Working Paper Series
ISSN 1177-777X

EXTRACTING CORPUS-SPECIFIC KNOWLEDGE BASES FROM WIKIPEDIA

David Milne, Ian H. Witten and David M. Nichols

Working Paper: 03/2007
June 2007

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Extracting Corpus Specific Knowledge Bases from Wikipedia

David Milne   Ian H. Witten   David M. Nichols

Department of Computer Science, University of Waikato
Private Bag 3105, Hamilton, New Zealand
+64 7 838 4021

{dnk2, ihw, dmn}@cs.waikato.ac.nz

ABSTRACT

Thesauri are useful knowledge structures for assisting information retrieval. Yet their production is labor-intensive, and few domains have comprehensive thesauri that cover domain-specific concepts and contemporary usage. One approach, which has been attempted without much success for decades, is to seek statistical natural language processing algorithms that work on free text. Instead, we propose to replace costly professional indexers with thousands of dedicated amateur volunteers—namely, those that are producing Wikipedia. This vast, open encyclopedia represents a rich tapestry of topics and semantics and a huge investment of human effort and judgment. We show how this can be directly exploited to provide WikiSauri: manually-defined yet inexpensive thesaurus structures that are specifically tailored to expose the topics, terminology and semantics of individual document collections. We also offer concrete evidence of the effectiveness of WikiSauri for assisting information retrieval.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content analysis – *Linguistic processing, Thesauri.*

H.3.3 [**Information Storage and Retrieval**]: Search and Retrieval – *search process, query formulation.*

General Terms

Algorithms, Design, Experimentation.

Keywords

Wikipedia, Data Mining, Thesauri, Disambiguation, Semantic Relatedness, Query Expansion.

1. INTRODUCTION

For all of their popularity and usefulness, modern search engines exhibit very little intelligence. At their core they are simply pattern matchers; they take a sequence of characters and find documents that repeat it. This is perfectly adequate if one recognizes the limitations and knows enough of what is being searched for to form an effective pattern. But many require more. They expect search engines to know what is out there and help them get to it; like the knowledgeable old man behind the counter at the local library. Unfortunately that old librarian is a deceptively complex creature. Even the slickest search engines cannot emulate his ability to interpret hazy requests, make suggestions, prompt for clarification, and ferret out the required documents. Such functionality requires a deep understanding of the information available; of the topics involved and how they relate to each other. This paper aims to provide the foundation by which this understanding can be obtained.

This is by no means a new goal; knowledge-based information retrieval has been pursued for decades. The first approaches focused on augmenting documents with human defined knowledge: thesauri, taxonomies, metadata and categories. This approach is wonderfully accurate, but fundamentally incapable of coping with the current deluge of information. And so we turned to computers and complex algorithms to replace slow, expensive indexers. The goal is to have computers that are capable of truly understanding written text, and organizing it as a human expert would. This is a hugely complex task; a milestone in computing that we may not reach for many decades.

This paper offers a shortcut: a way to provide knowledge bases automatically, without expecting computers to replace expert human indexers. Essentially we replace the professionals with thousands or even millions of amateurs instead: with the growing community of contributors who form the core of Web 2.0. In the next section we describe this recent explosion of public contribution, and specifically how one of its children—Wikipedia—is being used to leapfrog over the current limitations of natural language processing. In Section 3 we present our own techniques for exploiting Wikipedia to provide manually-defined yet cheap knowledge bases that are specifically tailored to individual document collections and those seeking knowledge from them. Section 4 offers concrete evidence of the effectiveness of these structures for assisting information retrieval, in the form of a detailed user study.

2. WEB 2.0 AND WIKIPEDIA

The term Web 2.0 was coined in 2004 to describe a change in the way the internet functioned.[9] This fuzzy concept defies definition, but if there is one thread that unites the myriad services behind Web 2.0, it is *public contribution*. Under the Web 2.0 model, services eschew the traditional separation between consumer and producer: *YouTube* enables any film-buff to share their directing efforts; those perusing *flickr* galleries are encouraged to publish their own photos; *digg* and *del.icio.us* users

CIKM'07, November 6–9, 2007, Lisbon, Portugal.

looking for interesting websites can guide others to their own favorites.

To some, Web 2.0 represents a profound empowerment of the individual. They see it as removing the blinkers from mainstream media: news is no longer dictated by syndicated stations; movies are no longer the exclusive domain of big-budget studios; music is freed from the pressure of big-name record labels. Creative people are given a direct conduit to the public. The middle man and 'the man' are taken out of the process.

Others view the entire movement with skepticism[5], and liken it to the seductive but empty promises of communism and the misguided idealism of the hippie movement. Far from freeing creativity, they see Web 2.0 as endangering it. Traditional mainstream media—for all its trappings and commercialism—is a filter. It ensures that we are only exposed to informed reporters and talented artists, rather than the rambling opinions and doodles of Joe Average. If Web 2.0 gives everyone a voice, they wonder if anything of worth will be heard.

Regardless, there is one concrete lesson to take from Web 2.0 that couldn't be further from hippie idealism: it has revealed an enormous workforce that was previously overlooked. It is often remarked that we have more information available now than we know what to do with; a problem that Web 2.0 has undoubtedly exacerbated. But perhaps it also offers the solution: a vast workforce that is capable of understanding and reasoning with the information it encounters. The most direct examples of this are social book-marking services such *digg* and *del.icio.us*, which rely on thousands of willing indexers and reviewers. In the absence of truly literate and intelligent search engines, these web citizens have stepped in and collectively expended an enormous effort to tag, organize and rank information. To reverse the old hippie mantra: *power from the people.*

It would be unwise to rely on the volunteer public as the exclusive portal to the world's knowledge. A quick glance through the offerings of *del.icio.us* reveals rampant bias towards the tastes of its contributors: a young, technology-oriented crowd with a taste for in-jokes. But still, they could teach search engines a lesson or two. We mean this very literally: we aim to take the knowledge and judgment expended by Web 2.0 users and eject it directly into search engines—hopefully allowing them to exhibit the same powers of reasoning.

2.1 Wikipedia

We focus on Wikipedia: arguably the largest and best known example of Web 2.0 public collaboration. Wikipedia was launched in 2001 with the goal of building free encyclopedias in all languages. Today it outstrips all other encyclopedias in size and coverage, and is one of the most visited sites on the web. Out of more than three million articles in 125 different languages, one-third are in English, yielding an encyclopedia almost ten times as big as the Encyclopedia Britannica, its closest rival.

Wikipedia's success is due to its editing policy. By using a collaborative wiki environment it turns the entire world into a panel of experts, authors and reviewers [6]. Anyone who wants to make knowledge available to the public can contribute an article. Anyone who encounters an article is able to correct errors, augment its scope, or compensate for bias. The result of this freedom of editing is the largest, fastest growing, most up to date encyclopedia in existence.

Even better, Wikipedia is an open source project that makes itself easily available for research. Its entire content can be obtained in the form of database dumps that are released sporadically, from several days to several weeks apart[1]. A daunting amount of data is released here; hundreds of gigabytes of text and images, and hundreds of megabytes of statistics and links. For those who are interested in working with this data (and in particular its link structure) we recommend WikipediaMiner[2], a toolkit that we have developed which wraps this information in an easy to use, well documented java API.

2.2 Wikipedia and Natural Language Processing

What makes Wikipedia particularly attractive as a knowledge base is that it represents a vast domain-independent pool of manually defined terms, concepts and relations. We are not the first to recognize and take advantage of this. In the last few years there has been a raft of papers documenting Wikipedia's discovery as a source of semantic knowledge and a promising tool for natural language processing.

Pure algorithmic natural language processing is typically very brittle [4]. The statistics, trends and rules it relies on may hold for specific tasks and in specific situations, but cracks appear when one moves into a new, unforeseen domain or task. Consequently there has been much work to enhance NLP algorithms with human-defined common-sense and world knowledge. This typically takes the form of painstakingly created ontologies and lexical databases like WordNet.

According to Ruiz-Casado et al. [11], Wikipedia articles can be easily and accurately matched to entries in these lexical resources; they advocate the use of Wikipedia to extend them. Wikipedia is gaining support as a fully-fledged semantic resource in its own right. Advocates site many reasons to consider Wikipedia as not just equivalent but perhaps superior—domain-independent coverage, constant maintenance, exponential growth, multilingualism—and their results are certainly promising. A standout example is Explicit Semantic Analysis [3], which uses Wikipedia to provide measures of semantic relatedness that far outstrip those generated from any other resource.

Wikipedia was never intended to be used as a knowledge base for NLP and turning it into a viable one is no trivial task. *DBPedia*[3] is a community based effort that uses both manual and automatic data extraction to construct onlogologies from Wikipedia's templates. Ponzetto [10] aims to do the same entirely automatically; and has so far succeeded in inducing a subsumption (is-a) hierarchy over Wikipedia's network of categories. In either case, the goal is to generate formal knowledge bases that describe the world; the topics (people, places, objects, ideas…) that exist and how they all fit together.

Our own goal is subtly different: we aim to construct knowledge bases that are focused to the task of organizing and facilitating retrieval within individual document collections. Given a set of documents, we want to produce a knowledge base that describes only the terms, topics and relations that are relevant to it; only

[1] http://en.wikipedia.org/wiki/Wikipedia:Database_download

[2] http://wikipedia-miner.sourceforge.net

[3] http://www.dbpedia.org

those that will be useful and relevant to those seeking information from those documents.

3. EXTRACTING FOCUSED KNOWLEDGE BASES FROM WIKIPEDIA

We have investigated the extraction of corpus-specific knowledge bases from Wikipedia before: in [8] we showed that Wikipedia could provide a viable alternative to Agrovoc: a well known thesaurus for the agricultural domain. Interestingly we also showed that all the necessary information could be obtained efficiently from only the skeleton structure of Wikipedia (page titles and hyperlinks) without complex processing of the text within it. Unfortunately we also showed that actually obtaining thesauri from Wikipedia would not be trivial; there are many problems that prevent the terms and relationships defined in Wikipedia from being easily matched to those found in thesauri. The remainder of this section describes our first attempts in overcoming these problems and, to our knowledge, the first complete process for automatically extracting thesauri from Wikipedia. In keeping with our previous research, all of the work described in this section uses only the skeleton structure of Wikipedia, rather than its full content.

The basis of our technique is to use Wikipedia's articles as the building blocks of the thesaurus, and its skeleton structure of hyperlinks to tell us which blocks we need and how these should fit together. Each article describes a single concept; its title is a succinct, well-formed phrase that resembles a term in a conventional thesaurus—and we treat it as such. Concepts are often referred to by multiple terms, in which case a thesaurus groups the preferred term with multiple alternatives. e.g. *money* might be grouped with *cash, currency,* and *legal tender*. Wikipedia does much the same thing with redirects; psudo-articles that exist only to connect an alternative title of an article with the preferred one. In [8] we showed that these redirects match the synonymy described in thesauri almost perfectly. Thus we see that Wikipedia is already nicely segmented into individual concepts which link to the terms by which they can be referred. All that remains for building thesauri is to identify the relevant concepts for our needs, and to place these in a structure that allows navigation between related concepts. We will tackle each of these tasks in Sections 2.2.2 and 2.2.3 respectively, but first it is necessary to describe a step that both of these others depend on; that of extracting measures of semantic relatedness between Wikipedia articles.

3.1 Extracting Semantic Relatedness Measures

Semantic relatedness measures are quantifications of the relatedness between terms or concepts. One might say that *cash* and *currency* is 100% related, or *currency* and *bank* are 85% related. Such measures are useful for a wide range of tasks in natural language processing, including—as we will soon see—the task of building thesauri.

Semantic relatedness is a fuzzy, subjective measure, and obtaining it is a deceptively complex task. For example it is difficult to quantify the relationship between *love* and *sex*; the measure is muddied by the ambiguity of the terms (are we talking of interpersonal relationships or biology?), and our own attitudes towards the concepts these terms represent. Despite this subjectivity, people are capable of creating semantic relatedness measures fairly consistently. In [2], Finkelstein et al. had at least 13 participants individually define semantic relatedness measures for over 350 term pairs. The measures were surprisingly consistent; the average correlation between an individual's judgments and those of the whole group was 79%.

This consistency suggests that the process could be automated, and there have been many attempts to do so. As with humans, these techniques must have background knowledge about the terms and concepts involved, and can be taxonomized by where this background knowledge is obtained. Up until a 2005 there were only two resources available: either statistical analysis of large corpora, or hand-crafted lexical structures such as thesauri. In either case the results were poor: the best technique was Latent Semantic Analysis [1], whose measures have only a 56% correlation with manual judgments.

It is likely the background knowledge that is the limiting factor; corpora are unstructured and imprecise, and thesauri are limited in scope and scalability. These limitations are the motivations behind several new techniques which infer semantic relatedness from the structure and content of Wikipedia. We have already described how Wikipedia forms a rare combination of both scale and structure, the strengths of both unstructured corpora and hand-crafted thesauri. A new technique called ESA [3] capitalizes on this to provide measures with a 75% correlation to manual judgments; far above any other automatically generated measure and a mere 3% below the average human.

We do not use ESA, despite its impressive accuracy, for two reasons. Firstly, the task of constructing WikiSauri we only require a measure of relatedness between disambiguated Wikipedia articles. ESA provides relatedness measures between raw, potentially ambiguous terms; a more difficult problem that requires a more sophisticated solution. Secondly, ESA relies the full text of Wikipedia articles, which goes against our objective of using only the skeleton structure of titles and hyperlinks.

Thus we use our own measure of semantic relatedness, which is described in detail in [7]. It quantifies the strength of the relation between two Wikipedia articles by weighting and comparing the links found within them. As we uncovered in [8], these links vary greatly in usefulness, so we weight each link by the probability of it occurring. Links are considered less significant for judging the similarity between articles if many other articles also link to the same target; e.g. the fact that two articles both link to *science* is much less significant than if they both link to a specific topic such as *atmospheric thermodynamics*.

Our most accurate measure was obtained by constructing and comparing vectors of these link weights, in much the same way as *tf-idf* and other weights are commonly compared in the vector space model. This resulted in measures with a correlation of 72% with manual judgments; not far off the bar set by ESA. Unfortunately we found that this measure had a distinct bias towards more obscure articles, which was especially destructive for constructing thesauri. Consequently we use a less accurate measure, which involved summing the weights of the links that were common to both articles. This results in a 59% correlation with manual judgments.

When comparing these to previous measures, it should be noted that we consider the semantic relatedness between articles, while others tackle the more sophisticated problem of comparing raw terms. Our measures can easily be adapted to do the same (through disambiguation of terms to articles) but this causes substantial drops in accuracy: to 45% for the vector based measure and 52% for the less sophisticated one.

3.2 Disambiguating unrestricted text

The first step in constructing a thesaurus is to decide which topics it should contain. In this paper we are interested in providing corpus specific thesauri; we want a thesaurus that encompasses all of the topics discussed within a collection of documents. Thus we must work through each document in turn, identifying the significant terms, and matching these to individual articles in Wikipedia. To our knowledge, this is the first approach for disambiguating unrestricted text to Wikipedia articles.

To lift the significant terms from their surrounding prose, we parse the text into a tree of it's grammatical components. From this we can identify nouns and noun-phrases; the terms that refer to people, places, and other concepts of interest. Figure 1 shows an example sentence being parsed into several noun phrases. We then need to identify candidate concepts for these terms in Wikipedia. For *airfield* this is trivial; there exists only one entry with this as its title: a redirect that points to the correct concept *aerodrome*. No match can be found for the other noun phrases: *decrepit plane* and *patched wings.* These must be broken up into their components, of which *plane* and *wings* are the only nouns. Here we have an example of ambiguity; *plane* might refer to propeller-driven aircraft, a theoretical surface of infinite area and zero depth, or a tool for flattening wooden surfaces. We must gather all of these potentially relevant concepts. Wikipedia's use of redirects and disambiguation pages means this can be done efficiently using only page titles and links; the full process is described in [7].

Selecting the correct sense for a term from a list of candidates is a task known as disambiguation. The sheer scale of Wikipedia makes this especially important; it can yield a vast number of senses for many terms. For example, the term *Jackson* could, according to Wikipedia, refer to over 50 different locations and over 100 different people. If we were to include all of these in the thesaurus, it would quickly become bloated and unfocused.

As with any other approach, we use the context surrounding a term to disambiguate it. In the example shown in Figure 1, we have certain clues surrounding *plane* that makes the indented sense obvious to the reader. Our automatic approach uses unambiguous and previously disambiguated terms as clues, and our previously defined semantic relatedness measures to reason with them. In our example, *aerodrome* is the only unambiguous concept. If we compare all of the candidate articles for *plane* with this, we have a clear winner; *fixed-wing aircraft* has a much higher semantic relatedness, and is thus identified as the correct sense. *Wing* would be disambiguated in much the same way, but now we would have two unambiguous concepts to compare it to; both *fixed-wing aircraft* and *aerodrome*.

This approach breaks down when the context is insufficient; when there are no unambiguous terms, or if several candidate senses are equally valid. The later case is illustrated when disambiguating *wing;* the senses *flight appendage* and *airforce unit* relate equally well to the context available. In this case we take a cascading approach: if there is not enough information at the sentence level to disambiguate a term, we cascade up to the surrounding paragraph and use all of it as context. Our example might go into describe other characteristics of the old plane—its droning engine or whirling propeller—thus providing more context to work with. Similarly, if any terms are left ambiguous at the paragraph level they are cascaded up to be compared with concepts extracted from the entire document. Finally, if any terms are left ambiguous at the document level (a rare occurrence) we give up; we simply include all of the equally likely candidate senses.

After processing an entire collection of documents in this way, we are left with a list of disambiguated concepts that are discussed within it: the topics and terminology of our corpus-specific thesaurus. We can expect this vocabulary to be rather broad, since it will contain every topic discussed within the documents; even those mentioned only in passing. A professional indexer might object to this, and instead restrict the thesaurus to only those topics that are central to the intent of the collection. We chose to weight the topics instead, so that the most significant ones can be emphasized and the rest remain available. In the process of identifying topics, we produce two useful measures for this purpose. The first is the traditional *tf-idf* weight, which is based on the assumption that a significant topic for a document should occur many times within it, and be useful in distinguishing the document from others. The second is based on the assumption that a significant topic should relate strongly to other topics in the document; here we use the average semantic relatedness measure between a topic and all of the others identified. Using these measures we weight each topic's significance to the documents in

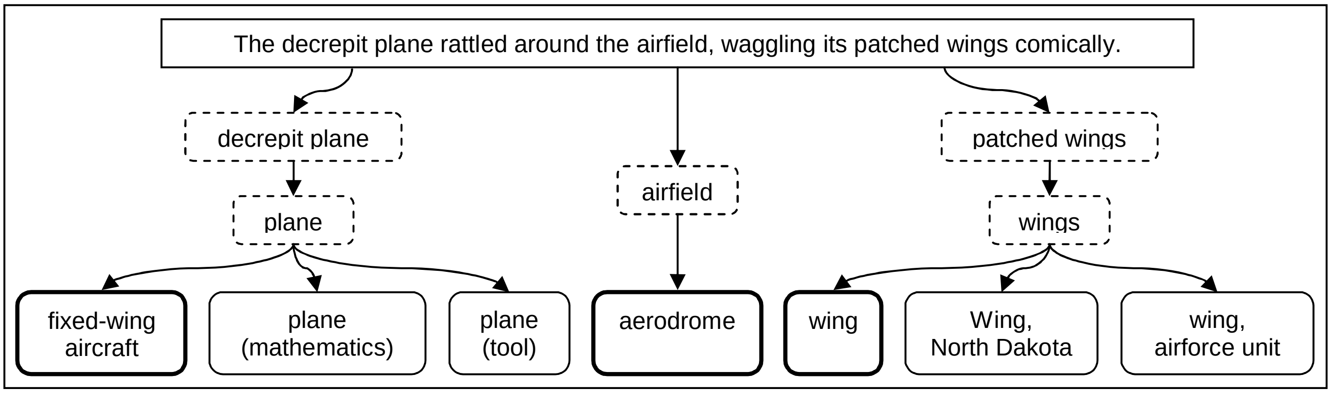

Figure #: Disambiguating unstructured text to Wikipedia articles which they are found and, by aggregating these, it's significance for the collection as a whole.

3.3 Identifying relations between concepts

This section is concerned with identifying semantic relations between the relevant Wikipedia topics identified in the previous section. In [8] we discussed the remarkable similarities between Wikipedia's network of articles, redirects, and categories, and the structure of traditional thesauri. Wikipedia's redirects correspond to synonymy (USE-USE FOR relations); its tree-like network of categories encodes hierarchical relations (BT-NT), and inter-article links correspond to flat relations (RT). If these promising theoretical similarities between thesauri and Wikipedia were accurate, then identifying relations would be trivial.

Unfortunately, the same work revealed they were not; with the exception of redirects, the relations described by Wikipedia's structure do map accurately to those in traditional thesauri. Hierarchical and flat relations are not cleanly separated as the structure would suggest, but are intermingled in both category and article links. Consider the example shown in Figure ?: the article links for *aircraft* point to broader (*craft* and *flight*), related (*aerodynamics*) and narrower topics (*rockets*, *balloons*, and *gliders*). Additionally, many direct links are irrelevant and need to be discarded, while other additional relations need to be inferred from chains of links.

Our process for mining useful semantic relations from Wikipedia's link structure starts with identifying candidate relations from article and category links. Again, consider the example of *aircraft* shown in Figure ?. If we follow its article links we would reach all of the articles on the left side of the diagram; *craft, flight, balloons, etc.* Any of these articles that are relevant to our documents (as described in the previous section) are candidate related topics for *aircraft.* If we follow its category links we would reach *aviation* (another potentially related topic), and, oddly, *aircraft*. Articles and categories often paired in this way; both describe the same topic. Such pairs can easily be identified, they always link to each other, and share the same name either directly or via redirects. Equivalent categories provide more links to related topics; *aircraft* belongs to the category *vehicles*, and is the parent of several more specific categories, including *concept aircraft* and *aircraft components*. Categories also contain descriptive text that commonly links to yet more articles, in this case *airworthiness* and *powered lift*.

From [8], we can expect such article and category links to yield ~69% of the BT-NT and ~66% of RT relations that are described in a typical thesaurus. We could increase this by examining chains of links (e.g. links from *aviation* or *aircraft components*) but this has not yet been attempted; partly because the current process already yields a large number of links (over ## for the *aircraft* example). Too many: from [8] we can expect only 16% precision for category links and 5% precision for article links.

Clearly the vast majority of links we extract do not relate to those found in traditional thesauri. Precision is likely improved through the intersection with a domain-specific corpus; were relations are discarded if they are not relevant to the source documents. It could also be improved through the use of our previously defined semantic relatedness measure by defining a minimum strength below which all relations are discarded. It is debatable whether such precision is even desirable, however. Wikipedia has far more contributors than any thesaurus; and it is plausible that they are able to produce a more richly-connected but equally valid structure. Consequently we do not implement a relationship threshold. Instead, we include all potential relationships in our thesaurus, and weight them by their strength. In this way, we can emphasize the strongest relations, but also make weaker, potentially less accurate relations available.

Ideally, the next step would be to categorize the relations by type; as broader, narrower or related if following the ISO 2788 standard for thesauri, or as meronomy, hyponymy, etc if implementing a more sophisticated ontology. Wikipedia provides many resources for this task: from Figure # it is clear that categories should be useful, and there are other elements in Wikipedia that we have not yet touched on: most notably its infoboxes and templates which explicitly tag links by their type. We have not yet attempted to exploit these resources. As will be shown in Section #, it is useful merely to know that two topics are related, without knowing the type.

4. CASE STUDY: QUERY EXPANSION

In this section we provide concrete evidence of the effectiveness of WikiSauri for enhancing information retrieval. While they could be applied to searching in many ways—we elaborate on this in Section 5—we focus on the task of query expansion. Here users initial queries are enhanced with additional terms and phrases, with the goal of improving recall of documents without sacrificing precision. To this end we developed and evaluated Koru: a knowledge-based search engine that uses WikiSauri to recognize and evolve user's queries.

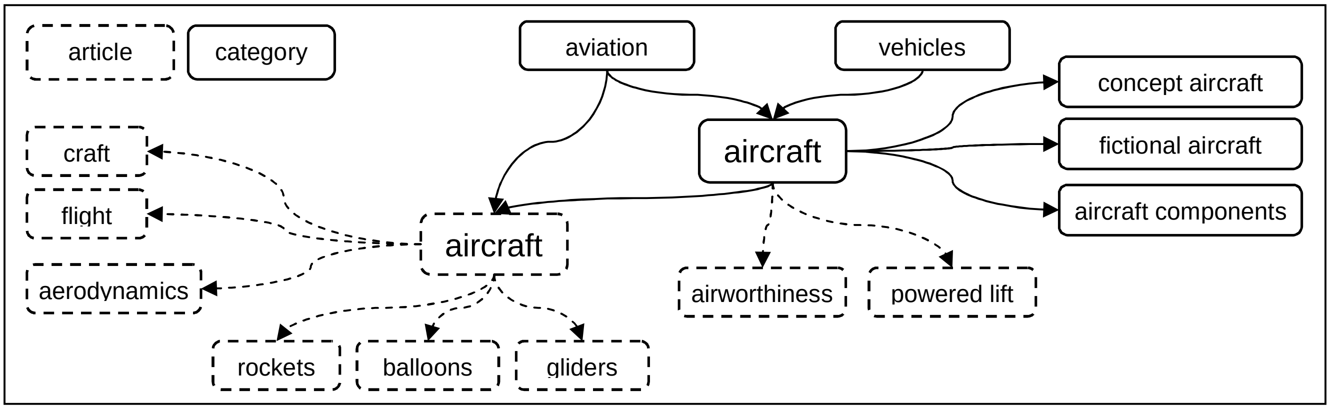

Figure #: Related categories and articles for *aircraft*

4.1 Koru

The Koru topic browsing system is a search engine in which users can progressively work towards the information they seek.[4] It exhibits an understanding of the topics involved in both queries and documents, allowing them to be matched more accurately by evolving queries both automatically and interactively.

These facilities depend on an ability to recognize the topics involved in users queries. Sophisticated entity extraction is not used: instead the words and consecutive sequences of words in the query are checked against entries in an appropriate WikiSaurus. The only “intelligence” in the process is embodied in the knowledge-base and the techniques used to generate it. Because this is exceptionally comprehensive, relates specifically to the document collection, and is backed up by a resource that excels in describing contemporary language, we anticipate that most queries that are valid for the collection will be recognized, even when non-technical terminology or slang is used.

The mere act of recognizing query terms offers an immediate benefit: the ability to identify multi-word terms such as *concept aircraft* and *powered lift*. Users of traditional interfaces are expected to encase these within quotes, but they rarely bother to do so. Koru performs this tedious task automatically, thus improving the precision of the results returned.

Larger benefits are gained through the inclusion of synonyms; different ways of talking about the same topics. If a user enters the term *George Bush*, for example, then Koru will—in the same breath—match formal documents that talk of *George W. Bush*, news stories that mention *President Bush*, and blogs about *Dubya*, *Shrubya* or *Baby Bush*.

More – some examples

4.2 Evaluation

This section describes a user study which aimed to evaluate Koru and its underlying data structure for their ability to facilitate and improve information retrieval. Again, the study is described in more detail elsewhere. Here we focus on whether the topics, terminology and semantics extracted from Wikipedia make a conclusive, positive difference in the way users locate information. This was measured by pitting Wikisaurus-based query expansion against traditional keyword search.

4.2.1 Procedure

Twelve participants were observed as they interacted with the two systems. All were experienced knowledge seekers; graduate or undergraduate computer scientists with at least 8 years of computing experience, and all use Google and other search engines daily. Each user was required to perform 10 tasks (one of which is shown in Table 1) 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.

4.2.2 Tasks and Data

Our tasks and documents were selected to preserve authentic query behavior while allowing detailed measurements of their effectiveness. We sourced both from the TREC 2005 HARD track, 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. Take the example shown in Figure ##: it is a task that forces users to think carefully about their query terms, and is 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. For each task, approximately ## relevance judgments are made; in which a document is identified as strongly relevant, weakly relevant, or irrelevant. This allows us to measure the effectiveness of every query issued; once we identify the appropriate task, we can deduce exactly which documents the user intended to retrieve.

The ACQUAINT text corpus is a large one—about 3GB uncompressed. It was impractical to create a WikiSaurus for the entire collection because the process has not been optimized. Instead we used a subset of the corpus: only stories from Associated Press, and only those mentioned in the relevance judgments for the 10 tasks. The result is a collection of approximately 1200 documents concerning a wide range of topics. This was used throughout the experiments.

A thesaurus was created automatically for this document collection, based on a snapshot of Wikipedia released on June 3, 2006. The full content and revision history at this point occupy 40 GB of compressed data. We use only the link structure and basic statistics for articles, which consume 500 MB (compressed).

	Wikipedia	**WikiSaurus**
topics	1,110,000	20,250
terms	2,250,000	57,276
relations	28,750,000	366,384

Ambiguous document terms	
according to Wikipedia	8504
according to WikiSaurus	3026

Polysemous document topics	
according to documents	2026
according to Wikipedia	6798
according to WikiSaurus	8722

Table 2: Details of Wikipedia and the extracted thesaurus

Details of the information available in Wikipedia at this time, and of the thesaurus that was produced, are shown in Table 2. While processing the 1200 documents about 18,000 terms were encountered that matched at least one article in Wikipedia. These are candidates for inclusion in our thesaurus. Including multiple matches yields 20,000 distinct topics—about 2% of those available in Wikipedia.

The disambiguation techniques described in Section 3 greatly reduce the number of multiple matches but do not eliminate them entirely: 47% of terms are ambiguous according to Wikipedia, but this shrank to 17% in the final thesaurus. This residual ambiguity is understandable. Documents in the collection used to derive the thesaurus are not restricted to any particular domain, so terms may

[4] Koru and its evaluation are described in more detail in a companion paper submitted to the same conference

well have several valid senses. As an example, the news stories talk of *Apple Corporation's* business dealings and the theft of Piet Mondrian's painting of an *apple* tree.

The full vocabulary of the thesaurus is almost three times larger than the number of topics, because many topics were referred to by multiple terms. 10% of the concepts are polysemous (have multiple meanings) within the document collection itself: e.g. one document talks of *President Bush* and also mentions *George W. Bush.* A further 33% were made so with the addition of Wikipedia redirects: e.g. Wikipedia adds the colloquialisms *Dubya, Shubya* and *Baby Bush* even though these are never mentioned in our (relatively formal) documents. In this context polysemy is highly desirable, for it increases the chance of query terms being matched to topics and increases the extent to which these are automatically expanded.

The thesaurus was a richly connected structure, with each topic relating to an average of 18 others. As a comparison, Agrovoc,[5] a manually-produced and professionally-maintained thesaurus of comparable size, contains just over two relations per topic on average.

4.2.3 Results

Our central question is whether the knowledge base provided by the thesaurus 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 3 records a significant gain in the recall, precision, and F-measure, averaged over all documents encountered using the topic browsing system. This means that the WikiSaurus based system returned better documents than the traditional one.

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

Table 3: Performance of tasks

The greatest gains are made in recall: the proportion of available relevant documents that the system returned. This can be directly attributed to the automatic expansion of queries to include synonyms. Normally gains made in recall are offset by a drop in precision: the inclusion of more terms causes more irrelevant documents to be returned. This was not the case. Table 3 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, though it is not statistically significant. This can plausibly be attributed to recognition of multi-word terms, which users of traditional interfaces are supposed to encase within quotes. We consistently reminded participants of this syntax when familiarizing themselves with the keyword search interface. Despite this, these expert Googlers did not once use quotation marks, even though they would have been appropriate in 53% of the queries that were issued. The new performs this often overlooked task reliably and automatically.

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

One major difference was the number of queries issued: 338 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.

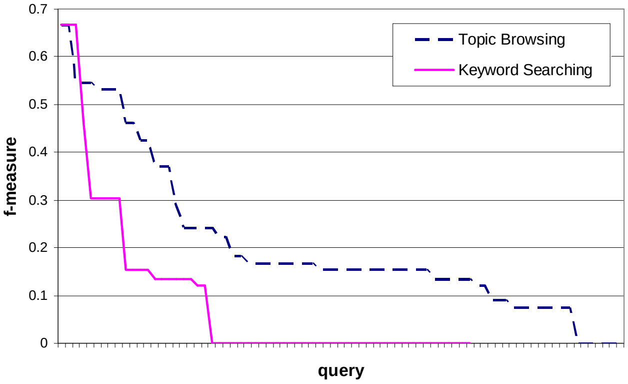

Figure 2: Performance of individual queries

Figure 2 indicates that the additional queries are being issued out of convenience rather than necessity. It plots the F-measure of individual queries issued using the two systems against query rank. The starting point is the same; initial queries are equally good on both systems. A difference soon emerges, however. The performance of keyword searches degrades much more sharply than topic-based ones. This clearly shows that the additional queries issued using Koru are not compensating for any deficiency in performance.

Next we investigate whether it is easier for users to arrive at effective queries when assisted by the knowledge-based approach. In assessing queries we take account of the number of users who made them. 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.

[5] http://www.fao.org/Agrovoc

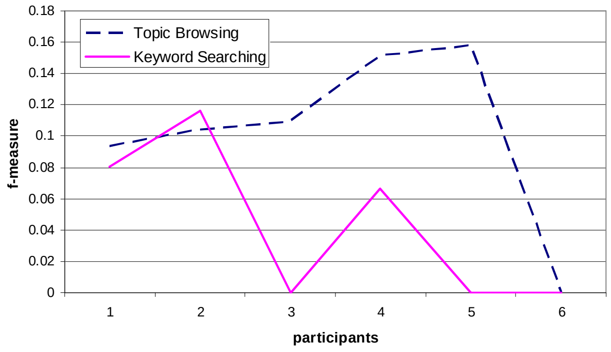

Figure 3: Average performance of queries grouped by number of participants who issued them

Figure 3 plots the average F-measure of queries against the number of participants that issued them. At the left are queries issued by only one participant; at the right are ones issued by five and six participants. For the sake of clarity, we have discarded one of the tasks for which the appropriate query terms were particularly easy to obtain. 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.

5. DISCUSSION AND CONCLUSION

That road is paved with librarians,
bushwhackers, scouts with string
through the labyrinths of information.[6]

This paper is, in some ways, an admission of defeat. We admit that—for now—even our slickest systems cannot match the humble reader’s ability to understand, distill, and share the information they encounter. Until massive gains are made in artificial intelligence and natural language processing, the road to information is best paved with human labor. This couldn’t be more against the grain of computer science research. Here people are doing the job of computers, when we have always striven for the reverse. But this is what the times offer us: Web 2.0 gives us a massive labor force, and this research is about how we can best put it to use.

This is not without its challenges. We must accept and deal with the flaws that are inherent in working with volunteers; that there is a limit to how much work can be expected from them, and that they are biased towards the information that interests them. Consequently our goal is not to burden Web 2.0 users with the job of Google and the like; they simply aren’t suited to the task. Instead we aim to take the effort they have already expended and use it to enhance the search engines; mixing the intelligence of people with the scalability and impartiality of machines.

To this end we developed techniques for applying the volunteer driven Wikipedia to the task of describing documents. By intersecting this vast domain-independent pool of manually defined terms, concepts and relations with individual document collections, we construct machine readable knowledge-bases that are suited to those who seek knowledge from the documents. These structures, which we call WikiSauri, are automatically extracted (cheap) and yet largely manually defined (accurate). In addition to this, WikiSauri have all the advantages offered by the resource from which they are obtained: constant maintenance, coverage of swiftly changing domains, and reflection of contemporary language and interests.

We have demonstrated the effectiveness of WikiSauri for improving information retrieval through query expansion. Our intuition that Wikipedia could provide a thesaurus that matched both documents and queries has so far been borne out: We have tested it with a varied domain-independent collection of documents and retrieval tasks, and it was able render assistance to almost all of the queries issued to it. This assistance made it easier for users to issue effective queries and resulted in significant improvements to the documents they were presented with.

Query expansion was only a case study. We have many other plans for using WikiSauri to facilitate information retrieval. As detailed maps of the topics contained within documents, we see them as highly promising for document tagging, summarization and clustering. As a map of how these topics relate to each other we also expect them to be highly useful for facilitating exploratory search. As always, the end goal is to emulate that helpful old librarian in his ability to guide users to the information they need. Of course, if we ever achieve this we will have to admit that we cheated: that it was done not through complex algorithms but by exploiting the efforts of that old man and his peers. But it’s the results that count.

6. REFERENCES

[1] Deerwester, S., Dumais, S. Furnas, G. Landauer, T. and Harshman, R. Indexing by latent semaintic analysis. *JASIS 41*(6), 1990.

[2] Finkelstein, L., Gabrilovich, Y.M., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. Placing search in context: The concept revisited. *ACM TOIS 20*(1), 2002.

[3] Gabrilovich, E. and S. Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, *Proc. of IJCAI’07*

[4] Gabrilovich, E. and S. Markovitch. Overcoming the Brittleness Bottleneck using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, *Proc. of AAAI’06*

[5] Keen, A. Web 2.0 Is Reminiscent Of Marx. *The Weekly Standard*, 02/15/2006

[6] Leuf, B. and W. Cunningham (2001) *The Wiki Way*. Addison Wesley Longman

[7] Milne, D. Computing Semantic Relatedness using Wikipedia Link Structure. *Proc. of NZCSRSC’07*

[8] Milne, D., Medelyan, O. and Witten, I. H. Mining Domain-Specific Thesauri from Wikipedia: A case study. *Proc. of WI’06*

[9] O’Reilly, T. (2005). What is Web 2.0? Design patterns and business models for the next generation of software.

[6] Excerpt from "Why I Am in Love with Librarians" by Julia Alvarez.

Retrieved May 18, 2007, from *http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html*

[10] Ponzetto, S.P. Creating a Knowledge Base From a Collaboratively Generated Encyclopedia. *Proc. of HLT-NAACL '07*

[11] Ruiz-Casado, M. and Alfonseca, E. and Castells, P. Automatic Assignment of Wikipedia Encyclopedic Entries to Wordnet Synsets. *Proc. of AWIC’05*