified as a certain type (e.g. *person*, *book*, *organization*). There are 20M facts asserted about these entities, of the form *bornIn(person, year)*, *locatedIn(object, region)*, and other relation types. These have been manually assessed as having an accuracy of 95%. YAGO is freely available for download, and can be queried online.

DBpedia[26] aims to construct an entirely new ontology from Wikipedia, as opposed to augmenting an existing one (Bizer et al. 2009). The English language version contains 2.9M entities and 190M RDF triples. The scale is impressive, but somewhat misleading. The resource is essentially a rough dump of Wikipedia's structure, with limited cleanup. To our knowledge, the relations within it have never been formally evaluated. A quick manual inspection reveals that large sections have limited value. For instance, 60% of the RDF triples are directly obtained from internal links between articles, which, as Section 3.5.3 explains, require extensive filtering before they can be used to indicate semantic relations. 15% of triples are taken directly from infoboxes. Of those, the most common relation (over 10%) is the uninformative formatting flag *wikiPageUsesTemplate*. Amongst remaining relations many obvious redundancies not identified as such, e.g. *placeOfBirth* and *birthPlace*, *dateOfBirth* and *birthDate*.

[25] *http://www.mpi-inf.mpg.de/~suchanek/downloads/yago*

[26] *http://dbpedia.org*

Freebase, an extensive ontology released open source by the Metaweb foundation, was described in Section 2.2.1. It derives knowledge not just from Wikipedia, but a wide array of sources including MusicBrains, the Notable Names Database, and direct crowd sourcing. It has never been formally evaluated, but its relations appear fairly accurate on cursory examination (the resource can be freely browsed online). Its impressive size (Table 3.2), combined with its apparent accuracy, suggests that content is gathered from already highly formal resources—that the resource is essentially a centralized Semantic Web. If automated mining of Wikipedia is a major source, then the algorithms involved must be significantly more advanced than the work described in Section 3.4.4, or rely heavily on manual refinement.

3.5 Extracting domain-specific thesauri

Wikipedia is a potential source of domain-specific, technical thesauri. This section compares its raw structural elements with Agrovoc, a thesaurus created and maintained by the UN Food and Agriculture Organization (FAO) to organize and provide efficient access to its document repository. Agrovoc is a substantial thesaurus, with approximately 28,000 terms (17,000 descriptors and 11,000 non-descriptors), organized vertically by 16,000 hierarchical relations, and connected horizontally by 27,000 associative relations. The following subsections give details of the analysis and measure Wikipedia's coverage of Agrovoc's concepts and relations.

This experiment was conducted in 2006, and only considered structural features that seemed relevant at the time. It is based on a version of Wikipedia released on June 3, 2006. That contains approximately 1M articles (the equivalent of Agrovoc descriptors) and 1M redirects (non-descriptors). The link anchors described in Section 3.3.4 are not considered, even though they are another viable source of non-descriptors. Wikipedia's articles are organized into 120,000 categories, which yield just over 3M hierarchical relations. There are 29M links between the articles.

Ideally, the experiment should be repeated with a newer version of Wikipedia (it has since tripled in size), using the full suite of structural features available for deriving thesauri described in Sections 3.4.1–3.4.3.

3.5.1 Comparison strategy

To compare concept coverage between the two resources, superficial differences—e.g., *process recommendations*, *recommended processes* and *processing recommendations*—are ignored. Terms in both structures are case-folded, stripped of punctuation, and

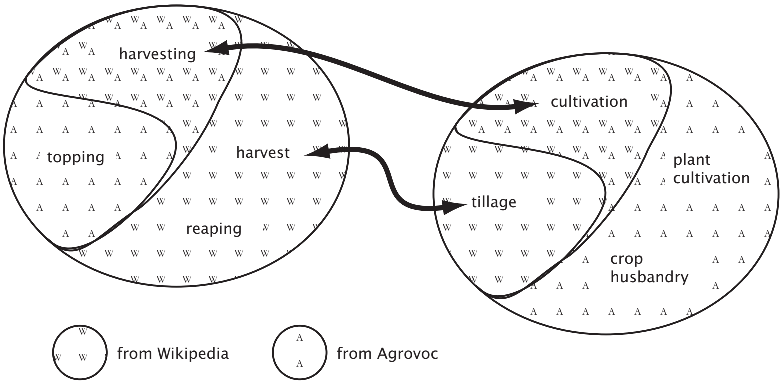


Figure 3.5: Comparing relations between Wikipedia and Agrovoc

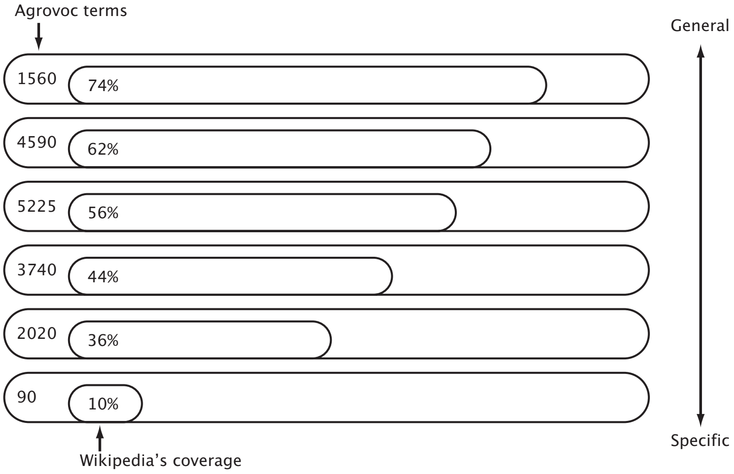


Figure 3.6: Wikipedia's coverage of general and specific Agrovoc concepts

stemmed using the Porter stemmer (Porter 1980) before comparison. In addition, stopwords are removed and word order within each phrase is normalized alphabetically.

Both Wikipedia and Agrovoc choose one primary term to represent each concept and support them with synonyms. The actual choice of descriptor is somewhat arbitrary, and differences between the two resources are common, particularly for concepts that can be described either with a scientific term or an everyday expression: Agrovoc tends towards the former, while Wikipedia prefers the latter. The comparisons that follow consider a concept to be matched between the two resources if the descriptor or any non-descriptors in one resource match (after casefolding, etc.) any of the descriptors/non-descriptors in the other.

Within each resource, only descriptors are referred to when connecting two concepts via hierarchical or associative relations. Figure 3.5 demonstrates how the choice of descriptor is ignored in the relation comparisons that follow. Here both resources are considered to express the same relation, even though Agrovoc connects *harvesting* to *cultivation* and Wikipedia connects *harvest* to *tillage.*

3.5.2 Coverage of concepts

Wikipedia covers approximately 50% of Agrovoc's concepts. The vast majority of those found in the former but not the latter are irrelevant because they lie outside Agrovoc's domain. More interesting—and concerning—are the Agrovoc concepts that Wikipedia fails to cover. Cursory examination indicates that these are generally scientific terms or highly specific multi-word phrases, such as *margossa*, *bursaphelenchus* and *flow cytometry cells*.

Concepts in Agrovoc can be stratified into groups according to where they occur within the thesaurus hierarchy. Figure 3.6 shows how the overlap varies across levels, and how Wikipedia's coverage of Agrovoc degrades as concepts become more specific. Not surprisingly, Wikipedia provides better coverage of more general terms.

One-third of the terms that were matched by Wikipedia were ambiguous according to the more general purpose resource; they match multiple articles. For example, the Agrovoc term *viruses* relates to separate articles for *biological viruses* and *computer viruses*.

3.5.3 Coverage and accuracy of relations

Figure 3.7a shows that Agrovoc's synonymy relations are covered particularly well by Wikipedia: only 6% are absent. Wikipedia's redirect structure is responsible for most of this, covering 75% of Agrovoc's synonymy relations. 19% of related term pairs that Agrovoc deems equivalent are encoded in Wikipedia through other links. In these cases—e.g. *aluminium foil* → *shrink film*, *spanish west africa* → *rio de oro*—Agrovoc judges two concepts to be "near enough" in that they do not require separate entries, whereas Wikipedia splits them.

Figure 3.7b analyses Agrovoc's hierarchical relations. Wikipedia covers 69% of them, but only 25% appear in the category structure. The remaining 44% are found in hyperlinks between articles. Many missing relations are due to the two resources describing hierarchies at different levels of detail: e.g. the Agrovoc relation *oceania* → *american samoa* is described in Wikipedia as the lengthy chain *oceania* → *oceanian countries* → *oceanic dependencies* → *american samoa*. Coverage increases significantly when implicit chains of relations are considered.

A full 84% of the relations in Wikipedia's category structure are absent from Agrovoc's hierarchy. Many are implicitly encoded (through chains of relations), while others are

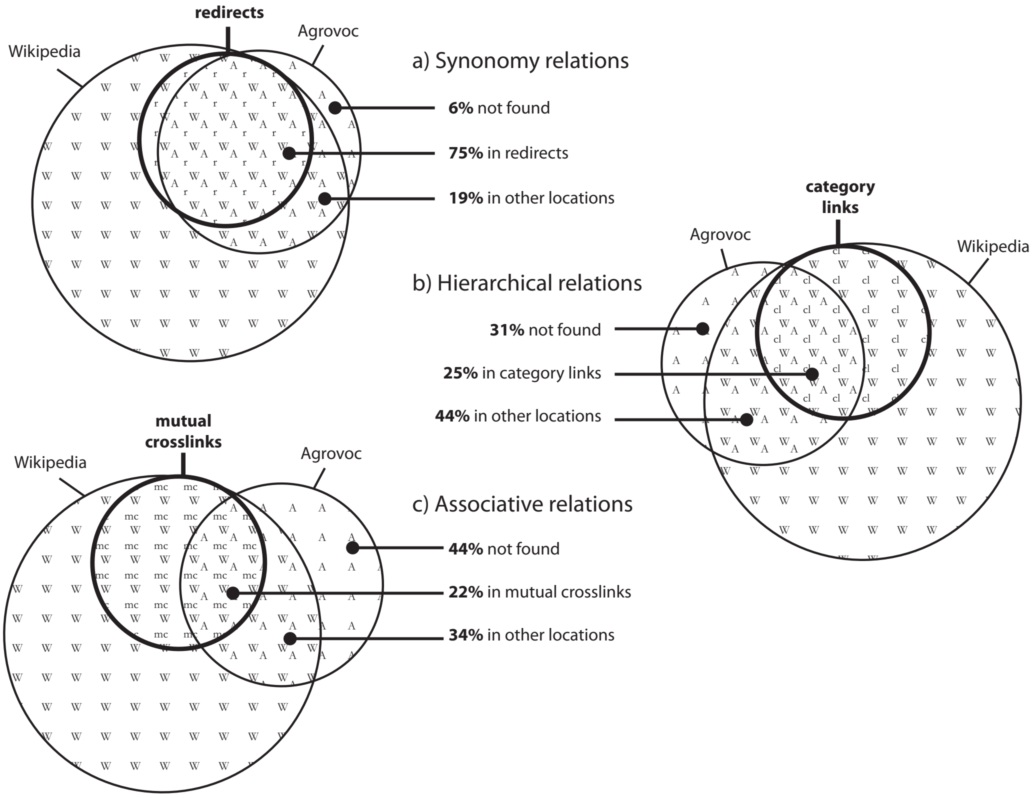


Figure 3.7: Wikipedia's coverage of Agrovoc relations

irrelevant to Agrovoc's domain and were included in the comparison because of lack of disambiguation. For example, Wikipedia contains several senses for the ambiguous term *power*, one of which relates to *sociology*. Agrovoc is concerned only with *electrical power* and not *personal empowerment*, and therefore does not make the same connection. Other relations may represent a useful increase in connectivity over Agrovoc.

Figure 3.7c depicts associative relations, of which Wikipedia covers 56%. Mutual links between articles were expected to match these relations closely. However, only 22% were found in this way; the remaining 34% were found within one-way links or the category structure. Much of the missing 44% are encoded implicitly in Wikipedia: for example, Agrovoc's relation *gene transfer* → *gene fusion* is present because both terms are siblings under the Wikipedia category *genetics*.

There are many mutual cross-links in Wikipedia that do not correspond to relations in Agrovoc. Many—e.g. *human* ↔ *ape* and *immune system* ↔ *lymphatic system*—are perfectly valid and relevant relations that do not appear in Agrovoc, even implicitly. Other cross-links describe relations that Agrovoc leaves implicit—e.g., all siblings (defined by hierarchical relations) are implicitly associatively related. Other mismatches are caused by a lack of sense disambiguation when terms are compared across the two resources.

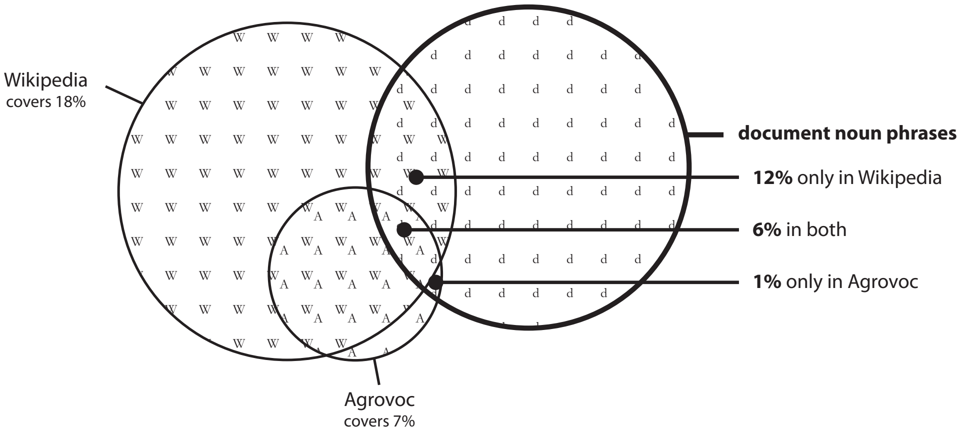


Figure 3.8: Wikipedia's and Agrovoc's coverage of document concepts

3.5.4 Corpus coverage

This section investigates Wikipedia relevance for a domain-specific document collection—that is, how well it covers the collection's terminology. The corpus in question is a set of 780 agricultural documents taken from the Food and Agriculture Organization's repository—i.e., the texts that Agrovoc was specifically constructed to support. All documents are full text (not abstracts) and have been professionally indexed with at least three Agrovoc terms. These terms form a small subset of Agrovoc (9.3%), but were manually chosen by the indexers as particularly relevant for the corpus. Encouragingly, Wikipedia's coverage grows from 50% of the full Agrovoc (from Section 3.5.2) to 72% of index terms. Wikipedia would be an adequate controlled vocabulary for tagging these documents. Coverage is still incomplete, however: Wikipedia missed important terms such as *yield forecasting*, *sediment pollution* and *land economics*. In most cases, more general Wikipedia concepts—e.g. *forecasting*, *pollution*, *economics*—are available.

Index terms form a small sample of relevant Agrovoc entries. To gain a more detailed view, noun phrases were automatically extracted from the documents using the OpenNLP toolkit for linguistic analysis. Figure 3.8 shows a three-way comparison between Agrovoc, Wikipedia, and the extracted noun phrases. Most phrases are not found in either source, which is unsurprising; it merely indicates that most noun phrases are not suitable thesaurus terms, syntactically or semantically. The terms found in either structure, however, can be assumed to represent valid concepts mentioned in test documents.

Wikipedia covers approximately three times as many document concepts as Agrovoc. Many of these—e.g. *high school*, *aztec religion*, and *asean free trade area*—probably lie outside Agrovoc's intended domain. They are, however, distinct concepts that are

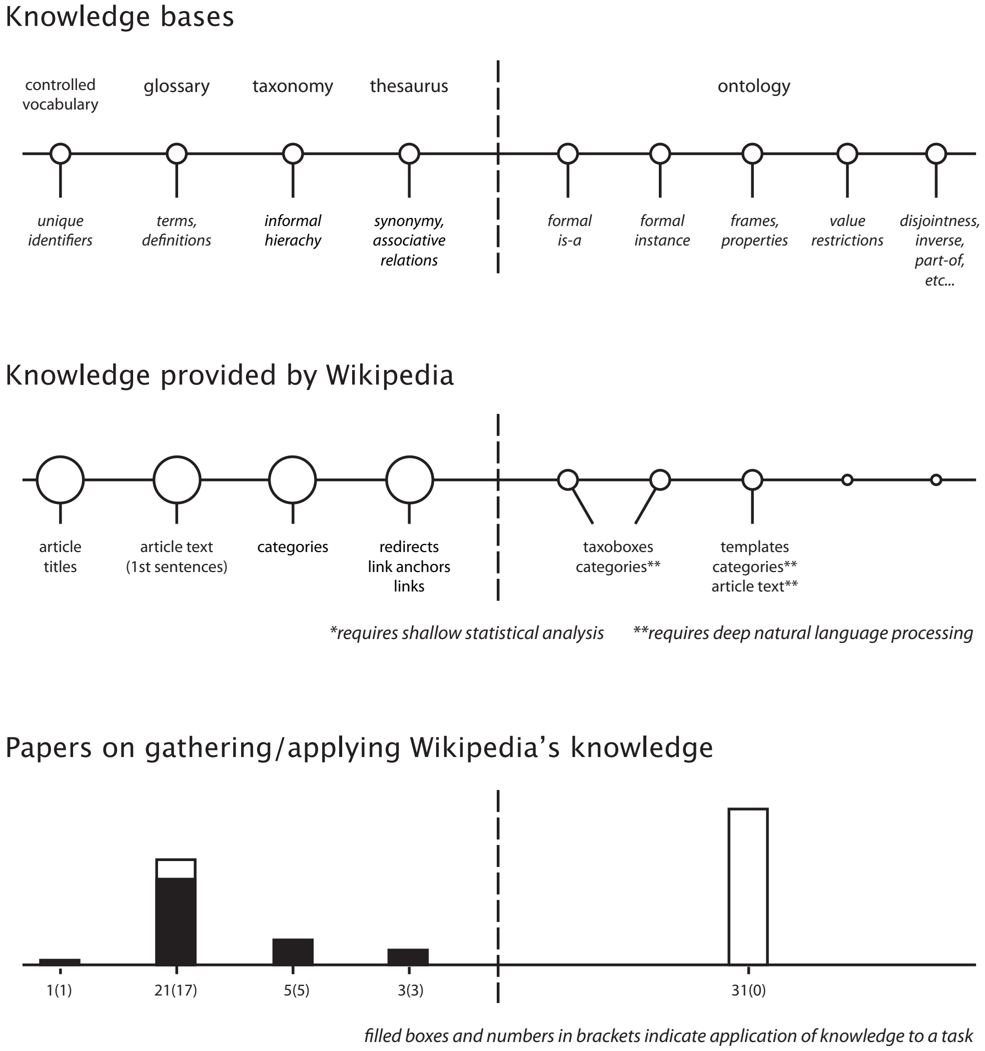

Knowledge bases
controlled vocabulary
glossary
taxonomy
thesaurus
ontology
unique identifiers
terms, definitions
informal hierachy
synonymy, associative relations
formal is-a
formal instance
frames, properties
value restrictions
disjointness, inverse, part-of, etc...
Knowledge provided by Wikipedia
article titles
article text (1st sentences)
categories
redirects link anchors links
taxoboxes categories**
templates categories** article text**
*requires shallow statistical analysis
**requires deep natural language processing
Papers on gathering/applying Wikipedia's knowledge
1(1)
21(17)
5(5)
3(3)
31(0)
filled boxes and numbers in brackets indicate application of knowledge to a task

Figure 3.9: Extracting knowledge from Wikipedia

mentioned in the corpus. It would be valuable for users to have them described and navigable. In terms of concept coverage, Wikipedia is substantially better suited to describing this document collection than Agrovoc.

3.6 Discussion

This section revisits the discussion made in Section 2.4, regarding the kinds of knowledge base that are suited to interactive information retrieval. The previous discussion raised doubts about the practicality of heavyweight ontologies for supporting exploration, particularly in the short term. For now, it would seem better to focus on simpler, less formal structures and on human-readable rather than machine-readable knowledge. These are easier to obtain, and—for the majority of strategies surveyed in Section 2.3—more directly applicable.

As Figure 3.9 illustrates by revisiting the McGuinness (2005) knowledge base spectrum, the point becomes stronger when made specific to mining and applying Wikipedia. If one looks at what Wikipedia provides directly at a large scale, there is a clear bias towards the left side of the spectrum. Controlled vocabularies, glossaries and taxonomies are directly encoded (by article titles, first sentence definitions, and the category hierarchy respectively) at a large scale. Synonymy is encoded extensively in redirects and anchors, to a high level of accuracy. The remaining component of human-readable thesauri—associative relations—is only roughly catered for by inter-article links. Fortunately, as Chapter 5 will demonstrate, these relations can be accurately and cheaply approximated by considering links in aggregate rather than individually.

As one moves across the bar into ontologies, the extraction of knowledge becomes more challenging. Each step moves further away from what Wikipedia explicitly describes in large volumes. Clean *is-a* hierarchies, for example, are only found within a few templates and taxoboxes. Section 3.4.4 described several attempts to mine the category hierarchy, but this is a challenging task for which, so far, only 100,000 relations have been extracted at 88% f-measure. Thirty times that many hierarchical associations were manually defined in the raw category structure in 2006, and Wikipedia has since doubled in size.

Most efforts to mine typed relations and properties from Wikipedia revolve around its templates, either using them directly or as training data for mining the text. Little of the work described in Section 3.4.4 extends to retrieving facts that are not explicitly described by templates, or types of relations that are not repeatedly invoked within them. The coverage of templates is growing but still incomplete. Even if they were applied universally to all Wikipedia articles—if the efforts of Wu and Weld (2006) achieved full fruition—they would yield limited connections. They are intended to provide succinct fact-sheets rather than exhaustive descriptions, and consequently contain only a fraction of the knowledge captured in human-readable form within article text, or informally within the categories and inter-article links.

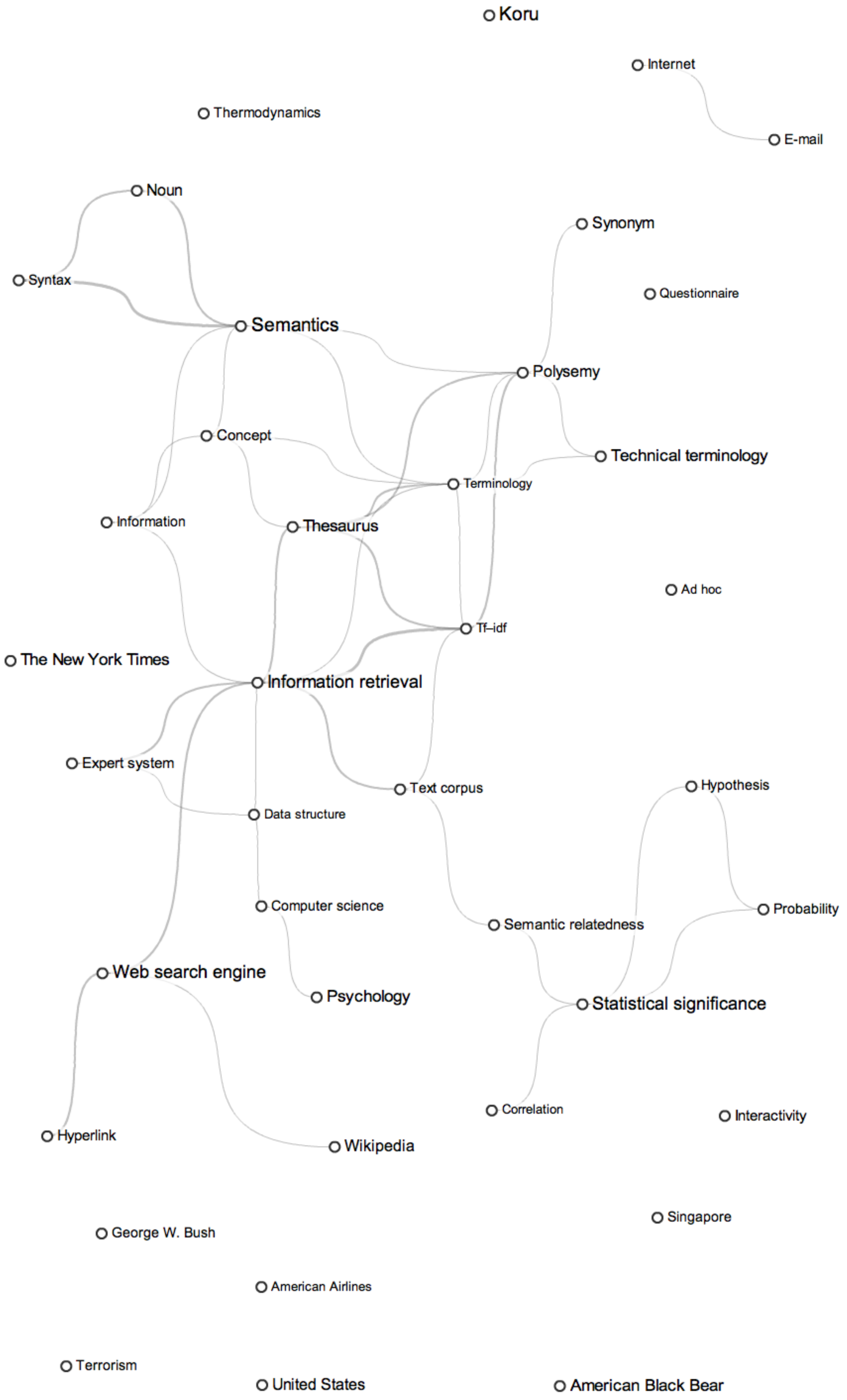
A further concern is that, even if the inherently difficult problem of extracting facts from Wikipedia's text were solved to perfection, the underlying source is controversial and vulnerable to inaccuracy. Ontologies mined from the text will inherit factual errors directly. Lesser structures, which deal in loose associations rather than hard facts, and seek only to organize other information sources rather than resolve information needs directly (the metadata vs. data issue raised in Section 2.4), do not face the same problem.

The bottom of Figure 3.9 plots the number of papers that mine Wikipedia's structure and content. This chart has been constructed by gathering all relevant papers that have been encountered during the course of this investigation—excluding our own—and manually classifying them according to the expressiveness they require from Wikipedia. Few papers mention the relevant knowledge bases—controlled vocabularies, taxonomies, etc—by name, and consequently the classifications are somewhat loose. For example, Gabrilovich and Markovitch (2006) is grouped under *glossaries* because it uses Wikipedia's article titles and textual content to gather features for text categorization. Wang et al. (2007c) is grouped under *taxonomies and thesauri* even though it addresses the same task, because it makes use of the links between categories and articles. A full list of papers, organized under these headings, is available in Appendix D.

Many papers treat Wikipedia as a glossary, or at least as a large corpus in which concept terms are associated with relevant text. That property alone is enough for it to be usefully applied to many tasks. This group also includes papers that use Wikipedia as a tagged corpus in which named entities are explicitly identified by the link markup. These efforts are classified under glossaries rather than thesauri, because the links are not used to indicate any form of semantic relation between source and target.

Most papers aspire to the right end of the spectrum, and consider Wikipedia to be a source of ontologies. There is a distinct valley between this last group and the one that treats Wikipedia as a corpus/glossary. Those who see Wikipedia as a source of structured knowledge generally leap ahead to deriving machine-readable relations from it. Comparatively few papers make direct use of the millions of manually defined human-readable relations it contains.

Almost all papers on the left of Figure 3.9 actively apply the knowledge they gather to some task—e.g. text categorization (Gabrilovich and Markovitch 2006), co-reference resolution (Strube and Ponzetto 2006), entity tagging (Mihalcea and Csomai 2007), etc. Those on the right do not. This imbalance does not imply that formal knowledge bases are inherently less useful than lesser structures. As Section 2.4 discussed, researchers have so far focused on knowledge extraction and representation, rather than putting the resulting structures to use. Given the scale of the resources that have been built recently—for example, Freebase contains more than 50M entities and more than 430M assertions—it may be time for that focus to shift.


Koru
Internet
Thermodynamics
E-mail
Noun
Synonym
Syntax
Questionnaire
Semantics
Polysemy
Concept
Technical terminology
Terminology
Information
Thesaurus
Ad hoc
Tf–idf
The New York Times
Information retrieval
Expert system
Hypothesis
Text corpus
Data structure
Probability
Computer science
Semantic relatedness
Web search engine
Psychology
Statistical significance
Correlation
Interactivity
Hyperlink
Wikipedia
Singapore
George W. Bush
American Airlines
Terrorism
United States
American Black Bear

4. Koru: a prototype search engine powered by Wikipedia

The previous two chapters exposed gaps in the way information seeking is currently supported, and showed how they might be addressed by exploiting external knowledge-bases in general, and by Wikipedia specifically. They explained our rationale for increasing the interactivity of search interfaces, and for using Wikipedia as a knowledge base to support this interaction.

This chapter describes a prototype system—called Koru—that was developed in 2006 to gain insight into supporting interactive retrieval, given the knowledge that could be obtained from Wikipedia without deep natural language processing. The first section of this chapter describes Koru's architecture, and broadly maps out a plan for applying Wikipedia to the retrieval process. Section 4.2 elaborates on this plan by describing Koru's interface, with specific examples of the kind of interaction it aims to promote. Section 4.3 describes the knowledge base that underpins the interaction, and explains how it was extracted automatically from Wikipedia. Section 4.4 gives examples of how Koru was used in practice by experimental subjects, and Section 4.5 evaluates the system by pitting it against traditional keyword search. The last section takes a critical look at both Koru and its evaluation. It identifies strengths and shortcomings that guide the remainder of the thesis.

4.1 Architecture

Figure 4.1 outlines Koru's architecture. The process begins at the right with a predefined collection of textual documents, such as a set of news articles or scientific papers. These documents are automatically cross-referenced with the relevant Wikipedia articles, using a procedure that is described in Section 4.3. Assuming that there is a reasonable overlap, the resulting structure captures the semantics of the document collection: the topics it discusses, the relations between them, and the terms that denote them. It closely resembles traditional thesauri, except that it is neither domain-specific nor domain-independent but instead customized to the particular documents being indexed.

Access to the Wikipedia-derived, collection-specific thesaurus affords many opportunities for improved interaction with the document collection. Compared to traditional full-text indexing, Koru exhibits an understanding of the topics the documents discuss, rather than the words they contain. Because Wikipedia excels in describing

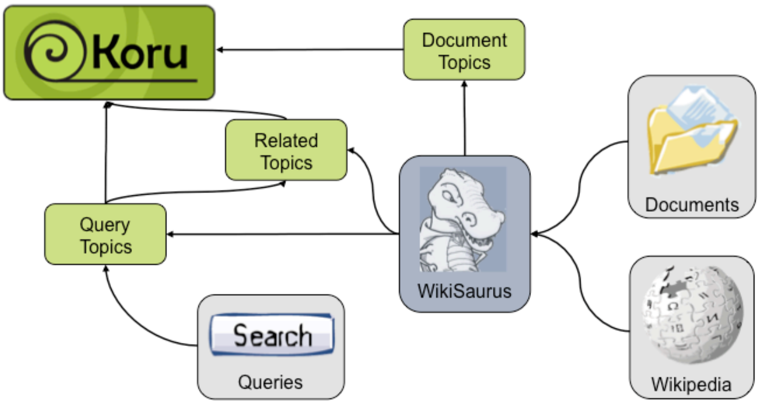


Figure 4.1: Architecture of a Wikipedia-powered search engine

contemporary concepts using contemporary language, queries that are valid for the collection are likely to be covered by the thesaurus, even when non-technical terminology or slang is used. This provides opportunities for improving how documents and queries are matched to each other automatically, or allowing users to navigate from one related query to the next as their information needs change (Section 2.3.1).

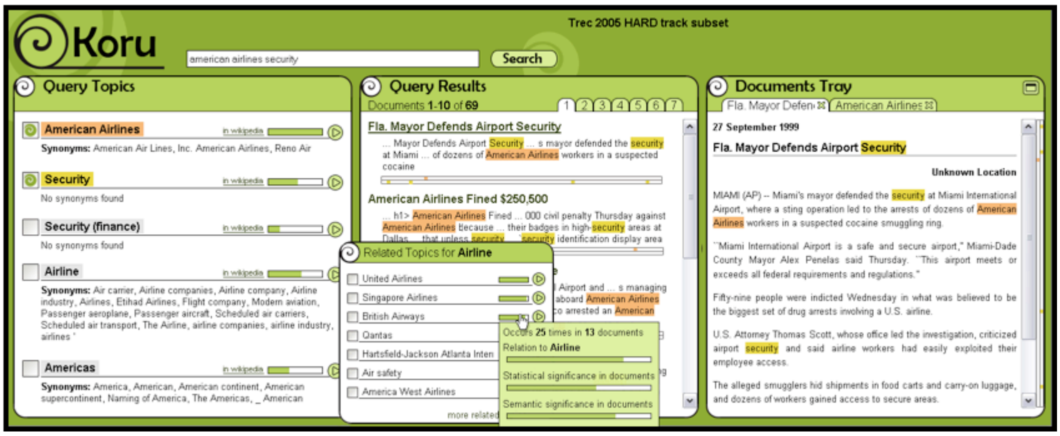
4.2 *Interface*

Koru's interface is implemented using the AJAX framework (Crane et al. 2005), or a combination of HTML, JavaScript and XML. Consequently, Koru's software requirements are minimal (the end user needs only a standard web browser) but the potential for smooth, uninterrupted interaction is great (the system can retrieve data—e.g., search results, documents, and query suggestions—at any point, without requiring the user to navigate to another page).

The Koru interface is illustrated in Figure 4.2. The upper area is a classic search box in which the user has entered the query *american airlines security*. Below are three panels: query topics, query results, and the document tray. What the figure does not convey is that to avoid clutter not all the panels are visible at any given time. There are three possible configurations, which relate to three stages of expected user behaviour:

1. Building an appropriate query. This involves adding and removing phrases until the query and corresponding results satisfies the user's information need. At this stage two panels are visible: query topics and query results (the leftmost two panels in Figure 4.2).

2. Browsing the document list. Once a suitable list of documents is returned, the user must determine the most relevant ones and judge whether they warrant further study. At this point the panels in Figure 4.2 slide across so that only the rightmost two—query results and document tray—are visible.



Synonyms: American Air Lines, Inc. American Airlines, Reno Air

No synonyms found

No synonyms found

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Fla. Mayor Defends Airport Security

... Mayor Defends Airport Security ... s mayor defended the security at Miami ... of dozens of American Airlines workers in a suspected cocaine

American Airlines Fined $250,500

... h1> American Airlines Fined ... 000 civil penalty Thursday against American Airlines because ... their badges in high-security areas at Dallas ... that unless security ... `security identification display area

Occurs 25 times in 13 documents

27 September 1999

Fla. Mayor Defends Airport Security

Unknown Location

MIAMI (AP) -- Miami's mayor defended the security at Miami International Airport, where a sting operation led to the arrests of dozens of American Airlines workers in a suspected cocaine smuggling ring.

``Miami International Airport is a safe and secure airport,'' Miami-Dade County Mayor Alex Penelas said Thursday. ``This airport meets or exceeds all federal requirements and regulations.''

Fifty-nine people were indicted Wednesday in what was believed to be the biggest set of drug arrests involving a U.S. airline.

U.S. Attorney Thomas Scott, whose office led the investigation, criticized airport security and said airline workers had easily exploited their employee access.

The alleged smugglers hid shipments in food carts and carry-on luggage, and dozens of workers gained access to secure areas.

Figure 4.2: Koru, with topics and articles related to *american airlines security*

3. *In-depth reading of a chosen document*. Having located a worthy document, the user then devotes time to actually reading the relevant sections. Here only the document tray is needed. It expands to fill the entire window, because anything else is a distraction.

4.2.1 The query topics panel

The first panel, query topics, provides users with a summary of their query and a base from which to evolve it. The panel lists every significant topic extracted from the query and assigns to each. The colour key is consistent throughout the interface. Query topics are identified without requiring any special syntax: in Figure 4.2 *American Airlines* has been identified as a single phrase even though the user did not surround it by quotes. Sophisticated entity extraction is unnecessary: words and word sequences are simply checked against a vocabulary of terms that have been previously extracted from Wikipedia.

In the event that terms cannot be matched to entries in the thesaurus interaction does not break down. Unmatched terms are still listed as topics and incorporated into the query. If the query contains overlapping phrases that each match a thesaurus term, the overlapping words are assigned to the topic that is most relevant to the document collection (see Section 4.3.3).

For the given query, consulting the Wikipedia-derived thesaurus yields the five topics *American Airlines*, *Security*, *Security (finance)*, *Airline* and *Americas*. The last is recognized because the thesaurus contains a use-for link from *America* to the preferred term *Americas*. Non-preferred synonyms for each term are listed below that term. For example, the topic *Airline*'s synonyms include *air carrier*, *airline company,* and *scheduled air transport*. These are used internally to improve queries (see Section 4.4) and presented to the user to help them understand the sense of the topic. The user can also learn more about a topic by clicking the adjacent link, which displays the relevant Wikipedia article in full.

Query terms are often ambiguous and relate to multiple entries in the knowledge base. By *security*, for example, the user could also mean property pledged as collateral for a loan, which appears in Figure 4.2 as *Security (finance)*. Each sense is included, and ranked according to the likelihood that it is a relevant, significant topic for the current query. This likelihood, displayed initially as a horizontal bar next to the topic and elaborated on in a tooltip, is calculated as a function o f the topic's statistical and semantic significance within the document collection. Section 4.3.3 explains how these weights are obtained.

Only the top-ranked topics that cover all the query terms—*American Airlines* and the first meaning of *security*—are used for retrieval, as indicated by the checkboxes to the left of the topics. This can be overridden manually. For example, it is useful for *Airlines* and *Americas* to appear separately—even though they are not included in Koru's default interpretation of the query—in case the user was interested in all airlines that operated in the U.S., rather than the specific company.

Each topic recognized in the query can be investigated in isolation by using it as a starting point for browsing the thesaurus. In Figure 4.2 the user has chosen to expand topics related to *Airline*. They have clicked the triangle to the right of that term, which brings up a menu of related topics. They can then investigate further topics of interest, such as *Singapore Airlines* and *British Airways*. Any of these topics could be incorporated into the query by clicking the appropriate checkbox. As with alternate senses, these topics are ranked according to their expected usefulness. Rankings are elaborated on in tooltips, as shown for *British Airways* in Figure 4.2. This is calculated in the same way as before, except that the strength of the relation to the parent topic—in this case, *Airline*—is also taken into consideration.

4.2.2 The query results panel

The second panel in Figure 4.2, query results, presents the outcome of the query as a series of document surrogates. The results are obtained by issuing the synonym-expanded query (see Section 4.4) to a Lucene index of the document collection with case folding and stemming enabled. The document surrogates resemble those found in typical search engines like Google, and consist of a title and a series of snippets that reflect the document's relationship to the query. Query topics, including synonyms, are highlighted within both titles and snippets for ease of identification.

The only unconventional addition is an overview of how topics are distributed throughout the document, which is presented graphically underneath each snippet using tilebars (Hearst 1995). These represent the entire content of the document as a horizontal bar from left (beginning of document) to right (end). Different bars relate to different query topics, in this case *American Airlines* (upper bar) and *Security* (lower bar). Colour-coded points appear along the bar to represent locations where the query terms and synonyms are found. These simple maps can give detailed insights into the relevance of a document. For example, it is apparent that *security* is relevant throughout the first document in Figure 4.2, but *American Airline*s is mentioned only once. That occurrence is close to a mention of *security*, so the document likely discusses the security of American Airlines, but only in passing. From this purely spatial information the user can make an informed decision about whether the document is worth opening.

4.2.3 The document tray

The third panel in Figure 4.2 shows the document tray, which allows the reader to collect documents they wish to peruse. More significantly, its purpose is to facilitate efficient reading by helping users identify relevant sections of a document and navigate between them. These sections are identified using the same information that made the document itself relevant: the query terms used to locate it. Term occurrences are easily seen because they are highlighted according to the colours defined in the query topics panel. Interesting patterns of highlights are likely to indicate sections and paragraphs that should be read.

These highlights can easily be missed, however, because most documents are too large to be viewed without scrolling. Consequently tilebars are again supplied to provide an overview of how terms are distributed throughout the document. These tilebars are oriented vertically, and appear on the right-hand side of the standard scrollbar, with a direct mapping to it (they look rather thin in Figure 4.2). If the scrollbar slider is moved alongside a cluster of points in the tilebar, the highlights that these points represent are visible in the document. Users can jump directly to a particular highlight by clicking the appropriate spot in the tilebar.

4.3 Extracting a relevant thesaurus

To support a collection of documents, Koru uses Wikipedia to provide a comprehensive thesaurus that is customized to that collection. As explained in detail in Section 3.3.1, Wikipedia is nicely segmented into individual concepts and the terms that can denote them. Each Wikipedia article serves to describe a single concept, and its title serves as a succinct descriptor. Non-descriptors (or synonyms) are provided by “redirects”: pages that exist only to connect an alternative title of an article with the preferred one.[27] All that remains for building a corpus-specific thesaurus is to create links that allow navigation between related concepts, and to identify the subset of concepts that are relevant to the documents at hand. These tasks are tackled in Sections 4.3.1 and 4.3.2 respectively.

4.3.1 Identifying relations between concepts

Wikipedia defines an extensive network of categories that encode hierarchical relations, and millions of hyperlinks between articles that correspond to flat relations. These relations are needed to provide Koru's exploratory search facilities.

Unfortunately the connections in Wikipedia do not map accurately to those in traditional thesauri. Referring back to Section 3.5.3, categories yield BT/NT relations with only 16% precision and relations obtained from article hyperlinks are even worse. Consequently, Koru gathers all relations from article and category links and assigns weight so that only the strongest are emphasized. Moreover, hierarchical and flat relations are not cleanly separated as the structure would suggest, but are intermingled in both category and article links. This is why the Koru interface merely identifies related topics without attempting to specify the nature of the relationship.

Weighing or quantifying relations between topics is a well-established task, known in the literature as measuring semantic relatedness. In Koru, the strength of a relation between two Wikipedia articles is quantified by comparing the links found within them. If articles *a* and *b* both contain a link to a third article *c*, this suggests that they are related in some way. The strength of the relation between *a* and *b* depends on the number of links they have in common, and also on the popularity of the link targets. The fact that two articles both link to *science*, for example, is much less informative than if they both link to the specific topic *atmospheric thermodynamics*. The rarity *r* of a link target *c* is defined as:

$$r(c) = \log\left(\frac{|W|}{|C|}\right)$$

where *C* is the set of all articles that link to *c* and *W* is the set of all articles in Wikipedia. This produces larger scores for targets that are rarely linked to.

The semantic relatedness *sr* of two articles *a* and *b* is defined as:

[27] As explained in Section 3.3.4, link anchors encode the same language phenomenon as redirects. Unfortunately, we did not think to use them at the time this prototype of Koru was developed.

$$sr(a,b) = \sum_{x \in (A \cap B)} r(x)$$

where *A* is the set of all articles that *a* links to, and *B* is the set of all articles that *b* links to. In other words, we gather the links *a* and *b* make and retain the targets that are common to both sources. The relatedness score is 0 if there are no common targets. Otherwise, the relatedness score is the sum of the rarity scores of the common targets.

4.3.2 Disambiguating unrestricted text

One of the major problems with using Wikipedia as a knowledge base is that the vast majority of it will be irrelevant for any given collection of documents. Unless these are somehow culled, users will be swamped with extraneous topics and suggestions.

To identify the relevant concepts, each document is processed during the initial indexing stage to identify significant terms and match them to individual Wikipedia articles. To lift terms from their surrounding prose, the text is parsed to identify nouns and noun phrases using the Stanford Parser (Klein and Manning 2003). Article titles and redirects are preprocessed to remove disambiguation notes—e.g. *Plane (mathematics)* and *Plane (woodworking tool)* both revert to *Plane*—and matched to the noun phrases via stemming (Porter 1980), case-folding and stop-word removal.

Whenever a document term or phrase matches to a single article title or redirect, the corresponding concept is added to the thesaurus. When terms match multiple titles or to titles of disambiguation pages—as in the example above—then disambiguation must be performed. The scale of Wikipedia makes disambiguation crucial. For example, the term *Jackson* covers more than 50 different locations and more than 100 different people. If all senses were included in the thesaurus, it would become bloated and unfocused.

Each ambiguous document term is resolved by selecting the sense that relates most strongly to the topics detected around it. The contextual relatedness *cr* of a sense *s* is given by:

$$cr(s) = \sum_{c \in C} sr(s,c)$$

where *C* is the set of all contextual topics detected in the surrounding text and *sr* is the semantic relatedness formula defined in the previous section. Only unambiguous terms are used as context, to avoid making this a circular problem. Initially only topics detected within the surrounding sentence are used, and the sense with the strongest weight is added to the thesaurus. If there are no unambiguous terms to compare against, or if the weights of the top senses are within 10% of each other, then a cascading approach is used. If a sentence contains insufficient information to disambiguate a term, the entire surrounding paragraph is used as context, and if the paragraph contains insufficient context the entire document is used. Sentence and paragraph detection is performed using the Stanford parser. It is rare that a term remains ambiguous at the document level, but if so, all top candidate senses—those whose weights are within 10% of each other—are included in the thesaurus.

4.3.3 Weighting topics, occurrences and relations

The association between a topic and a document in which it is found is weighted within the thesaurus, so it can be determined whether a document is largely about a topic or merely mentions it in passing. Two separate weights are used to determine this.

The first weight is TF×IDF, which is based on the assumption that a significant topic for a document should both occur many times within it and be useful in distinguishing the document from others. For a topic *t* in a document *d*, this is defined as:

$$TF \times IDF = \frac{freq(t,d)}{\sum_i freq(t_i,d)} \times \log\left(\frac{|D|}{|D_t|}\right)$$

where *freq(t,d)* is the occurrence count of all terms in *d* that have been resolved to *t*, *D* is the set of all documents, and D_t is the set of all documents containing *t*. The first component in the expression corresponds to term frequency: the *TF* part of *TF×IDF*. It normalizes the occurrences of *t* in *d* by the occurrences of all terms in *d*. The second component corresponds to inverse document frequency (*IDF*). It is larger for topics that occur in fewer documents.

The second weight is simply the average semantic relatedness of a topic to all others identified within the document. It is based on the assumption that a significant topic should relate strongly to other topics in the document.

The weights are averaged across all documents to determine the significance of topics within the collection as a whole. These aggregated weights are used in several places within the Koru interface shown in Figure 4.1. The senses in the query panel, for example, are ranked by their combined TF×IDF and average relatedness scores, and this allowed *American Airlines* to be automatically selected over *America* and *Airlines*, and *Security* over *Security (finance)*. The related topics shown for *Airlines* in Figure 4.2 are ranked—as the tooltip elaborates on for *British Airways*—by a combination of their relatedness to the query topic, and corpus-wide TF×IDF and average relatedness scores.

4.4 Koru in action

Koru was evaluated with a user study in which participants performed tasks for which the relevant documents had been manually identified. The tasks, documents and relevance judgments were obtained from the 2005 TREC HARD track (Allan 2005), which pits retrieval techniques against each other on the task of high-performance retrieval through user interaction. The tasks were specifically engineered to encourage a high degree of interaction.

To give a flavour of Koru in action, Table 4.1 shows three of the TREC tasks, along with information about the initial querying behaviour of a few different users for each task.

Example 1: Black Bear Attacks

It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

User 1 Query: *black bears humans*
Topics: *American Black Bear*, *Human*
Query issued: (*American Black Bear* OR *Black Bear* OR *Ursus americanus*)
AND (*Human* OR *All Humankind* OR *Everybody* OR *Homo Sapien* OR
Human Being OR *Human Kind OR Human species* OR *Humanity* OR *Man*)

User 2 Query: *black bear man*
Same results as above

User 3 Query: *black bear behaviour*
Topics: *American Black Bear, Behaviour*
Query issued: (*American Black Bear* OR *Black Bear* OR *Ursus americanus*)
AND (*Behavioural, Behaviours, Behaviour, Behavioural, Behaviours*)

Example 2: Email Abuse

The availability of E-mail to many people through their job or school affiliation has allowed for many efficiencies in communications but also has provided the opportunity for abuses. What steps have been taken by those bearing the cost of E-mail to prevent excesses?

User 1 Query: *email abuse*
Topics: *E-mail, Abuse*
Query issued: (*E Mail* OR *E-Mail* OR *Electronic Mail* OR *E-mail account* OR *Internet mail*)
AND (*Abuse* OR *Abused* OR *Abusive* OR *Maltreatment* OR *Mistreatment*)

User 2 Query: *email abuse employees*
Topics: *E-mail, Abuse, Employment*
Query issued: *as above AND*
(*Employment* OR *Employ* OR *Employee* OR *Employer* OR *Job*)

Example 3: Hubble Telescope

Identify positive accomplishments of the Hubble telescope since it was launched in 1991. Documents are relevant that show the Hubble telescope has produced new data, better quality data than previously available, data that has increased human knowledge of the universe, or data that has led to disproving previously existing theories or hypotheses. <further qualifications omitted>.

User 1 Query: *Hubble telescope achievements*
Topics: *Hubble Space Telescope* (*achievements* not in thesaurus)
Query issued: (*Hubble Space Telescope* OR *Hubble Telescope*) AND *achievements*

User 2 Query: *Hubble telescope universe expansion*
Topics: *Hubble Space Telescope, Universe, Hubble's law*
Query issued: (*Hubble Space Telescope* OR *Hubble Telescope*) AND *Universe*
AND (*Hubble's law* OR *Cosmological redshift* OR *Expansion of space* OR
Expansion of the Universe OR *Hubble Flow* OR *Expansion*)

Table 4.1: Example retrieval tasks, queries, and topics identified

These tasks require the user to think carefully about their query terms, and are unlikely to be satisfied by a single query or document.

The TREC tasks are paired with the AQUAINT text corpus, a collection of newswire stories from the Xinhua News Service, the New York Times News Service, and the Associated Press Worldstream News Service. The thesaurus that was used throughout was generated using the method described in Section 4.3.

In the first example in Table 4.1, User 1 types the query *black bear humans*. Koru identifies four topics: *American Black Bear*, *Human*, *Bear*, and *Black (people)* (only the first two are shown in the table). The first two cover all terms in the query, and are checked by default in the interface. The query that Koru issues to the back-end search engine contains two clauses AND'd together (one for each topic). The first has three OR'd components and the second has nine, corresponding to synonyms of the topics. Koru places each of these components between quotation marks before passing them to the Lucene search engine, so that they are treated as phrases. This generates a fairly complex query, such as a librarian might issue, from the user's simple three-word input—including some non-obvious synonyms.

User 2 types *black bear man*, which yields precisely the same results. User 3 types *black bear behaviour*, which yields a different query. Notice incidentally how Koru caters for spelling variants and plural forms. Many related topics can be obtained by clicking beside each search topic (as for *Airline* in Figure 4.2). Examples are *Alaska* and *West Virginia* for the topic *American Black Bear*, *Civilization* for *Human*, and *Psychology* and *Brain* for *Behaviour*.

The second example in Table 4.1 concerns email abuse. User 1 simply types these two words as the initial query. Each of these terms is recognized as a topic, and behind the scenes Koru automatically expands them to embrace synonyms and alternate forms. User 2 adds the word *employees,* which is also recognized as a topic in itself, resulting in a lengthy 3-term query.

In the third example, which is about the Hubble telescope, User 1 types *Hubble telescope achievements*. The first two words are identified as the topic *Hubble Space Telescope*; the word *achievements* is not recognized as a topic at all because it does not appear as a term in the thesaurus. Nevertheless it is still added to the query, along with the expansions of the first topic. User 2 introduces *universe expansion* into the query. Quite fortuitously, the word *expansion* is related in the thesaurus to *Hubble's law* because Wikipedia redirects it to that article: no other senses of *expansion* made it into the thesaurus.

4.5 Evaluation

This section describes an evaluation of how well Koru and its underlying data structure facilitates and improves information retrieval. Of particular interest is whether the topics, terminology and semantics extracted from Wikipedia make a conclusive, positive difference in the way users locate information, which is measured by pitting the new knowledge-based topic browsing technique against traditional keyword search. Koru's usability is also of interest: whether it allows users to apply the knowledge found in Wikipedia to their retrieval process easily, effectively and efficiently. This is assessed by observing participants closely as they interact with the system to perform the tasks.

4.5.1 Evaluation procedure

To provide a baseline for comparison, a second version of Koru was created that provides as much of the same functionality as possible without using a thesaurus, and whose interface is otherwise identical. This allows a clean comparison of the new system with traditional keyword search. The baseline system simply omits the query topics panel in Figure 4.2. Tilebars were omitted from both systems to further reduce interference in the comparison. While they can be of assistance in both topic browsing and keyword searching, they are not a fundamental component of either. The Wikipedia links placed beside each topic were also omitted, to focus participants on using Koru rather than browsing an external knowledge source.

Subjects. Twelve participants were observed as they interacted with the two systems. All were experienced knowledge seekers. All were graduate or undergraduate computer scientists with at least 8 years of computing experience who use Google and other search engines daily. Sessions typically lasted for 1½ hours, and were conducted in a controlled environment with video and audio recording and an observer present. Data was also collected from questionnaires and system logs.

Each user was required to perform ten tasks (of which Table 4.1 shows three) by gathering the documents they felt were relevant. Half the users performed five tasks using Koru in one session and the remaining five using the traditional search interface in a second session; for the other half the order was reversed to counter the effects of bias and transfer learning. For each task, approximately 750 relevance judgments are provided by TREC in which a document is identified as strongly relevant, weakly relevant, or irrelevant.

Document collection. The ACQUAINT text corpus that was used for the experiments is large—about 3GB uncompressed. It was impractical to create a thesaurus for the entire

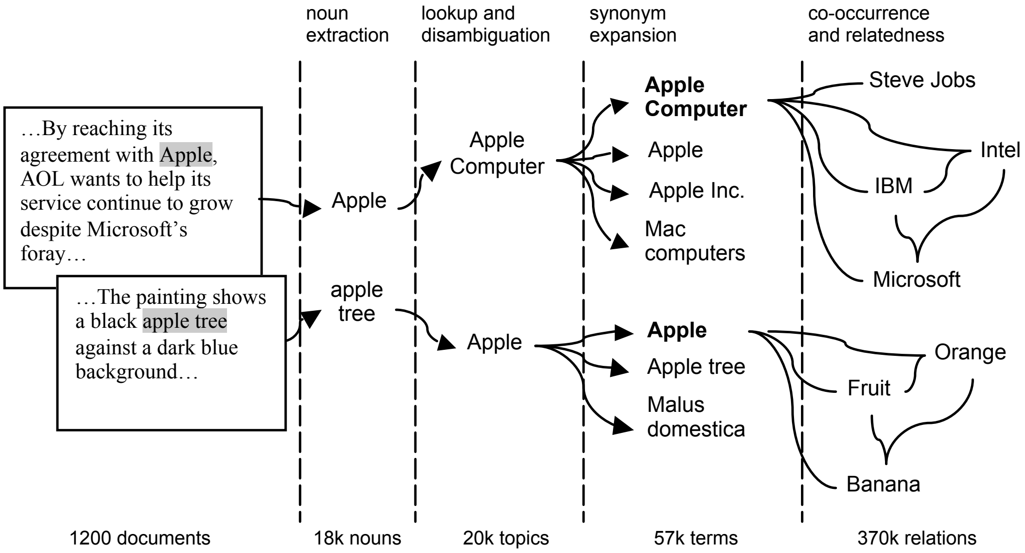


Figure 4.3: Excerpt and statistics from the automatically constructed thesaurus

collection because the process of detecting and disambiguating Wikipedia topics within the documents was not optimized. Instead a subset was selected: only stories from Associated Press that were mentioned in the relevance judgments for the 10 tasks. The result is a collection of approximately 1200 documents concerning a range of topics. This was used throughout the experiments.

Thesaurus. A thesaurus was created automatically for this document collection using a snapshot of Wikipedia released on June 3, 2006. This dump contains approximately 1M articles and a further 1M redirects, or a total vocabulary of 2M terms against which document phrases can be matched. Figure 4.3 illustrates the thesaurus building process, with examples and quantities of the documents, terms, topics, and relations that were involved. 1,200 documents were processed, and approximately 18,000 distinct nouns were identified that could be matched to at least one article in Wikipedia. The topics they match to are candidates for inclusion in the thesaurus. The final thesaurus contains about 20,000 distinct topics, because some of the terms still correspond to multiple topics even after the disambiguation algorithms were invoked. This residual ambiguity is understandable. Documents in the collection used to derive the thesaurus are not restricted to any particular domain, so terms may well have several valid senses. As an example, one of the news stories talks of *Apple Corporation's* business dealings while another mentions Piet Mondrian's painting of an *apple* tree.

The full vocabulary of the thesaurus (57,000) is almost three times larger than the number of topics (20,000). Some of this polysemy is expressed by different terms within the document collection itself: e.g. one document talks of *President Bush* and also mentions

	Keyword searching	Topic browsing
Recall	43.4%	51.5%
Precision	10.2%	11.6%
F-measure	13.2%	17.3%

Table 4.2: Performance of tasks

George W. Bush. More synonyms were obtained from Wikipedia redirects: e.g., Wikipedia adds the colloquialisms *Dubya, Shubya* and *Baby Bush* even though these are never mentioned in the (relatively formal) documents. In this context polysemy is desirable, because it increases the chance of query terms being matched to topics and increases the extent to which they are automatically expanded.

The thesaurus is a richly connected structure, with 370,000 relations. These are pairs of topics that are found together within at least one document, and have non-zero semantic relatedness between them. By this measure each topic relates—with varying degrees—to an average of 18 others. As a comparison, the Agrovoc thesaurus (discussed in Section 2.2.2) is of comparable size and contains just over two relations per topic on average.

4.5.2 Results

The two systems, Koru and the traditional interface, were compared on the basis of overall task performance, detailed query behaviour, and questionnaires that users filled out. In the discussion below, Koru is referred to as *topic browsing* and the traditional interface as *keyword searching* because this characterizes the essential difference between the two. Koru identifies topics based on the user's query and encourages browsing; the traditional interface provides plain keyword searching.

Task performance: The first question is whether the knowledge base provided by Wikipedia is relevant and accurate enough to make a perceptible difference to the retrieval process. The most direct measure of this is whether users perform their assigned tasks better when given access to the knowledge-based system. Examination of the documents encountered during the retrieval experience shows that this is certainly the case. Table 4.2 records gains in the recall, precision, and F-measure, averaged over all documents encountered using the topic browsing system. This means that the new interface returned more appropriate documents than the traditional one.

The greatest gains are made in recall: the proportion of available relevant documents that the system returned. This can be most directly attributed to the automatic expansion of queries to include synonyms. Typically, gains made in recall are offset by a drop in precision, because the inclusion of more terms causes more irrelevant documents to be returned. This was not the case. Table 4.2 shows no decrease in precision, which attests to the high quality of the Wikipedia redirects from which the additional terms were obtained. Indeed there is even a slight gain. This could plausibly be attributed to recognition of multi-word terms, which users of traditional interfaces are supposed to encase within quotes. The participants were all experienced Googlers and were consistently reminded of this syntax when familiarizing themselves with the keyword search interface. Despite this, they did not once use quotation marks, even though they would have been appropriate in 53% of the queries that were issued. The new system performs this often-overlooked task reliably and automatically.

Successful topic browsing depends on query terms being matched to entries in the knowledge base. This is typically a bottleneck when using manually defined structures. It is difficult to obtain an appropriate thesaurus to suit an arbitrary document collection, and any particular thesaurus is unlikely to include all topics that might be searched for. Furthermore, specialist thesauri adopt focused, technical vocabularies, and are unlikely to speak the same language as people who are not experts in the domain— the very ones who require most assistance when searching. Koru does not seem to suffer the same problems. For 95% of the queries issued it was able to match all terms in the query (the term *achievements* in Example 3 of Table 4.1 is a typical exception).

4.5.3 Query behaviour

The TREC tasks were specifically selected to encourage user interaction, and participants were invariably forced to issue several queries in order to perform each one. There were significant differences in query behaviour between the two systems.

One major difference was the number of queries issued: 338 distinct queries on the topic browsing system vs. 274 for keyword searching. This did not correlate to an increase in time spent using Koru, despite its unfamiliarity and greater complexity. Participants were always encouraged to spend 5 minutes on each task regardless of the system used. There are two possible reasons for the increase: Koru either encourages more queries by making their entry more efficient, or requires more queries because they are individually less effective.

Figure 4.4 indicates that the additional queries are being issued out of convenience rather than necessity. Queries issued by all participants are divided into two groups, one for each interface. Then each group was sorted by F-measure, and the F-measure plotted against rank. The figure shows that for both topic browsing and keyword searching the best queries had the same F-measure—in other words, the best queries are equally good on both systems. As rank increases a difference soon emerges, however: the performance

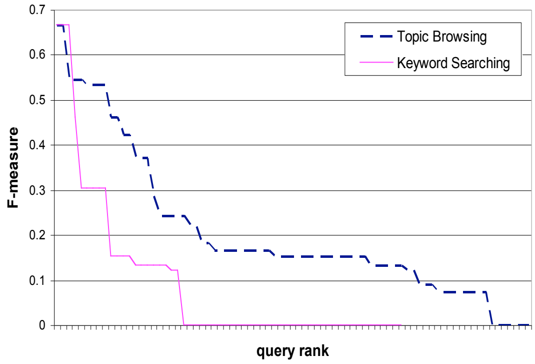


Figure 4.4: Performance of individual queries in order of rank

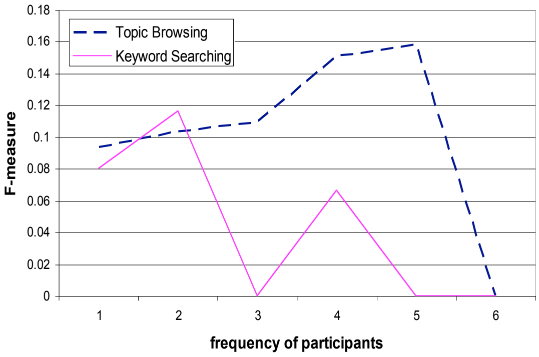


Figure 4.5: Performance of queries grouped by participant frequency

of keyword searches degrades much more sharply than topic-based ones. In general, the *n*th best query issued when topic browsing is appreciably better (on average) than the *n*th best query issued when keyword searching, for any value of *n*.

This clearly shows that the additional queries issued using Koru are not compensating for any deficiency in performance—for Koru's performance is uniformly better. Instead, it probably reflects the way in which Koru presents the individual topics that make up queries. These are automatically identified and presented to the user, and can be included or excluded from the query with a click of the appropriate checkbox. Each click issues a new query, whether or not that is the user's intention.

Several participants modified their search behaviour to take advantage of this feature. They initially issued large, overly specific queries and then systematically selected combinations of the individual terms that were identified. To illustrate this, suppose a user issued a query similar to that in Figure 4.2 (*american airlines security*) but with additional terms related to security such as *baggage check, terrorism,* and *x-rays*. This is a poor initial query because few documents will satisfy all topics. But it forms a base for several excellent queries (e.g. *baggage check* and *terrorism,* or *baggage check* and *x-rays*), which in Koru can be issued with a few mouse clicks.

The ability to quickly reformulate queries was greatly appreciated by participants; just under half listed it as one of their favourite features. The only way to emulate this behaviour manually in the traditional interface is either by time-consuming reformulation (hence fewer queries issued) or by using Boolean syntax (which even the expert Googlers in this study avoided).

The next point to investigate is whether it is easier for users to arrive at effective queries when assisted by the knowledge-based approach. To do so, Figure 4.5 plots the average F-measure of queries against the number of participants that issued them. At the left are

	Topic	Keyword	Neither
Relevance and usefulness	75%	25%	0%
Ease of navigation	8%	67%	25%
Clarity of structure	42%	42%	16%
Clarity of content	8%	42%	50%
Overall preferred	67%	33%	0%

Table 4.3: Comparative questionnaire responses

queries issued by only one participant; at the right are ones issued by five and six participants. For the sake of clarity, one of the tasks has been discarded because appropriate query terms were particularly easy to obtain. A good query issued by many participants is a matter of common sense, whereas one issued by a lone individual is likely to be a product of expert knowledge or some nugget of encountered information. For topic-based queries, performance climbs as they become more common—in other words common queries perform better on average than idiosyncratic ones. This is reversed for keyword searching. Participants were able to arrive at effective queries much more consistently when Koru lent a hand.

The gains are almost exclusively due to automatic query expansion and topic identification. Koru also enables interactive browsing of the topic hierarchy, but participants rarely bothered to use this facility—and even more rarely did such browsing yield additional query topics. In part this was due to users being put off by inaccuracy in the relations that were offered, but typically users felt that the relations were accurate. A more fundamental problem is that even topics that are closely related to a query topic are often irrelevant to the query as a whole. Consider the second example of Table 4.1, for which most participants issued the query *email abuse.* Most of the related topics Koru presented for *email* (*browsers, internet, AOL,* etc) and *abuse* (*rape, child abuse, torture,* etc) are perfectly valid but completely irrelevant to the task.

4.5.4 Questionnaire responses

Each participant completed three separate questionnaires, which solicit their subjective impressions of the two systems. After each session they completed a questionnaire that asked for their impressions of the interface used in that session (Koru or the traditional interface). The third questionnaire was completed at the conclusion of the second session and asked for a direct comparison between the two interfaces, to compare topic browsing and keyword searching directly.

The results of the final questionnaire are shown in Table 4.3 (e.g. the first row corresponds to the question *which of the two systems was more relevant and useful to your needs?).* The final question asked participants to name their preferred system overall: two-thirds chose the topic browsing system. Other questions indicate that the main reason for this was relevance and usefulness: in other words the additional functionality that Koru offers is relevant to user needs and produces useful results for their queries. In the words of one participant:

> *The (topic browsing) system provides more choices for users to search for information or documents they need.*

This was somewhat offset by Koru's additional complexity; unsurprisingly, participants felt that the simpler, more familiar system was easier to navigate and use. Simplicity was the reason cited by all participants who chose keyword searching over topic browsing. Several participants took pains to indicate that the difference was marginal. There was no mention of Koru being cumbersome or confusing, just more complex.

> *Not much navigation required (for keyword searching). Topic browsing was very easy to navigate as well.*
>
> *(Keyword searching is) more minimal. I didn't use the topic browsing stuff anyway.*

The last remark alludes to Koru's presentation of related topics. This feature was barely used and needs substantial revision. Many participants found it promising however, and two went so far as to list it as their favourite feature.

> *The three different parts (topics, list of articles, one article) are very easy to understand and easy to use. Only the related topics are not so easy to find.*

The sliding three-panel layout was found to be useful and easy to understand. The remainder of the topic browsing system appeared ergonomic and intuitive for users: there were no other frustrations sited in the surveys and almost all users discovered Koru's full range of features without instruction.

4.6 Discussion

One of the key hypotheses of this thesis is that Wikipedia can provide a knowledge base that is well suited to open-domain information retrieval. This study certainly supports that hypothesis, given that almost all of the queries issued in Koru were matched to thesaurus entries. It seems that Wikipedia contains the topics discussed within these documents even though they belong to no particular domain, and describes them in terminology that is familiar to the study participants.

Deriving the knowledge base from Wikipedia was a complicated process that is only partially evaluated in this study. The comparisons between Wikipedia and Agrovoc in Section 3.5.3 provide enough evidence that Wikipedia's terminology and synonymy relations can be used directly, but leaves two significant problems to be investigated: gathering relations between topics, and identifying when they are discussed within documents. This chapter sketched out approaches for both problems, but gives little insight into how these algorithms performed.

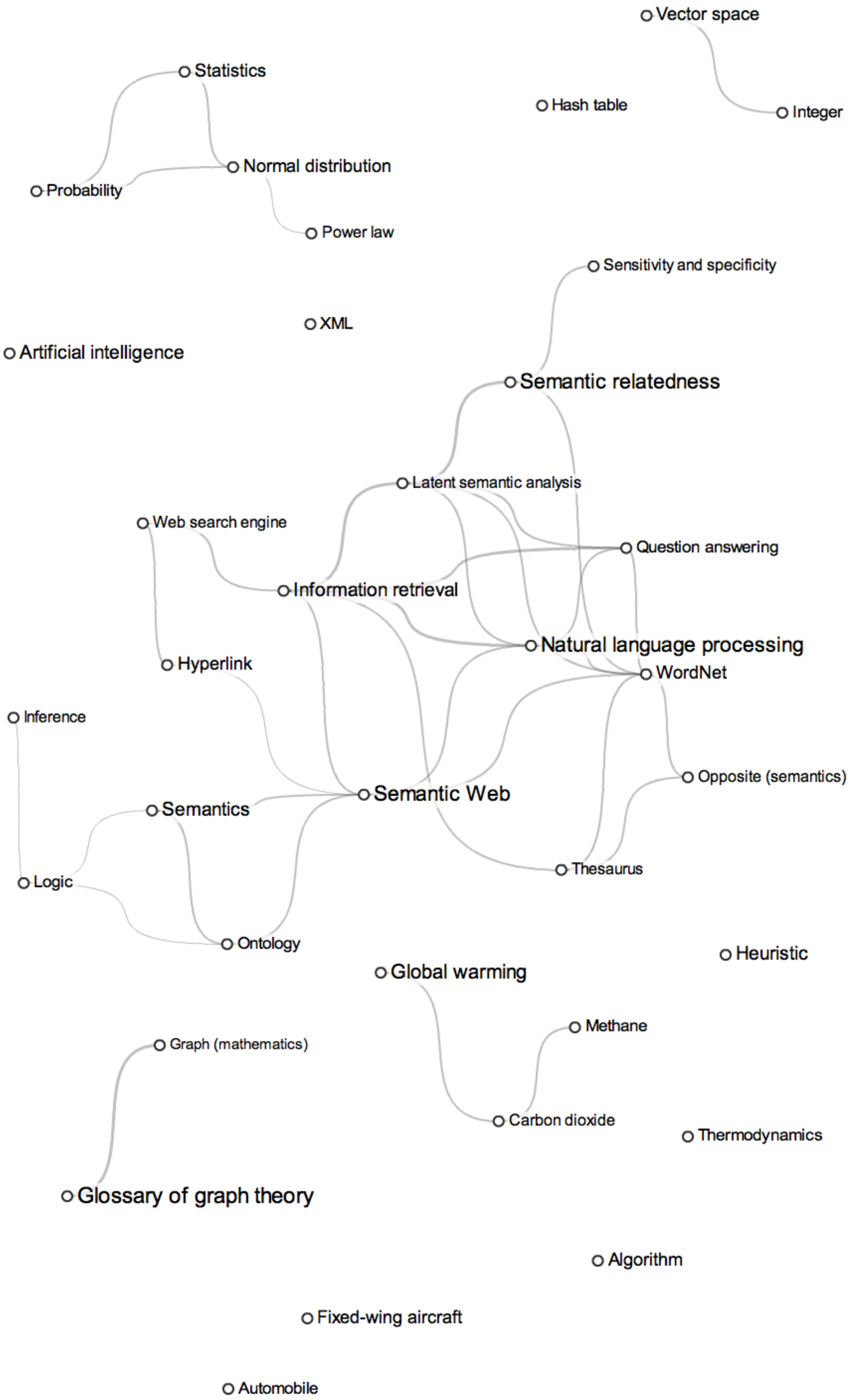
The task of gathering semantic relations from Wikipedia is an extremely complex one; a survey by Medelyan et al. (2009) lists more than 40 papers on the topic. Here it is simplified by only measuring the strengths of relations rather than deciding explicitly what—and how—topics are related. Even so, the approach presented is rather ad hoc. It does not draw on related work, and has not received direct evaluation. Consequently there is little to say about how it performed, what its shortcomings are, and to what extent these contribute to Koru's poor support for interactive query expansion. The problem of measuring relatedness with Wikipedia is explored more thoroughly in Chapter 5.

Detecting topics within documents is an even more complicated problem. This chapter has presented only a sketchy, ad hoc approach to the task. It involves entity extraction, disambiguation, link detection and topic indexing, for which there is much related work that has not been drawn on. Does the presented approach extract the correct terms and disambiguate them to the correct articles? Do the weights given to articles accurately represent their importance within individual documents and the collection as a whole? Chapter 6 describes more sophisticated approaches to detection, disambiguation and weight assignment, and evaluates them much more directly.

Turning now to the interface, it is encouraging that participants were able to use the system and discover its features without prompting. The sliding three panel layout was a success; participants found it intuitive, despite its unfamiliarity. There was no excessive switching, which indicates that the three views provided by this layout correspond well to the workflow of users. It was also encouraging to see users alter their search strategies significantly to take advantage of the topic panel, effortlessly issuing more queries and gathering better documents as a result.

The automatic expansion of queries was worthwhile, but not particularly significant. The study compares a somewhat unsophisticated approach against entirely unassisted retrieval, and ignores decades of research on automatically augmenting queries. Automatic query expansion will not be pursued further, despite the gains reported here.

Instead the focus is on improving the interactivity of the search engine to allow system and user to work together more effectively. It is therefore particularly disappointing that Koru's exploratory search facilities were barely used. There are many possible reasons. Were the relations expressed in the thesaurus too inaccurate to be useful? Is there a mismatch between concepts that are related to query terms (the relations expressed in the thesaurus), and those related to the retrieval task at hand? Could the suggestions be better presented or better explained, or are there more effective ways for users to incorporate them into their search processes? Did the evaluation procedure discourage exploration, or render it unnecessary? These questions are addressed in Chapter 7.


Vector space
Statistics
Hash table
Integer
Normal distribution
Probability
Power law
Sensitivity and specificity
XML
Artificial intelligence
Semantic relatedness
Latent semantic analysis
Web search engine
Question answering
Information retrieval
Natural language processing
Hyperlink
WordNet
Inference
Opposite (semantics)
Semantics
Semantic Web
Logic
Thesaurus
Ontology
Heuristic
Global warming
Methane
Graph (mathematics)
Carbon dioxide
Thermodynamics
Glossary of graph theory
Algorithm
Fixed-wing aircraft
Automobile

5. Quantifying relatedness

This thesis is about how Wikipedia can be applied to information retrieval without deep natural language processing or artificial intelligence. Chapter 3 explained many useful elements that could be extracted easily from Wikipedia, but one glaring omission was a structured, machine-readable representation of how the articles—and the concepts they represent—relate to each other. Although articles are organized hierarchically and cross-reference each other extensively, the links they make do not translate cleanly to the typed relations found within thesauri and ontologies. Although infoboxes express exactly the kinds of relations required for ontologies, their use is far from consistent or comprehensive. Most of the research surrounding Wikipedia is concerned with the complex problems of extending the infobox structure and turning haphazard links and categories into structured knowledge (Section 3.4.4).

This chapter investigates a simpler task. It aims to measure semantic relations as quantities rather than explicitly defining what—and how—topics are related. For example, consider the relation between *Wikipedia* and *encyclopaedia*. Is Wikipedia a legitimate encyclopaedia? As the debate between Nature and Encyclopaedia Britannica (Section 3.2.3) illustrates, this is a prickly question. It is difficult for a person to answer, let alone an algorithm. The algorithms described in this chapter sidestep the issue—and many other difficulties in natural language processing—by defining relations as numbers: if 0% means no relation, and 100% means synonymous, then Wikipedia and encyclopaedia might reasonably—and less controversially—be defined as 80% related.

The remainder of this chapter is structured as follows. The first section explains our interest in relatedness measures by hypothesizing about the kinds of knowledge bases and retrieval systems they can provide. Section 5.2 investigates how people quantify relatedness manually, while Section 5.3 surveys related work in automatically generating relatedness measures. Section 5.4 describes WLM: our own approach for generating relatedness measures with Wikipedia. This is evaluated and compared against alternative approaches in Section 5.5. The chapter concludes with a brief discussion of the strengths and weaknesses of the new approach.

5.1 Relatedness and information retrieval

One of the key ideas of this thesis is that semantic relatedness measures can provide effective knowledge for information retrieval. This may sound like a step backwards

- **Global warming** is the increase in the average temperature of the atmospheric and oceanic temperatures.
- Atmospheric and oceanic temperatures are largely driven by the **greenhouse effect**, which responds to varying concentrations of **greenhouse gasses**.
- **Greenhouse gases** include **Carbon Dioxide** (9-26%) and **Methane** (4-9%).
- Concentrations of both **Carbon Dioxide** and **Methane** have significantly increased in recent years.
- **Carbon Dioxide** is largely produced by burning **fossil fuels.**
- **Methane** is largely produced during the refinement of **fossil fuels** and by **agriculture.**
- **Fossil fuels** are primarily used for **power generation**, followed by **transportation**.

Figure 5.1: Some key facts surrounding *global warming*

compared to the vision of Semantic Web (discussed in Section 2.2.4). The only knowledge base that relatedness measures can possibly provide is a weighted graph of concepts and relations. The RDF framework (Manola et al. 2004) is much more likely to yield search engines and retrieval agents that think for themselves.

To illustrate why semantic relatedness measures are relevant, let us consider the question, *what are the causes of global warming*? Figure 5.1 shows some of the pertinent facts. Consider the extent to which retrieval systems need to understand these facts to provide a reasonable answer.

At one extreme is the ambitious end-goal of question answering, retrieval agents and the Semantic Web, where the machine takes responsibility for understanding the available information. Try to imagine hand crafting an ontology that captures the facts in Figure 5.1 to such a level of specificity that the entire argument—one that places blame on power generation, transportation and agriculture—could be inferred automatically. It would require a substantial effort, and this is an incomplete answer to just one question among many. Many of the relations involved are as open to debate as the connection between *Wikipedia* and *encyclopaedia*. Imagine the sophistication involved in automatically building the knowledge base, inferring the answer from it, and constructing a concise, cogent response. The state of the art in artificial intelligence and natural language processing is simply not adaptable, robust or knowledgeable enough to answer such complex, open-ended questions, unless it puts severe restrictions on the domain and the questions that can be asked. It may not be for decades.

At the other extreme is the kind of retrieval system that is common today. Current search engines make it easy to locate documents that mention *global warming* together with *cause*, but leave it to the user to sift through them. The work that is left—synthesis, sense making and all of the other processes discussed in Section 2.1—will be significant unless

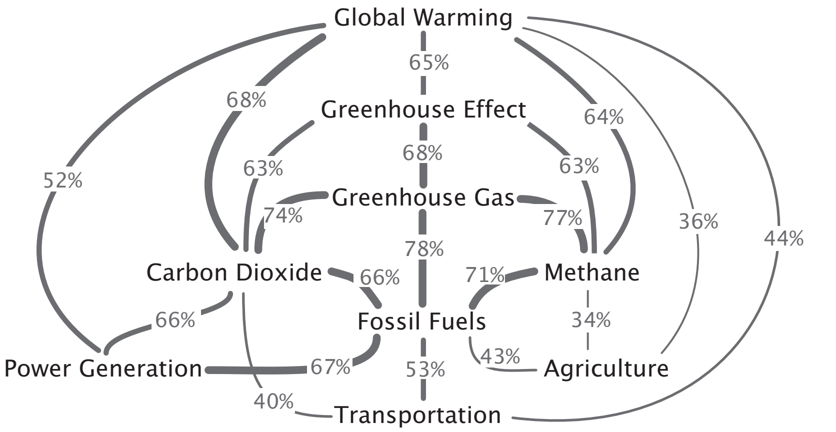


Figure 5.2: Topics and relatedness measures relevant to *global warming*

there is a single document that concisely collates the various causes; something that cannot be relied on in a universe of infinite questions.

The two systems described above are at opposite ends of a spectrum, where one places all responsibility on the machine, and the other places it on the user. How can relatedness measures and other "semantic-light" techniques provide a practical medium between them? Figure 5.2 shows what the machine's understanding of Figure 5.1 is reduced to in a knowledge base constructed with relatedness measures. This identifies the relevant concepts and quantifies the connections between them, e.g. global warming is 68% related to Carbon Dioxide, 64% related to Methane, 52% related to power generation, and 44% related to transportation.

The graph is a far cry from the sophisticated ontologies required for inference. With it, the machine can have no knowledge of how Carbon Dioxide and other gasses contribute to the greenhouse effect, or what produces them. It does know enough, however, to bring the relevant concepts to the user's attention and emphasize Carbon Dioxide over Methane, or power generation over transportation and agriculture. It would not be difficult to automatically locate texts within Wikipedia to explain what the concepts are, and how they are connected to each other; these entities are manually tagged with link markup. With access to the concepts, connections and explanatory texts, the user can efficiently reconstruct the logic of how these concepts are relevant to the original question and formulate a sensible, well-informed answer.

The retrieval system described above is not responsible for reasoning about the information, but merely for providing tools to explore it efficiently. Intuitively, this seems like the most appropriate distribution of responsibilities: searching, gathering, extracting and measuring are all tedious to perform manually but relatively easy to automate. The opposite is true for reasoning and inference.

5.2 Measuring relatedness manually

If ten different people were asked to quantify the relatedness between *Carbon Dioxide* and *Global Warming*, one might reasonably expect ten different answers. How different would they be? This is an important question: if they varied wildly, then the task of generating measures would be hopelessly subjective and impossible to automate.

Table 5.1 summarizes several experiments that explore how people generate relatedness measures manually. The first experiment, by Rubenstein and Goodenough (1965), involved 65 pairs of terms and a total of 51 participants. Each participant was independently given a randomly shuffled deck of cards. On each card was a pair of terms. They were asked to sort the cards in order of the relatedness between the terms on them, and annotate each with a numerical value from 0.0 (entirely unrelated) to 4.0 (completely synonymous). 15 of the participants attended another session—two weeks earlier—and performed the same task with 48 pairs of terms, including 36 from the other session. The high correlation (85%) between the annotations in each session indicates that each individual participant was reasonably consistent in the measures they produced. This experiment produced a dataset of 65 term pairs and their average human-defined relatedness scores, which has been used to evaluate semantic relatedness algorithms ever since.

Miller and Charles (1991) sampled 30 term pairs from Rubenstein and Goodenough's dataset, to evenly cover low, intermediate, and high levels of similarity (as defined by the previous experiment). When these term pairs were annotated by 38 participants, the mean scores had a correlation of 97% with the previous ones. The same 30 term pairs were taken by Resnik (1995) and annotated by 10 participants. The mean scores have a 96% correlation with Miller and Charles' dataset. The average correlation between each

Source	term pairs	judges per pair	correlation between judges	correlation between datasets
Rubenstein and Goodenough (1965)	65	51	-	
Miller and Charles (1991)	30	38	-	97%
Resnik (1995b)	30	10	90%	96%
Finkelstein et al. (2001)				94%
set 1	153	13	80%	
set 2	200	16	73%	

Table 5.1: Datasets for evaluating semantic relatedness measures

individual participant and the previous dataset is 88%, and the average between each individual and the remainder of the group (via leave-one-out resampling) is 90%. The three experiments— Rubenstein and Goodenough (1965), Miller and Charles (1991) and Resnik (1995b)—are remarkably consistent, given the decades which separate them.

Finkelstein et al. (2001) produced a larger but slightly less controlled dataset called WordSimilarity 353. In one experiment they asked 13 participants to independently assign relatedness scores (between 0 and 10) to 153 term pairs. A second experiment asked 16 participants to annotate 200 term pairs. Participants worked in their own time without observation, and did not explicitly compare across term pairs by pre-sorting them. Consequently the resulting annotations are not as consistent as in previous experiments. The correlation of each individual's scores against the mean (again using leave-one-out resampling) ranges from 68% to 86% (80% on average) for the first experiment, and 50% to 81% (73% average) for the second. The first set of term pairs contains Miller and Charles' dataset. The correlation of individuals' measures to this dataset range from 64% to 96% (88% on average), and the mean of all participants has a 94% correlation.

Confusingly, the literature contains at least three different terms that are subtly different but often used interchangeably: *semantic relatedness*, *semantic similarity*, and *semantic distance*. Budanitsky and Hirst (2006) provide a good explanation of the differences. They consider relatedness to be more general than similarity, because the latter is restricted to synonymy and hyponymy, while the former encompasses meronymy, antonymy, functional associations, and many other kinds of relations. Thus, as shown in Table 5.2, *global warming* is related, but not similar, to *carbon dioxide*. Semantic distance is roughly the inverse of relatedness (*global warming* and *carbon dioxide* have high relatedness and low distance), except that antonyms (e.g. *warm* vs. *cold*) are considered strongly related *and* semantically distant. Of the three, semantic relatedness suits our purposes best. Returning to the example in Figure 5.2, the related topics have only a loose functional association with global warming, and antonyms (e.g. opposing theories to global warming) would certainly be relevant.

Judging by the instructions given to participants, the experiments in Table 5.1 are roughly

Term 1	**Term 2**	**Semantic relatedness**	**Semantic similarity**	**Semantic distance**
Global warming	Carbon Dioxide	high	low	low
Carbon Dioxide	Methane	high	high	low
Warm	Cold	high	low	high

Table 5.2: Examples of semantic relatedness, similarity and distance

split so that the first three focus on semantic similarity; only Finkelstein et al. generalizes to semantic relatedness (Agirre et al. 2009). This partially explains why the first three agree experiments with each other more closely than the last, because it is easier for people to agree on more strict relationships.

5.3 Existing approaches for measuring relatedness

Making judgments about the relatedness of different terms is a routine yet deceptively complex task. To perform it, people draw on an immense amount of background knowledge about the concepts the terms represent. When asked to what extent *Power Generation* is related to *Global Warming*, for example, one needs to know what the words represent and much of the logic described in Figure 5.1 to come up with a reasonable number.

Any attempt to measure semantic relatedness automatically must also involve consulting some source of knowledge—although not necessarily with the same sophistication as we do, otherwise a retrieval system based on these measures would be no less complex and brittle than full inference. The various techniques for doing so can be roughly categorized by the kinds of knowledge they use.

Table 5.3 compares several knowledge sources and techniques by how well they perform against the Finkelstein et al. (2001) WordSimilarity-353 collection. For the first two entries, background knowledge is obtained from lexical resources. WordNet, Roget's Thesaurus and various dictionaries have been used for this purpose (Budanitsky and Hirst 2006). These hand-built structures are expensive to create and maintain. As a result they have comparatively small vocabularies of terms for which they can provide comparisons, and thus have limited applications for open-domain information retrieval. They are also quite sparsely connected—Boyd-Graber et al. (2006) describe WordNet's limitations in this area—because the indexers and domain experts who construct them are careful to represent only the most clear, concrete semantic relations. This allows them to measure semantic similarity impressively well; e.g., Resnik (1999) achieves 77% correlation with Miller and Charles and 78% with Rubenstein and Goodenough. Unfortunately the emphasis of quality over quantity does not seem conducive to measuring relatedness. Both Jarmarz (2003) and Strube and Ponzetto (2006) implement many techniques and achieve only 35% correlation (at best) over the Finkelstein et al. dataset.

Corpus-based approaches obtain background knowledge by performing statistical analysis of large untagged document collections. One of the more well-known techniques is Latent Semantic Analysis (Landauer et al. 1998), which relies on the tendency for related words to appear in similar contexts. As shown in Table 5.3, this approach and other corpus-based techniques perform as well or even slightly better than those using lexical structures. They also offer the same vocabulary as the documents upon which they are built. These approaches could sensibly be used to support open-domain retrieval.

Systems that rely on either lexical resources or corpora trail far behind the performance of humans. The former resources are limited in vocabulary and the types of relations they define, while the latter are unstructured and contain much ambiguity. These limitations are the motivation behind several new techniques that infer semantic relatedness from Wikipedia. With millions of articles and an extensive network of cross-references, portals and categories, Wikipedia's rare combination of scale and structure makes it an attractive resource for this task.

Strube and Ponzetto (2006) were the first to compute measures of semantic relatedness using Wikipedia. Their approach—WikiRelate—took familiar techniques that had previously been applied to WordNet and modified them to suit Wikipedia. Their best results—the 0.48 correlation shown in Table 5.3—come from an adaption of Leacock and Chodorow's (1998) path-length measure, which takes into account the depth within WordNet at which the concepts are found. WikiRelate's implementation does much the same for Wikipedia's hierarchical category structure. This results in modest gains to accuracy.

Gabrilovich and Markovitch (2007) achieve extremely accurate results with Explicit Semantic Analysis (ESA), a technique that is somewhat reminiscent of the vector space model widely used in information retrieval. Instead of comparing vectors of term weights to evaluate the similarity between queries and documents, they compare weighted vectors of the Wikipedia articles related to each term. The name of the approach stems from the way these vectors are comprised of manually defined concepts, as opposed to the

Relatedness measure	**Reference**	**Correlation with humans**
Thesaurus based		
WordNet	(Jarmasz 2003, Strube and Ponzetto 2006)	0.21-0.35
Roget	(Jarmasz 2003)	0.55
Corpus based		
Latent Semantic Analysis (LSA)	(Finkelstein et al. 2002)	0.56
Wikipedia based		
WikiRelate	(Strube and Ponzetto 2006)	0.19-0.48
Explicit Semantic Analysis (ESA)	(Gabrilovich and Markovitch 2007)	0.75

Table 5.3: Performance of existing semantic relatedness measures
(from Gabrilovich and Markovitch 2007)

mathematically derived contexts used by Latent Semantic Analysis. The result is a measure that approaches the accuracy of humans. Additionally, it provides relatedness measures for any stretch of text: unlike WikiRelate, there is no restriction that the input be matched to article titles.

5.4 The Wikipedia link-based measure (WLM)

One of the main contributions of this thesis is a new approach for measuring semantic relatedness called the Wikipedia Link-based Measure or WLM. Like ESA and WikiRelate, it uses Wikipedia to provide structured world knowledge about the terms of interest. WLM is unique in that it does so using the hyperlink structure of Wikipedia rather than its category hierarchy or textual content.

There are two reasons why a new measure of relatedness was developed rather than simply using ESA or WikiRelate. First, all of the applications for semantic relatedness within this thesis (which are scattered throughout Chapters 5, 6 and 7) require a measure between Wikipedia concepts rather than between terms. This is a significantly simpler problem, since the elements being compared are manually disambiguated and clearly defined. Simpler solutions will suffice.

The second, more significant reason is speed. The applications that follow often require thousands of measures to be gathered in an interactive setting. Any delays will directly affect the user's impression of the system. It would not be practical to calculate these measures in advance. There are more than 10^{13} possible pairs of Wikipedia concepts. Assuming the relatedness of each pair could be calculated in a millisecond, it would require more than 100,000 machine days to build the complete weighted graph. Consequently efficiency is paramount, because the relatedness measures must be gathered on demand.

ESA is expensive because it makes use of Wikipedia's full textual content. It also cannot be efficiently adapted to measure relatedness between articles rather than terms. WikiRelate is more promising because it can trivially be stripped down and simplified to measure the relatedness of articles, and is cheap if restricted to using only category relations. Unfortunately its accuracy is comparatively poor.

The new approach forms a compromise between the speed and simplicity of WikiRelate and the accuracy of ESA by using the hyperlink structure of Wikipedia rather than its category hierarchy or textual content. For accuracy, Wikipedia provides hundreds of millions of links between articles, as Figure 5.3 illustrates with only a small sample—just 0.34%—of the links available for determining how *automobiles* are related to *global*

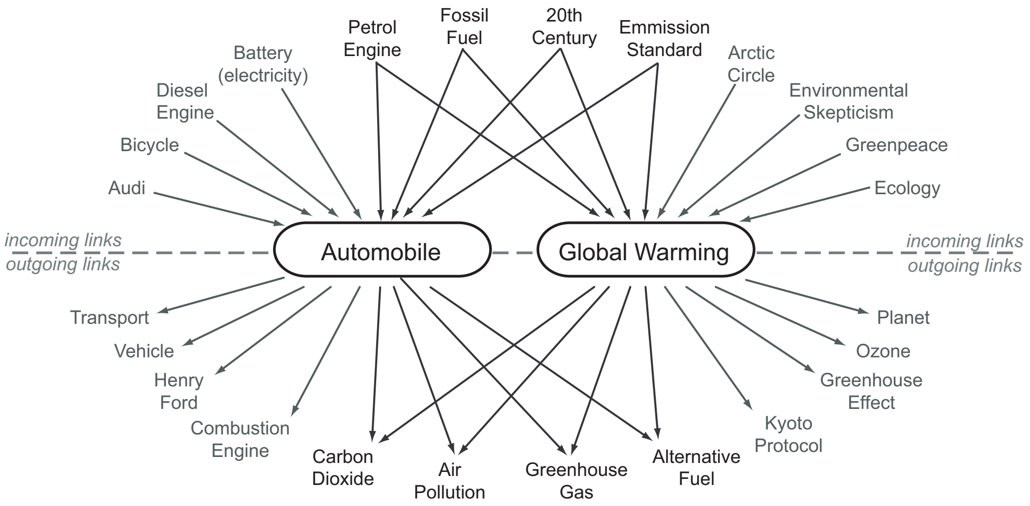


Figure 5.3: Measuring the relatedness of *Automobile* and *Global Warming*

warming. WikiRelate relies on a much smaller network, and is unlikely to draw a connection in this example because the categories involved (starting with *automobiles* and *vehicles* for the first topic, and *climate change*, *environmental issues with energy*, and *weather hazards* for second) remain distinct high into the category tree. For efficiency, the links can be reduced to a fraction of the 20Gb of raw text analysed by ESA.

The remainder of this section describes the approach and the various options that were tested. It also assesses these individual components to identify the best ones and define the final algorithm. Assessment of the algorithm as a whole—and comparison with related work—is left for Section 5.5. The testing reported in this section is only for development purposes, and was done over a subset of 50 randomly selected term pairs from the Finkelstein et al. dataset to avoid overfitting the algorithm to the data.

5.4.1 Measuring relatedness between articles

Like WikiRelate, WLM starts as a measure of relatedness between Wikipedia articles rather than terms. It incorporates two measures: one based on the links extending out of each article; the other on the links made to them. These correspond to the bottom and top halves of Figure 5.3.

The first measure is defined by the angle between the vectors of links found within the two articles of interest. These are almost identical to the vectors of TF×IDF values (described in Section 4.3.3) used extensively within information retrieval. The only difference is that the values used are link counts weighted by the probability of each link occurring, instead of term counts weighted by the probability of the term occurring. This probability is defined by the total number of links to the target article over the total number of articles. Formally, if *s* and *t* are the source and target articles, then the weight *w* of the link *s* →*t* is:

$$w(s \rightarrow t) = \log\left(\frac{|W|}{|T|}\right) \text{ if } s \in T\text{, 0 otherwise}$$

where *T* is the set of all articles that link to *t*, and *W* is the set of all articles in Wikipedia. In other words, the weight of a link is the inverse probability of any link being made to the target, or 0 if the link does not exist. Thus links are considered less significant for judging the similarity between articles if many other articles also link to the same target. The fact that two articles both link to *art* is much less significant than if they both link to a specific topic such as *abstract impressionism*.

The link weights are used to generate vectors to describe each of the two articles of interest. The set of links considered for the vectors is the union of all links made from either of the two source articles. The remainder of the approach is exactly the same as in the vector space model: the similarity of the articles is given by the angle (cosine similarity) between the vectors. This ranges from 0° if the articles contain identical lists of links to 90° if there is no overlap between them. In practice the angles are normalized and inverted to fall between 0 (entirely unrelated) and 1 (synonymous).

The second measure is modelled after the Normalized Google Distance (Cilibrasi and Vitanyi 2007), which uses the Google search engine to obtain and compare pages that mention the terms of interest. Pages that contain both terms indicate relatedness, while pages with only one of the terms suggest the opposite. WLM adapts this to use Wikipedia's links rather than Google's search results. Formally, the measure is:

$$sr(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where *a* and *b* are the two articles of interest, *A* and *B* are the sets of all articles that link to *a* and *b* respectively, and—as before—*W* is the entire Wikipedia. This formula returns 0 if the pages are completely related, tending to infinity as they get less related, but is undefined if there is nothing in common. In practice, it is normalized to match the previous measure (where 0=unrelated and 1=synonymous) by truncating large or undefined values to 1 and inverting.

In- and out-links are treated differently because they follow different distributions. The number of links made from an article is bounded by article length, and follows a normal distribution. The number of links made to an article is bounded only by the number of articles available, and roughly follows a power law; the average article receives about 25 links, but some receive hundreds of thousands. *United States*, for example, receives 280K in-links. WLM is modelled after two formulas that were designed for similar situations: the formula for out-links is modelled after TF×IDF, where documents are compared by the terms found within them, and the formula for in-links is modelled after the Google distance inspired formula, where terms are compared by the web pages on which they are found.

All of the datasets described in Section 5.2 compare terms rather than articles. To evaluate WLM, it must be extended to automatically select the appropriate articles for each term pair. This problem is tackled in the next section. First, a rough comparison of the two link-based approaches can be made by manually disambiguating the Finkelstein et al. subset to obtain a ground truth of *article* pairs—as opposed to *term* pairs—and manually defined measures of relatedness between them. Table 5.4 shows how the two link-based measures correlate with this new dataset, and clearly identifies *Google Distance* as the more accurate measure. It also shows that modest gains can be made by taking the average of the measures. This combination is used in the remainder of this chapter.

5.4.2 Measuring relatedness between terms

This section explains how WLM's measure between articles can be extended to compare terms and phrases. This extension is necessary to provide comparison with the alternative techniques described in Section 5.3, even though all of the applications within this thesis do not require it. Unfortunately this forces WLM to deal with two language phenomena: ambiguity and synonymy.

Ambiguity is the tendency for terms to relate to multiple concepts: for example *plane* might refer to a fixed-wing aircraft, a theoretical surface of infinite area and zero depth, or a tool for flattening wooden surfaces. The correct sense depends on the context of the term to which we are comparing it to; consider the relatedness of *plane* to *wing*, *plane* to *axis*, and *plane* to *chisel*.

Synonymy is the tendency for concepts to be known by multiple names: a *plane* may also be referred to as *fixed wing aircraft, airplane* or *aeroplane*. It should be possible to

Relatedness measure	**Correlation with humans**
TF×IDF inspired (out-links)	0.66
Google Distance inspired (in-links)	0.72
combined (average)	0.74

Table 5.4: Performance of relatedness measures (manual disambiguation)

navigate to the appropriate article (and thus obtain the same relatedness measure) with any of these synonyms.

As Section 3.4.1 explained, Wikipedia's raw structure provides three easily processed means of mapping terms or surface forms to concepts: specifying titles of articles, assigning redirects (alternative titles) to articles, and using different terms within the links that are made to an article. Appendix C.3.2 describes how these can be combined into a graph of terms and concepts that encodes both ambiguity (terms can map to many concepts) and synonymy (concepts can map to many terms).

All 95 distinct terms in the Finkelstein et al. subset are found in this graph. All but one of the terms was ambiguous, and had to be resolved by selecting one candidate article to represent each term. There are several options to perform such disambiguation automatically, which are assessed in Table 5.5 across the Finkelstein et al. subset.

One method is to use the most common sense of each term. For example, when making a judgment between *Israel* and *Jerusalem,* one would consider only the nation and its capital city. The obscure but strong connection between two townships in Ohio with the same names would be completely ignored. WLM defines the commonness of a sense as the number of times the term is used to link to it: e.g. 95% of *Israel* anchors link to the nation, 2% to the football team, 1% to the ancient kingdom, and 0.1% to the Ohio township. As shown in Table 5.5, merely selecting the *most common pair* of concepts performs fairly well.

Another approach is to use the two terms involved to disambiguate each other. For example, when identifying the relatedness of *jaguar* and *car* it makes sense to use *car* to determine that we are talking about the automobile manufacturer *Jaguar Cars Ltd*, rather than the species of cat. This amounts to selecting the two candidate senses that most closely relate to each other. As shown in Table 5.5, selecting the *most closely related pair* of senses performs slightly better than the most common sense heuristic, but is marred by the number of obscure senses available (as in the *Jerusalem* vs. *Israel* example given above). For efficiency and accuracy's sake, WLM only considers articles that

Relatedness measure	**Correlation with humans**
most common pair	0.68
most closely related pair	0.69
highest (commonness + relatedness)	0.74
sequential decision	0.75
final relatedness measure	0.78

Table 5.5: Performance of relatedness measures (automatic disambiguation)

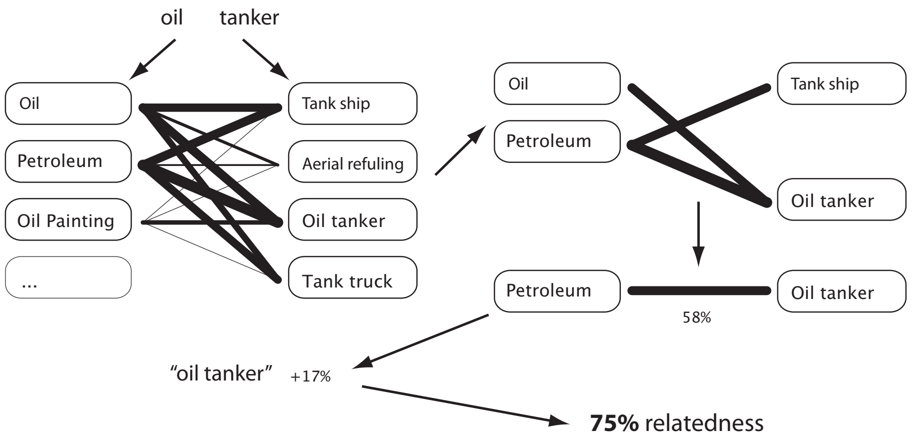


Figure 5.4: Measuring the relatedness of *oil* and *tanker*

receive at least 1% of the anchor's links. This theoretically leaves up to 100 candidates to be examined, but in practice the distribution of links for each anchor follows the power law, meaning that the vast majority are made to a handful of candidates. In the sample, the largest number of candidates examined for a term was 26.

The best results are obtained when both commonness and relatedness are considered. Evenly weighting the candidate senses by these variables and choosing the pair with the highest weight—*highest (commonness + relatedness)*—gives exactly the same results as with manual disambiguation (Table 5.4, row 3). Surprisingly, this can be improved upon by making a simple *sequential decision*, which first groups the most related pairs of candidates (within 40% of the most related pair) and then chooses the most common pair. This makes subtly different choices from those made manually. Given the term *environment* and the context *ecology*, for example, WLM selects *ecosystem* as the representative article rather than *natural environment*, and consequently obtains a more accurate relatedness score.

Finally, as a minor amendment, WLM also considers the case where two words are closely related because they belong together in the same phrase: e.g. *family planning* is a well-known phrase, and consequently *family* and *planning* are given a high semantic relatedness by humans even though their respective concepts are semantically distant. To identify these cases WLM simply concatenates the terms and consults the anchor vocabulary. If there is a match, the frequency with which the anchor is used is normalized (by taking its log and dividing by 30; a constant established over the development data) and added to the relatedness score of the original terms. This gives the *final relatedness measure*, which has a correlation of 0.78 with manual judgments over the Finkelstein et al. sample.

Figure 5.4 provides one last example to summarize the WLM algorithm, by measuring the relatedness of *oil* and *tanker*. The first step is to consult Wikipedia's anchor vocabulary to find out what these words could refer to. Three senses are considered for *oil* (more are available, but these fall below the sense probability threshold of 1%) and four for *tanker*. Relatedness measures are calculated for all possible combinations of these senses, using the combined measure, which averages the TF×IDF inspired measure (out-links) and the Google distance inspired measure (in-links).

A sequential decision is made to select the best pair of senses. First, all relations within 80% of the strongest relation are retained. This leaves the three relations shown on the top right of the Figure: *Petroleum→Oil tanker* (58%), *Oil→Oil tanker* (48%) and *Petroleum→Tank ship* (47%). Of these, the pair with the highest combined sense probability—*Petroleum* (32%) and *Tank ship* (74%)—is used represent the original two terms. The relatedness between these two concepts (58%) is boosted by 17% because “oil tanker” is a common collocation (it is used as a link anchor 190 times). This yields the final relatedness score of 64%.

5.5 Evaluation

This section describes WLM's evaluation against the datasets described in Section 5.2. The version of Wikipedia used to obtain the measures was released on November 20, 2007. At this point it contained approximately 13GB of uncompressed XML markup. This relates to just under two million articles, which constitute the various concepts for which semantic relatedness judgments were available. It provided over five million distinct anchors, titles and redirects, which define the vocabulary of terms by which concepts can be accessed.

Table 5.6 compares WLM with its two main competitors—WikiRelate and ESA—by their correlation with manually defined judgments. Only the best measures obtained by the different approaches are shown. It should be noted that the results were obtained with different snapshots of Wikipedia, which may affect performance.

There is a consistent trend across all datasets: WLM is better than WikiRelate but worse

Dataset	**WikiRelate**	**ESA**	**WLM**
Rubenstein and Goodenough (1965)	0.45	0.73	0.70
Miller and Charles (1991)	0.52	0.82	0.64
Finkelstein et al. (2001)	0.49	0.75	0.69
Weighted average	*0.49*	*0.76*	*0.68*

Table 5.6: Performance of relatedness measures against all datasets

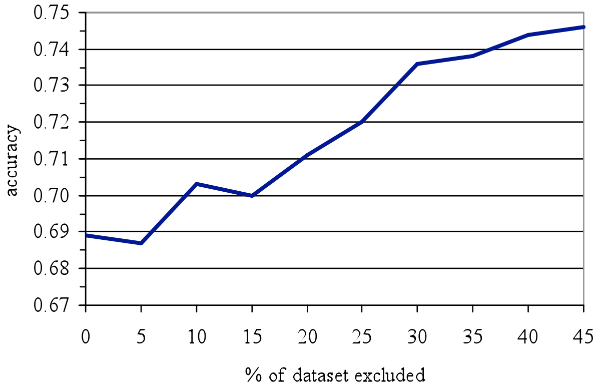


Figure 5.5: Accuracy of WLM with weakly defined terms excluded

than ESA. The final row in the table combines the results across the three datasets, with correlations weighted by the size of each dataset. This shows WLM outperforming WikiRelate by 0.19, and in turn being outperformed by ESA by 0.08. The third row in Table 4 shows the performance of the algorithms over the WordSimilarity-353 collection. The accuracy of 0.69 for WLM can be directly compared to the results in Table 5.3, which were obtained from the same dataset. It outperforms all others except ESA by at least 0.13.

It is interesting to note the drop in WLM's performance between the Finkelstein et al. sample used in the previous section (0.78) and the full dataset used here (0.69). Much of the drop may be due to over-fitting the algorithm to the sample. Analysis of the results, however, reveals another reason: WLM differs most from the ground truth when the terms being compared cannot be resolved to suitable Wikipedia articles. For example, there is no article for the concept *defeat*; the anchor points only to specific military encounters. These cases are common in the full dataset but were, by chance, excluded from the sample.

Figure 5.5 plots the performance of WLM as successively more pairs are discarded from the Finkelstein et al. collection so that only the most well-defined terms are considered. Anchor frequency is used as a simple indicator of how well a term is defined; if a term is not used to make a sufficient number of links, it is considered problematic. It is likely that ESA's performance would remain constant here, since it does not distinguish between terms used as anchors and those that appear in plain text. Thus Figure 5.5 shows how WLM's performance approaches the benchmark of 0.75 set by ESA when the terms involved are well defined as individual articles in Wikipedia.

5.6 Discussion

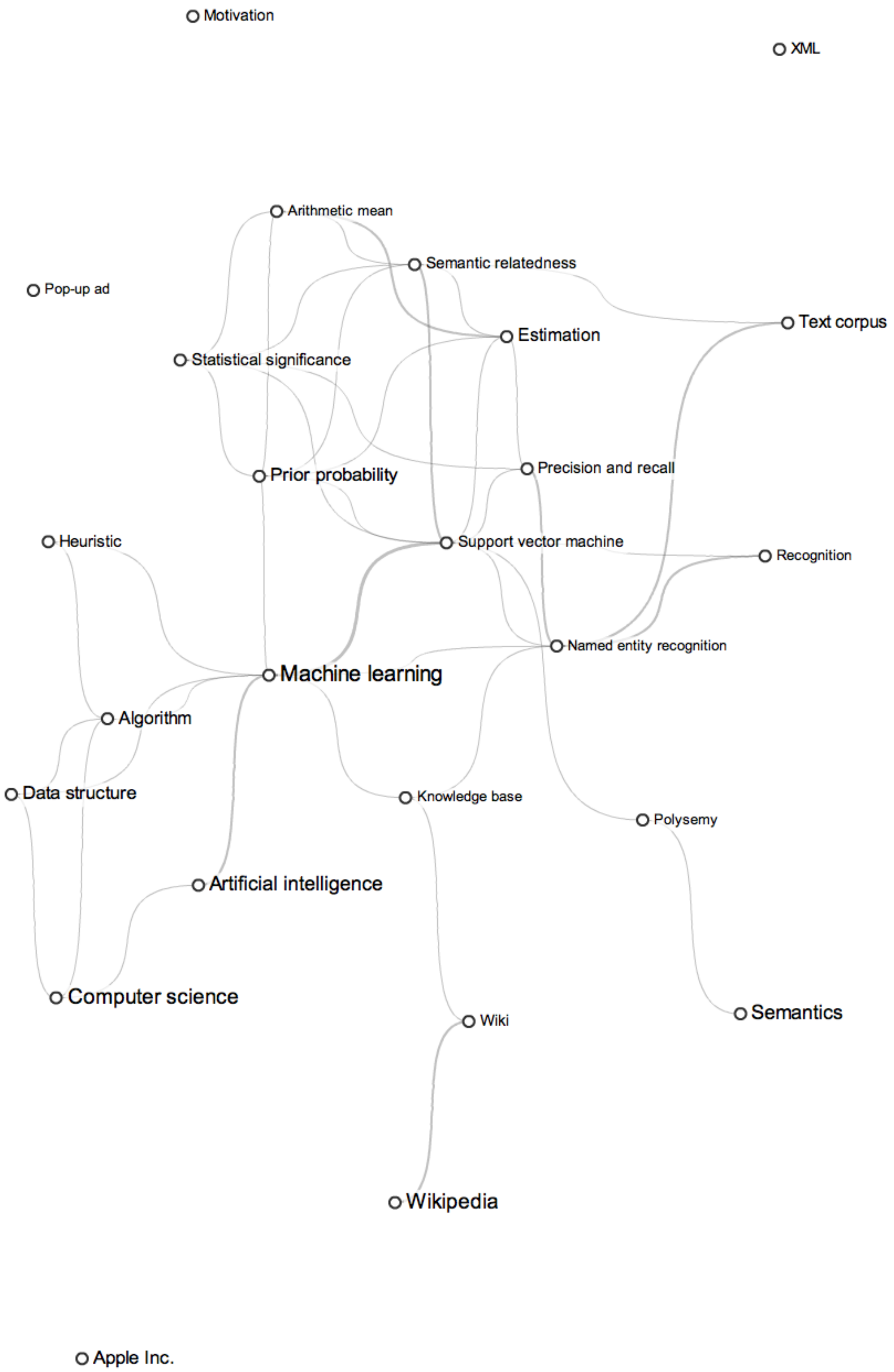
The previous section clearly identifies ESA as the best of the semantic relatedness measures derived from Wikipedia. It is less brittle than WLM and WikiRelate, because it only requires that articles mention the terms involved. WLM, however, achieves competitive levels of accuracy when the terms involved correspond to distinct Wikipedia articles. Given Wikipedia's sheer size and rate of growth, one can expect this to hold true whenever the terms represent topics which one could reasonably write an article about. This is the case for most applications in the literature, which deal primarily with topics: named entities (Bunescu and Paşca 2006, Cucerzan 2007); key phrases (Medelyan et al. 2008); or entries in existing ontologies (Medelyan and Legg 2008) and thesauri (Ruiz-Casado et al. 2005a). In these applications, WLM can be expected to compete with the state of the art. Unfortunately, current datasets—and the evaluation conducted here—focus on common nouns rather than entities.

The advantage of WLM over ESA is that it requires far less data and resources. To obtain measures from ESA, one must preprocess a vast amount of textual data; 25Gb as of January 2010. Each term must be matched to the articles in which it is found, and each of the resulting lists of articles must be weighted, sorted, and pruned. One assumes (given the sorting requirement) that this is a log-linear at best. By comparison, WLM requires only the link structure of Wikipedia (1Gb) and the statistics of how often terms are used to refer to different concepts (360Mb). All of this information can reasonably be cached to memory. No preprocessing is required other than to extract this information from Wikipedia's XML dumps. This is a straight-forward task that can be achieved in linear time (assuming constant hash-table operations).

ESA is able to determine the relatedness of texts of any length as easily as individual terms, by gathering and merging the lists of articles related to each word. WLM and WikiRelate are not so easily extended: they require an additional step—which is tackled in the next chapter—to discover the topics mentioned in the texts. This requirement may well turn out to be an advantage, however, because the techniques would then be comparing collections of concepts and topics rather than words. This highlights a fundamental difference between ESA and the other two approaches: that it disregards stop-words and word order. It considers, for example, *wind break* and *break wind* to be the same thing. It is unclear how much this affects the overall accuracy of the three techniques, as the datasets upon which they have so far been compared are restricted to individual words.

The use of semantic relatedness datasets has been a limitation common to all of the experiments described in this chapter. The number of word pairs these sets provide is admittedly small, and may not be sufficient for fine-grained comparison. An additional concern is the focus these datasets maintain on common nouns: our intended applications for WLM focus on named entities. One option would be to compare the connections these techniques produce to the relations found in existing thesauri, but such experiments would obviously be biased towards the techniques that rely these structures. Testing against amalgamations of traditionally-generated thesauri—as Curran (2004) advocates—would be an improvement, but would still penalize automated algorithms and crowd sourcing for their ability to provide greater connectivity.

Perhaps the best way to test semantic relatedness measures is to put them to use. In this respect the utility of the Wikipedia-derived measures has been well demonstrated. WikiRelate has been successfully applied to co-reference resolution (Strube and Ponzetto 2006). ESA—or at least something very much like it—has been applied to document categorization (Gabrilovich and Markovitch 2006) and relevance feedback (Egozi et al. 2008). The chapters that follow rely extensively on WLM. Chapter 6 applies it to topic disambiguation and detection, with experiments that involve millions of relatedness comparisons. Chapter 7 uses it for interactive query expansion, and exposes the measures to direct inspection from users.


Motivation
XML
Arithmetic mean
Semantic relatedness
Pop-up ad
Text corpus
Estimation
Statistical significance
Precision and recall
Prior probability
Heuristic
Support vector machine
Recognition
Named entity recognition
Machine learning
Algorithm
Data structure
Knowledge base
Polysemy
Artificial intelligence
Computer science
Semantics
Wiki
Wikipedia
Apple Inc.

6. Linking free text to structured knowledge

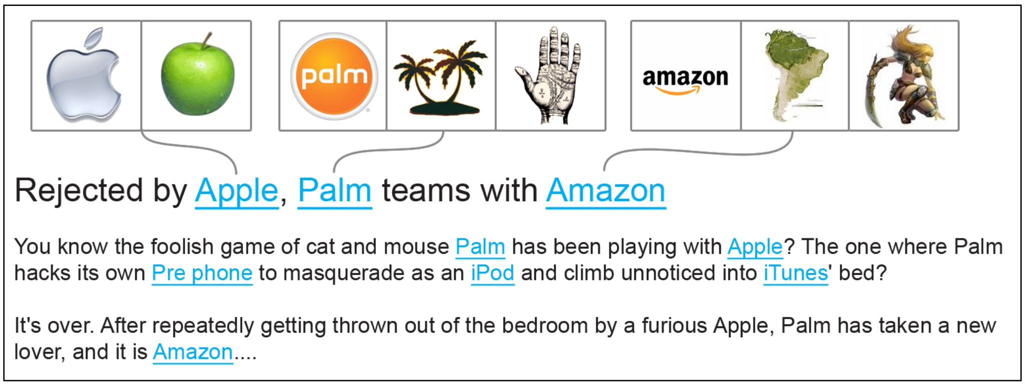
This thesis aims to apply Wikipedia to information retrieval within arbitrary document collections. One of the challenges that must be addressed is to somehow connect the resources—Wikipedia and the given documents—together. To do this, it is necessary to detect Wikipedia entities when they are mentioned within unstructured text. This task was introduced along with the prototype Koru system in Chapter 4, but was only given a cursory glance: an ad hoc algorithm was presented without justification, evaluation, or reference to related work. This chapter isolates the task of cross-referencing documents with Wikipedia and investigates it much more systematically.

The following section surveys related work and explains how cross-referencing documents can be broken down into two subtasks: detecting terms that are suitable link anchors, and disambiguating terms that would otherwise link to multiple topics. Each subtask is tackled separately—and evaluated against manually defined ground truth—in Sections 6.2 and 6.3. In both cases we present new, uniquely machine-learned approaches in which Wikipedia is used not only as a source of information to point to, but also as training data for how best to create links. Section 6.4 describes a third evaluation, in which news stories are cross-referenced and then judged by human participants. The chapter concludes with a discussion of the implications, which go far beyond enriching documents with explanatory links. The techniques described here can provide structured knowledge about any unstructured fragment of text, and are therefore applicable to a wide variety of tasks.

6.1 Cross-referencing documents with Wikipedia

Wikipedia's articles are peppered with hundreds of millions of links. These connections explain the topics being discussed and provide an environment where serendipitous encounters with information are commonplace. Anyone who has browsed Wikipedia has likely experienced the feeling of being happily lost, browsing from one interesting topic to the next and encountering information that they would never have searched for.

The automatic construction of these links—knows as "wikification"—has emerged as an interesting research problem in recent years. Escalating maintenance issues in Wikipedia is one source of motivation (Huang et al. 2009b). More broadly, any piece of text can potentially be enhanced through wikification. Take the example shown in Figure 6.1, where a short news excerpt has been augmented with links to Wikipedia. The destination articles



Rejected by Apple, Palm teams with Amazon

You know the foolish game of cat and mouse Palm has been playing with Apple? The one where Palm hacks its own Pre phone to masquerade as an iPod and climb unnoticed into iTunes' bed?

It's over. After repeatedly getting thrown out of the bedroom by a furious Apple, Palm has taken a new lover, and it is Amazon....

Figure 6.1: A document marked with relevant Wikipedia concepts

provide explanations of the concepts involved, and opportunities to explore them in more detail. Csomai and Mihalcea (2007) have shown that augmenting educational documents with such links can improve the quality of the knowledge readers acquire and reduce the time they need to gain it. The benefits do not just extend to human readers. Any task that is currently addressed with bags of words—indexing, clustering, retrieval, and summarization to name a few—could use the links to draw on a vast structured knowledge base that provides valuable machine-readable semantics about the words. These far-ranging implications and applications of wikification are discussed in detail in Section 6.5.

Cross-referencing documents with Wikipedia presents two fundamental problems. The first, *detection,* involves identifying significant terms and phrases within a document from which links should be made. In Figure 6.1, for example, an algorithm must somehow identify *amazon, apple* and *palm* as significant, but *cat, masquerade* and *bed* as not. The second problem, *disambiguation*, involves deciding where those links should be made to. An algorithm must identify that *amazon*, *palm*, and *apple* should link to the respective organizations, rather than *Amazon River*, *Palm Tree*, or *Apple* (the fruit). The various solutions to these two problems are presented in Section 6.1.2. First, however, we explain how they can be evaluated.

6.1.1 Evaluating cross-referencing algorithms

To evaluate approaches for automatic wikification, one needs some form of ground truth; some corpus of documents that has been manually annotated with links to the relevant Wikipedia articles. Fortunately, every single article in Wikipedia has been annotated in exactly this way. Most evaluations of wikification take existing Wikipedia articles, strip them of all links, and attempt to restore the links automatically.

Mihalcea and Csomai (2007) show how Wikipedia's markup can be used to isolate the two problems of detection and disambiguation, which they evaluate separately. The former is evaluated by expecting the system to automatically identify the same vocabulary of anchors that is marked up in the original articles. The latter is tested by informing the system of the correct anchor vocabulary and expecting it to automatically identify their original destinations. The details of thier system are described in Section 6.1.2.

Much of the research on wikification take place within the framework of the Link-the-Wiki track (Trotman and Geva 2006), which was introduced to the *Initiative for the Evaluation of XML Retrieval* (INEX) conference in 2006. This competition does not merely aim to link documents to Wikipedia, but to completely integrate them into the resource. Given a new document, participants must not only identify relevant outgoing links from it (wikification), but also locate other relevant documents from which to add incoming links. Since 2008 Link-the-Wiki has additionally aimed to support *focused link discovery* by locating the best entry points—the locations in the target document where the reader is most likely to start reading—for each link. The problems of in-link and entry point detection will be ignored in this chapter, however, because they are not part of the task it addresses.

INEX performs both automatic and manual assessment (Huang et al. 2009c). Automatic evaluation is done in a similar fashion to Mihalcea and Csomai (2007), but with a much larger set of test topics. INEX Link-the-Wiki evaluations take a large subset of articles (6600 in 2008, 5000 in 2009) and orphan them by removing all incoming and outgoing links. Algorithms are then judged by the accuracy with which they are able to add the links back. Michalcea and Csomai, in contrast, used only 85 articles for testing. This does not appear to be a significant limitation, however: INEX participants have found that results differ little when testing against their full data set or with a subset of 90 articles (Geva 2007).

At INEX, participants are only expected to locate up to 50 out-links and 250 in-links for each document, which seems problematic: the top-ranked 50 automatically generated out-links are compared against the first 50 links found in the article, so that an algorithm's score is based not only on how well it ranks potential links but also on how well it balances between choosing highly ranked links and ones that appear early. It is not clear how the 250 ground truth in-links are selected, but similar difficulties seem inevitable: popular articles often receive thousands of in-links, and sampling from these will inevitably cause correct recommendations to be erroneously evaluated as irrelevant.

INEX treats in- and out-link recommendations as inherently ranked, so results take the form of precision vs. recall curves and mean average precision (MAP) scores. This differs from Mihalcea and Csomai (2007), who treat both automatic and ground truth links as unranked sets and consequently report only precision, recall and f-measure. Best entry points are not assessed automatically, because Wikipedia links are rarely annotated with them (it is possible to point to a specific section in an article, but the vast majority of links do not).

Manual assessment in INEX is performed by the participants, using a handpicked subset (50 articles in 2008, 33 in 2009) of orphaned articles. A pool is constructed for each orphaned article, which contains all automatically recommended links plus the original 50 in-links and 250 out-links that were used as ground truth for automatic assessment. In 2008 each pool contained between 400 and 1700 links. Every link is judged manually using an interactive system that presents users with documents, links, and the link destinations. Participants first work through the topic article and mark all irrelevant anchors. They then work through each retained anchor, and mark those that link to irrelevant targets. At the same time they evaluate in-links, and entry points for both in- and out-links. The organizers of INEX 2008 calculate that approximately 4–6 man-hours were spent evaluating each topic, or 200–300 hours in total. Only the out-links are of interest here. Assuming that the three components of a link (its anchor, destination, and best entry point) require equal time to evaluate, and that in- and out-links also require the same time, approximately 22–33 hours would have been spent evaluating the anchors and destinations of out-links—the task that is of interest in this chapter.

This investment of manual judgment is admirable, but it should be noted that there is no redundancy: only one person inspected each link. The decision of whether a link should or should not be made is complex (Section 3.3.4 describes some of Wikipedia's exhaustive guidelines regarding linking) and inherently subjective. It is unclear what criteria the INEX evaluators had for making their decisions, whether they had the same criteria, and to what extent they agree with each other.

These limitations mean that it is disappointingly unclear which "ground truth" is more valid: the original Wikipedia links or those vetted by INEX evaluators. The first set is manually defined,[28] while the latter is automatically generated and then manually pruned. The former have been built and inspected by thousands of people, but primarily by casual readers who were presumably more interested in using the links than evaluating them. Many of the readers would not have had the expertise or time to correct the links. The latter set was evaluated more directly, but only a single person inspected each link. The INEX evaluators were all computer scientists while the documents spanned a host of

[28] Huang et al. (2009b) state that many of Wikipedia's links are automatically generated, but there is little evidence to support their claim (see Section 3.3.4).

disjoint topics from *Heteronormativity* to the *Mau Mau Uprising*. They may not have had the domain-specific knowledge required to make informed decisions.

The validity of these two data sets is an important question, because they differ strikingly from each other. Huang et al. (2009b) illustrated the difference by comparing corrected versions of Jenkinson et al. (2008) and Geva (2007)—two approaches considered representative of all participants of INEX—against the two ground truths. They found Jenkinson et al's approach to be superior when evaluating against Wikipedia's original links, but comparison against the links vetted by INEX participants shows no significant difference between the approaches. More alarmingly, the experiment compared Wikipedia's original links against the vetted ones, and found that these supposedly ideal examples fared no better than either of the automatic systems. From this, Huang et al. (2009b) conclude that using Wikipedia's existing links as ground truth is entirely unsound. We argue that the experiment casts equal suspicion on the ground truth that INEX's participants provide, given the limitations described above. Additionally, Section 6.1.2 explains that both of the automatic systems involved perform astonishingly well, which means either that the difficulty discerning between them and ground truth is entirely unsurprising, or something has gone awry with the experiment.

6.1.2 Existing approaches

One of the first attempts to automatically cross-reference documents with Wikipedia is the Wikify system developed by Mihalcea and Csomai (2007). This treats *detection* and *disambiguation* as two entirely separate problems. To detect the terms and phrases from which links should be made, Mihalcea and Csomai's most accurate approach is based on link probabilities obtained from Wikipedia's articles. Formally, the link probability of a phrase is defined as the number of Wikipedia articles that use it as an anchor, divided by the number of articles that mention it at all. The detection approach is to gather all n-grams for a document and retain those whose link probability exceeds a certain threshold. When tested on Wikipedia articles, the resulting anchor vocabularies matched the original markup with a precision of 53% and a recall of 56%.

To disambiguate the detected phrases and choose the most appropriate destination, Wikify's best approach extracts features from the phrase and its surrounding words (the terms themselves and their parts of speech), and uses a Naïve Bayes classifier to compare them with training examples obtained from the entire Wikipedia. When run over anchors obtained from Wikipedia articles, this is able to match the manually defined destinations with a precision of 93% and a recall of 83%. However, it requires a large preprocessing effort, because the entire text of all Wikipedia articles must be tagged for parts of speech.

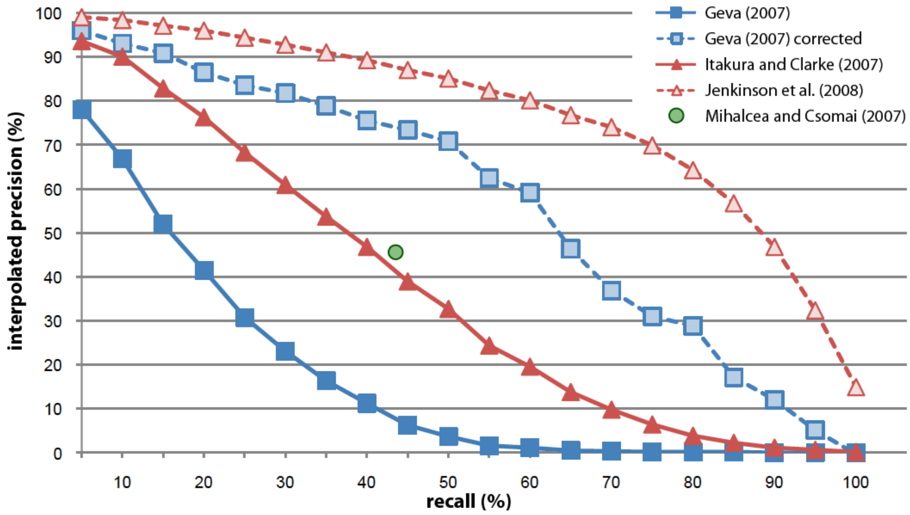


Figure 6.2: Performance of existing wikification techniques on INEX 2008 data

Wikify's two stages of detection and disambiguation were evaluated individually, but the combined result when both operated together was not reported. However, one can estimate Wikify's overall accuracy by assuming that disambiguation performance is constant across all terms and combining recall and precision across the two steps. This yields 50% precision and 46% recall.

Other existing approaches to wikification are part of the INEX Link-the-Wiki track, which was described in the previous section. According to Huang et al. (2009b), INEX participants fall into two broad categories: *page name analysis* and *anchor link analysis.* Figure 6.2 plots the precision and recall of four submissions at the 2008 competition, which exemplify the two basic approaches. The algorithms are described below.

Page name analysis (Geva 2007) involves searching documents for occurrences of article titles, including redirects. Matches are treated as potential links, which are ranked by emphasizing shorter titles over longer ones, with the exception of single-word terms (these are ranked last). The top 50 matches are used as links. Disambiguation is not necessary because article titles are unique, but using such a limited vocabulary causes many problems. The story in Figure 6.1, for example, would be a particularly troublesome. The corporations within it would only be known to the system as *Apple Inc.*, *Amazon.com* and *Palm.inc*, and therefore would not be detected. The approach achieved a mean average precision of 14% and ranked 13th at INEX 2008. A corrected version of this approach (shown in Figure 6.2) achieves vastly improved performance (53% mean average precision) with only minor modifications to case-folding, punctuation and tokenization. Unfortunately these modifications are described only briefly in Huang et al.

(2009b), and personal communication with the authors could not provide a satisfactory explanation for the extraordinarily large leap in performance.

Anchor link analysis, initially proposed by Itakura and Clarke (2007), is similar to the Wikify system described above. Rather than treat detection and disambiguation as separate problems, this approach merges the steps and uses Wikipedia's link statistics to solve both. It combines the probability that a particular phrase is used as a link (identical to Wikify's detection approach) with the probability of a particular article being the target destination of that link. Formally, the strength of an anchor/target pair is defined as the number of documents that link from the anchor phrase to the target article, divided by the number of documents in which the phrase appears. This approach, as Figure 6.2 illustrates, is a significant improvement upon the original page name analysis (Geva 2007) algorithm. It achieved a mean average precision of 34% at the 2008 Link-the-Wiki competition, and placed first in 2007.

The similarities between this approach and the Wikify system described previously provide a second opinion about its accuracy. Unlike Wikify, Itakura and Clarke's algorithm does not address the problem of disambiguation. For any given anchor, it will always select the same target article—the most common destination for the given anchor—regardless of context. Mihalcea and Csomai (2007) use this approach as a baseline for their disambiguation evaluation, and report that it achieves 87% precision and 78% recall. Combining this result with the detection phase yields an estimated performance of the Itakura and Clarke algorithm: approximately 46% precision and 43% recall. As one would expect, these figures are similar to the 2008 Link-the-Wiki experiment in Figure 6.2, which shows 38% precision at 45% recall. The difference can reasonably be attributed to implementation details (tokenization, case-folding, etc) and the datasets involved.

Jenkinson et al. (2008) was the best approach seen in the 2008 Link-the-Wiki track. It achieved a mean average precision of 73%—more than double that of its nearest rival. The score is surprising, however, because the algorithm is almost identical to Itakura and Clarke's and makes only minor modifications to the handing of capitalization and punctuation. From the description of the algorithm and the two independent experiments described above, one would expect Jenkinson et al. to achieve between 40% and 50% precision at 45% recall. Figure 6.2 shows that it instead achieves 87% precision at this point. Personal communication with Jenkinson et al. could not provide an explanation as to why the minor modifications result in such vastly improved performance.

Both Jenkinson et al. (2008) and the version of page name analysis found in Huang et al. (2009b) are conservative re-implementations of existing systems—Itakura and Clarke

(2007) and Geva (2007) respectively. Personal communication with the authors could not provide any explanation as to why the new implementations provide such remarkable improvements over their predecessors. Given the level of confusion and inconsistency surrounding these two algorithms and their evaluation, the remainder of this chapter treats Mihalcea and Csomai (2007) as the current state-of-the-art. It is certainly the most sophisticated approach.

6.1.3 Topic indexing

The problem of topic indexing is closely related to wikification. Here the aim is to identify the most significant topics; those that the document was written about (Maron 1977). These index topics can be used to summarize the document and organize it under category-like headings. Wikipedia is a natural choice as a vocabulary for obtaining index topics, since it is broad enough to be applicable to most domains. To use Wikipedia in this way, one must go through much the same process as wikification: first detect the significant terms being mentioned, and then disambiguate them to the appropriate topics. The only difference is an additional stage where the most important topics are identified.

Medelyan et al. (2008) make these similarities very clear in their approach to topic indexing with Wikipedia, and even reuse Wikify's approach for detecting significant terms. They differ in how they disambiguate terms, however, gaining similar results much more cheaply by balancing (a) the commonness (or prior probability) of each sense and (b) how the sense relates to its surrounding context. This approach is explained in Section 6.2.1, where it is improved upon by weighting context terms and using machine learning to balance commonness and relatedness.

6.1.4 Named entity recognition

Named Entity Recognition is a well-known problem in Information Extraction, in which systems annotate documents with occurrences of proper nouns (such as people, places, and organizations) and numeric expressions (e.g., dates, figures and monetary amounts). Nadeau and Sekine (2007) provide a recent survey.

This work aims for more exhaustive annotation than wikification or topic indexing, but each annotation is less exact; entities need not be resolved or disambiguated to specific concepts. At most, these systems classify entities under broad headings. For example, the competitions held as part of the Message Understanding Conference (MUC) require items to be tagged as an *organization*, *person,* or *location* (Chinchor and Robinson 1997).

Wikipedia's contributions to entity recognition are two-fold. Firstly, it provides an extensive vocabulary of entities and surface forms for these algorithms to check against and background information about entities that can be used to classify them (Kazama and Torisawa 2007a, Dakka and Cucerzan 2008). Secondly, its link markup provides an extensive corpus of documents in which entities have been manually annotated. Although Wikipedians are selective about the annotations they make (Section 3.3.4), their efforts can be manipulated and aggregated to provide good training corpora for entity tagging systems. Most state-of-the-art systems are machine-learned and rely heavily on expensive hand-annotated documents. Their performance varies greatly when they are trained and tested against different corpora. Nothman et al. (2009) demonstrate that training with Wikipedia can make them less brittle.

6.2 Learning to disambiguate links

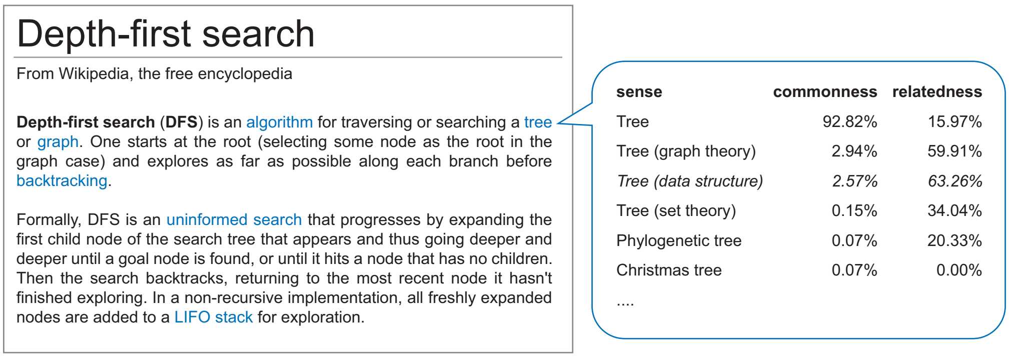
This section describes and evaluates a new approach to disambiguating terms that occur in plain text, so that they can be linked to the appropriate Wikipedia article. It seems odd to cover this problem first when the techniques described previously tackle the task of *detection*—recognizing terms that should be linked—before deciding where they should link to. This reflects one of the key differences of the new approach: it uses disambiguation to inform detection; thus this stage must be described first.

The new algorithm applies machine learning and uses the links found within Wikipedia articles for training. For every link, a Wikipedian has manually selected the correct destination to represent the intended sense of the anchor. There are millions of links, and each one represents several training instances. The connection between an anchor term and its chosen destination gives a positive example, while the remaining possible destinations provide negative ones. Figure 6.3 demonstrates this with the anchor *tree*: there are 26 possible senses (20 more than are shown in the table on the right). Only one sense is a positive example. The remaining 25 are negative.

Mihalcea and Csomai (2007) made exhaustive use of this vast source of training data; using it in its entirety. The processing effort required must be immense, particularly considering that many of the features require expensive, language dependant parsing to determine parts of speech. In contrast, the new approach is language independent, requires only a fraction of the training data, and all its features can be calculated cheaply.

6.2.1 Balancing commonness and relatedness

The two primary features used by the new algorithm are commonness (i.e. prior probability of a sense) and relatedness (the extent to which it connects to the surrounding context). With a few modifications, these features are the same as those used by Medelyan et al. (2008). The main contribution of the new algorithm is the way in which machine learning



Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hasn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	**commonness**	**relatedness**
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	*2.57%*	*63.26%*
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.00%
....		

Figure 6.3: Disambiguating *tree* using surrounding unambiguous links as context

is used to combine these features, so that the balance can be adjusted from document to document. The previous work instead used a fixed heuristic, determined in advance.

To review, the commonness of a sense is defined by the number of times it is used as a destination in Wikipedia for the particular anchor in question: Figure 6.3 shows that 93% of *tree* anchors link to the woody plant, 3% to the type of graph, and 3% to the computer science concept. The algorithm is predisposed to select the first of these senses rather than the more obscure ones, which go all the way down to *The Trees*, a song by the British rock band *Pulp*.

As Figure 6.3 demonstrates, the most common sense is not always the best one. Here *tree* clearly refers to one of the less popular senses—the hierarchical data structure—because it is surrounded by computer science concepts. These cases can be identified by comparing each possible sense with its surrounding context. This is a cyclic problem because these terms may also be ambiguous. Fortunately in a sufficiently long piece of text one generally finds terms that do not require any disambiguation at all, because 90% of Wikipedia's link anchors are only ever used to link to a single article. There are four unambiguous links in the text of Figure 6.3, including *algorithm*, *uninformed search* and *LIFO stack*. These unambiguous links can be safely used as context to disambiguate ambiguous ones.

Each candidate sense and context term is represented by a single Wikipedia article. Thus the problem is reduced to selecting the sense article that has most in common with all of the context articles; the same problem that was tackled in Chapter 5. Comparison of articles is facilitated by the Wikipedia Link-based Measure described in Section 5.4, which measures the semantic similarity of two Wikipedia pages by comparing their incoming and outgoing links. For the sake of efficiency the disambiguation algorithm (and the link detection system that follows) only considers the links made to each article. As Table 5.4 demonstrates, the in-link based (*Google-distance inspired*) measure is only marginally less accurate than the combination of both in- and out-links. The algorithm must make a vast number of comparisons, and this small sacrifice allows all the necessary information to be stored in memory. The relatedness of a candidate sense is the weighted average of its relatedness to each context article, where the weight of each comparison is defined below.

One of the main differences between this work and Medelyan et al. is that we do not consider all context terms to be equally useful. The word *and*, for example, has zero value for disambiguating other concepts, and yet Wikipedia contains an article about conjunctions to describe it. Mihalcea and Csomai's *link probability* feature helps to identify such cases; there are millions of articles that mention *and* but do not use it as a link. Weighting context terms according to this feature emphasizes those that are most likely a priori—ones that are almost always used as a link within the articles where they are found, and always link to the same destination.

A second difference derives from the fact that many of the context terms will not relate to the central thread of the document. For example, Figure 6.3 includes the term *goal,* which would provide good context in an article on *football*, but is unhelpful in this case. These potentially confusing outliers can be identified by their average semantic relatedness to all other context terms, using the measure described previously: *goal* does not relate to anything else in the document.

The two variables—link probability and relatedness—are averaged to provide a weight for each context term. This is then used when calculating the weighted average of a candidate sense to context articles.

To balance commonness and relatedness, it makes sense to consider how good the context is. If it is plentiful and homogenous—if the document has a clear theme—then relatedness becomes very telling. In Figure 6.3, for example, the most common sense of *tree* is entirely irrelevant because the document is clearly about computer science. However, if *tree* is found in a general document with ambiguous or confused context, then the most common sense should be chosen. By definition, this will be correct in most cases. Thus the final feature—context quality—is given by the sum of the weights that were previously assigned to each context term. This takes into account the number of terms involved, the extent to which they relate to each other, and how often they are used as Wikipedia links.

These features—commonness, relatedness, and context quality—are used to train a classifier that can distinguish valid senses from irrelevant ones. It does not actually choose the best sense for each term. Instead it considers each sense independently, and produces a probability that it is valid. If strict disambiguation is required, one can simply choose the sense that has the highest probability. If more than one sense may be useful,

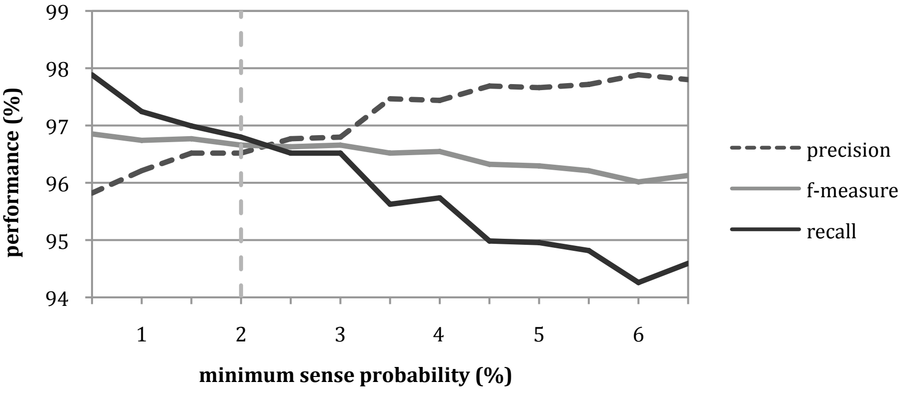


Figure 6.4: Disambiguation performance vs. minimum sense probability

one can gather all senses that have a higher probability of being valid than not. These options are evaluated in Section 6.2.3.

6.2.2 Configuration and attribute selection

Configuring the disambiguation classifier involves setting one parameter and identifying the most suitable classification algorithm. This parameter specifies the minimum probability of senses that the algorithm will consider. As illustrated earlier with the *tree* example, terms often have extremely unlikely senses that can be safely ignored. The distribution follows the power law: the vast majority of links are made to just a few destinations and there is a long tail of unlikely senses. *Jackson*, for example, has 230 senses, of which only 31 have more than a 1% chance of occurring. If all these are considered they must each be compared to all the context terms. Many unnecessary comparisons can be avoided by imposing a threshold below which all senses are discarded. This has the added advantage of increasing precision, since the discarded senses are unlikely to be relevant, but it decreases recall. Figure 6.4 plots this trade-off over the development dataset, and identifies 2% as a sensible probability threshold that balances the two metrics.

The Weka workbench[29] was used for all of the machine-learning experiments described in this chapter. Several classification algorithms were tested, and the results are shown in Table 6.1. were provided by the Weka. As one would expect, Naïve Bayes performs worst. There are dependencies between the features that lead this classifier astray. Surprisingly, C4.5 (Quinlan 1993) outperforms the more sophisticated Support Vector

[29] Weka is an open-source suite of data mining tools, and is available for download from *www.cs.waikato.ac.nz/ml/weka*

Machine, and consequently is used in the remainder of the chapter. Feature selection makes no difference, and bagging improves the classifier by only 0.3%.

6.2.3 Evaluation

To evaluate the disambiguation classifier, 11,000 anchors were gathered from 100 randomly selected articles and disambiguated automatically. Table 6.2 compares the result with three baselines. The first chooses a *random sense* from the anchor's list of destinations. Another always chooses the *most common sense.* The final baseline is the heuristic approach developed by Medelyan et al. (2008)

Having the classifier choose what it considers to be the *most valid sense* for each term outperforms all other approaches. The key differences between this and Medelyan et al. are the use of machine learning and the weighting of context. These provide a 76% reduction in error rate. The classifier never gets worse than 88% precision on any of the documents, and for 45% of documents it attains perfect precision. Recall is never worse than 75%, and perfect for 14% of documents. Recall can be increased by allowing the classifier to select *all valid senses* rather than just the most valid one for each anchor. Unfortunately this causes precision to degrade and makes for slightly lower overall performance. Consequently strict disambiguation is used throughout the remainder of this chapter.

Mihalcea and Csomai's best disambiguation technique had an f-measure of 88%. Direct comparison may not be fair, however, since their disambiguation approach was evaluated on an older version of Wikipedia. One could argue that the task gets more difficult over time as more senses (Wikipedia articles) are added, in which case it is encouraging that the new approach (which was run on newer data) yields better results. On the other hand disambiguation may well be getting easier over time. The baseline of simply choosing the most common senses has improved since Mihalcea and Csomai's experiments, which shows that common senses are becoming more and more dominant. Consequently any algorithm that is trained and tested on the newer documents—particularly one that uses commonness as a feature—will inherently have a higher accuracy. In any case, the new

	recall	**precision**	**f-measure**
Naïve Bayes	96.6	95.0	95.8
C4.5	96.8	**96.5**	96.6
Support Vector Machine	96.5	96.0	96.3
Feature selected C4.5	96.8	**96.5**	96.6
Bagged C4.5	**97.3**	**96.5**	**96.9**

Table 6.1: Performance of classifiers for disambiguation over development data

approach is competitive and has the distinct advantage of not requiring parsing of the text. This significantly reduces the resources required and, in principle, provides language independence. Additionally, the system requires much less training data (500 articles vs. the entire Wikipedia). On a modest desktop machine (with a 3Ghz Dual Core processor and 4Gb of RAM) the new disambiguator was trained in 13 minutes and tested in 4, after spending another 3 minutes loading the required summaries of Wikipedia's link structure and anchor statistics into memory.

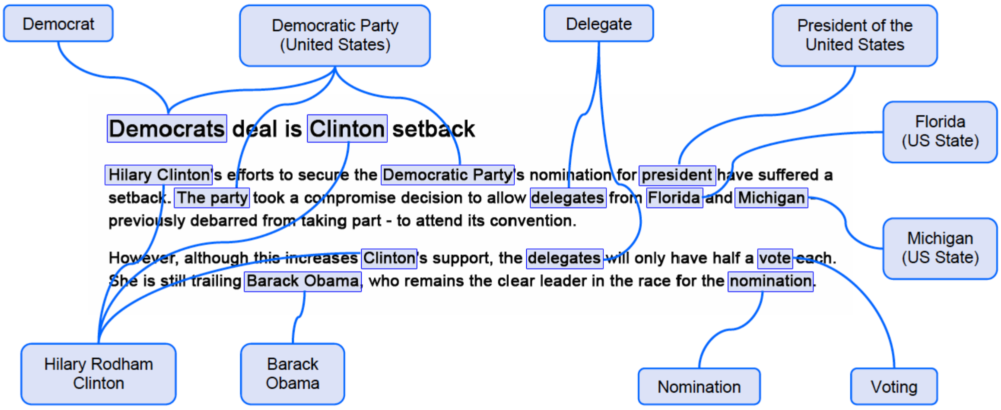
This evaluation can also be considered as a large-scale test of the Wikipedia link-based semantic relatedness measure described in Chapter 5. In this experiment, the testing phase alone involved more than two million comparisons in order to weight context articles and compare them to candidate senses. When these operations were separated out from the rest of the disambiguation process they were performed in 3 minutes (a rate of about 11,000 every second) on the above-mentioned machine.

6.3 Learning to detect links

This section describes a new approach to link detection. The central difference between this and the systems of Mihalcea and Csomai (2007) and Itakura and Clarke (2007) is that Wikipedia articles are used to learn what terms should and should not be linked, and the context surrounding the terms is taken into account when doing so. The previous approaches, in contrast, rely exclusively on link probability. If a term is used as a link for a sufficient proportion of the Wikipedia articles in which it is found, they consider it to be a link whenever it is encountered in other documents—regardless of context. This will always make mistakes, no matter what threshold is chosen. No matter how small a term's link probability is, if it exceeds zero there is (by definition) some context in which has been used as a link. Conversely, no matter how large the probability is, if it is less than one there is some context in which it should not be used a link. Consequently these systems will always discard relevant links and retain irrelevant ones, regardless of the chosen threshold. The new system yields better results by using link probability as just one feature among many.

	recall	**precision**	**f-measure**
Random sense	56.4	50.2	53.1
Most common sense	92.2	89.3	90.7
Medelyan et al. (2008)	92.3	93.3	92.9
Most valid sense	95.7	**98.4**	**97.1**
All valid senses	**96.6**	97.0	96.8

Table 6.2: Performance of disambiguation algorithms over final test data



Democrats deal is Clinton setback

Hilary Clinton's efforts to secure the Democratic Party's nomination for president have suffered a setback. The party took a compromise decision to allow delegates from Florida and Michigan previously debarred from taking part - to attend its convention.

However, although this increases Clinton's support, the delegates will only have half a vote each. She is still trailing Barack Obama, who remains the clear leader in the race for the nomination.

Figure 6.5: Associating document phrases with appropriate Wikipedia articles

6.3.1 Machine-learning for link detection

The link detection process starts by gathering all n-grams (up to 15-grams, the longest sequence of words used as an anchor in Wikipedia) in the document and retaining those whose link probability exceeds a very low threshold. This threshold—the value of which is established in the next section—is only intended to save time by discarding phrases that have effectively no chance of being useful. All the remaining phrases are disambiguated using the classifier described in the previous section. As shown in Figure 6.5, this results in a set of associations between terms in the document and the Wikipedia articles that describe them, which is obtained without any form of part-of-speech analysis. Sometimes, as is the case with *Democrats* and *Democratic Party,* several terms link to the same concept if that concept is mentioned more than once. Sometimes, if the disambiguation classifier found more than one likely sense, terms may point to multiple concepts. *Democrats*, for example, could refer to the *Democratic Party (United States)* or to any proponent of democracy (*Democrat*).

These automatically identified Wikipedia articles provide training instances for a classifier. Positive examples are the articles that were manually linked to, while negative ones are those that were not. Features of these articles—and the places where they were mentioned—are used to inform the classifier about which topics should and should not be linked. The features are as follows.

Link probability. The prior probability that a given term is used as a link anchor has proven to be a useful statistic. It forms the basis of both the Mihalcea and Csomai (2007) and Itakura and Clarke (2007) algorithms. Because each training instance may involve several candidate link locations (e.g. *Hilary Clinton* and *Clinton* in Figure 6.5), there are multiple link probabilities. These are combined into two separate features: the average and the maximum. The former is expected to be more consistent, but the latter may be more indicative of links. For example, the text *Democratic Party* has a much higher link probability than *the party*. As a matter of style, this document only refers to it once by its proper name. The fact that it was important enough to be referred to in full is a strong indication of link-worthiness, but this is lost when the probabilities are averaged.

Relatedness. Intuitively, one would expect that topics relating to the central thread of the document are more likely to be linked. *Clinton*, *Obama*, and the *Democratic Party* are more likely to be of interest to the reader than *Florida* or *Michigan*. Recall that the system has already gone to some lengths to obtain a relatedness score between each topic and its surrounding context, in order to disambiguate them. This provides a relatedness feature with no further computation. However, since the semantic relatedness comparisons are very cheap, this is augmented with a second feature: the average relatedness between each topic and all of the other candidates.

Disambiguation confidence. The disambiguation classifier described earlier does not just produce a yes/no judgment as to whether a topic is a valid sense of a term; it also gives a probability or confidence in this answer. This is used as a feature to help prune terms that are likely to have been misinterpreted. As with link probability, there may be multiple confidence values for each instance because several different terms may be disambiguated to the same topic. These are again combined as average and maximum values, for the same reasons.

Generality. Readers are more likely to be interested in specific topics that they may not know about, rather than general ones that require little explanation. The generality of a topic is defined as the minimum depth at which it is located in Wikipedia's category tree. This is calculated beforehand by performing a breadth-first search starting from the *Fundamental* category that forms the root of Wikipedia's organizational hierarchy.

Location and Spread. The remaining features are based on the locations where topics are mentioned; that is, the n-grams from which they were mined. *Frequency* is an obvious choice, since the more times a topic is mentioned the more important and link-worthy it is. Another is *first occurrence* because, as observed by David et al. (1995), topics mentioned in the introduction of a document tend to be more important. Significant topics are also likely to occur in conclusions, so *last occurrence* is also used. Finally the distance between first and last occurrences, or *spread*, is used to indicate how consistently the document discusses the topic. These last three location-based features are all normalized by the length of the document in words.

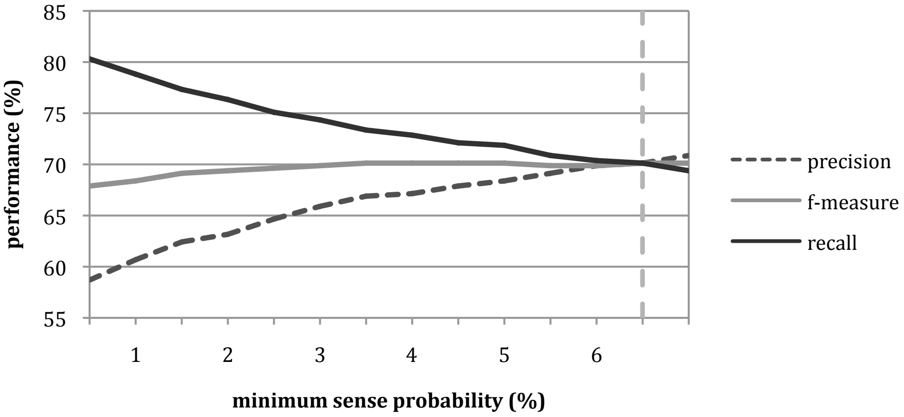


Figure 6.6: Link detection performance vs. minimum link probability

6.3.2 Training and configuration

As with the disambiguation classifier, three different sets of Wikipedia articles were set aside for training, configuration and evaluation. The same 500 articles used to train the disambiguation classifier are used for training here. This is done to reduce the number of disambiguation errors, because they directly affect the quality of training. As described earlier, terms must be disambiguated to the appropriate articles before they can be used as training instances. If a valid link were disambiguated incorrectly, many of its features would indicate a valid link, but the instance would be a negative example. Reusing the training data reduces the chance of these confusing examples occurring.

Likewise, configuration is done on the same 100 articles used to configure the disambiguation classifier. The only variable to configure is the initial link probability threshold used to discard nonsense phrases and stop words. This variable sets up a tradeoff with speed and precision on one side and recall on the other, since a higher threshold means that only the most likely instances are inspected, but risks discarding valid links. Figure 6.6 plots this tradeoff, and identifies 6.5% link probability as the point where precision and recall are balanced.

Despite the reuse of training data, the disambiguation classifier described in Section 3 performed quite poorly when used as part of the wikification classifier. It became very

	recall	**precision**	**f-measure**
Naïve Bayes	70.2	70.3	70.2
C4.5	**77.6**	72.2	74.8
Support Vector Machine	72.5	**75.0**	73.7
Bagged C4.5	77.3	72.9	**75.0**

Table 6.3: Performance of classifiers for link detection over development data

accepting, considering not just one or two senses to be valid for each term, but five or six. This is because the disambiguator was trained on links, but is being used here on raw text. In training, the context was restricted to manually defined anchors, but here it is mined from all unambiguous terms that have a link probability exceeding the initial threshold. The problem was resolved by modifying the disambiguation training to take these other unambiguous terms into account. This was achieved by taking all unambiguous n-grams with a link probability greater than 6.5%, and adding their destination articles to pool of context concepts described in Section 6.2.1. The resulting classifier was 1% worse (f-measure) when disambiguating links, but behaves more consistently when incorporated into the wikifier.

Table 6.3 lists the various classifiers that were tested. Naïve Bayes performs reasonably well, because the features are fairly independent. Again, C4.5 outperforms Support Vector Machine overall, although the latter attains higher precision. The evaluation described in the next section uses bagged C4.5 in order to gain the best results.

6.3.3 Evaluation

Evaluation of the link detector was performed over an entirely new randomly selected subset containing 100 Wikipedia articles. Ground truth was obtained by gathering the 9,300 topics that these articles were manually linked to. The articles were then stripped of all markup and handed to the link detector, which produced its own list of link-worthy topics for each article. This evaluation is only concerned with identifying the correct topics that should be linked to, not the exact locations from which these links should be made. This is consistent with Mihalcea and Csomai's work (which compared vocabularies of anchors, but not their locations) and the file-to-file evaluations at INEX.

The result is shown in Table 6.4, where recall, precision, and f-measure are all approximately 74%. There is a marked drop in performance between disambiguating links and detecting them, but this is to be expected. Deciding where a link should point to is far less subjective than deciding whether the link should be made at all. The time required is also significantly increased, even though many of the features are carried over from the disambiguator. The link detector was trained in 37 minutes, and tested (while simultaneously performing disambiguation) in 8 minutes.

	recall	**precision**	**f-measure**
Mihalcea and Csomai (2007) - *estimate*	46.5	49.6	48.0
New wikification algorithm	**73.8**	**74.4**	**74.1**

Table 6.4: Performance of wikification (detection and disambiguation) algorithms

Wikify's two stages of detection and disambiguation were evaluated individually, but Mihalcea and Csomai did not report the combined result when both operated together. In our approach the two stages are inseparable, which makes comparison difficult. As explained in Section 6.1.2, one can estimate Wikify's overall accuracy by assuming that disambiguation performance is constant across all terms, and combining recall and precision across the two steps. Table 6.4 compares this estimate against the new system, and shows a dramatic improvement. Recall is increased by 59%, precision by 50%, and overall f-measure by 54%. As with the previous experiment, a limitation of this comparison is that the two link detection approaches were developed and evaluated on different versions of Wikipedia. The difference between the datasets, however, is unlikely to make it any easier on the new system. The version of Wikipedia used here is much larger and probably more challenging than the one used by Michalcea and Csomai.

6.4 Wikification in the wild

All experiments described up until this point have treated Wikipedia as both training ground and proving ground. Even though training and testing sets have been kept separate, it is still reasonable to wonder whether the process works as well (or at all) on documents that are not obtained from Wikipedia. This section aims to address such concerns by applying our techniques to new documents, and subjecting the resulting link mark-up to manual evaluation.

6.4.1 Experimental data

The test set for this experiment is a subset of 50 documents from the AQUAINT text corpus: a collection of newswire stories from the Xinhua News Service, the New York Times, and the Associated Press. Documents were randomly selected from the last of these providers, restricting selection to short documents (250-300 words) to avoid overtaxing the attention spans of the human evaluators.

Another collection of 500 Wikipedia articles were used as training. The original intention was to use exactly the same set as in previous experiments, but unfortunately the difference in size between these verbose encyclopaedic articles and the short news stories produced a classifier that identified few link-worthy topics. A new training set was created by gathering all of the Wikipedia articles of the same length as the newswire stories, and selecting those that contained the highest proportion of links. The resulting classifier identified 449 link-worthy topics within the 50 newswire stories, an average of 9 links per document.

6.4.2 Participants and tasks

Mechanical Turk—a crowd sourcing service hosted by Amazon—was used to gather willing participants to inspect the wikified news stories. This service provides what Barr and Cabrera (2006) describe as *artificial artificial intelligence*; a way for human judgment to be easily incorporated into software applications. From the perspective of the people who develop these applications—who are known as *requestors*—the process is a function call where a question is asked and the answer is returned. What makes this system unique is the thousands-strong crowd of human contributors—or *workers*—who wait at the receiving end of the calls. These people identify the tasks they are interested in, submit their responses, and (pending review) receive payment for their efforts.

For our purposes, Mechanical Turk provided the means to conduct a labour-intensive experiment under strict time constraints, without having to gather participants ourselves. Naturally this raises some concerns about whether the anonymous workers could be trusted to invest the required effort and give well-considered responses. Even more alarming, it is possible for Mechanical Turk tasks to be done by automated "bots" created to gather funds for unscrupulous would-be workers (Howe 2006). Several checks were implemented to identify and reject low-quality responses and unqualified or poorly motivated participants. These are discussed in the following sections, which describe the two different types of tasks that the workers performed.

Evaluating detected links

The first type of task was used to evaluate the links that the system produced. For each of the 449 different links, the evaluator was given the text of the news article with one automatically generated link within it. The link was presented with a popup box that contained the first sentence of the relevant Wikipedia article. This allowed both the context of the link and its intended destination to be taken in at a glance. The participant was given the following options to specify whether the link was valid:

- No, <*link anchor*> is not a plausible location for a link.
- No, <*link anchor*> is a plausible location, but the link doesn't go to the right Wikipedia article.
- Kind of, <*link anchor*> is a plausible link to the correct Wikipedia article, but the article isn't helpful or relevant enough to be worth linking to.
- Yes, <*link anchor*> is a plausible link to the correct Wikipedia article, and this article is helpful and relevant.

Only the last option indicates that the link was detected correctly. The other three identify the different reasons why the algorithm made a mistake. The first indicates that a term or phrase should not have been considered as a candidate; the second identifies a candidate that was disambiguated incorrectly; and the third indicates a candidate that should have been discarded in the final selection stage. It should be noted that judging the helpfulness and relevance of a link is subjective. To do so, participants were asked to put themselves in the shoes of someone who was genuinely interested in the story, and judge whether the linked Wikipedia article would be worthy of further investigation.

The experiment is very similar to the manual assessments conducted at INEX. One major difference is that here each task was performed by three different people, to cope with subjectivity and verify individual responses. Another difference is that the participants used are anonymous workers, which raises some concerns. To ensure they were real people (rather than bots), tasks were paired with predetermined codes that had to be submitted alongside the answer. To ensure that participants gave well-considered answers, each code was only made available after the worker had spent at least 30 seconds inspecting the link and its surrounding context. An additional check was to only accept workers who had gained a high reputation from other requestors by having at least 90% of their responses to previous tasks accepted and rewarded. After rejecting and returning invalid submissions, responses were eventually gathered from 88 different people, who evaluated an average of 15 and a maximum of 156 links each. They spent an average of 1.5 minutes on each link, giving a total of 36 man-hours of labour. This is a similar investment of manual labour as at INEX (approximately 20 to 30 hours for file-to-file out-link assessment), but here it is spent gaining multiple assessments on smaller documents.

Identifying missing links

A second type of task was created to identify the links that the algorithm should have detected but failed to. In each of the 50 tasks (one for each document) the evaluator was given the news story with all of the detected links clearly identified. Again each link could be clicked to reveal a popup box that summarized the intended destination. Participants were asked to list any additional Wikipedia topics that they felt should be linked to, by supplying both the phrase where the link should start from and the URL of the Wikipedia article it should go to. They were asked not to add every single concept that was mentioned, since this is not what wikification aims to do. Instead they were instructed to only choose articles that were relevant for the news article, and were ones that readers would likely to want to investigate further.

The same checks were implemented as before to ensure that the answers were genuine and well-considered. Due to the increased difficulty and subjectivity of these tasks, each was conducted by five different participants, and the minimum time spent on them was increased to five minutes. After rejecting and returning invalid submissions, responses

correct	76.4%
incorrect (wrong destination)	0.9%
incorrect (irrelevant and/or unhelpful)	19.8%
incorrect (unknown reason)	2.9%

Table 6.5: Accuracy of the automatically detected links

were eventually gathered from 29 different people, who evaluated an average of 8.6 and a maximum of 35 documents each. In total they invested 47 man-hours of labour, or an average of 11 minutes per person per document. The INEX manual assessments do not have an equivalent task for identifying missing links; links are only considered if they are found within the original article or generated by one of the automatic systems.

6.4.3 Results

As is to be expected for subjective tasks, there was some disagreement between the evaluators. In the case of the first group of tasks this was unfortunately exacerbated by ambiguity. When an evaluator encountered a link that they felt was irrelevant, they had two equally valid responses available: they could say that the location of the link was implausible, or that the Wikipedia article it pointed to was unhelpful. This issue was resolved by combining the responses into a single option: that the link was irrelevant and/or unhelpful. Following this, 57% of the links received a unanimous decision from all three evaluators. Almost all of the remaining links received a two-vs.-one vote, for which the majority decision was considered correct. 3% of the links received different responses from all evaluators. Because there is only one possible response that indicates a valid link, these were judged to be incorrect—for an unknown reason.

Table 6.5 shows the results. The precision of the algorithm is 76%, meaning that 34% of the links were incorrect. Almost all the mistakes were due to incorrect candidate identification or selection, with only four links identified as being incorrectly disambiguated. As mentioned earlier, about 3% of the links were judged differently by all evaluators, and thus the reason for their rejection could not be identified.

For the second type of task, the evaluators identified just fewer than 400 distinct Wikipedia articles that they felt were worthy of linking to. This equates to around 8 additional links per document. Because of the subjectivity of the task, the participants did not entirely agree on the articles that were to be added. The majority (53%) of additional links were only identified by one of the participants. 17% were identified by two participants, 13% by three, another 13% by four, and only 4% were unanimously considered to be missing by all five participants. To compile the diverse opinions into coherent judgments, the majority (at least 3) of the participants were required to identify a link before it was considered link-worthy. This produced 117 links that the algorithm should have added to the documents, but did not.

The results of both sets of tasks were used to correct the original automatically-tagged articles and thereby generate ground truth. The four links that were identified as pointing to the wrong article were manually corrected by the authors. All remaining invalid links were simply discarded, and the missing links that were identified by the majority of participants were added. The result is a new corpus containing only manually-verified links, which is available online.[30]

Comparison of the original (automatically tagged) articles with this manually-verified corpus reveals the performance of the topic detector. As mentioned previously, precision is 76%; slightly better than when the system was tested on Wikipedia articles (Table 6.4). Recall is 73%; just one point worse than in the previous experiment. F-measure is 75%. Overall the figures are remarkably close to those obtained when the system was evaluated against Wikipedia articles, which indicates that algorithm works as well "in the wild" as it does on Wikipedia.

6.5 Discussion and implications

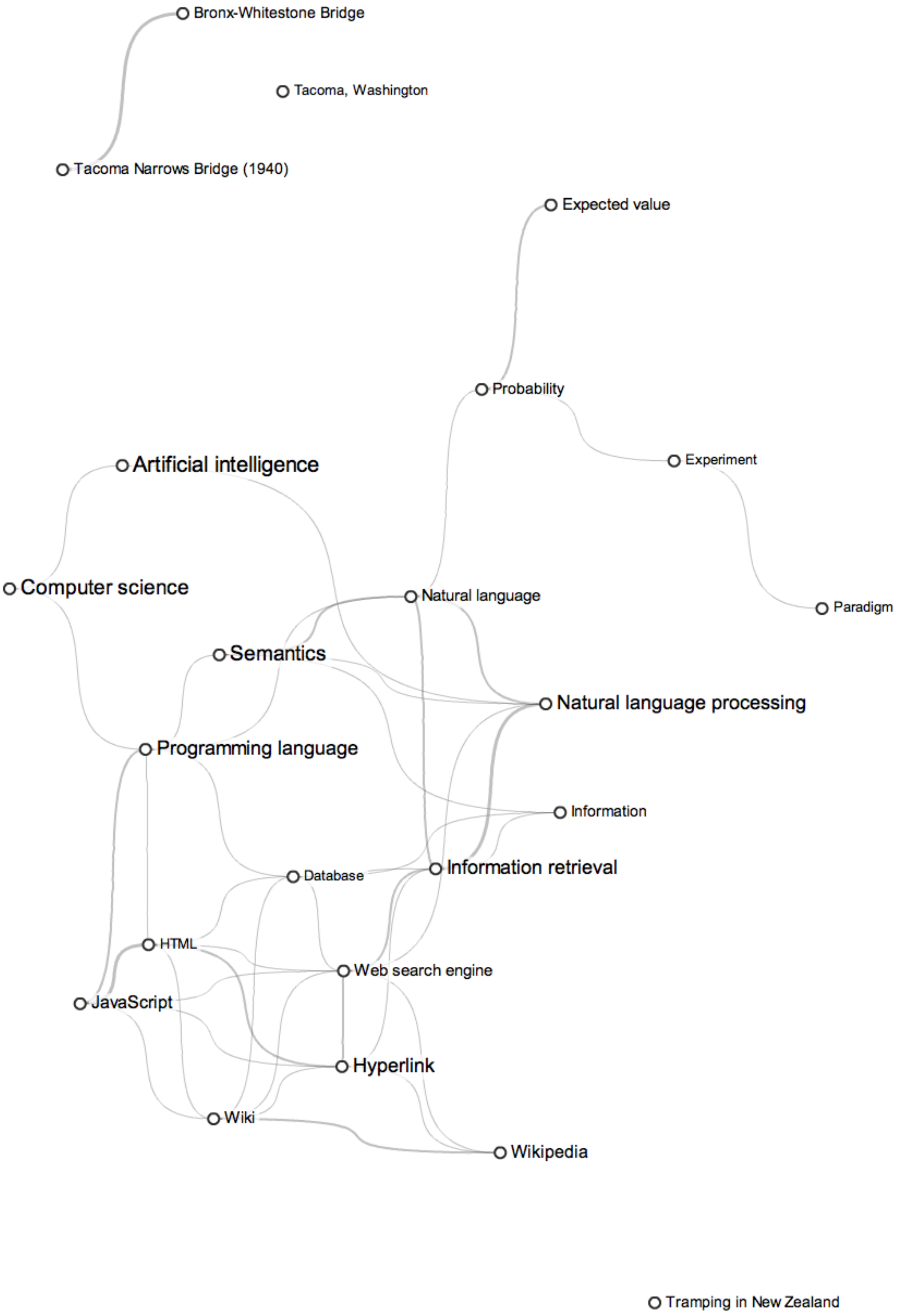
This chapter has described a new algorithm that disambiguates terms to their appropriate Wikipedia articles, and determines those that are most likely to be of interest to the reader. It is easy to imagine applications for it, such as adding explanatory links to news stories or educational documents, or detecting missing links in Wikipedia articles and smoothing the process for contributing to them. However, this barely scratches the surface of potential applications.

In essence, this is a tool that can cross-reference documents with one of the largest knowledge bases in existence. It can provide structured knowledge about any unstructured document, because it can represent texts as graphs of the concepts they discuss. To illustrate this, each chapter of this thesis begins with a graph of relevant topics, which have been automatically extracted using the algorithm described here. Without any form of manual cleanup, the Wikipedia articles that our algorithm considers most link-worthy have been added to the visualization. The size of each node corresponds to the algorithm's confidence in its prediction. The relations between them have been identified automatically by the WLM algorithm discussed in Chapter 5—again, the strongest relations have been added without cleanup. The layout is performed by a

[30] The manually verified and corrected corpus of wikified news articles is available at *www.nzdl.org/wikification*

modified version—developed for the Hōpara visualization described in Chapter 7—of the force-directed graph provided by the Prefuse Flare toolkit. This is the only part of the process that involves manual input: a minimal amount of rearranging was done for the sake of readability.

The graphs are not perfect. The one for this chapter is missing key concepts such as *wikification* and *disambiguation*. Nevertheless, they provide a clear sense of what each chapter is about. They resolve ambiguity, so none of the graphs mistakenly refer to *infrared* (from *IR*) or *neuro-linguistic programming* (from *NLP*). They do the same for polysemy, so it wouldn't matter if this chapter talked about *entity extraction*, *named-entity recognition*, *entity recognition* or *named-entity detection*; they are all roughly the same thing. By navigating the relationships of meaning between the topics, one can identify the threads of discussion; Chapter 3 has a thread surrounding *Wikipedia*, and another surrounding *Natural Language Processing*. All of this adds up to a machine-readable representation that has far-ranging applications. Any task that is currently addressed using the bag of words model, or with knowledge obtained from less comprehensive knowledge bases, could benefit from using the techniques described here to draw upon Wikipedia topics and their relations instead.


Bronx-Whitestone Bridge
Tacoma, Washington
Tacoma Narrows Bridge (1940)
Expected value
Probability
Experiment
Artificial intelligence
Computer science
Natural language
Paradigm
Semantics
Natural language processing
Programming language
Information
Information retrieval
Database
HTML
Web search engine
JavaScript
Hyperlink
Wiki
Wikipedia
Tramping in New Zealand

7. Augmenting retrieval over Wikipedia

This chapter describes the development and evaluation of a new search engine called Hōpara. Like Koru—the central component of this thesis—the purpose of this system is to investigate the utility of Wikipedia's structure and vocabulary for supporting and augmenting information retrieval. In Koru, the investigation requires Wikipedia's structure to be cross-referenced with the documents being searched. As the previous chapter explained, this task is inherently subjective and difficult to automate. Our algorithms to address it are imperfect, and can potentially have a large effect on the utility of the retrieval system. This chapter aims to sidestep the cross-referencing problem by only supporting retrieval within a document collection that has been manually connected to Wikipedia's structure: namely, Wikipedia itself. Hōpara is a search engine for Wikipedia.

Hōpara aims to make Wikipedia easier to explore by working on top of the encyclopaedia's existing link structure. It abstracts away from document content and allows users to navigate the resource at a higher level. It utilizes semantic relatedness measures to emphasize articles and connections that are most likely to be of interest, visualization to expose the structure of how the available information is organized, and lightweight information extraction to explain itself. The chapter is structured as follows: before describing the details of the system, we first discuss the challenges involved in navigating Wikipedia, and survey the work that has been undertaken to improve it. Sections 7.2 and 7.3 then describe Hōpara's interface and the algorithms behind it, respectively. The system is evaluated with a formal user study in Section 7.4. The final section discusses the limitations and implications of the experiment, and focuses on the lessons that can be applied to Koru and the broader investigation.

7.1 Browsing Wikipedia

Wikipedia is one of the most visited information resources on the planet, and represents a useful test bed for information retrieval research. It is large scale and open domain, and presents real challenges to the millions of people who use it to solve their information needs. This section provides a brief survey of how Wikipedia currently supports information retrieval, and the attempts that have been made to improve it.

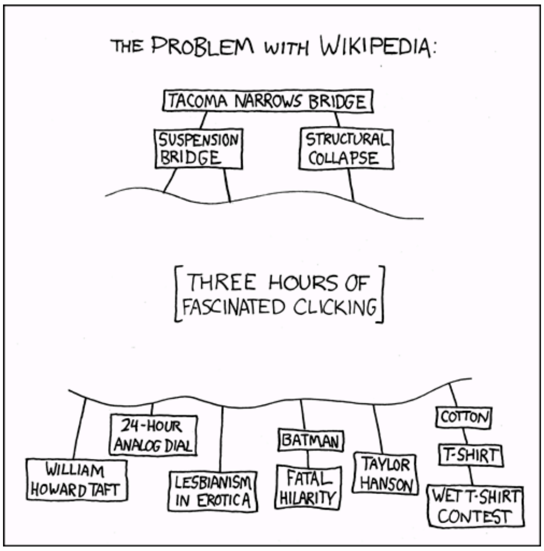


Figure 7.1: The problem with Wikipedia (from Munroe 2007)

7.1.1 Existing navigation features

Wikipedia's editors have implemented many initiatives to help corral and organize its content. Table 7.1 lists these features, and the proportion of Wikipedia that they help to organize or provide access to as of August 2009. These features were described in Section 3.3 with a focus on how they provide structured knowledge; this section concentrates on how they support retrieval. Unfortunately, to our knowledge there has never been a comprehensive usability study of Wikipedia and how its users go about locating information within it. Consequently this section is limited to discussing and quantifying the navigation features. They are revisited in Section 7.4.2, where our own usability study provides some insight into how they are actually put to use.

The most widespread and widely used navigation feature by far is its network of inter-article links. Anyone who has browsed the encyclopaedia has likely experienced the feeling of serendipitous exploration and discovery, following links from one interesting topic to the next and encountering information that they would never have searched for explicitly. Anywhere you go, there are attractive, even seductive options to go elsewhere.

Are Wikipedia's links too much of a good thing? The abundance of available paths can easily cause distraction and disorientation. This is humorously and succinctly illustrated in Figure 7.1, where cartoonist from the online web comic *xkcd* enters Wikipedia with specific information need: he wants to find out about the *Tacoma Narrows Bridge*, which famously twisted itself apart in mild winds in 1940. His forays are initially focused and relevant, but soon degrade into three wasted hours (at least in regards to the initial information need) of aimless wandering.

What does this say for hypertext in general? Wikipedia represents a preview of what large corpora would look like if automatic hypertext generation (Green 1999, Alfonseca et al. 2007) were able to match the accuracy of humans and gained widespread acceptance. It is extraordinarily large and densely interlinked, and almost every link has been constructed manually. Hypertext has long been championed as a way to improve navigation in large document collections, and Wikipedia is an excellent example of a large-scale, exhaustively linked resource. Yet information seekers still have difficulty navigating it. Scale introduces new challenges. At some point a hypertext system grows to such an extent that links alone are not enough to navigate the documents; something more is needed to navigate the links (Kim and Hirtle 1995).

All of the remaining navigational elements take the form of meta-pages: pages that have been constructed to describe and organize others. Categories, for example, are pages that organize articles hierarchically. As explained in Section 3.3.5, this produces a large directional graph with broad, general pages at the top and narrow, specific pages at the bottom. As Table 7.1 shows, these categories have become widespread since they were first introduced in 2006. From 96% of articles in Wikipedia, it is possible to navigate to one or more parent articles to generalize an investigation (to move, for example, from an article about *Computer Science* to a category of the same name, and from there up to *Applied Sciences* or down to *Algorithms*, *Data Structures* and other related topics).

List pages are articles that do not provide informative content on their own, but instead organize lists of links to other pages (Section 3.3.6). They are similar in function to categories, but are flat rather than hierarchical, and allow links to be explicitly ordered, grouped, and explained with free text. *List of Algorithms*, for example, organizes algorithms by purpose, such as *searching* and *sorting.* Table 7.1 shows that lists are not as widespread as categories: there are 53K different list pages (so users have a 2% chance of encountering them when searching for articles), and 40% of articles are organized (listed) by them. Lists are somewhat difficult to identify automatically. The number of instances in the table was identified by performing a breadth first search of the category

	instances	**invocations**	**articles referred to**
Inter-article links	68M	-	2.8M (91%)
Categories	550K	-	3.0M (96%)
List pages	53K	-	1.2M (40%)
Portals	1K	-	82K (3%)
Navboxes	27K	1.2M	430K (14%)
Infoboxes	4K	1.2M	440K (14%)

Table 7.1: Coverage of navigational elements in Wikipedia

hierarchy to gather all descendant categories of *Category:Lists* that include *list*, *index,* or *outline* in their titles, and gathering all of the pages that that belong to these. The article coverage statistics were calculated by gathering all distinct articles that these list pages link to.

Portals are intended to provide a hub or home page for a particular field (Section 3.3.7). The *Computer Science* portal, for example, showcases a couple of high-quality articles, organizes links to some of the most important articles into groups like *Hardware*, *Programming Paradigms*, and *Computer Scientists*, and lists some of the broadest, most important categories. As Table 7.1 indicates, the coverage of portals is sparse. There are less than a thousand, and less than 3% of Wikipedia's articles are linked to by them.

Templates are designed to allow content to be duplicated or consistently formatted across more than one page (Section 3.3.8). A Navbox is a specific type of template that is designed to provide lists or tables of links for navigating between groups of related pages. There are 27K different navboxes, which are invoked by 1.2M articles. In other words, one has a 40% chance of encountering them when browsing. These provide access to 14% of articles.

Infoboxes are another type of template, which have received much attention from the research community for their ability to express typed relationships. This is a by-product, however; their primary purpose is to provide factsheets or summaries of a topic, and to facilitate navigation between articles that share a class/instance or object/property relationship. There is a *Programming Language* infobox, for example, that expresses properties common to programming languages, such as *designed by*, *paradigm* (e.g. functional, object oriented) and *typing discipline* (strong or weak, static or dynamic). The infobox provides easy access to articles explaining what these properties and values are. Table 7.1 shows that the coverage of infoboxes is limited, despite their high profile in the research community. The 4K infoboxes are invoked by 1.2M articles, so one has a 40% chance of encountering them. They provide access to 14% of Wikipedia.

Overall, Wikipedia's adoption of navigation tools has been fairly cautious. All of the features described above are carefully, manually crafted. Wikipedia's reliance on manual labour has been a boon for computer science researchers (for whom manually-defined semantics are always in short supply) but may not be the best way forward for the resource itself. Wikipedia's volunteer workforce is almost inevitably geared towards content creation rather than content management. People are easily motivated and readily equipped to share what they know; but fewer people have the librarian bent; the willingness and technical skill to organize someone else's work. As a consequence the coverage of navigation features—with the exception of categories and inter-article links—is patchy and sparse and will probably remain so indefinitely. To move forward, it seems necessary to find new ways to efficiently make use of the effort that has been expended already. This thesis is about how Wikipedia and its structure can support new interfaces for information retrieval, but there is much potential for the flow to be reversed.

7.1.2 Alternative interfaces

Because Wikipedia's content is freely available, there have been several attempts to provide alternative means of accessing it. The most well-known venture of this kind is the PowerSet search engine,[31] which uses the Freebase ontology (Section 2.2.1) to augment Wikipedia. This commercial system is intended to be a showcase of natural language processing and artificial intelligence. It focuses on query interpretation (i.e. question answering and fact finding) and summarization. The only information seeking facilities it provides are oriented towards navigating within individual articles, rather than navigating across the resource. Consequently the system has limited relevance to this investigation.

The remaining systems that repackage Wikipedia are all examples of information visualization. As explained in Section 2.3.6, visualization has been more successful in supporting analysis—exposing and communicating general trends, patterns and outliers in data—rather than retrieval. This same imbalance is seen in the visualization work that is specific to Wikipedia.

On the *analysis* side, Wikipedia provides many interesting properties to visualize. Holloway et al. (2007) investigate the semantic coverage and bias of Wikipedia by visualizing its categories and authors. Chris Harrison has produced several (unpublished) visualizations, including timelines of article popularity[32] and graphs of categories and their connections.[33] Bruce Herr, Todd Holloway, and Katy Borner represent Wikipedia's topic coverage as large mosaics.[34] Visualization of Wikipedia's edit history—the flow of contributions, revisions and reverts that make up each collaboratively produced article—has been particularly fruitful. History flows (Viégas et al. 2004), tilebars (Gawryjolek and Gawrysiak 2007), chromograms (Wattenberg et al. 2007) and timelines (Nunes et al.

[31] *http://www.powerset.com/*

[32] *http://www.chrisharrison.net/projects/wikitop*

[33] *http://www.chrisharrison.net/projects/clusterball*

[34] *http://www.scimaps.org/maps/wikipedia*

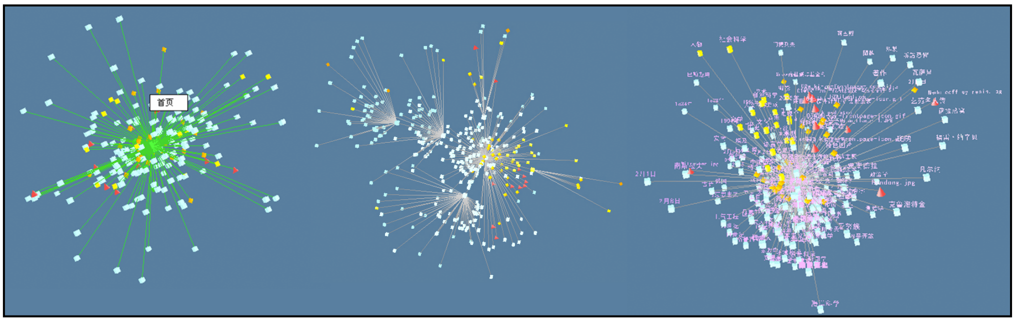


Figure 7.2: A selection of Wikivis visualizations

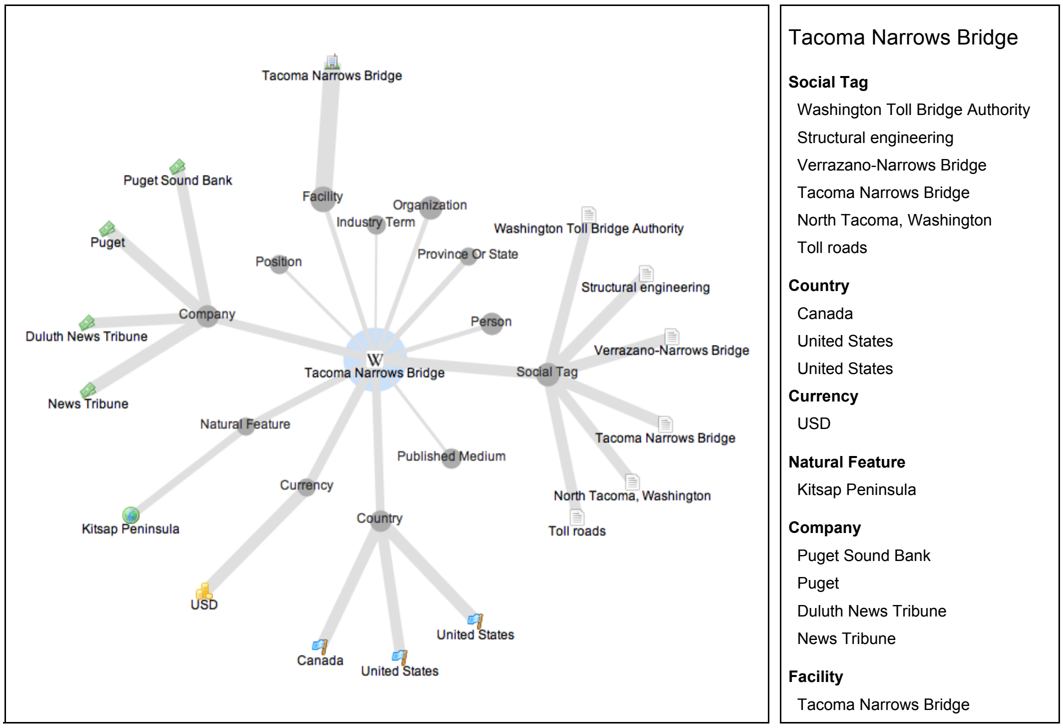
2008) have all been used to expose vandalized, contentious or under-developed articles and assist with conflict resolution.

Visualization of Wikipedia for the purposes of *retrieval* has been much less fruitful. To our knowledge there are only two relevant studies. The first, by Biuk-Aghai (2006), experiments with various layouts for visualizing articles and the connections between them as 3D graphs. Unfortunately the resulting system—WikiVis—received no evaluation. Anecdotally, its visualizations (Figure 7.2 shows several typical examples) are illegible at first glance, and their utility for searching is questionable.

The second relevant system, developed by Hirsch et al. (2009a), is ThinkPedia. Figure 7.3a shows the visual component of this system. Not shown is a search box (in which the user has entered *Tacoma Narrows Bridge)*, a list of closely related queries (e.g. *Tacoma Narrows Bridge Collapse*), and a frame containing the entirety of the relevant Wikipedia article. The interface uses the professional visualization toolkit ThinkMap,[35] and is consequently much cleaner and more legible than WikiVis. Nodes and relations are clearly labelled and can be readily identified. Relations are grouped so that similar topics (e.g. the companies *Puget Sound Bank* and *Duluth News Tribune*) are collocated spatially.

However, the same properties—legibility and spatial organization—can be achieved trivially without visualization. Figure 7.3a communicates the same information as Figure 7.3b. It seems reasonable to expect the visualization to convey something more, as payoff for the unfamiliarity and technical difficulties it causes—additional software requirements, browser navigation and bookmarking issues, etc. The visualization is also much less space-efficient than the list-based alternative, and scales less easily. A slider is provided to expand the graph (to see more of the *Company*, *Person* or *Facility* groups, etc), but this causes layout issues (overlapping labels and edges) and necessitates panning. Additionally, many of the suggestions ThinkPedia makes are of questionable

[35] *http://thinkmap.com*



Tacoma Narrows Bridge

Social Tag
Washington Toll Bridge Authority
Structural engineering
Verrazano-Narrows Bridge
Tacoma Narrows Bridge
North Tacoma, Washington
Toll roads

Country
Canada
United States
United States

Currency
USD

Natural Feature
Kitsap Peninsula

Company
Puget Sound Bank
Puget
Duluth News Tribune
News Tribune

Facility
Tacoma Narrows Bridge

a) Original interface b) Alternative interface

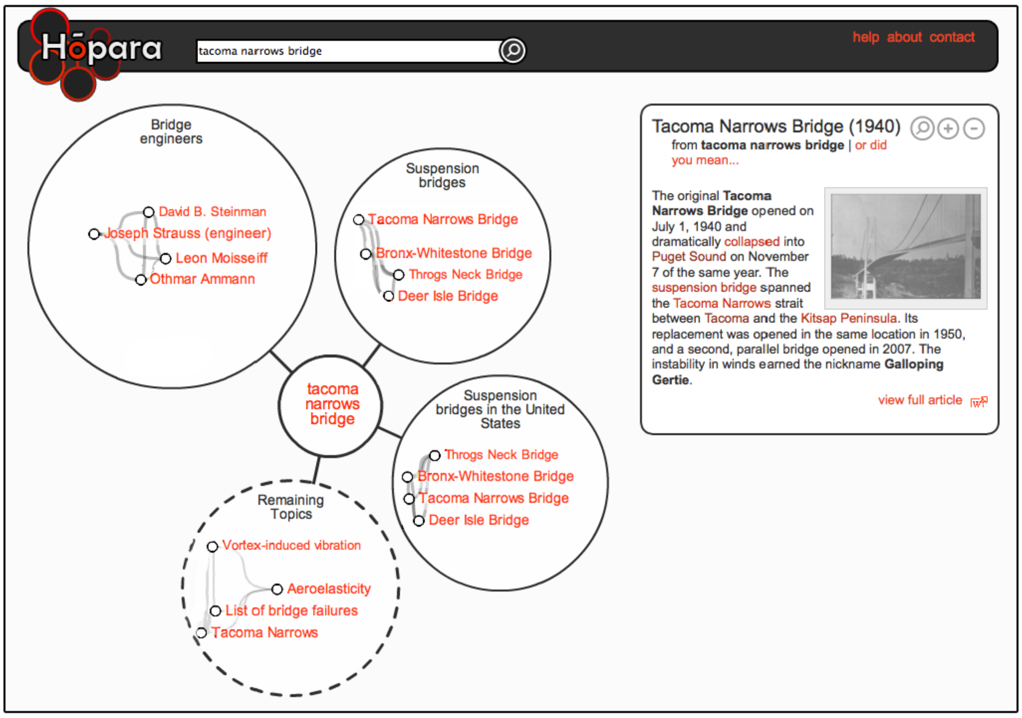
Figure 7.3: A ThinkPedia visualization of topics related to *tacoma narrows bridge*

use. Some are irrelevant (e.g. *USD* has no particular significance, and *Puget* links to a French commune). Others are redundant (e.g. *United States* is listed twice and is extremely general, *Tacoma Narrows Bridge* is listed three times) or incomplete (relevant natural features such as the *Tacoma Narrows Straight* are not listed, even when the graph is fully expanded).

Unfortunately, ThinkPedia has never been a formally evaluated. In anecdotal studies, users found the tool to be “beautiful” and “fun”, but also “disordered” and “frustrating” to search with (Hirsch et al. 2009a). The developers have had more success in constructing similar interfaces to more structured resources, such as the semantic database Freebase and the corporate wiki Confluence (Hirsch et al. 2009b).

The primary metaphor for both WikiVis and ThinkPedia is to visualize a subset of Wikipedia as a graph, where the nodes are concepts and the edges between them indicate relatedness. Both systems must draw connections between articles, and both systems abandon Wikipedia’s manually defined inter-article links in favour of automatically generated connections. WikiVis uses co-authorship analysis to identify related articles, while ThinkPedia uses the SemanticProxy web service.[36] One of the key ideas of this

[36] *http://semanticproxy.com*



Tacoma Narrows Bridge (1940)

from **tacoma narrows bridge** | or did you mean...

The original **Tacoma Narrows Bridge** opened on July 1, 1940 and dramatically collapsed into Puget Sound on November 7 of the same year. The suspension bridge spanned the Tacoma Narrows strait between Tacoma and the Kitsap Peninsula. Its replacement was opened in the same location in 1950, and a second, parallel bridge opened in 2007. The instability in winds earned the nickname **Galloping Gertie**.

view full article

Figure 7.4: The Hōpara search engine, with topics related to *tacoma narrows bridge*

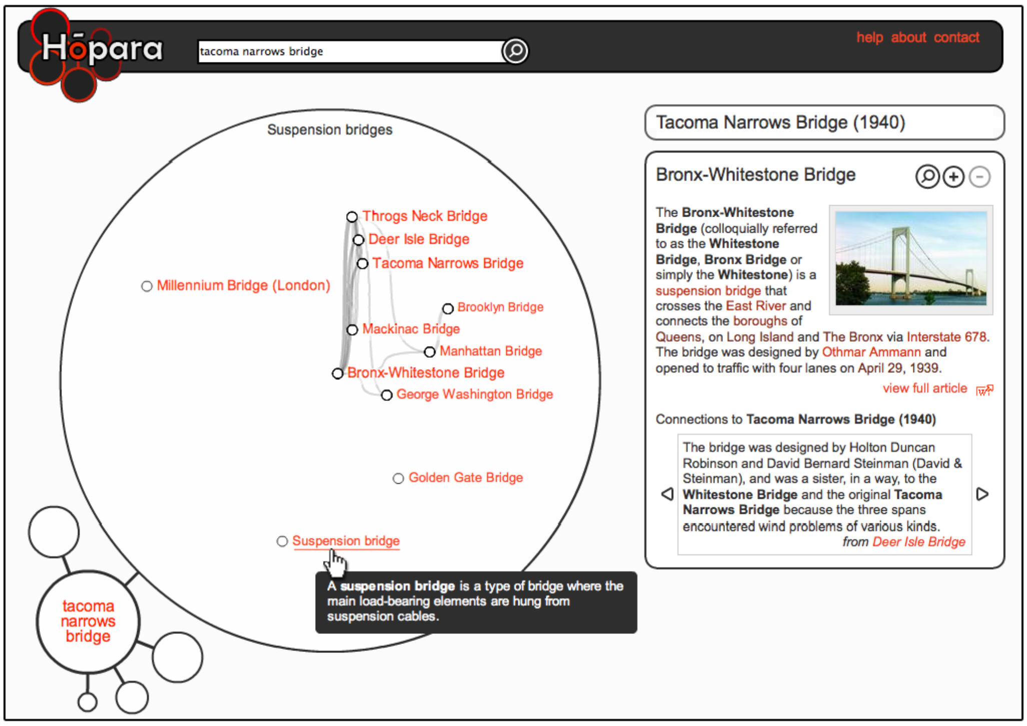
chapter is to instead use the connections provided directly by Wikipedia to help readers navigate from one article to the next.

7.2 Hōpara

The Māori who traversed the Pacific Ocean to discover New Zealand were skilled navigators, and Hōpara is their word for exploration. The Hōpara system, shown in Figure 7.4, is a new search engine that applies semantic relatedness, information extraction and visualization to Wikipedia.[37] The key idea behind Hōpara is to take Wikipedia's existing features and use them to facilitate interactive retrieval—to make the resource easier to explore without requiring its volunteer editors to expend further effort.

The upper area of the Hōpara interface is a classic search box where the user has entered *Tacoma Narrows Bridge*; the cartoonist's query from Figure 7.1. Below and to the left of the search box is a visualization of the related topics: the engineers involved, other similar bridges, and some of the engineering concepts connected to its demise. To the right is an extract of the relevant article.

[37] The reader is strongly encouraged to try using Hōpara for themselves. It exhibits high degrees of interactivity and animation that cannot be easily expressed in text. The system is available online at *http://www.nzdl.org/hopara*



A suspension bridge is a type of bridge where the main load-bearing elements are hung from suspension cables.

Tacoma Narrows Bridge (1940)

Bronx-Whitestone Bridge

The **Bronx-Whitestone Bridge** (colloquially referred to as the **Whitestone Bridge**, **Bronx Bridge** or simply the **Whitestone**) is a suspension bridge that crosses the East River and connects the boroughs of Queens, on Long Island and The Bronx via Interstate 678. The bridge was designed by Othmar Ammann and opened to traffic with four lanes on April 29, 1939.

view full article

Connections to **Tacoma Narrows Bridge (1940)**

The bridge was designed by Holton Duncan Robinson and David Bernard Steinman (David & Steinman), and was a sister, in a way, to the **Whitestone Bridge** and the original **Tacoma Narrows Bridge** because the three spans encountered wind problems of various kinds.

from Deer Isle Bridge

Figure 7.5: The Hōpara search engine, focusing on *suspension bridges*

The query is ambiguous: Wikipedia contains an article about the bridge that famously collapsed in 1940, and another about its replacement. When queries have multiple interpretations, the system selects one sense automatically but makes the remaining senses available with the link entitled *or did you mean*. The system also allows for synonymy: e.g. the query *galloping gertie* (the bridge's nickname) takes the user to the same result. The process by which senses and synonyms are identified and resolved is described in Section 7.3.

The visualization on the left of the interface displays the user's query in the centre, and four categories or groupings of suggestions surrounding it. Within each grouping is a small graph whose nodes are topics and edges are semantic relations between them. Larger topics, such as *Leon Moisseif* (the project's lead engineer) are more strongly related to the query. An edge between two topics indicates that they are semantically related to each other, and thicker edges indicate stronger relations.

Only four groupings of topics can be shown at a time. Moving clockwise from the top left corner of Figure 7.4, the first three represent the categories containing the strongest, most relevant suggestions. Their size represents their expected value to the searcher. The fourth grouping, indicated with a dotted outline, shows the best topics that did not belong to the top categories. If this is clicked, the visualization smoothly rotates to reveal the remaining, less relevant categories. Thus the system can scale to show many categories of topics, without panning or zooming.

Similarly, only the four best topics are shown within each grouping. If more are available, then scalability is achieved by allowing categories to be expanded. If the user clicks *Suspension Bridges* in Figure 7.4, for example, then the system smoothly animates to the layout shown in Figure 7.5. This dedicates much more space to the category. Layout within a category is based on a force-directed graph that encourages related topics to be clustered together spatially, and others (such as the general topic *Suspension Bridge,* or *Millennium Bridge*—the only one not located in the U.S.) to be separated.

Mousing over any topic link reveals a tooltip containing the first sentence of the article in question, as shown for *Suspension Bridge* in Figure 7.5. The user can click on any topic to open a box on the right side of the interface, as shown for *Tacoma Narrows Bridge (1940)* in Figure 7.4 and *Bronx-Whitestone Bridge* in Figure 7.5. The box contains the first paragraph and an image extracted from the article, and a link to Wikipedia. It also contains sentence snippets to explain how the topic relates to the original query. In Figure 7.5 this reveals that the *Bronx-Whitestone Bridge* and the *Deer Isle Bridge* faced similar design issues as *Galloping Gertie,* which explains why Hōpara emphasizes them over the other bridges. The methods for judging the relevance of suggestions and extracting the snippets to explain them are described in Section 7.3.

On the top right corner of the *Bronx-Whitestone Bridge* topic box in Figure 7.5 is a set of three buttons that control how it can be investigated further. The first explores it as a new query. The second adds it to the current query, to explore things that relate to or connect both bridges. The third button removes it from the query—it is disabled in Figure 7.5 because *Bronx-Whitestone Bridge* is not part of the current search.

Multi-topic queries can also be built directly in the search box. For example, the query *bridge failure* is automatically recognized as two distinct topics: *Bridge* and *Structural failure.* The interaction is much the same as for the mono-topic query in the figures, except that the visualization on the left is narrowed to contain only suggestions that relate to (or bridge between) both query topics (e.g. *Catastrophic failure*, *Structural design*), and multiple connection snippets are shown within each topic box (because there are multiple query topics to connect to).

Care has been taken to minimize the negatives introduced by the system's technical requirements. Most of the interface is built using standard HTML elements under the AJAX framework. The visualization itself is implemented using the Prefuse Flare toolkit,[38] and is seamlessly integrated into the page. It requires only a browser with JavaScript and Flash enabled (both are almost ubiquitous). Unlike many similar systems, browser navigation (history, bookmarking, etc.) is preserved.

7.3 Mining Wikipedia's structure

The backend of Hōpara is built upon the efforts described in previous chapters. Searching is provided by indexing article titles, redirects and the anchor texts found within inter-article links. As Section 3.3.4 explained, anchors are particularly useful for this because they encode both synonymy and ambiguity. For example, Hōpara knows that *Galloping Gertie* and *Original Tacoma Narrows Bridge* are synonymous because both are used as anchors to the same article. It also knows that *Tacoma Narrows Bridge* is ambiguous and which sense is most likely of interest, because the anchors containing this phrase go to two different locations; 43% to *Galloping Gertie* and 57% to its replacement.

Resolving multi-topic queries, such as *suspension bridge failure,* is a similar problem as detecting and disambiguating links in free text, which was addressed in Chapter 6. Here detection is simpler because every match between an n-gram and an anchor is used, unless the n-gram is a stopword or is entirely subsumed by a longer match. Disambiguation is hampered because there is little or no context. If a query contains just one ambiguous term, then the sense with the highest prior probability is chosen. If there is more than one term, then the *sequential decision* disambiguation strategy described in Section 5.4.2 is used.

To suggest related topics, Hōpara gathers all articles that link to the query articles or are linked by them. The resulting topics are ranked by how strongly they relate to all query topics using the semantic relatedness measure described in Chapter 5. Only those that are sufficiently related are used, to avoid sending the user off topic. To review, the relatedness measure works by comparing any two Wikipedia articles by their incoming links. The formula (repeated from Section 5.4.1) is:

$$relatedness(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$$

where *a* and *b* are the two articles of interest, *A* and *B* are the sets of all articles that link to *a* and *b* respectively, and *W* is set of all articles in Wikipedia. The topic suggestion process probably sounds expensive and slow—links to and from query topics are gathered as suggestions, and links to these are in turn used for ranking—but is sufficiently responsive in

[38] *http://flare.prefuse.org*

practice. Approximately 11,000 relatedness measures can be made per second on a 3GHz dual core machine.

Having gathered and ranked the suggested topics, the next task is to organize them. Each article in Wikipedia is manually tagged with one or more categories. Hōpara gathers the relevant categories for the suggestions and ranks them by the total strength (relatedness to query topics) of the top five suggestions within them. Categories containing fewer than three suggestions are discarded, and the remaining uncategorized suggestions form a *Miscellaneous* group, which is presented last.

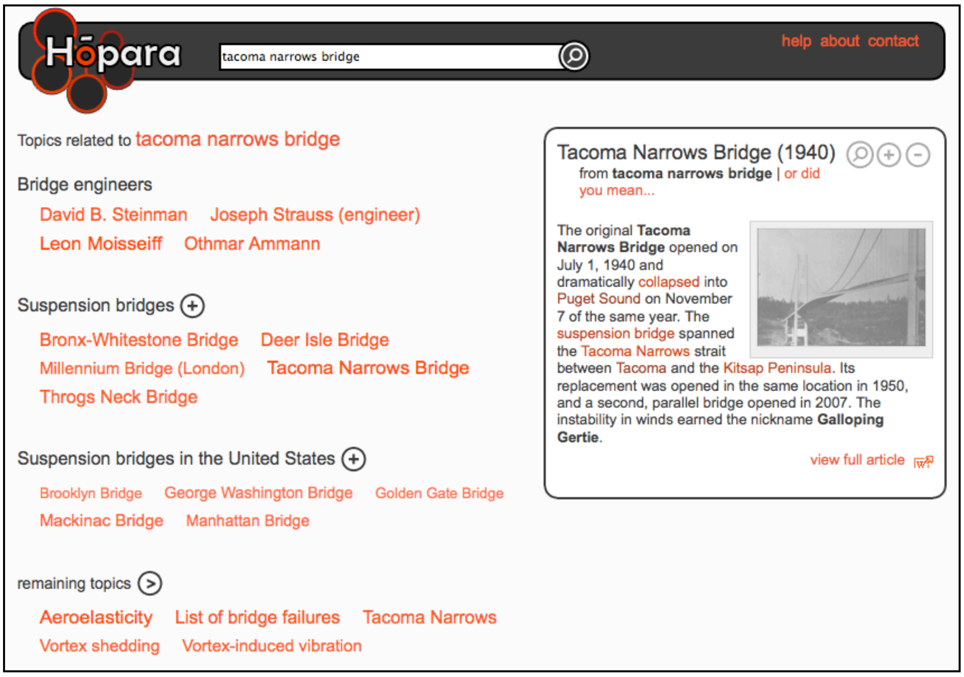
To explain connections between two topics, say *a* and *b,* Hōpara gathers sentences in *a* that mention *b*, sentences in *b* that mention *a*, and sentences in other articles that mention both. A fortunate side effect of the relatedness measure is that all the articles that could possibly contain these sentences are known in advance: this is the set $A \cap B$ in the formula above. The exact locations of links to *a* and *b* within these articles are identified at run-time using regular expressions, and sentence boundaries surrounding them are identified at the same time with a simple rule-based tagger. Both link occurrences and sentence boundaries could be indexed beforehand, but this does not seem to be necessary in practice: they are cheap enough to locate on the fly.

7.4 Evaluation

A user study was undertaken to evaluate Hōpara and its underlying algorithms for their ability to facilitate exploratory search. The study compares three systems: the incumbent Wikipedia interface, the full Hōpara system described in the previous section, and a baseline system that packages the same functionality in a more familiar interface.

Wikipedia's support for exploration is already better than most information sources. The various features described in Section 7.1.1—extensive inter-article links, portals to organize articles thematically, a category network to organize them hierarchically, and various templates and list pages to group related articles together—add up to a robust, challenging baseline.

As explained in Section 2.3.6, information visualization for the purposes of retrieval has had limited success in the past. It is easy to construct visualizations that do more harm than good. To be successful, they have to provide some kind of payoff while minimizing the increased complexity and unfamiliarity they cause. To isolate the effects of Hōpara's interface, a baseline system was constructed to provide the same functionality without visualization. As Figure 7.6 demonstrates, it is identical to the



Hōpara

tacoma narrows bridge

help about contact

Topics related to tacoma narrows bridge

Bridge engineers

David B. Steinman Joseph Strauss (engineer)
Leon Moisseiff Othmar Ammann

Suspension bridges ⊕

Bronx-Whitestone Bridge Deer Isle Bridge
Millennium Bridge (London) Tacoma Narrows Bridge
Throgs Neck Bridge

Suspension bridges in the United States ⊕

Brooklyn Bridge George Washington Bridge Golden Gate Bridge
Mackinac Bridge Manhattan Bridge

remaining topics ⊳

Aeroelasticity List of bridge failures Tacoma Narrows
Vortex shedding Vortex-induced vibration

Tacoma Narrows Bridge (1940)
from **tacoma narrows bridge** | or did you mean...

The original **Tacoma Narrows Bridge** opened on July 1, 1940 and dramatically collapsed into Puget Sound on November 7 of the same year. The suspension bridge spanned the Tacoma Narrows strait between Tacoma and the Kitsap Peninsula. Its replacement was opened in the same location in 1950, and a second, parallel bridge opened in 2007. The instability in winds earned the nickname **Galloping Gertie**.

view full article

Figure 7.6: A tag-based interface to Hōpara

system described previously (Figure 7.4) in every way, except that the graphs of related topics (the circular categories) are each replaced with tag clouds; alphabetically ordered lists where the size of each item indicates its relatedness to the query. Comparison between the full and baseline systems provides direct insights into whether Hōpara's visualization provides a positive difference.

For the remainder of the chapter, the three systems—Wikipedia, Hōpara, and the tag-cloud baseline—will be referred to as *wiki*, *vis* and *tag* respectively.

7.4.1 Subjects and tasks

Twelve participants were observed as they interacted with the three systems. All were experienced searchers participating in a graduate level computer science course. All but one was a regular visitor to Wikipedia. Sessions typically lasted for one and a half hours, and were conducted in a controlled environment with video and audio recording and an observer present. Data was also collected from questionnaires and system logs.

Each user performed the three tasks shown in Table 7.2 by gathering the Wikipedia articles they felt were relevant. Rather than repeat the tasks used in Section 4.5, new tasks were created to suit Wikipedia. These were modelled after Kules and Capra (2008) to be open-ended, multifaceted, and provide imaginative context for the participant to relate to. Performing a task amounted to using one of the systems, building a list of relevant article titles, and supporting each selected article with a short sentence (either copied from the system or typed freely) to explain its relevance. Each task was performed on a different system and the order in which the systems were used was staggered to counter the effects of bias and transfer learning. Participants were given as much time as they needed to familiarize themselves with the system (typically 5-10 minutes) before being given a task, but were specifically asked to spend 15 minutes performing each one.

7.4.2 Results

This section compares the three systems—*wiki*, *tag* and *vis*—on the basis of the behaviour observed, objective measures of task performance and subjective impressions gathered from questionnaires.

User behaviour

The exploratory search tasks were specifically selected to encourage user interaction, and participants were invariably forced to issue several queries in order to perform each one. This section takes a close look at how participants used the various navigation features provided by the *wiki*, *tag* and *vis* systems in order to complete these tasks.

Inter-article hyperlinks provide the primary navigation feature for all three systems. Participants were reluctant to follow links within the *wiki* system. Figure 7.7 shows that wiki users opened an average of only 10 articles per task and most of these came directly

a) Walking around the South Island Imagine you are a tourist travelling in the South Island of New Zealand. This area is world-famous for its natural beauty; a great venue for a hiker! What hiking trails and parks should you know about? Which landmarks would you like to see, and which areas would you particularly want to visit?
b) Windy Jazz Imagine you are a trumpet player who typically gets stuck in the back of a 50-piece orchestra. You'd like to play in a smaller group, with more room for expression. What about playing Jazz instead? Which wind instruments are commonly used in Jazz, and who are the famous musicians who play them? Are there specific types of jazz where wind instruments are particularly common?
c) Keeping New Zealand green Imagine writing environmental policies for the New Zealand Government. New Zealand may have a clean green image, but many of its iconic species are under threat. Which New Zealand species have been made extinct recently, and which are endangered? What threatens them? Which organizations and public figures are addressing these problems, and what projects have they undertaken?

Table 7.2: Exploratory search tasks

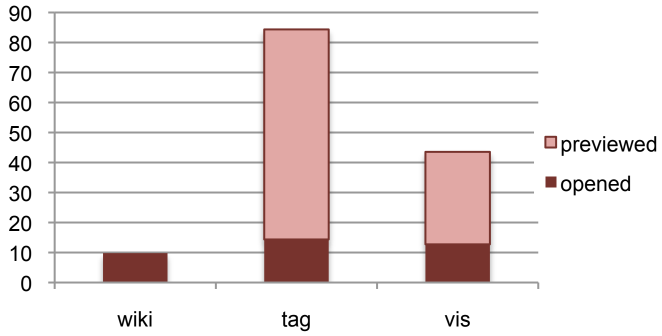


Figure 7.7: Topics opened and previewed per task with *wiki*, *tag* and *vis*

from search results, rather than via browsing. This is because following a link in *wiki* is a comparatively heavyweight activity; it causes the entire page to change, and destinations are often lengthy and can require a distracting second or two to load. Opening a link in *tag* and *vis* is much less likely to cause disorientation or delay because only the right half of the interface is altered and only a paragraph or two of new text is added. Additionally, good relevant links can be somewhat hard to find in the *wiki* system; they are scattered within lengthy prose, so that users were often seen rapidly—and somewhat aimlessly—scrolling through articles to find them. In *tag* and *vis* they are isolated, organized, visually weighted and generally given a lot of exposure. Another advantage of the *tag* and *vis* systems is that they allow link destinations to be previewed before being committed to. Figure 7.7 shows that users of *tag* and *wiki* made extensive use of the tooltip previews, and where thus exposed to many more articles (the total height of the bars) and were more informed about the links they did choose to follow.

Wikipedia's hierarchical category structure seems like a promising resource for browsers, since it allows the scope of investigations to be adjusted easily. From almost any location in Wikipedia it is possible to navigate up to broader, more general topics, or down to more specific ones. Unfortunately the users of *wiki* rarely encountered them. Most did not know of the categories until they were told, despite being regular visitors. Even after they were consistently reminded of the feature while familiarizing themselves with the interface, *wiki* users rarely navigated to categories when performing the task.

List pages are another useful feature of Wikipedia that group related pages together. Table 7.1 demonstrates that these are far less widespread than categories, but surprisingly they were much more likely to be used in this study. They proved to be particularly useful for the *Walking around the South Island* task. There are several list pages (*Tramping in New Zealand*, *New Zealand tramping tracks, National parks of New Zealand*) that address the task directly.

The marked difference in use between categories and lists within the *wiki* system is somewhat surprising. As Section 7.1.1 explained, one has a much greater chance of encountering the former, and yet participants made more use of the latter. This was not because the lists were more useful: there are many valuable categories, such as *Hiking and tramping tracks in New Zealand*, *Jazz horn players* and so on. Instead, accessibility issues are the likely cause. In *wiki*, category links are hidden at the bottom of articles and are easily overlooked, and do not appear in search results by default. We predict that the usability of *wiki* could be easily improved by giving categories greater exposure, as *tag* and *vis* do.

There was no significant usage of portals, navboxes or infoboxes within the study. The odds of them being encountered and providing the desired connections are low (Table 7.1). Consequently the *tag* and *vis* systems did not suffer despite ignoring them completely. The lack of use of infoboxes should be of particular concern to the research community surrounding Wikipedia. As Section 3.3.8 explained, these components are seen as extremely valuable because of their ability to capture structured knowledge. They are the basis behind almost all of the ontologies that have been extracted from Wikipedia (Section 3.4.4). Unfortunately they do not seem to provide the connections information seekers need (at least within this small-scale study).

Task performance

One limitation of this study (and other investigations of exploratory search) is that the answers to tasks—which are deliberately ambiguous and open-ended—could not be determined in advance. Therefore there is no ground truth against which to judge the topics gathered by each participant.

One simple measure of task success is the number of Wikipedia concepts (articles) gathered. A good system should allow participants to locate more relevant concepts in the same time. Figure 7.8a shows that users gathered more with the *tag* and *vis* systems than *wiki*, indicating that in this respect Hōpara is an improvement. However, the numbers are quite low across all systems.

Another measure is the extent to which participants agree with each other. Complete agreement is unlikely because the tasks are open-ended and participants have limited time to follow the available leads. Nevertheless, a recommendation is more likely to be correct if multiple participants agree on it. It would not be fair to measure agreement across systems, because two of them (*tag* and *vis*) return the same results and would consequently gain an unfair advantage. Instead Figure 7.8b measures the agreement of participants within each combination of task and system. The first column, which

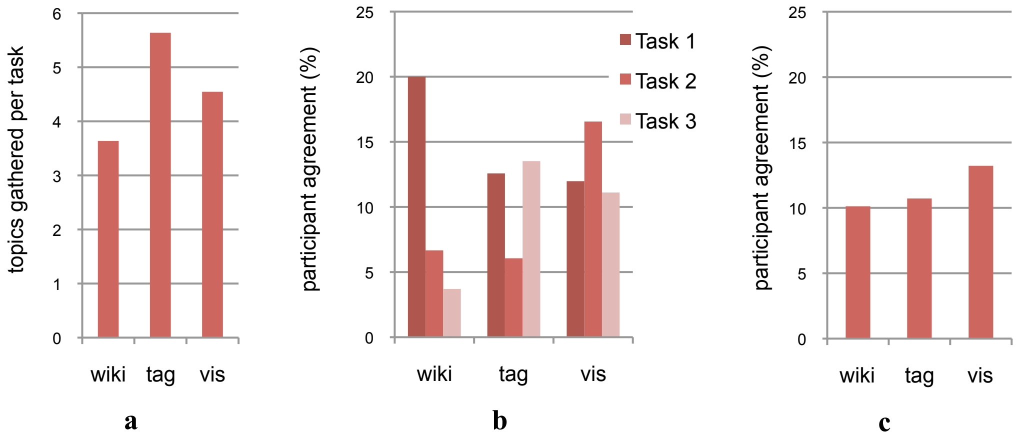


Figure 7.8: Performance of tasks with the *wiki*, *tag* and *vis* systems

measures the agreement between the four participants who used *wiki* to perform the *Walking around the South Island* task, is a clear outlier. The other two tasks were performed much more poorly with this system. This is because Wikipedia contains a couple of articles (*Tramping in New Zealand* and *New Zealand tramping tracks*) that address the first task directly by organizing lists of links to all of the relevant articles. Answers to the other tasks are scattered throughout Wikipedia with no meta-articles to organize them, and were therefore significantly more difficult to answer. The other two systems are able to answer tasks relatively consistently, whether relevant meta-articles were available or not.

Figure 7.8c averages agreement measures across tasks to provide a single measure per system. It identifies *vis* as the system for which participants were most consistent with each other.

Questionnaire responses

Each participant was given four questionnaires to record their subjective impressions of the systems. Three post-task questionnaires were used to capture usability issues, most of which were minor and will not be discussed in this thesis. Of more interest is the final questionnaire, in which participants compared the three systems directly according to various criteria such as quality of search results and ease of navigation. The ratings were given using the triangles shown in Figure 7.9, which allow all three systems to be compared simultaneously. For example, a mark in the centre indicates that all three systems are equivalent. A mark at the top corner indicates a strong preference for *wiki*, and half way between top and centre indicates a moderate preference. A mark below the centre rates *wiki* as worse than the other two. Each dot within each triangle in Figure 7.9 is the rating made by a single participant. The larger circle represents the average rating

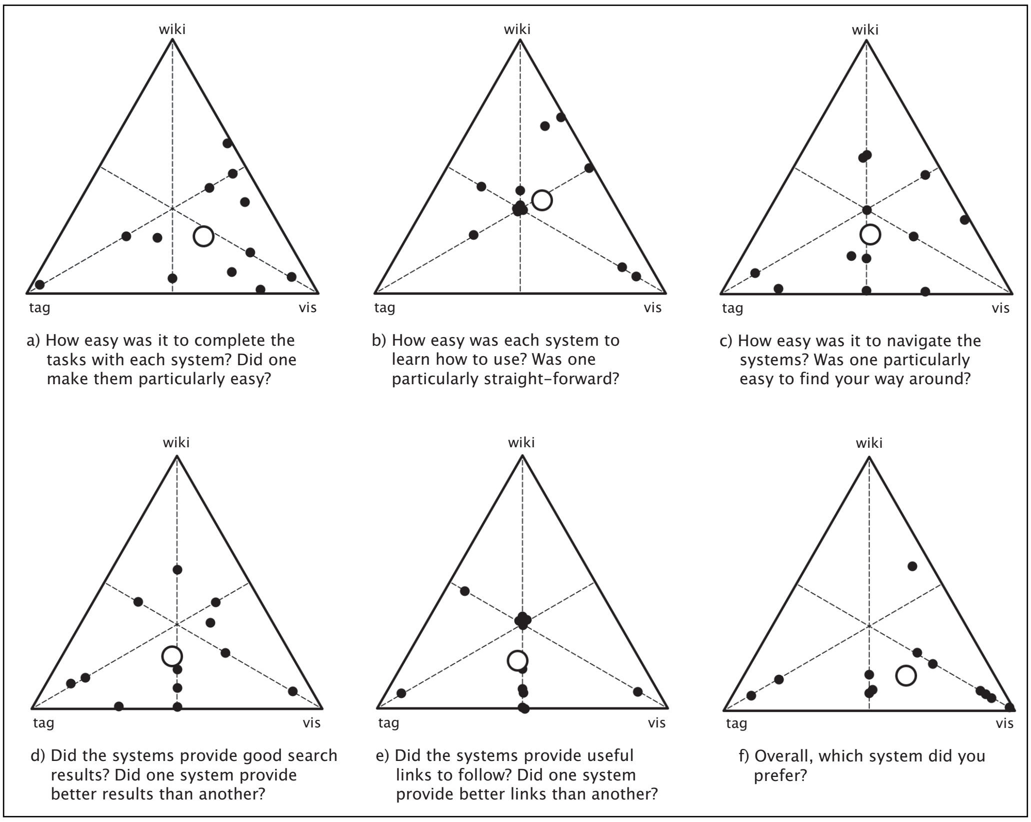


Figure 7.9: Subjective comparisons of the *wiki*, *tag* and *vis* systems

for all participants. Each rating was explained and justified by participants either orally or by typing free text.

The first question (Figure 7.9a) asks whether one of the systems made tasks particularly easy to complete. Most of the responses are scattered towards the *vis* system.

> *It was easiest with vis, 2nd was tag, It wasn't very easy with Wikipedia. When I type in the query [in tag or vis], it takes the quote out of the article to show how it is related to the query. [In wiki] if I wanted that, I would have to do a ctrl-f and search for the words I am looking for. Between vis and tag, the circles just made everything clearer.*

Figure 7.9b asks whether one system was particularly easy to learn. Half the responses are clustered around the centre of the triangle, indicating that there was little difference between the systems. Two participants chose Wikipedia as easier to learn, and cited its familiarity and similarity with traditional search engines and web pages. Another two participants had a strong preference for the *vis* system, because they felt the visualization more clearly communicated how the available information was organized and connected.

> *They are all very easy to use.*

> *I think wiki is easier to learn… …it is familiar to me, more like other web search.*

One advantage of vis is that it gives a clearer picture of how things are related.

Participants had strong but conflicting preferences when asked whether one system was easier to navigate than the others (Figure 7.9c). Two participants chose the wiki system, again because of its simplicity and familiarity. Others chose against it, and felt that the links they needed to follow were hidden within lengthy prose. The tooltips and connection snippets provided by *tag* and *vis* made it easier for participants to judge the relevance of links. Those who chose against *vis* pointed out that it does not clearly indicate when categories can or cannot be expanded, or how many categories are available.

Wiki has a classic search box. Most people know how to use that.

Wikipedia wasn't very easy to navigate around. I couldn't tell which of the links were useful or not. I had to click on them and go there and have a read, then go back and everything. But with vis and tag I could tell.

The connections snippets made it a bit clearer the effect of following that link.

Figure 7.9d asks whether one system provided better search results than another. Three participants chose *wiki*, because it gave them more flexibility in how queries were specified. In *tag* and *vis*, query terms must be matched to anchor texts. This forced users to think carefully about which keywords to use, and sometimes made for brittle searching; slight variations in spelling or pluralisation could have a large effect on the results. Additionally, both *tag* and *vis* did not behave as users expected when queries involved classes of topics, e.g. the query *jazz musicians* returns topics that are related to both *Jazz* and *Musician* but are not *Jazz Musicians.*

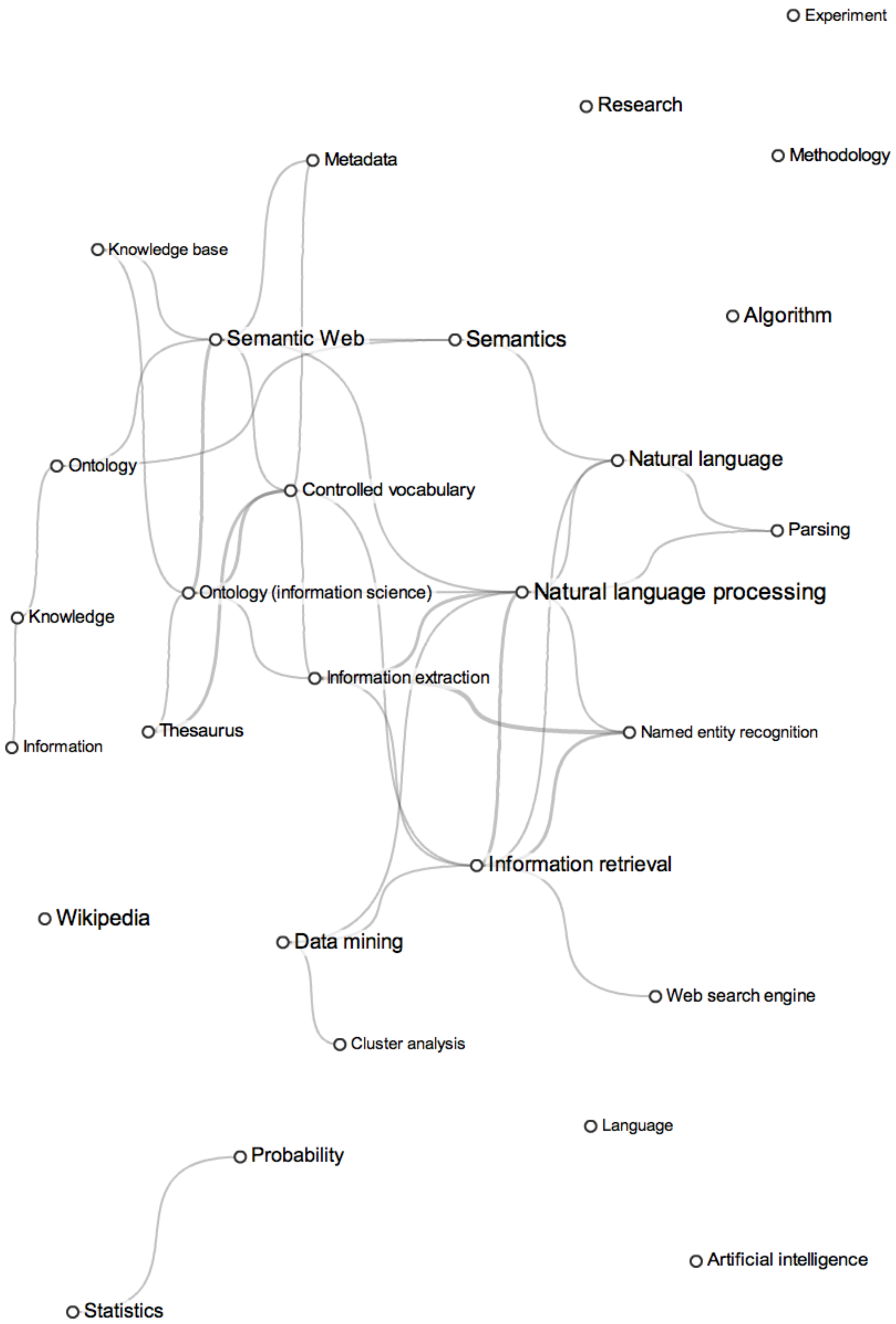
Hard to say. I think wiki provides better results. Depends on whether I clearly know the keyword to use. Most of the time when we search for stuff, we don't know.

Figure 7.9e asks about the usefulness of the links provided by the three systems. One third of participants felt there was no significant difference between them. The remaining participants chose against *wiki*, again citing that the links they needed were hidden within lengthy articles. They also appreciated the way *tag* and *vis* organized the suggested links into categories, and explained them with sentence snippets.

I could easily spend all day following links on all the systems. The categorisation of links for tag and vis allowed me to cut the chaff from the wheat easily.

[I choose tag and vis], just because it is easy to see the connection. I don't have to read [laugh] or think more.

Finally, Figure 7.9f asks which system was better overall. Only one participant chose *wiki*—they felt that some of the features in *tag* and *vis* were too unfamiliar and awkward to use. Of the remaining participants, only two chose *tag* over *vis*—they liked Hōpara's functionality, but preferred it when packaged in a more traditional interface. Overall, ¾ of participants indicated that *vis* was at least as good as the other two systems, and ½ identified it as the best system.


Experiment
Research
Methodology
Metadata
Knowledge base
Algorithm
Semantic Web
Semantics
Natural language
Ontology
Controlled vocabulary
Parsing
Ontology (information science)
Natural language processing
Knowledge
Information extraction
Thesaurus
Named entity recognition
Information
Information retrieval
Wikipedia
Data mining
Web search engine
Cluster analysis
Language
Probability
Artificial intelligence
Statistics

8. Conclusions

This thesis has investigated how Wikipedia's structure and content can be applied to make interaction with information retrieval systems more effective.

We are by no means the first to recognize Wikipedia's potential to provide machines with background knowledge. In the last few years it has generated a great deal of interest among computer scientists, notably in the fields of natural language processing and artificial intelligence. Most efforts to mine knowledge from it focus on extracting formal relations suitable for automated inference. Underlying this work is an unwritten expectation that Wikipedia must be carefully and exhaustively rendered machine-readable before machines can put it to widespread use.

This thesis advocates a more direct approach to applying Wikipedia. It makes the following claim:

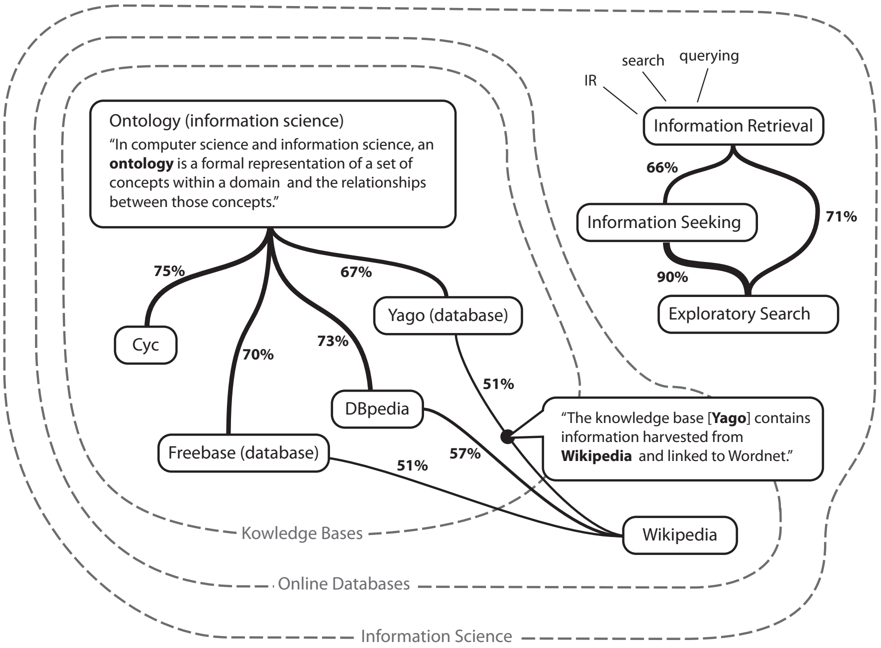
> *Wikipedia's structure can be applied effectively to open-domain information retrieval of textual documents without deep natural language processing or artificial intelligence.*

The following sections tackle different aspects or sub-claims of this central hypothesis, to explain how they have been developed and tested over the course of this investigation.

8.1 Extracting knowledge from Wikipedia

Our first sub-claim states that a significant amount of linguistic, semantic and encyclopaedic knowledge is defined explicitly in Wikipedia's structure, and can be captured without sophisticated extraction techniques. The implication is that a vast, constantly updated knowledge base suitable for supporting information retrieval is hidden just below the surface, and can be obtained with minimal natural language processing.

The validity of this claim depends on the type of knowledge base one requires. Formal, machine-readable knowledge is not easily obtained from Wikipedia. Although the relevant structural features have generated a great deal of research interest, they are either sparsely used or too informal to be mined without sophisticated, NLP-intensive cleanup (Section 3.4.4). Efforts to recruit Wikipedia's human editors to provide tidier, more expressive markup—e.g. Völkel et al. (2006)—have failed to gain traction. As a result, significant advances in natural language processing and information extraction will be needed before Wikipedia yields accurate ontologies on a large scale.



"In computer science and information science, an **ontology** is a formal representation of a set of concepts within a domain and the relationships between those concepts."

"The knowledge base [**Yago**] contains information harvested from **Wikipedia** and linked to Wordnet."

Figure 8.1: A Wikipedia-derived knowledge base

In contrast, Wikipedia encodes almost every element of human-readable knowledge bases—controlled vocabulary, glossaries, synonymy and informal hierarchical relations—directly on a large scale. The only missing component of thesauri—associative relations—is poorly encoded in the raw structure (Section 3.5.3), but can be approximated with reasonable accuracy using shallow statistical analysis of article text or inter-article links (Chapter 5). Section 2.3 argues that there is a strong precedent—stronger than the support for formal ontologies—for applying these informal, human-oriented resources to interactive information retrieval.

Exactly what kind of knowledge base can we obtain from Wikipedia without deep natural language processing? Figure 8.1 provides a visual description. As of January 2010, it contains more than 3M concepts, like *Information Seeking* and *Exploratory Search*. They have been manually associated with a total of 9.5M distinct English labels (e.g., *IR*, *search* and *querying*) and 8.4M translations (e.g. *Information Retrieval* translates to *recherche d'information* in French and 資訊檢索 in Chinese). There are 500K ambiguous labels, which have all been explicitly connected to their possible senses along with prior probability statistics (e.g. *ontology* has a 84% chance of referring to the philosophical study and 16% to knowledge base).

Every concept is supported by a succinct definition—as Figure 8.1 demonstrates for *Ontology*—and a larger body of relevant text (400 words on average) that has been used effectively to inform algorithms about the meaning of concepts (see Section 8.3). Articles are organized under an average of 3.4 parent categories each, and 96% are organized under at least one category. These are organized in turn under other categories to form a manually defined hierarchy 16 levels deep.

Rather than describing explicit associative relations between topics, our Wikipedia-derived knowledge base allows any concept to be efficiently compared against any other. It could be viewed as a complete graph, with approximately 10^{13} weighted edges. The edge weights are derived automatically, but correlate strongly with human-defined measures of relatedness. The semantic relations these edges denote are not typed, but human-readable explanations are made efficiently accessible for strongly and moderately weighted edges (weak edges are of little interest, because they are unlikely to represent useful semantic relations). The structure is a glossary not just of concepts, but of relations as well.

The general properties of the knowledge base—as opposed to the specific structural features described above—were described in Section 3.2. *Scale* and *adaptability* are impressive compared to traditional, expert driven efforts. *Accuracy* is obviously a concern, and detailed peer-reviewed evaluations of Wikipedia are difficult to come by. However, the resource is resilient to blatant attacks, and more subtle errors are of limited concern: the structure described above is based on loose associations, and is unlikely to be affected by isolated faults within Wikipedia's prose. The same argument eases concerns about *bias of opinion*, which must also be subtle in order to survive Wikipedia's editorial process and is consequently lost in the translation to structured knowledge. *Bias of coverage* is more of an issue, given Wikipedia's distinct leaning towards domains that capture the interest of its contributors: e.g., current events and domains that evolve rapidly (technology, politics, entertainment, etc.). However, if knowledge is a scarce resource then we argue that it makes sense to concentrate on exactly these areas: they are more relevant to "everyman" and thus more likely to occur in web search. Wikipedia's bias towards general and introductory concepts is similarly practical: searchers feel most confused and disoriented during their first forays into unfamiliar domains (Section 2.1), where preliminary overviews are more useful than details.

8.2 Connecting Wikipedia to textual documents

Our second sub-claim is that Wikipedia provides sufficient training data to allow existing data mining techniques to accurately detect and disambiguate Wikipedia topics when they are mentioned in plain text. This implies that documents could be annotated against the resource automatically, on a grand scale.

Annotation is a significant roadblock for the Semantic Web and other efforts to apply structured knowledge to web-scale search. Berners-Lee’s (1998b) canonical vision places the responsibility for the task squarely on people: “Instead of asking machines to understand people's language, it involves asking people to make the extra effort.” Cimiano et al. (2004), however, raise concerns about the sheer amount of labour involved. They question whether manual efforts would ever be sufficient, or even yield enough data to train automated systems: “Here, one encounters a vicious circle where there is no Semantic Web because of a lack of metadata, and there are no metadata because there is no Semantic Web that one could learn from.”

Using Wikipedia as a knowledge base sidesteps this impasse. Every single article is a manually constructed example of how to annotate documents against it. In Chapter 6, this abundance of training data allowed us to achieve state of the art performance for topic detection and disambiguation with a few simple, easily calculated features. The ability to measure relatedness between arbitrary topics (described above) was central to both tasks. There was no need for language dependant tagging of parts of speech, expensive analysis of the surrounding prose or extensive training—e.g. Mihalcea (2007). As a result, the algorithms are language independent and economical enough to be used in real-time applications, such as annotating search results on the fly, or incorporated into the workflows of news publishers, bloggers and other web-based sources of information (where close supervision is needed and delays would be tedious).

Our disambiguation algorithm detects the correct Wikipedia article to represent a term or phrase with an f-measure of 97% (Section 6.2). Our detection algorithm emulates the decisions Wikipedia’s human editors make when deciding what should and should not be linked, with an f-measure of 75% (Section 6.3). These figures remain constant whether the algorithms are tested against Wikipedia-derived ground truth or separately annotated news articles (Section 6.4). The connections produced are well suited to augmenting documents with additional explanatory information—as demonstrated in a user study by Csomai and Mihalcea (2007) of a similar but less accurate algorithm. They are also likely to provide good support for berrypicking (Bates 1989) because the links they are modelled after are manually created with the explicit purpose of helping Wikipedia’s users navigate from one topic to the next, and these topics can be intuitively used as query components (see Section 8.3).

The key limitation of this work is that it seeks only to replicate the decisions made by Wikipedia’s editors. There are many reasons why Wikipedians construct links (see Section 3.3.4), but they boil down to recommending next steps for readers to take: similar decisions were made when participants were asked to annotate news stories by putting themselves in the shoes of someone who was genuinely interested in each story, and deciding which Wikipedia articles would be most worthy of further investigation (Section 6.4). This is a somewhat quirky and subjective form of annotation, which is likely the root cause behind Huang et al.'s (2009b) objections to using Wikipedia's links as training data (discussed in Section 6.1.1). The model that these links provide, while well suited to the interactive information seeking applications described above, will not fit everyone's purposes.

Two more conventional variants of annotation are topic indexing and named entity recognition. The former attempts to detect only the most important topics covered by a document, to be used as subject headings, tags, and index terms (Section 6.1.3). The latter aims to detect every named entity—every person, place, organization, etc—mentioned within documents, and is useful for many natural language processing applications, including enhancing document representations for clustering, categorization, and related tasks (Section 6.1.4).

Wikipedia is a promising resource to index documents against. Few other controlled vocabularies can compete with its sheer breadth. Additionally, the features we use for link recommendation—relatedness, prior probabilities, etc.—are directly applicable to the task. In fact, our machine-learned link detector could be trivially adapted to topic indexing, by simply feeding it the appropriate training data, such as was gathered by Medelyan et al. (2008). We have not attempted this, but Medelyan (2009) applied many of the same features (including our relatedness measure) with great success: she found her algorithm—Maui—to be more consistent with human indexers than they were with each other. Wikipedia does not provide training data directly for this problem, but fortunately it appears that little is needed: Maui requires only a handful of manually annotated examples. Accurate, automated indexing of documents against Wikipedia is well within our grasp.

The work is not so easily adapted to named entity recognition. Simply lowering the thresholds of the link recommendation and indexing systems described above—having them return topics that do not resemble links or key topics—is not sufficient. These machine-learned systems depend on building clear models of what is desired, and return nonsensical results if asked to return topics that do not fit that model. Our own algorithm, for example, returns the *Depend* brand of adult nappies (albeit with only a 9% probability of being a link) in response to the preceding sentence. Fortunately, named entity recognition is a more widely investigated problem than either topic indexing or link recommendation, and consequently there is a great deal of related work to draw on. Current state-of-the-art systems achieve f-measures of 89% or more (Florian et al. 2003).

Additionally, Wikipedia's link structure can be adapted to provide large tagged corpora that are well suited to training these systems (Dakka and Cucerzan 2008, Nothman et al. 2009). Large-scale automated recognition of Wikipedia entities seems entirely plausible, given the maturity of the research and the availability of quality training data.

8.3 Applying Wikipedia to the retrieval process

The result of the previous two claims is a large-scale, domain-independent knowledge base that can be easily connected to any textual document collection. All that is needed to confirm the central thesis claim is to demonstrate the utility of this structure for information retrieval over the documents to which it is connected. Thus, our final claim is that Wikipedia-derived knowledge, having been extracted and connected appropriately to documents, can be applied to the information seeking process in ways that are helpful and intuitive to users.

Obviously, this claim depends greatly on the knowledge base Wikipedia provides. The only semantic relations we were able to obtain at a large scale are fuzzy, informal, and oriented towards human users. Are they still useable by search engines? Chapter 2 demonstrated a strong precedent for applying informal knowledge bases to search: they have been applied usefully to query expansion, indexing, categorization, clustering, search engine personalization and adaptive hypermedia. In contrast, there are few applications for which more formalized knowledge has been demonstrated as directly applicable. This same pattern is apparent when turning specifically to Wikipedia. Attempts to mine formal relations from the resource are widespread, but so far there have been few examples in which the resulting ontologies are put to use. In contrast, most efforts to mine lesser structures are immediately able to apply the knowledge they gather.

We do not imply that formal knowledge is inherently less useful than informal knowledge. We merely point out that the latter has not been applied as widely thus far. This is likely due to the pragmatics of conducting research: those who focus on the formal knowledge bases must invest much greater effort to obtain them, and consequently have had limited opportunity to apply them. In this case, we would encourage researchers to shift their focus from gathering knowledge to putting previous efforts to use. The resources they have built have scaled up remarkably in recent years. Freebase, for example, now weighs in with more than 50M topics and 440M assertions. How much more knowledge needs to be gathered, before we have enough to do something compelling and useful with it?

Some of the most successful applications of Wikipedia require little structure or formalism from it. The continuing work of Gabrilovich, Markovich and Egozi provides good examples: these researchers essentially treat Wikipedia as a verbose glossary, where each concept is supported by a lengthy document. Without consulting any other structural elements, they have made significant contributions to document categorization (Gabrilovich and Markovitch 2005), document similarity metrics (Gabrilovich and Markovitch 2007) and relevance feedback (Egozi et al. 2008). These are all challenging tasks that have resisted attempts to apply background knowledge or concept-based document representations in the past.

The above applications keep Wikipedia hidden behind the scenes. They are effective, but not what we had in mind for this investigation. Instead, we aimed for a knowledge base that could withstand close inspection from users and be transparently applied to the information seeking process. This ambition for more direct interaction makes for a challenging investigation, for two reasons. Firstly, it makes it difficult to separate from the two previous sub-claims, because the utility of the knowledge base depends directly on how accurately it is built and connected to documents. Secondly, interactive systems obviously cannot be evaluated without involving human participants, which makes gathering conclusive quantitative measures extremely difficult. Practices for consistent evaluation and effective comparison of interactive information retrieval systems are still in development (White et al. 2008).

These complications came into play with the evaluation of Koru in Chapter 4. This experiment supports our intuition that Wikipedia is broad enough to suit open-domain search: when tested against a heterogeneous collection of documents and retrieval tasks, it was able to recognize and lend assistance to almost all queries issued to it. This recognition of query topics had a positive effect on user behaviour, and both allowed and motivated them to issue more queries and gather better documents as a result. However, our expectation that Wikipedia's semantic relations would provide an effective platform for interactive query expansion and navigation was not borne out. Was this because of the way this functionality was presented to users, the quality of the knowledge base and semantic relations behind it, the accuracy with which this structure was connected to documents, or did the experimental methodology discourage participants from exploring? All of these factors are likely relevant, but impossible to separate within this study.

The evaluation of Hōpara in Chapter 7 was a cleaner experiment. It focused closely on the utility of Wikipedia-based semantic relatedness for supporting interactive query expansion and exploration, and the methods by which these were presented to users. Subjective preference for the new systems was very high, with only one of the twelve participants choosing the incumbent interface to Wikipedia over them. The semantic relatedness measures provided sensible recommendations for query expansion, and the extraction of sentence snippets to explain these recommendations was a popular feature. The use of visualization was intuitive and offered concrete advantages over the more traditional, simpler alternative. Unfortunately, evaluation remains a weak point. The small number of participants and tasks, and the artificiality of the laboratory-based experiment meant that we were not able to provide conclusive, objective proof of Hōpara's utility; only of users' subjective feelings about it. In future we plan to conduct longer-term ethnographic studies, where participants are not given artificial tasks but instead have their own reasons to use the system. This should provide deeper insights into how information seeking is performed with Wikipedia and Hōpara, and yield stronger evidence of their relative strengths and weaknesses.

We had always intended to conclude this investigation by revisiting the Koru prototype, and incorporating all of the lessons learned since its development. Unfortunately time did not permit. For now, Hōpara offers good opportunities to focus on supplying relevant query recommendations and presenting them in an intuitive manner. Once this work as been verified and refined with the ethnographic studies described above, we intend to feed these features back into Koru, along with the link recommendation and topic indexing work described above, to provide similar support to arbitrary document collections. This would again provide new opportunities for large-scale, long-term evaluations, and new insights into how Wikipedia can support interactive information retrieval over a range of document collections and domains.

The algorithms behind these end-user systems provide broader opportunities for future work. Currently there are many tasks—indexing, clustering, categorization and summarisation to name a few—for which simple bag-of-words models and word overlap provide robust baselines that researchers struggle to improve upon. Our techniques for efficient, language-independent topic detection, disambiguation and comparison have the potential for extraordinarily broad application, because they could allow these algorithms to draw on Wikipedia's vast network of concepts and relations and gain an understanding—of sorts—of what these words mean. Initial forays have been made to apply this work to document summarization (Nastase et al. 2009) and clustering (Huang et al. 2009a), but there are many more opportunities to explore.

8.4 Closing remarks

Wikipedia's contributors have laboured for countless hours. The have not produced an ontology, or anything that can be directly applied to hard AI. They have not broken the knowledge acquisition bottleneck. They have, however, created an exceptionally large tapestry of concepts and informal relations, handcrafted for the explicit purpose of helping information seekers navigate effectively. They have also provided millions of examples of how to connect this structure to textual documents.

Similarly, this thesis has not pushed the boundaries of computational linguistics, natural language processing or knowledge representation, as many of the researchers working with Wikipedia have attempted. It has instead demonstrated how the resource can be exploited directly, with minimal computational effort. It has provided efficient, language independent algorithms for topic detection, disambiguation and comparison, which can be applied to a wide array of problems. It has also explored how these algorithms can be used to enhance end-user search applications. This progress is due to our philosophy of using Wikipedia as directly as possible—as source of loosely organized human-readable knowledge—rather than attempting to turn it into something more formal.

The thesis began with a question posed to Socrates. We close it with a quote that is commonly attributed to him:

> *Employ your time in improving yourself by other men's writings, so that you shall gain easily what others have laboured hard for.*

References

Adafre, S.F., Jijkoun, V. and de Rijje, M. (2007). Fact Discovery in Wikipedia. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,* Silicon Valley, CA. pp. 177–183

Adar, E., Skinner, M. and Weld, D.S. (2009). Information arbitrage across multi-lingual Wikipedia. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining,* Barcelona, Spain. pp. 94–103.

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M. and Soroa, A. (2009). A Study on Similarity and Relatedness using Distributional and WordNet-based Approaches. In *Proceedings of the 10th Annual Conference of the North American Chapter of the Association for Computational Linguistics* Boulder, CO. pp. 19–27.

Ahlberg, C. and Shneiderman, B. (1994). Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Boston, MA. pp. 313–317.

Alani, H., Chandler, P., Hall, W., O'Hara, K., Shadbolt, N. and Szomszor, M. (2008). Building a Pragmatic Semantic Web. *IEEE Intelligent Systems, 23*(3), 61–68.

Alfonseca, E., Rodríguez, P. and Pérez, D. (2007). An approach for automatic generation of adaptive hypermedia in education with multilingual knowledge discovery techniques. *Computers & Education, 49*(2), 495–513.

Allan, J. (2005). HARD track overview in TREC 2005: High accuracy retrieval from documents. In *Proceedings of the 14th Text Retrieval Conference,* Gaithersburg, MD. pp. 51–67.

Andrews, S. (2007). Wikipedia uncovered. *PC Authority* Retrieved 12 Febuary 2010, from http://www.pcauthority.com.au/Feature/93908,wikipedia-uncovered.aspx/1

Angeletou, S., Sabou, M. and Motta, E. (2008). Semantically Enriching Folksonomies with FLOR. In *Proceedings of the 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web,* Tenerife, Spain. pp. 65–79.

Anonymous. (2005). Can you trust Wikipedia? *The Guardian* Retrieved 12 February 2010, from http://www.guardian.co.uk/technology/2005/oct/24/comment.newmedia

Anonymous. (2006). The word: Common sense *New Scientist* Retrieved 12 February 2010, from http://www.newscientist.com/article/mg19025471.700-the-word-common-sense.html

Antoniou, G. and Van Harmelen, F. (2004). *A semantic web primer*. Boston, MA: The MIT Press.

Armstrong, T.G., Moffat, A., Webber, W. and Zobel, J. (2009). Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management,* Hong Kong, China. pp. 601–610.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference,* Busan, South Korea. pp. 715–728.

Auer, S. and Lehmann, J. (2007). What Have Innsbruck and Leipzig in Common? Extracting Semantics from Wiki Content. In *Proceedings of the 4th European Conference on The Semantic Web,* Innsbruck, Austria. pp. 503–517.

Banerjee, S. and Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing* Mexico City, Mexico. pp. 136–145.

Banerjee, S., Ramanathan, K. and Gupta, A. (2007). Clustering short texts using Wikipedia. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Amsterdam, The Netherlands. pp. 787–788.

Banko, M., Etzioni, O. and Center, T. (2008). The tradeoffs between open and traditional relation extraction. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio. pp. 28–36.

Barr, J. and Cabrera, L.F. (2006). AI gets a brain. *ACM Queue, 4*(4), 24–29.

Bates, M. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review, 13*(5), 407–424.

Bates, M. (1990). Where should the person stop and the information search interface start? *Information Processing and Management, 26*(5), 575–591.

Belkin, N. (1996). Intelligent information retrieval: Whose intelligence. In *Proceedings of the 5th International Symposium for Information Science,* Konstanz, Germany. pp. 25–31.

Berger, A.L. and Mittal, V.O. (2000). OCELOT: a system for summarizing Web pages. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Athens, Greece. pp. 144–151.

Berners-Lee, T. (1998a). Semantic Web Road map. Retrieved March 29, 2010, from http://www.w3.org/DesignIssues/Semantic.html

Berners-Lee, T. (1998b). What the Semantic Web can represent. Retrieved 8 April 2010, from http://www.w3.org/DesignIssues/RDFnot.html

Berners-Lee, T. and Fischetti, M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. San Francisco, CA: Harper.

Bhole, A., Fortuna, B., Grobelnik, M. and Mladenic, D. (2007). Extracting named entities and relating them over time based on wikipedia. *Informatica, 4*(4), 463–468.

Biuk-Aghai, R.P. (2006). Visualizing co-authorship networks in online wikipedia. In *Proceedings of the International Symposium on Communications and Information Technologies,* Bangkok , Thailand. pp. 737–742.

Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009). DBpedia – A Crystallization Point for the Web of Data. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 7*(3), 154–165.

Blohm, S. and Cimiano, P. (2007). Using the Web to Reduce Data Sparseness in Pattern-Based Information Extraction In *Proceedings of the Knowledge Discovery in Databases,* Warsaw, Poland. pp. 18–29.

Boyd-Graber, J., Fellbaum, C., Osherson, D. and Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the 3rd International WordNet Conference,* Jeju Island, Korea. pp. 29–36.

Bray, H. (2004). One great source -- if you can trust it. *The Boston Globe* Retrieved 12 February 2010, from http://www.boston.com/business/technology/articles/2004/07/12/ one_great_source____if_you_can_trust_it

Britannica. (2006). Fatally Flawed: Refuting the recent study on encyclopedic accuracy by the journal Nature. Retrieved 12 February 2010, from http://corporate.britannica.com/britannica_nature_response.pdf

Brooks, H. (1987). Expert systems and intelligent information retrieval. *Information Processing and Management, 23*(4), 367–382.

Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction, 11*(1–2), 87–110.

Budanitsky, A. and Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics, 32*(1), 13–47.

Bunescu, R. and Paşca, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy. pp. 9—16.

Buscaldi, D. and Rosso, P. (2006). A Bag-of-Words Based Ranking Method for the Wikipedia Question Answering Task In *Proceedings of the 7th Workshop of the Cross-Language Evaluation Forum* Alicante, Spain. pp. 550–553.

Bush, V. (1945). As We May Think. *Atlantic Monthly, 176*(1), 101—108.

Califf, M.E. and Mooney, R.J. (1999). Relational learning of pattern-match rules for information extraction. In *Proceedings of the National Conference on Artificial intelligence,* Orlando, FL. pp. 328–334.

Catone, J. (2008). Thinkbase: Mapping the World's Brain. *ReadWriteWeb* Retrieved March 23, 2010, from http://www.readwriteweb.com/archives/ thinkbase_mapping_the_worlds_brain.php

Chai, J.Y. and Biermann, A. (1997). The use of lexical semantics in information extraction. In *Proceedings of the ACL Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications,* Madrid, Spain. pp. 61–70.

Chen, C. and Yu, Y. (2000). Empirical studies of information visualization: a meta-analysis. *International Journal of Human-Computer Studies, 53*(5), 851–866.

Chernov, S., Iofciu, T., Nejdl, W. and Zhou, X. (2006). Extracting semantic relationships between wikipedia categories. In *Proceedings of the 1st Workshop on Semantic Wikis,* Budva, Montenrego.

Chi, E.H., Pirolli, P., Chen, K. and Pitkow, J. (2001). Using information scent to model user information needs and actions and the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Seattle, WA. pp. 490–497.

Chinchor, N. and Robinson, P. (1997). MUC-7 named entity task definition. In *Proceedings of the 7th Message Understanding Conference,* Washington DC, WA.

Choi, Y. (2008). Making Faceted Classification more acceptable on the Web: A comparison of Faceted Classification and ontologies. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1–5.

Cilibrasi, R.L. and Vitanyi, P.M.B. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering, 19*(3), 370–383.

Cimiano, P., Handschuh, S. and Staab, S. (2004). Towards the self-annotating web. In *Proceedings of the 13th international Conference on World Wide Web,* New York, NY. pp. 462–471.

Clauson, K.A., Polen, H.H., Boulos, M.N.K. and Dzenowagis, J.H. (2008). Scope, completeness, and accuracy of drug information in Wikipedia. *The Annals of Pharmacotherapy, 42*(12), 1814–1821.

Correa, P., Correa, A. and Askanas, M. (2006). *Wikipedia: a techno-cult of ignorance*. Concord, Canada: Akronos Publishing.

Crane, D., Pascarello, E. and James, D. (2005). *Ajax in action*. Greenwich, CT: Manning Publications

Csomai, A. and Mihalcea, R. (2007). Linking educational materials to encyclopedic knowledge. *Frontiers in Artificial Intelligence and Applications, 158*, 557–559

Cucerzan, S. (2007). Large-Scale Named Entity Disambiguation Based on Wikipedia Data. In *Proceedings of the Empirical Methods in Natural Language Processing,* Prague, Czech Republic.

Culotta, A., McCallum, A. and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting,* New York, NY. pp. 296–303.

Curran, J.R. (2004). *From distributional to semantic similarity.* PhD Thesis, University of Edinburgh, Edinburgh, Scotland.

Cutting, D.R., Karger, D.R., Pedersen, J.O. and Tukey, J.W. (1992). Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th ACM SIGIR Conference on Research and Development in Information Retrieval,* Copenhagen, Denmark. pp. 318–329.

Dakka, W. and Cucerzan, S. (2008). Augmenting wikipedia with named entity tags. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing,* Hyderabad, India.

David, C., Giroux, L., Bertrand-Gastaldy, S. and Lanteigne, D. (1995). Indexing as problem solving: A cognitive approach to consistency. In *Proceedings of the ASIS Annual Meeting,* Medford, NJ. pp. 49–55.

Davidov, D., Gabrilovich, E. and Markovitch, S. (2004). Parameterized generation of labeled datasets for text categorization based on a hierarchical directory. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Sheffield, England. pp. 250–257

Dee, J. (2007). All the news that's fit to print out. *The New York Times* Retrieved February 12 2010, from http://www.nytimes.com/2007/07/01/magazine/01WIKIPEDIA-t.html

Denoyer, L. and Gallinari, P. (2006). The Wikipedia XML Corpus. *SIGIR Forum, 40*(1), 64–69.

Dinet, J., Favart, M. and Passerault, J. (2004). Searching for information in an online public access catalogue (OPAC): the impacts of information search expertise on the use of Boolean operators. *Journal of Computer Assisted Learning, 20*(5), 338–346.

Egozi, O., Gabrilovich, E. and Markovitch, S. (2008). Concept-based feature generation and selection for information retrieval. In *Proceedings of the 23rd National Conference on Artificial intelligence,* Chicago, IL. pp. 1132–1137.

Einbinder, H. (1964). *The myth of the Britannica*. New York, NY: Grove Press.

Fallis, D. (2008). Toward an epistemology of Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(10), 1662–1674.

Fayyad, U.M., Grinstein, G.G. and Wierse, A. (2002). *Information visualization in data mining and knowledge discovery*. San Francisco, CA: Morgan Kaufmann.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web,* Hong Kong. pp. 406–414.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems, 20*(1), 116–131.

Finkelstein, S. (2006). I'm on Wikipedia, get me out of here. *The Guardian* Retrieved 12 February 2010, from http://www.guardian.co.uk/technology/2006/sep/28/wikipedia.web20

Florian, R., Ittycheriah, A., Jing, H. and Zhang, T. (2003). Named Entity Recognition through Classifier Combination. In *Proceedings of the 7th Conference on Natural Language Learning,* Edmonton, Canada. pp. 168–171.

Fluit, C., Sabou, M. and Van Harmelen, F. (2003). Ontology-based information visualization *Visualizing the Semantic Web: XML-based Internet and Information Visualization* (pp. 36–48). London, England: Springer-Verlag.

Gabrilovich, E. and Markovitch, S. (2005). Feature generation for text categorization using world knowledge. In *Proceedings of the 18th International Joint Conference on 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. In *Proceedings of the National Conference on Artificial Intelligence,* Boston, MA. pp. 1301–1306.

Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence,* Hyderabad, India. pp. 1606–1611.

García, E. and Sicilia, M. (2003). User Interface Tactics in Ontology-Based Information Seeking. *PsychNology, 1*(3), 242–255.

Gauch, S., Chaffee, J. and Pretschner, A. (2003). Ontology-Based Personalized Search and Browsing. *Web Intelligence and Agent Systems, 1*(3), 219–234.

Gawryjolek, J. and Gawrysiak, P. (2007). The Analysis and Visualization of Entries in Wiki Services. In *Proceedings of the 5th Atlantic Web Intelligence Conference,* Fontainebleau, France.

Geva, S. (2007). GPX: Ad-hoc queries and automated link discovery in the wikipedia. In *Proceedings of the Initiative for the evaluation of XML retrieval (INEX),* Dagstuhl, Germany. pp. 404–416.

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies-an etymological note. *Journal of Documentation, 59*(1), 7–18.

Giles, J. (2005). Internet encyclopedias go head to head. *Nature, 438*(7070), 900–901.

Golder, S.A. and Huberman, B.A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198–208.

Granitzer, M., Kienreich, W., Sabol, V., Andrews, K. and Klieber, W. (2004). Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *Proceedings of the IEEE Symposium on Information Visualization,* Austin, TX. pp. 127–134.

Green, S.J. (1999). Building hypertext links by computing semantic similarity. *IEEE Transactions on Knowledge and Data Engineering, 11*(5), 713–730.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Norwell, MA: Kluwer Academic Publishers.

Hafner, K. (2007). Seeing Corporate Fingerprints in Wikipedia Edits. *The New York Times* Retrieved 12 February 2010, from http://www.nytimes.com/2007/08/19/technology/19wikipedia.html

Hara, T., Ito, M. and Nishio, S. (2008). Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management,* Napa Valley, CA. pp. 817–826.

Hearst, M.A. (1995). TileBars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Denver, CO. pp. 59–66.

Hearst, M.A. (2006). Clustering versus faceted categories for information exploration. *Communications of the ACM, 49*(4), 59–61.

Hearst, M.A. (2009). *Search User Interfaces*. New York, NY: Cambridge University Press.

Hearst, M.A. and Pedersen, J.O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th ACM SIGIR Conference on Research and Development in Information Retrieval,* Zurich, Switzerland. pp. 76–84.

Herbelot , A. and Copestake, A. (2006). Acquiring Ontological Relationships from Wikipedia Using RMRS. In *Proceedings of the Workshop on Web Content Mining with Human Language Technologies,* Athens, GA.

Hirsch, C., Hosking, J. and Grundy, J. (2009a). Interactive Visualization Tools for Exploring the Semantic Graph of Large Knowledge Spaces. In *Proceedings of the Workshop on Visual Interfaces to the Social and the Semantic Web,* Sanibel Island, FL.

Hirsch, C., Hosking, J., Grundy, J., Chaffe, T., MacDonald, D. and Halytskyy, Y. (2009b). The Visual Wiki: A New Metaphor for Knowledge Access and Management. In *Proceedings of the 42nd Hawaii International Conference on System Sciences,* Waikoloa, Hawaii.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems, 22*(1), 89–115.

Holloway, T., Božicevic, M. and Börner, K. (2007). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity, 12*(3), 30–40

Hotho, A., Staab, S. and Maedche, A. (2001). Ontology-based Text Clustering. In *Proceedings of the IJCAI Workshop on Text Learning: Beyond Supervision,* Seattle, WA. pp. 48–54.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired Magazine* Retrieved 12 February 2010, from http://www.wired.com/wired/archive/14.06/crowds.html

Huang, A., Milne, D., Frank, E. and Witten, I.H. (2009a). Clustering Documents Using a Wikipedia-Based Concept Representation. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining,* Bangkok, Thailand. pp. 628–636.

Huang, W.C., Trotman, A. and Geva, S. (2009b). The importance of manual assessment in link discovery. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Boston, MA. pp. 698–699.

Huang, W.C., Trotman, A. and Geva, S. (2009c). The Methodology of Manual Assessment in the Evaluation of Link Discovery. In *Proceedings of the 14th Australian Document Computing Symposium (ADCS 2009),* Sydney, Australia. pp. 110–115.

Itakura, K.Y. and Clarke, C.L. (2007). University of Waterloo at INEX2007: Adhoc and Link-the-wiki Tracks. In *Proceedings of the Initiative for the evaluation of XML retrieval (INEX),* Dagstuhl, Germany.

Jarmasz, M. (2003). *Roget's thesaurus as a lexical resource for natural language processing.* MSc Thesis, University of Ottowa, Ottawa, Canada.

Jenkinson, D., Leung, K.C. and Trotman, A. (2008). Wikisearching and Wikilinking. In *Proceedings of the Initiative for the evaluation of XML retrieval (INEX),* Dagstuhl, Germany.

Karger, D. and Schraefel, M. (2006). The pathetic fallacy of rdf. Retrieved 12 February 2010, from http://swui.semanticweb.org/swui06/papers/Karger/Pathetic_Fallacy.html

Kaser, O. and Lemire, D. (2007). Tag-Cloud Drawing: Algorithms for Cloud Visualization. In *Proceedings of the WWW'07 Workshop on Taggings and Metadata for Social Information Organization,* Banff, Canada.

Kazama, J. and Torisawa, K. (2007a). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* Prague, Czech Republic. pp. 698–707.

Kazama, J. and Torisawa, K. (2007b). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* Prague, Czech Republic. pp. 698–707.

Kim, H. and Hirtle, S.C. (1995). Spatial metaphors and disorientation in hypertext browsing. *Behaviour and Information Technology, 14*(4), 239–250.

Klein, D. and Manning, C.D. (2003). Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics,* Sapporo, Japan. pp. 423–430.

Klein, G., Moon, B. and Hoffman, R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems, 21*(4), 70–73.

Krötzsch, M., Vrandečić, D. and Völkel, M. (2005). Wikipedia and the semantic web-the missing links. In *Proceedings of the 1st International Wikimedia Conference (Wikimania),* Frankfurt am Main, Germany.

Krötzsch, M., Vrandečić, D. and Völkel, M. (2006). Semantic MediaWiki In *Proceedings of the 5th International Semantic Web Conference,* Athens, GA. pp. 935–942.

Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42*(5), 361–371.

Kules, B. and Capra, R. (2008). Creating exploratory tasks for a faceted search interface. In *Proceedings of the 2nd Workshop on Human-Computer Interaction,* Redmond, WA.

Kules, B. and Shneiderman, B. (2008). Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing and Management, 44*(2), 463–484.

Lam, S.T.K. and Riedl, J. (2009). Is Wikipedia growing a longer tail? In *Proceedings of the ACM International Conference on Supporting Group Work,* Sanibel Island, FL. pp. 105–114.

Landauer, T.K., Foltz, P.W. and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Leacock, C. and Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification *WordNet: An electronic lexical database* (pp. 265–283). Cambridge, MA: MIT Press.

Lee, H.M., Lin, S.K. and Huang, C.W. (2001). Interactive query expansion based on fuzzy association thesaurus for web information retrieval. In *Proceedings of the 10th IEEE International Conference on Fuzzy Systems,* Melbourne, Australia. pp. 724–727.

Legg, C. (2007). Ontologies on the semantic web. *Annual Review of Information Science and Technology, 41*, 407–451.

Lenat, D.B., Guha, R.V., Pittman, K., Pratt, D. and Shepherd, M. (1990). Cyc: Toward programs with common sense. *Communications of the ACM, 33*(8), 30–49

Li, Y., Luk, W.P., Ho, K.S. and Chung, F.L. (2007). Improving weak ad-hoc queries using wikipedia as an external corpus. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Amsterdam, The Netherlands pp. 797–798.

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th International Symposium on Online Journalism,* Austin, Texas.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM, 37*(7), 30–40.

Mandala, R., Takenobu, T. and Hozumi, T. (1998). The use of WordNet in information retrieval. In *Proceedings of the Coling-ACL '98 Workshop on usage of WordNet in Natural Language Processing Systems,* Montréal/Canada. pp. 31–37.

Mandala, R., Tokunaga, T. and Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval,* Berkeley, CA. pp. 191–197.

Mangis, C. (2005). Visual Thesaurus 3. *PC Mag* Retrieved March 23, 2010, from http://www.pcmag.com/article2/0,2817,1820529,00.asp

Manola, F., Miller, E. and McBride, B. (2004). RDF primer. *W3C recommendation* Retrieved 12 February 2010, from http://www.w3.org/TR/rdf-primer/

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge, England: Cambridge University Press.

Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM, 49*(4), 41–46.

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

Maron, M.E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science, 28*(1), 38–43.

McCool, R. (2005). Rethinking the Semantic Web, Part 1. *IEEE Internet Computing, 9*(6), 88–87.

McGuinness, D.L. (2005). Ontologies come of age *Spinning the Semantic Web: bringing the World Wide Web to its full potential* (pp. 171–192). Boston, MA: MIT Press.

McHenry, R. (2004). The Faith-Based Encyclopedia. *TCS Daily* Retrieved 12 February 2010, from http://www.tcsdaily.com/article.aspx?id=111504A

Medelyan, O. (2009). *Human-competitive automatic topic indexing.* PhD Thesis, The University of Waikato, Hamilton, New Zealand.

Medelyan, O. and Legg, C. (2008). Integrating CYC and Wikipedia: Folksonomy meets rigorously defined common-sense. In *Proceedings of the 1st Workshop on Wikipedia and Artificial Intelligence,* Chicago, IL.

Medelyan, O., Milne, D., Legg, C. and Witten, I.H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies, 67*(9), 716–754.

Medelyan, O., Witten, I.H. and Milne, D. (2008). Topic Indexing with Wikipedia. In *Proceedings of the 1st Workshop on Wikipedia and Artificial Intelligence,* Chicago, IL.

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics,* Rochester, NY. pp. 196–203.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management,* Lisbon, Portugal. pp. 233–242.

Miller, G.A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39–41.

Miller, G.A. and Charles, W.G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes, 6*(1), 1–28.

Minier, Z., Bodo, Z. and Csato, L. (2007). Wikipedia-based kernels for text categorization. In *Proceedings of the 9th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing,* Timisoara, Romania. pp. 157–164.

Morrison, P.J. (2008). Tagging and searching: search retrieval effectiveness of folksonomies on the World Wide Web. *Information Processing and Management, 44*(4), 1562–1579.

Muchnik, L., Itzhack, R., Solomon, S. and Louzoun, Y. (2007). Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Physical Review E, 76*(1).

Munroe, R. (2007). The problem with Wikipedia. *xkcd* Retrieved 12 February 2010, from http://xkcd.com/214/

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes, 30*(1), 3–26.

Nakayama, K. (2008). Extracting Structured Knowledge for Semantic Web by Mining Wikipedia. In *Proceedings of the 7th International Semantic Web Conference,* Karlsruhe, Germany.

Nakayama, K., Hara, T. and Nishio, S. (2007a). Wikipedia mining for an association web thesaurus construction. In *Proceedings of the 8th International Conference on Web Information Systems Engineering,* Nancy, France. pp. 322–334.

Nakayama, K., Hara, T. and Nishio, S. (2007b). Wikipedia: A New Frontier for AI Researches. *Japanese Society for Artificial Intelligence, 22*(5).

Nastase, V., Milne, D. and Filippova, K. (2009). Summarizing with Encyclopedic Knowledge. In *Proceedings of the 2nd Text Analysis Conference,* Gaithersburg, MA.

Nastase, V. and Strube, M. (2008). Decoding wikipedia categories for knowledge acquisition. In *Proceedings of the 23rd National Conference on Artificial intelligence,* Chicago, IL. pp. 1219–1224.

Nature. (2006). Encyclopaedia Britannica and Nature: a response. Retrieved 12 February 2010, from http://www.nature.com/press_releases/Britannica_response.pdf

Nguyen, D.P.T., Matsuo, Y. and Ishizuka, M. (2007a). Exploiting syntactic and semantic information for relation extraction from wikipedia. In *Proceedings of the IJCAI Workshop on Text-Mining & Link-Analysis,* Hyderabad, India.

Nguyen, D.P.T., Matsuo, Y. and Ishizuka, M. (2007b). Relation Extraction from Wikipedia Using Subtree Mining. In *Proceedings of the 22nd National Conference on Artificial Intelligence,* Vancouver, Canada. pp. 1414–1420.

Noruzi, A. (2007). Editorial. *Webology, 4*(2).

Nothman, J., Curran, J.R. and Murphy, T. (2008). Transforming Wikipedia into named entity training data. In *Proceedings of the Australasian Language Technology Association Workshop,* Hobart, Tasmania. pp. 124–132.

Nothman, J., Murphy, T. and Curran, J.R. (2009). Analysing Wikipedia and gold-standard corpora for NER training. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics,* Athens, Greece. pp. 612–620.

Nov, O. (2007). What motivates wikipedians? *Communications of the ACM, 50*(11), 60–64.

Nunes, S., Ribeiro, C. and David, G. (2008). Wikichanges-exposing wikipedia revision activity. In *Proceedings of the International Symposium on Wikis,* Porto, Portugal.

Osiriski, S. and Weiss, D. (2004). Conceptual clustering using lingo algorithm: Evaluation on open directory project data. In *Proceedings of the Intelligent Information Processing and Web Mining,* Zakopane, Poland. pp. 369–378.

Overell, S.a.R., S. (2006). Identifying and grounding descriptions of places. In *Proceedings of the SIGIR Workshop on Geographic Information Retrieval,* Seattle, WA. pp. 14–16.

Panciera, K., Halfaker, A. and Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM International Conference on Supporting Group Work,* Sanibel Island, FL. pp. 51–60.

Pantel, P. and Lin, D. (2001). DIRT - Discovery of Inference Rules from Text. In *Proceedings of the 7th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* San Francisco, CA. pp. 323–328.

Paynter, G.W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries,* Denver, CO. pp. 291–300.

Pirolli, P. and Card, S. (1999). Information foraging. *Psychological review, 106*(4), 643–675.

Pitkow, J., Schütze, 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.

Ponzetto, S.P. and Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence,* Vancouver, Canada. pp. 1440–1445.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program, 14*, 130–137.

Potthast, M., Stein, B. and Anderka, M. (2008). A wikipedia-based multilingual retrieval model. In *Proceedings of the 30th European Conference on IR Research,* Glasgow, Scotland. pp. 522–530.

Pratt, W., Hearst, M.A. and Fagan, L. (1999). A knowledge-based approach to organizing retrieved documents. In *Proceedings of the 16th National Conference on Artificial Intelligence,* Orlando, FL. pp. 80–85.

Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L. and Riedl, J. (2007). Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the ACM International Conference on Supporting Group Work,* Sanibel Island, FL. pp. 259–268.

Quinlan, J.R. (1993). *C4. 5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann.

Raimond, Y., Abdallah, S., Sandler, M. and Giasson, F. (2007). The music ontology. In *Proceedings of the 8th International Conference on Music Information Retrieval,* Vienna, Austria. pp. 417–422.

Reed, S.L. and Lenat, D.B. (2002). Mapping ontologies into cyc. In *Proceedings of the AAAI Workshop on Ontologies for the Semantic Web,* Palo Alto, CA. pp. 2–11.

Resnick, P. and Varian, H.R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56–58.

Resnik, P. (1995a). Disambiguating noun groupings with respect to wordnet senses. In *Proceedings of the 3rd ACL Workshop on Very Large Corpora,* Cambridge, MA.

Resnik, P. (1995b). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence,* Montreal, Canada. pp. 448–453.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence, 11*, 95–130.

Rho, S., Song, S., Hwang, E. and Kim, M. (2009). COMUS: Ontological and Rule-Based Reasoning for Music Recommendation System In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining,* Bangkok, Thailand. pp. 859–866.

Roget, P.M. (1852). *Roget's thesaurus of English words and phrases*. Burnt Mill, Harlow, Essex: Longman Group Limited.

Rubenstein, H. and Goodenough, J.B. (1965). Contextual correlates of synonymy. *Communications of the ACM, 8*(10), 627–633.

Ruiz-Casado, M., Alfonseca, E. and Castells, P. (2005a). Automatic assignment of Wikipedia Encyclopedic Entries to WordNet synsets. In *Proceedings of the Advances in Web Intelligence,* Lodz, Poland. pp. 380–386.

Ruiz-Casado, M., Alfonseca, E. and Castells, P. (2005b). Automatic Extraction of Semantic Relationships for WordNet by Means of Pattern Learning from Wikipedia. In *Proceedings of the Applications of Natural Language to Information Systems,* Alicante, Spain. pp. 67–79.

Ruiz-Casado, M., Alfonseca, E. and Castells, P. (2006). From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In *Proceedings of the 1st Workshop on Semantic Wikis,* Budva, Montenegro.

Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review, 18*(2), 95–145.

Sanger, L. (2005). The early history of Nupedia and Wikipedia: A Memoir *Open Sources 2.0: The Continuing Evolution* (pp. 307–338). Sebastopol, CA: O'Reilly Media, Inc.

Schiff, S. (2007). Know It All. *The New Yorker* Retrieved 12 February 2010, from http://www.newyorker.com/archive/2006/07/31/060731fa_fact

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

Schoop, M., Moor, A. and Dietz, J. (2006). The pragmatic web: a manifesto. *Communications of the ACM, 49*(5), 75–76

Schraefel, M.C., Wilson, M., Russell, A. and Smith, D.A. (2006). mSpace: improving information access to multimedia domains with multimodal exploratory search. *Communications of the ACM, 49*(4), 47–49.

Schütze, H. and Pedersen, J.O. (1997). A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management, 33*(3), 307–318.

Segond, F., Schiller, A., Grefenstette, G. and Chanod, J. (1997). An experiment in semantic tagging using hidden markov model tagging. In *Proceedings of the Workshop in Automatic Information Extraction and Building of Lexical Semantic Resources for NLP applications,* Madrid, Spain. pp. 78–81.

Shiri, A. and Revie, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology, 57*(4), 462–478.

Shirky, C. (2003). The semantic web, syllogism, and worldview. *Networks, Economics, and Culture.*

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the IEEE Symposium on Visual Languages,* Boulder, CO. pp. 336–343.

Shneiderman, B., Byrd, D. and Croft, W. (1997). Clarifying search: A user-interface framework for text searches. *D-Lib Magazine, 3*(1), 18–20.

Siegenthaler, J. (2005). A false Wikipedia "biography". *USA Today* Retrieved 12 February 2010, from http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm

Sleator, D. and Temperly, D. (1993). Parsing English with a Link Grammar. In *Proceedings of the 3rd International Workshop on Parsing Technologies,* Tilburg, The Netherlands.

Smith, L.C. (1976). Artificial Intelligence in Information Retrieval Systems. *Information Processing & Management, 12*(3), 189–222.

Spärck Jones, K. (1991). The role of artificial intelligence in information retrieval. *Journal of the American Society for Information Science, 42*(8), 558–565.

Strube, M. and Ponzetto, S.P. (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence,* Boston, MA. pp. 1440–1445.

Suchanek, F.M., Ifrim, G. and Weikum, G. (2006). Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. In *Proceedings of the Knowledge Discovery and Data Mining,* Philadelphia, PA.

Suchanek, F.M., Kasneci, G. and Weikum, G. (2007). Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web, 6*(3), 203–217.

Sullivan, D. (2007). Google Ramps Up Personalized Search. *Search Engine Land* Retrieved March 18, 2010, from http://searchengineland.com/google-ramps-up-personalized-search-10430

Summers, E. (2008). Uncool URIs. *The lcsh.info blog* Retrieved March 24, 2010, from http://lcsh.info/comments1.html

Sutcliffe, A. and Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers, 10*(3), 321–351.

Swartz, A. (2002). MusicBrainz: A semantic web service. *IEEE Intelligent Systems and their Applications, 17*(1), 76–77.

Thomas, C. and Sheth, A.P. (2007). Semantic Convergence of Wikipedia Articles. In *Proceedings of the EEE/WIC/ACM International Conference on Web Intelligence,* Silicon Valley, CA. pp. 600–606.

Toral, A. and Munoz, R. (2006). A proposal to automatically build and maintain gazetteers for Named Entity Recognition by using Wikipedia. In *Proceedings of the EACL Workshop on New Text: Wikis and blogs and other dynamic text sources,* Trento, Italy.

Trajkova, J. and Gauch, S. (2004). Improving ontology-based user profiles. In *Proceedings of the 7th RIAO Conference on coupling approaches, coupling media and coupling languages for information retrieval,* Avignon, France. pp. 380–389.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*(1), 433—460.

Viégas, F.B., Wattenberg, M. and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Vienna, Austria. pp. 575–582.

Viégas, F.B., Wattenberg, M., Kriss, J. and Van Ham, F. (2007a). Talk before you type: Coordination in Wikipedia. In *Proceedings of the Hawaii International Conference on System Sciences,* Waikoloa, Hawaii.

Viégas, F.B., Wattenberg, M. and McKeon, M.M. (2007b). The hidden order of Wikipedia. In *Proceedings of the Online Communities and Social Computing,* Beijing, China. pp. 445–454.

Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H. and Studer, R. (2006). Semantic Wikipedia. In *Proceedings of the 15th International Conference on World Wide Web,* Edinburgh, Scotland. pp. 585–594.

Voorhees, E.M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Pittsburgh, PA. pp. 171—180.

Waldman, S. (2004). Who knows? *The Guardian* Retrieved 12 February 2010, from http://www.guardian.co.uk/technology/2004/oct/26/g2.onlinesupplement

Wang, G., Yu, Y. and Zhu, H. (2007a). PORE: Positive-Only Relation Extraction from Wikipedia Text. In *Proceedings of the 6th International Semantic Web Conference,* Busan, Korea. pp. 580–594.

Wang, G., Zhang, H., Wang, H. and Yu, Y. (2007b). Enhancing Relation Extraction by Eliciting Selectional Constraint Features from Wikipedia. In *Proceedings of the Natural Language Processing and Information Systems,* Paris, France. pp. 329–340.

Wang, P., Hu, J., Zeng, H.J., Chen, L. and Chen, Z. (2007c). Improving text classification by using encyclopedia knowledge. In *Proceedings of the 7th IEEE International Conference on Data Mining,* Omaha, NE. pp. 332–341.

Wattenberg, M., Viégas, F.B. and Hollenbach, K. (2007). Visualizing activity on wikipedia with chromograms. In *Proceedings of the Human-Computer Interaction (INTERACT),* Rio de Janeiro, Brazil. pp. 272–287.

White, R. and Roth, R. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*: Morgan & Claypool Publishers.

White, R.W. and Marchionini, G. (2007). Examining the effectiveness of real-time query expansion. *Information Processing and Management, 43*(3), 685–704.

White, R.W., Marchionini, G. and Muresan, G. (2008). Editorial: Evaluating exploratory search systems. *Information Processing and Management, 44*(2), 433–436.

Widdows, D. (2004). *Geometry and meaning*. Palo Alto, CA: CSLI Publications.

Wilkinson, D.M. and Huberman, B.A. (2008). Cooperation and quality in wikipedia. In *Proceedings of the International symposium on Wikis,* Montreal, Quebec, Canada pp. 157–164.

Wu, F., Hoffmann, R. and Weld, D.S. (2008). Information extraction from Wikipedia: Moving down the long tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* Beijing, China. pp. 731–739.

Wu, F. and Weld, D.S. (2006). Autonomously semantifying wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management,* Lisbon, Portugal. pp. 41–50.

Wu, F. and Weld, D.S. (2008). Automatically refining the wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web,* Beijing, China pp. 635–644.

Yan, Y., Ishizuka, M. and Matsuo, Y. (2009). Unsupervised Relation Extraction by Mining Wikipedia Texts supported with Web Redundancy Information. In

Proceedings of the 23rd Annual Conference of the Japanese Society for Artificial Intelligence, Takamatsu, Japan.

Yang, C., Chen, H. and Hong, K. (2003). Visualization of large category map for Internet browsing. *Decision Support Systems, 35*(1), 89–102.

Yang, X.F. and Su, J. (2007). Coreference resolution using semantic relatedness information from automatically discovered patterns. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics,* Prague, Czech Republic. pp. 528–535.

Yee, K.P., Swearingen, K., Li, K. and Hearst, M.A. (2003). Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* Ft. Lauderdale, FL. pp. 401–408

Zesch, T. and Gurevych, I. (2007). Analysis of the wikipedia category graph for NLP applications. In *Proceedings of the HLT-NAACL Workshop on Graph-Based Algorithms for Natural Language Processing,* Rochester, NY. pp. 1–8.

Zhang, Q., Suchanek, F.M., Yue, L. and Weikum, G. (2008). TOB: Timely ontologies for business relations. In *Proceedings of the 11th International Workshop on Web and Databases,* Vancouver, Canada.

Ziegler, C., McNee, S., Konstan, J. and Lausen, G. (2005). Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web,* Freiburg, Germany. pp. 22–32.

Zirn, C., Nastase, V. and Strube, M. (2008). Distinguishing between Instances and Classes in the Wikipedia Taxonomy. In *Proceedings of the 5th European Semantic Web Conference,* Tenerife, Spain. pp. 376–387.

Appendix A. Publications and presentations

The following papers have been published during this investigation. They are listed here to provide self-contained explanations of the experiments that have been conducted, or to give details of work that was mentioned only in passing.

Milne, D. (2010) A link-based visual search engine for Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference*, Wellington, New Zealand.

This paper describes Hōpara and its evaluation. It overlaps significantly with Chapter 7.

Medelyan, O., **Milne, D.**, Legg, C. and Witten, I.H. (2009) Mining meaning from Wikipedia. *International Journal of Human-Computer Studies 67*(9), 716–754.

This paper surveys research on mining Wikipedia and applying it to natural langue processing, information extraction and information retrieval. It overlaps significantly with Chapter 3.

Milne, D. (2009) An open-source toolkit for mining Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference*, Auckland, New Zealand.

This paper describes the Wikipedia Miner toolkit. It overlaps significantly with Appendix C.

Huang, A., **Milne, D.**, Frank, E. and Witten, I. (2009) Clustering documents using a Wikipedia-based concept representation. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Bangkok, Thailand.

This paper applies the WLM semantic relatedness measure (Chapter 5) and detection/disambiguation algorithms (Chapter 6) to text clustering.

Nastase, V., **Milne, D.** and Filippova, K. (2009) Summarizing with encyclopaedic knowledge. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MA.

This paper applies the WLM semantic relatedness measure (Chapter 5) and topic detection/disambiguation algorithms (Chapter 6) to document summarization.

Milne, D. and Witten, I.H. (2008) Learning to link with Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management*, Napa Valley, CA.

This paper describes how to apply machine learning to detect and disambiguate Wikipedia topics when they are mentioned in textual documents, and received the CIKM 2008 best paper award. It overlaps significantly with Chapter 6.

Milne, D. and Witten, I.H. (2008) An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence*, Chicago, IL.

This paper describes and evaluates the WLM semantic relatedness measure. It overlaps significantly with Chapter 5.

Medelyan, O, Witten, I.H., and **Milne, D.** (2008) Topic Indexing with Wikipedia. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence*, Chicago, IL.

This paper applies the WLM semantic relatedness measure (Chapter 5) to topic indexing, and demonstrates the close similarities between this task and wikification (Chapter 6).

Milne, D., Nichols, D.M, and Witten, I.H. (2008) A competitive environment for exploratory query expansion. In *Proceedings of the Joint Conference on Digital Libraries*, Pittsburgh, PA.

This paper describes a game built around the Koru search engine (Chapter 4) that aims to teach users how to search, and gather examples of effective search strategies.

Medelyan, O. and **Milne, D.** (2008) Augmenting domain-specific thesauri with knowledge from Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference*, Christchurch, New Zealand.

This paper describes and evaluates an algorithm for automatically aligning thesauri with Wikipedia.

Milne, D., Witten, I.H. and Nichols, D.M. (2007). A Knowledge-Based Search Engine Powered by Wikipedia. In *Proceedings of the ACM Conference on Information and Knowledge Management*, Lisbon, Portugal.

This paper describes Koru and its evaluation. It overlaps significantly with Chapter 4.

Milne, D. (2007). Computing Semantic Relatedness using Wikipedia Link Structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*, Hamilton, New Zealand.

This paper describes a precursor to the WLM semantic relatedness measure (Chapter 5).

Milne, D., Medelyan, O. and Witten, I. H. (2006). Mining Domain-Specific Thesauri from Wikipedia: A case study. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong.

This paper compares Wikipedia to the domain-specific thesaurus Agrovoc. It overlaps significantly with Section 3.5.

Witten, I. H., Medelyan, O. and **Milne D.** (2006). Finding documents and reading them: Semantic metadata extraction, topic browsing and realistic books. In *Proceedings of the Russian Conference on Digital Libraries*, Suzdal, Russia.

This paper describes a range of techniques and strategies for augmenting digital libraries.

The following videos provide introductory descriptions of work that has been conducted during this investigation.

www.youtube.com/watch?v=NFCZuzA4cFc

This 50-minute tech talk describes the overall thesis investigation, and provides demonstrations of Koru (Chapter 4), the WLM relatedness measure (Chapter 5) and the topic detection/disambiguation algorithms (Chapter 6).

www.videolectures.net/cikm08_milne_ltlww

This 20-minute conference presentation describes the topic detection and disambiguation algorithms and their evaluation (Chapter 6).

Appendix B. Applications and resources

Several applications and resources were constructed as by-products of this investigation. The resource that is most likely to be of interest to researchers is the Wikipedia Miner Toolkit: an open-source suite of code for navigating and making use of the structure and content of Wikipedia. This is described separately in Appendix C.

B.1 Koru

www.nzdl.org/koru

Koru is a search engine that uses Wikipedia to organize documents, make queries more accurate, and suggest new topics for people to explore. Its development and evaluation is the focus of Chapter 4.

B.2 Hōpara

www.nzdl.org/hopara

Hōpara is a search engine that uses visualization, semantic relatedness measures and lightweight information extraction to make Wikipedia easier to explore. Its development and evaluation is the focus of Chapter 7.

B.3 Manually-disambiguated WordSim 353 collection

www.nzdl.org/wikipediaSimilarity

The original WordSimilarity 353 collection is a set of 353 term pairs, each associated with between 12 and 15 human-assigned similarity judgments. Section 5.4.1 describes an experiment in which the term pairs are manually disambiguated to provide a test set for evaluating measures of relatedness between Wikipedia articles (rather than terms). The result is available at the URL above.

B.4 Manually-verified corpus of wikified newswire stories

www.nzdl.org/wikification

Section 6.4 describes an experiment in which 50 newswire stories were automatically "wikified"—augmented with links to relevant Wikipedia articles—and then closely inspected by human evaluators. The result is a new corpus containing only manually verified links, which is available at the URL above.

Appendix C. The Wikipedia Miner Toolkit

This appendix describes the Wikipedia Miner toolkit,[39] a suite of code that was developed during this investigation. It is released open-source to allow developers and researchers to easily explore and draw upon the content of Wikipedia.

Although Wikipedia's content is already readily available,[40] it is challenging to extract and access useful information from it in a scalable and timely manner. As of January 2010, the English Wikipedia dump includes approximately 8M pages—not including revision history or background discussion. Its useful semantic features are buried under 25Gb of cryptic markup. We hope that Wikipedia Miner will simplify access to these features, and allow researchers to avoid re-inventing the wheel and instead invest their time on the novel aspects of their work.

The Wikipedia Miner toolkit consists of four main components: a set of PERL scripts to extract information from Wikipedia's XML dumps; a MySQL database for efficient, persistent access to the summarized data; a Java API to manage the database and provide programmatic access to it; and a suite of human- and machine-readable web services which provide a restricted range of functionality suitable for live web applications and users who would prefer not to host their own version of the toolkit.

C.1 PERL extraction scripts

The PERL scripts are responsible for extracting summaries such as the link graph and category hierarchy. All of the scripts scale in linear time, and can flexibly split the data where necessary in case of memory constraints. All but one of the summaries can be extracted within a day or two on modest desktop hardware. Only the link-likelihood statistics take longer, and are entirely optional; they are only used for detecting Wikipedia topics when they are mentioned in plain text (see Appendix C.3.5). The script to extract them requires approximately ten days, but can be shared across multiple machines. Finally, several pre-summarized versions of Wikipedia are available from the toolkit's website. The entire extraction process can be avoided unless one requires a specific edition of Wikipedia that we have not provided.

[39] Code, data and online demonstrations of the Wikipedia-Miner toolkit are available at *http://wikipedia-miner.sourceforge.net*

[40] Wikipedia's entire content is released every month or so as html and xml dumps at *http://download.wikimedia.org*

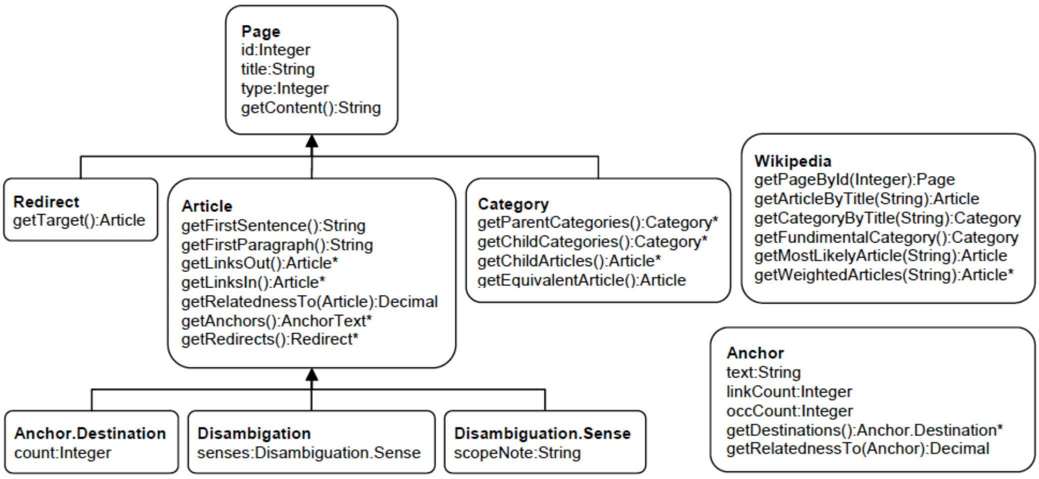

Page
id:Integer
title:String
type:Integer
getContent():String
Redirect
getTarget():Article
Article
getFirstSentence():String
getFirstParagraph():String
getLinksOut():Article*
getLinksIn():Article*
getRelatednessTo(Article):Decimal
getAnchors():AnchorText*
getRedirects():Redirect*
Category
getParentCategories():Category*
getChildCategories():Category*
getChildArticles():Article*
getEquivalentArticle():Article
Wikipedia
getPageById(Integer):Page
getArticleByTitle(String):Article
getCategoryByTitle(String):Category
getFundimentalCategory():Category
getMostLikelyArticle(String):Article
getWeightedArticles(String):Article*
Anchor
text:String
linkCount:Integer
occCount:Integer
getDestinations():Anchor.Destination*
getRelatednessTo(Anchor):Decimal
Anchor.Destination
count:Integer
Disambigation
senses:Disambiguation.Sense
Disambiguation.Sense
scopeNote:String

Figure C.1: A sample of classes in the Wikipedia-Miner toolkit

C.2 MySQL database

After running these scripts (or downloading the pre-prepared data), developers can construct programs to read the resulting summaries—which are simply delimited text files—directly. This does, however, require significant time and memory; the link-graph summary alone occupies more than 1Gb. Instead, the toolkit communicates with a MySQL database, so that the data can be indexed persistently and accessed immediately, without waiting for anything to load. Developers should not need to work with the database directly; it is maintained and accessed via the Java API described below.

C.3 Java API

The largest component of the toolkit is a documented Java API that abstracts away from the data to provide simplified, efficient access to Wikipedia. The following sections describe its most important functions.

C.3.1 Modelling Wikipedia

This section describes classes for modelling Wikipedia's structure and content, and explains how they correspond with the elements of a traditional thesaurus. These classes are shown in Figure C.1, along with their inheritance hierarchy and some selected properties and methods. A more complete description is available from the JavaDoc.[41]

[41] Full documentation of classes and methods is available at *http://wdm.cs.waikato.ac.nz:8080/doc*

Wikipedia itself is, of course, one of the more important objects to model. It provides the central point of access to most of the functionality of the toolkit. Among other things, here you can gather statistics about the encyclopaedia, or access the pages within it through iteration, browsing, and searching.

All of Wikipedia's content is presented on pages of one type or another. The toolkit models every *Page* as a unique id, a title, and some content expressed as MediaWiki markup. More specific functionality depends on the page type.

Articles provide the bulk of Wikipedia's informative content. Each *Article* describes a single concept or topic, and their titles are succinct, well-formed phrases that can be used as non-descriptors in ontologies and thesauri. For example, the article about domesticated canines is entitled *Dog*, and the one about companion animals in general is called *Pet*. Articles follow a fairly predictable layout, and consequently the toolkit can provide short and medium length definitions of concepts by extracting the first sentence and first paragraph from the content. Articles often contain links to equivalent articles in other language versions of Wikipedia. The toolkit allows the titles of these pages to be mined as a source of translations; the article about *dogs* links to (among many others) *chien* in the French Wikipedia, *haushund* in German, and 犬 in Chinese.

Redirects are pages whose sole purpose is to connect an article to alternative titles. These correspond to synonyms and other variations in surface form. The article entitled *dog*, for example, is referred to by redirects *dogs, canis lupus familiaris*, and *domestic dog*. Redirects may also represent more specific topics that do not warrant separate articles, such as *male dog* and *dog groups.* The toolkit allows redirects to be mined for their intended target, and articles to be mined for all of the redirects that refer to them.

Disambiguations are used to group the possible senses of an ambiguous term. However, as Section 3.3.3 explained, they are deceptively difficult to parse, and consequently the toolkit makes little use of them. Anchor texts (described below) provide a cleaner, more abundant, and more easily parsed alternative.

Almost all of Wikipedia's articles are organized within one or more *Categories*, which can be mined for hyponyms, holonyms and other broader (more general) topics. *Dog,* for example, belongs to the categories *domesticated animals*, *cosmopolitan species*, and *scavengers.* If a topic is broad enough to warrant several articles, the central article may be paired with a category of the same name: the article *dog* is paired with the category *dogs*. This equivalent category can be mined for more parent categories (*canines*) and subcategories (*dog breeds*, *dog sports*). Child articles and other descendants (*puppy*, *fear of dogs*) can also be mined for hypernyms, meronyms, and other more specific topics. All of Wikipedia's categories descend from a single root called *Fundamental.* The toolkit uses the distance between a particular article or category and this root to provide a measure of its generality or specificity. According to this measure, *Dog* has a greater distance than *carnivores*, which has the same distance as *omnivores* and a greater distance than *animals*.

The *Anchor* texts of links made to an article provide another source (in addition to redirects) of synonyms and other variations in surface form. The article about *dogs*, for example, has links from anchors like *canis familiaris*, *man's best friend,* and *doggy*. They also provide a cleaner alternative to disambiguations for identifying ambiguous terms: in other contexts, *dog* is used to link to the sign of the Chinese zodiac (狗), or to the broader biological family *canidae*. As Appendix C.3.2 explains, the *Anchor* class is the key component for searching Wikipedia.

C.3.2 Searching for concepts

The Wikipedia Miner toolkit indexes pages so that they can be searched efficiently. The concept-based search it provides is not full-text search. When searching for *dog*, for example, it will not return all articles that contain the term; it will instead return only articles that could reasonably be given that title—only the concepts that *dog* could reasonably refer to.

Wikipedia's raw structure provides three easily processed means of mapping terms or surface forms to concepts: specifying titles of articles, assigning redirects (alternative titles) to articles, and using different terms within the links that are made to an article. For the sake of straightforward searching, all are combined in one place within the toolkit: the association of *anchors* to *anchor senses*. An *anchor* represents the term or phrase that has been used to refer to articles. An *anchor sense* is a direct descendant of the *article* class.

Searching is a matter of instantiating the *anchor* class with the term or phrase of interest, and inspecting the *senses* to which it refers. Unknown terms will have no senses, unambiguous terms will have just one sense, and ambiguous terms will have multiple senses. In the last case, senses are ranked by how often the associations are made: 98% of associations for *dog* refer to the domesticated animal, 1% to the sign of the zodiac, and less than 1% to the biological family *canidae*.

By default, anchor texts are indexed and searched without modifying them in any way. They already encode many of the desired variations in letter case (*Dog* and *dog*), pluralism (*dogs*), and punctuation (*US* and *U.S.*), so automatic term conflation is often unnecessary and may introduce erroneous matches—returning *digital on-screen graphic* (or *DOG*) as a match to *dog*, for example. When modification is desirable, the toolkit provides several text processors—case-folders, stemmers, and punctuation cleaners—to re-index the anchor associations. It has been designed so that many different text processes can be maintained at one time, and quickly swapped in and out. It is also easy for users to develop and apply their own text processors.

C.3.3 Comparing terms and concepts

The toolkit includes the WLM algorithm for generating semantic relatedness measures, which quantify the extent to which different words or concepts relate to each other. According to the toolkit, *dog* is 100% related to *canis familiaris*, 47% related to *domestic animal*, and 19% related to *animal*. These measures have a wide range of applications—particularly for natural language processing and data mining—because they allow terms and concepts to be compared, organized, and reasoned with. The algorithm for generating these measures is described in Section 5.4.

The semantic relatedness measures are available via the *getRelatednessTo* methods defined by the *Article* class (for comparing concepts) and the *Anchor* class (for comparing terms). By default these methods use both in- and out-links read from the MySQL database, but these can be easily cached to memory when greater speed is required. Caching only the in-links provides vastly increased performance and only slightly decreased accuracy (see Section 5.4.1).

C.3.4 Building a thesaurus browser

This section demonstrates how to write a simple thesaurus browser using the Wikipedia Miner toolkit. The application described here, with code and truncated output displayed in Figure C.2, searches Wikipedia to locate different senses of the term *Dog*, and elaborates on the most likely one.

The first line of code creates a new instance of Wikipedia and connects it to a pre-prepared MySQL database. Line 2 queries this instance to get a list of articles that represent the different senses of *Dog*. The second argument of this method call is an optional text processor. As described in Section 2.2, the list of senses returned by this call is obtained by investigating all the times the word *dog* is used as a link anchor, title or redirect. The proportion of references that go to each of the candidate senses is used to sort them, so that the most likely sense—the domestic pet—appears first. Line 6 stores this sense in a new variable, and 7 outputs a short, plain-text definition.

Lines 8–10 output the different anchors that refer to the article about *Dogs*, which correspond to synonyms in a thesaurus. Lines 11–14 output the different links to

```
1.   Wikipedia w = new Wikipedia(“dbServer”, “dbName”) ;
2.   SortedVector<Article> ss = w.getWeightedArticles(“Dog”, null) ;
3.   System.out.println(“Senses for Dog:”)
4.   for(Article s: ss)
5.     System.out.println(“ –“ + s.getTitle()) ;
6.   Article dog = ss.first() ;
7.   System.out.println(dog.getFirstSentence()) ;
8.   System.out.println(“Synonyms: ”) ;
9.   for (AnchorText at:dog.getAnchorTexts())
10.    System.out.println(“ –“ + at.getText()) ;
11.  System.out.println(“Translations: “) ;
12.  HashMap<String, String> ts = dog.getTranslations() ;
13.  for (String lang:ts.keySet())
14.    System.out.println(“ –“ + lang + “, “ + ts.get(lang)) ;
15.  SortedVector<Article> rts = new SortedVector() ;
16.  for (Article rt:dog.getLinksOut()) {
17.    rt.setWeight(rt.getRelatednessTo(dog)) ;
18.    rts.add(rt, false) ;
19.  }
20.  System.out.println(“Related Topics: ”) ;
21.  for (Article rt: rts)
22.    System.out.println(“ –” + rt.getTitle()) ;
```

```
Senses for Dog:
 - Dog
 - Dog (zodiac)
 - Canidae
 ...
The dog (Canis lupis familiaris) is a
domestic subspecies of the wolf, a
mammal of the Canidae family of the
order Carnivora.
Synonyms:
 - Canis familiaris
 - Man’s Best Friend
 - Doggy
 ...
Translations:
 - Chien (fr)
 - Haushund (de)
 - 犬 (zh)
 ...
Related Topics:
 - Dog Breed
 - American Kennel Club
- Pet
 ...
```

Figure C.2: Java code and truncated output of a simple thesaurus browser

equivalent articles in other language versions of Wikipedia, which correspond to translations. The remaining code mines the links within the *Dog* article for related topics. Not all of these links represent useful semantic relations, so lines 15–19 sort the articles according to their semantic relatedness to *Dog*. Lines 20–22 output the result.

Of course, there are many more features of the toolkit that we could take advantage of. We could use the article’s redirects to gather more synonyms, and navigate the article and category links surrounding it to gather more related topics. These could be classified as broader, narrower, or associated topics by inspecting the hierarchical relationships between them, or clustered using semantic relatedness measures. However, we have already done much with just 22 lines of code. A little more work to provide input and selection facilities would yield a useful—and large-scale—thesaurus browser.

C.3.5 Building an HTML wikifier

For any given document, it is probable that Wikipedia knows something about the topics discussed within it and could add additional information. Wikipedia Miner includes algorithms to detect and disambiguate Wikipedia topics when they are mentioned in documents. The resulting connections can be used to give readers additional insight, present information seekers with opportunities to explore further, or provide retrieval systems with more informative document representations. Detailed descriptions of these algorithms are provided in Chapter 6. This section merely explains how to access them using the Java API.

The workflow to “wikify” documents is shown in Figure C.3. The process involves five main steps: preparing documents so that they can be processed cleanly; detecting all terms and phrases within them that could refer to topics; resolving ambiguous terms to clean up the mapping between terms and topics; predicting the link probability (or usefulness) of each topic; and finally marking up the document with references to these topics. The toolkit is intended to be modular, so a separate class handles each step.

The workflow begins in the same way as the thesaurus browser (Appendix C.3.4), by creating an instance of *Wikipedia* and connecting it to a pre-prepared MySQL database. Line 2 instantiates a *CaseFolder* text processor, which will be used to ignore case variations when searching for topics. Line 3 defines a file containing stopwords (e.g. *a, and, or*) that will also be ignored.

Lines 4–9 are needed to ensure timely access to the necessary data. For lengthy documents, the topic disambiguation and detection processes could easily involve thousands of term lookups and semantic relatedness comparisons. Access to the database becomes a significant bottleneck, so caching anchor associations (line 6) and in-links (line 7) to memory is strongly recommended. It is also desirable to cache page titles/types (line 8) and generality statistics (line 9) if space permits. Note that all methods for caching data can optionally accept a set of topic ids that are of interest; missing topics will not be cached and will not be available for detection. Line 5 specifies this as any article that receives at least three links from other articles.

Lines 10 and 11 specify and print the document that will be automatically augmented. This is a piece of HTML markup, which could easily be rendered invalid if altered incorrectly. Lines 12 and 13 specify and use a *DocumentPreprocessor* that will ensure that topics will not be detected from invalid parts of the marukup, such as within tags or between existing anchor links. This example uses a preprocessor designed for HTML. There is another for MediaWiki markup, and more can be developed for other languages.

Lines 14 and 15 specify the *Disambiguator* that will be responsible for resolving ambiguous terms. This is machine-learned, so it requires a model of what good and bad senses look like. Here the model is loaded from a file (there is one provided in the toolkit) but it could also be learned at this point by training the disambiguator with Wikipedia articles (see Section 6.2).

```
1.   Wikipedia wikipedia = new Wikipedia("dbServer", "dbName") ;
2.   TextProcessor textProcessor = new CaseFolder() ;
3.   File stopwordFile = new File("path/to/stopword.file") ;
4.   File dataDirectory = new File("path/to/summarized/data ") ;
5.   TIntHashSet ids = wikipedia.getDatabase().getValidPageIds(dataDirectory, 3, null) ;
6.   wikipedia.getDatabase().cacheAnchors(dataDirectory,  textProcessor, ids, 3,  null ) ;
7.   wikipedia.getDatabase().cacheInLinks(dataDirectory, ids,  null ) ;
8.   wikipedia.getDatabase().cachePages(dataDirectory, ids,  null ) ;
9.   wikipedia.getDatabase().cacheGenerality(dataDirectory, ids,  null ) ;
10.  String html = "<p>This piece of <em>HTML markup</em> has been automatically augmented with links to the
     relevant Wikipedia articles.</p>" ;
11.  System.out.println("Original markup:\n" + html) ;
12.  DocumentPreprocessor preprocessor = new HtmlPreprocessor() ;
13.  PreprocessedDocument doc = preprocessor.preprocess(html) ;
14.  Disambiguator disambiguator = new Disambiguator(wikipedia ,  textProcessor) ;
15.  disambiguator.loadClassifier(new File("path/to/disambiguation.model")) ;
16.  TopicDetector topicDetector = new TopicDetector(wikipedia, disambiguator, stopwordFile, true, false) ;
17.  Collection<Topic> allTopics = topicDetector.getTopics(doc, null) ;
18.  System.out.println("All detected topics:") ;
19.  for (Topic t:allTopics)
20.    System.out.println("\t" + t.getTitle()) ;
21.  LinkDetector linkDetector = new LinkDetector(wikipedia) ;
22.  linkDetector.loadClassifier(new File("path/to/detection.model")) ;
23.  SortedVector<Topic> goodTopics = linkDetector.getBestTopics(allTopics, 0.5) ;
24.  System.out.println("Topics that are probably good links:") ;
25.  for (Topic t:bestTopics)
26.    System.out.println("-" + t.getTitle() + "[" + t.getWeight() + "]" ) ;
27.  DocumentTagger tagger = new MyHtmlTagger() ;
28.  String newHtml = tagger.tag(doc, bestTopics,DocumentTagger.ALL) ;
29.  System.out.println("Augmented markup:\n" + newHtml) ;
```

```
Original markup:
  <p>This piece of <em>HTML markup</em> has been automatically augmented with links to the relevant
  Wikipedia articles.</p>

All detected topics:
   Article (publishing)   Augmented reality    Has Been            HTML             HTML element
   Hyperlink              Markup language      Relevance (law)     Wikipedia

Topics that are probably good links:
   HTML [0.87]                Wikipedia [0.78]                HTML element [0.66]
   Hyperlink [0.63]           Markup language [0.62]

Augmented markup:
  <p>This piece of <em><a href="http://www.en.wikipedia.org/wiki/HTML">HTML</a>
  <a href=" http://www.en.wikipedia.org/wiki/Markup language">markup</a></em> has been automatically
  augmented with <a href="http://www.en.wikipedia.org/wiki/Hyperlink">links</a> to the relevant
  <a href="http://www.en.wikipedia.org/wiki/Wikipedia">Wikipedia</a> articles.</p>
```

Figure C.3: Java code and output of an HTML wikifier

Line 16 specifies a *TopicDetector* that will gather terms and phrases from the document, match them to Wikipedia topics, and use the *Disambiguator* to discard irrelevant senses. The last two arguments specify that it will only return the best sense for each term (i.e. strict disambiguation) and will not return disambiguation pages as topics. This class also provides a first pass to discard topics that are extremely unlikely to be relevant, by ignoring stopwords and terms that are rarely used as links in Wikipedia (these options can be customized).

Lines 17–20 use the detector to gather and print the topics found within the document. A *Topic* is a descendant of the *Article* class, with additional statistics regarding where it is mentioned within the document, its relatedness to other detected topics, and all the other features described in Section 6.3.1.

The detector is a first-pass effort only, and is quite generous in what it will accept as a topic. The example in Figure C.3 lists several dubious items, including *Has Been* (a musical album from William Shatner). Lines 21 and 22 define a *LinkDetector* that separates relevant and irrelevant topics by replicating the decisions made by people who add links to Wikipedia articles (see Section 6.3). As with the *Disambiguator*, it requires training to build a model of what constitutes a valid link (or relevant topic). Here the model is loaded from a file.

Lines 23–26 filter the detected topics to include only those that are at least 50% likely to be a link, and output the result; a much cleaner list of topics. Lines 27–29 construct and use a *DocumentTagger* to insert the topics into the original markup. Note that one of the topics—*HTML Element*—has been discarded during the tagging process. It was obtained from the phrase "HTML markup" and overlaps with the *HTML* and *Markup* topics. The tagger must resolve such collisions, and chose the latter two topics because their average link probability outweighs that of *HTML Element*. As with the *DocumentPreprocessor*, this example uses a tagger designed for HTML. There is another suited to MediaWiki markup, and more can be developed for other languages.

C.4 Web services

The features described up to this point require a significant commitment from the user: one has to download the entirety of Wikipedia and spend days preprocessing it. Fortunately the toolkit provides a suite of web services that give more convenient access to much of its functionality.

There are four services: the *search* service for locating Wikipedia concepts, the *define* service for gathering definitions and icons for concepts, the *compare* service for gathering relatedness measures between terms or concepts, and the *wikify* service for augmenting documents. To access one of these directly, users can issue requests to

http://wdm.cs.waikato.ac.nz:8080/service?task=

followed by the name of the appropriate service (*search*, *define*, *compare* or *wikify*). By default these produce interactive web pages suitable for human users, but can be made to return more easily processed XML messages by appending *&xml* to the URL. This section focuses on this developer-oriented use of the services, because the human-readable interface is self-explanatory.

Note that the examples that follow gloss over many parameters and arguments; more detailed assistance can be obtained by replacing *&xml* with *&help* at any point.

C.4.1 Search

The *search* service returns details of pages and their connections in response to terms or ids. It is essentially a more complete version of the thesaurus browser example discussed in Appendix C.3.4.

Searching via terms produces either a list of candidate articles (if the term is ambiguous), or the details of a single article (if it is not). For example, appending *&term=Dog* to the request results in:

```
<SearchResponse term="dog">
  <SenseList>
    <Sense commonness="0.979" id="4269567" title="Dog">
      <FirstSentence>
        The <b>dog</b> (<i>Canis lupus familiaris</i>, ) is a domesticated <a
        href="service?task=search&id=185901&term=Subspecies">subspecies</a> of the <a
        href="service?task=search&id=33702&term=Gray Wolf">gray wolf</a>, a member of the <a
        href="service?task=search&id=6736&term=Canidae">Canidae</a> family of the order <a
        href="service?task=search&id=5221&term=Carnivora">Carnivora</a>.
      </FirstSentence>
    </Sense>
    <Sense commonness="0.007" id="277029" title="Dog (zodiac)">
      <FirstSentence>
        The <b>Dog</b>(<b>狗</b>) is one of the 12-year cycle of animals which appear in the <a
        href="service?task=search&id=21360689&term=Chinese zodiac">Chinese zodiac</a> related
        to the <a href="service?task=search&id=6966&term=Chinese calendar">Chinese
        calendar</a>.
      </FirstSentence>
    </Sense>
    ...
  </SenseList>
</SearchResponse>
```

Searching via ids is always unambiguous. For example, appending *&id=4269567* gives:

```
<SearchResponse term="Dog">
  <Article id="4269567" title="Dog">
    <FirstParagraph>...</FirstParagraph>
    <RedirectList>
      <Redirect id="6741" title="Canis familiaris"/>
      <Redirect id="215026" title="Dogs"/>
      ...
    </RedirectList>
    <AnchorList totalOccurrences="2764">
      <Anchor occurrences="2171" proportion="0.785" text="dog"/>
      <Anchor occurrences="226" proportion="0.082" text="dogs"/>
      <Anchor occurrences="37" proportion="0.013" text="canine"/>
      ...
    </AnchorList>
    <LanguageLinkList>
      <LanguageLink lang="xh" text="Inja"/>
      <LanguageLink lang="nn" text="Hund"/>
      <LanguageLink lang="gd" text="Cù"/>
      ...
    </LanguageLinkList>
    <CategoryList>
      <EquivalentCategory id="704389" title="Dogs"/>
      <Category id="6835657" title="Cosmopolitan species"/>
      <Category id="18064051" title="Scavengers"/>
    </CategoryList>
    <LinkOutList size="176">
      <LinkOut id="2835" relatedness="0.569" title="Afghan Hound"/>
      <LinkOut id="4765" relatedness="0.546" title="Basenji"/>
      <LinkOut id="5221" relatedness="0.515" title="Carnivora"/>
      ...
    </LinkOutList>
    <LinkInList size="2558">
      <LinkIn id="627" relatedness="0.282" title="Agriculture"/>
      <LinkIn id="681" relatedness="0.4" title="Aardwolf"/>
      <LinkIn id="737" relatedness="0.095" title="Afghanistan"/>
      ...
    </LinkInList>
  </Article>
</SearchResponse>
```

C.4.2 Define

The *define* service provides sentence snippets and image icons in response to article ids. Images are obtained from Freebase (see Section 2.2). The length and format of the snippet, the format of links found within it, and image sizes can all be altered by specifying additional parameters. For example, appending *&id=4269567&length=1 &format=1&getImages=true* results in:

```
<DefinitionResponse id="4269567" title="Dog">
  <Definition>
    The '''dog''' (''Canis lupus familiaris'', ) is a domesticated [[subspecies]] of the [[Gray Wolf|gray
    wolf]], a member of the [[Canidae]] family of the order [[Carnivora]]. The term is used for both
    [[feral]] and [[pet]] varieties. The domestic dog has been one of the most widely kept [[working
    dog|working]] and companion animals in human history.
  </Definition>
  <Image url="..."/>
</DefinitionResponse>
```

C.4.3 Compare

The *compare* service returns semantic relatedness measures between terms or article ids. It can additionally be prompted for details about how the comparison was made. For example, appending *&term1=dog&term2=flea&details=true* generates:

```
<RelatednessResponse relatedness="0.554" term1="dog" term2="flea">
  <Sense1 candidates="7" id="4269567" title="Dog">
    <FirstSentence>…</FirstSentence>
  </Sense1>
  <Sense2 candidates="2" id="77305" title="Flea">
    <FirstSentence>…</FirstSentence>
  </Sense2>
  <LinksIn>
    <SharedLinkList size="42">
      <SharedLink id="6678" title="Cat"/>
      <SharedLink id="1710257" title="Dog collar"/>
      <SharedLink id="4502053" title="Frontline (medicine)"/>
      …
    </SharedLinkList>
    <Link1List size="2516">
      <Link1 id="627" title="Agriculture"/>
      <Link1 id="681" title="Aardwolf"/>
      …
    </Link1List>
    <Link2List size="239">
      <Link2 id="4746" title="Plague (disease)"/>
      <Link2 id="14795" title="Infectious disease"/>
      …
    </Link2List>
  </LinksIn>
  <LinksOut>
    …
  </LinksOut>
</RelatednessResponse>
```

The example above compares ambiguous terms, but it is much more efficient to compare unambiguous ids instead; this should be done whenever possible. For example, appending *&ids1=4269567&ids2=77305;33702* produces:

```
<RelatednessResponse ids1="4269567" ids2="77305;33702">
  77305,4269567,0.554
  33702,4269567,0.654
</RelatednessResponse>
```

Every concept listed in *ids1* (in this case only *Dog*) is compared to every concept in *ids2* (*Flea, Grey Wolf*). The process is extremely fast, so hundreds or even thousands of comparisons can be requested in a single call—for clustering, visualization, and other such applications. For efficiency, the results are returned in a concise CSV format containing one line per comparison.

C.4.4 Wikify

The wikify service augments web pages or snippets of markup with links to the relevant Wikipedia articles. For example, appending *&source=<text snippet>* results in:

```
<WikifierResponse bannedTopics="" minProbability="0.5" repeatMode="2" sourceMode="0">
  <Source>
    This piece of '''MediaWiki markup''' has been automatically augmented with links to the relevant
    Wikipedia articles
  </Source>
  <Result documentScore="1.575371" outputMode="3">
    This piece of '''[[MediaWiki]] [[Markup language|markup]]''' has been automatically augmented
    with [[Hyperlink|links]] to the relevant [[Wikipedia]] articles
  </Result>
  <DetectedTopicList>
    <DetectedTopic id="323710" title="MediaWiki" weight="0.956120"/>
    <DetectedTopic id="5043734" title="Wikipedia" weight="0.775238"/>
    <DetectedTopic id="49547" title="Hyperlink" weight="0.634209"/>
    <DetectedTopic id="18910" title="Markup language" weight="0.622337"/>
  </DetectedTopicList>
</WikifierResponse>
```

The source argument can be a snippet of MediaWiki or HTML markup, or the URL of web page; the result is formatted accordingly. There are additional parameters to make the algorithm more or less restrictive in what it considers links, simplify the output (to return the augmented markup without the surrounding XML), handle repeat mentions of topics, alter link colors, and add javascript tooltips to links (only valid when processing URLs).

Appendix D. References for mining Wikipedia

The following references mine Wikipedia's structure or content for human-readable or machine-readable knowledge. They are grouped according to expressiveness and formality they aspire to, from simple controlled vocabularies to highly complex ontologies. Items in italics apply the knowledge they gather to some task, such as entity tagging or query expansion. The classifications are discussed in Section 3.6.

The following paper treats Wikipedia as a **controlled vocabulary**, by making exclusive use of article and category titles.

(Geva 2007)

The following **21** papers treat Wikipedia as a **glossary**, by using the textual content of pages in addition to titles. Papers that use Wikipedia as a tagged corpus (where link markup identifies entities) are also included.

(Banerjee et al. 2007)

(Biuk-Aghai 2006)

(Csomai and Mihalcea 2007)

(Denoyer and Gallinari 2006)

(Gabrilovich and Markovitch 2006)

(Itakura and Clarke 2007)

(Kazama and Torisawa 2007b)

(Mihalcea and Csomai 2007)

(Nothman et al. 2009)

(Ruiz-Casado et al. 2005a)

(Yang and Su 2007)

(Buscaldi and Rosso 2006)

(Bunescu and Paşca 2006)

(Cucerzan 2007)

(Egozi et al. 2008)

(Gabrilovich and Markovitch 2007)

(Jenkinson et al. 2008)

(Mihalcea 2007)

(Nothman et al. 2008)

(Potthast et al. 2008)

(Toral and Munoz 2006)

The following **5** papers treat Wikipedia as a **taxonomy**, by applying the category network in addition to the structural elements described above.

(Bunescu and Paşca 2006)

(Schonhofen 2006)

(Zesch and Gurevych 2007)

(Li et al. 2007)

(Strube and Ponzetto 2006)

The following **3** papers treat Wikipedia as a **thesaurus**, by using **inter-article links** (as an indication of relatedness rather than as entity tags) in addition to the structural elements described above.

(Minier et al. 2007)
(Overell 2006)
(Wang et al. 2007c)

The following **31** papers treat Wikipedia as an **ontology**, by using taxoboxes and infoboxes, proposing modifications to Wikipedia's markup, or performing natural language processing of other structural elements to obtain formal, machine-readable knowledge.

(Adafre et al. 2007)
(Auer et al. 2007)
(Auer and Lehmann 2007)
(Bhole et al. 2007)
(Blohm and Cimiano 2007)
(Chernov et al. 2006)
(Culotta et al. 2006)
(Herbelot and Copestake 2006)
(Krötzsch et al. 2005)
(Krötzsch et al. 2006)
(Medelyan and Legg 2008)
(Muchnik et al. 2007)
(Nakayama et al. 2007a)
(Nakayama et al. 2007b)
(Nakayama 2008)
(Nastase and Strube 2008)
(Nguyen et al. 2007a)
(Nguyen et al. 2007b)
(Ponzetto and Strube 2007)
(Ruiz-Casado et al. 2005b)
(Ruiz-Casado et al. 2006)
(Suchanek et al. 2006)
(Suchanek et al. 2007)
(Völkel et al. 2006)
(Wang et al. 2007a)
(Wang et al. 2007b)
(Wu and Weld 2006)
(Wu and Weld 2008)
(Wu et al. 2008)
(Yan et al. 2009)
(Zirn et al. 2008)