

# Computing Semantic Relatedness using Wikipedia Link Structure

David Milne

Department of Computer Science,  
The University of Waikato, Hamilton, New Zealand  
dnk2@cs.waikato.ac.nz

**Abstract.** This paper describes a new technique for obtaining measures of semantic relatedness. Like other recent approaches, it uses Wikipedia to provide a vast amount of structured world knowledge about the terms of interest. Our system, the Wikipedia Link Vector Model or WLVM, is unique in that it does so using only the hyperlink structure of Wikipedia rather than its full textual content. To evaluate the algorithm we use a large, widely used test set of manually defined measures of semantic relatedness as our bench-mark. This allows direct comparison of our system with other similar techniques.

**Keywords:** Wikipedia, Data Mining, Semantic Relatedness

## 1 Introduction

How related are “love” and “sex”? This is a delicate question. Any answer is bound to be subjective—and revealing. But what if we were to consult a dispassionate, objective computer? According to the techniques described in this paper, the answer would be a clinical 67%.

Making such judgments about the semantic relatedness of different terms is a routine yet deceptively complex task. To perform it, we draw not only on our attitudes and personal background, but also on an immense amount of background knowledge about the concepts that these terms represent. Any attempt to compute semantic relatedness automatically must do the same. Many techniques use statistical analysis of large corpora to provide this context. Others use hand-crafted lexical structures such as taxonomies and thesauri. In either case it is the background knowledge that is the limiting factor; for the former approach it is unstructured and imprecise, and for the later it is 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. With over a million articles and thousands of contributors, this massive online repository of knowledge is easily the largest, fastest growing encyclopedia in existence. With its extensive network of cross-references, portals and categories it also contains a huge amount of explicitly defined semantics. This rare combination of scale and structure makes Wikipedia an attractive resource for this work and for other NLP applications.

This paper describes a new technique, the Wikipedia Link Vector Model (or WLVM), which calculates semantic relatedness between terms using the links found within their corresponding Wikipedia articles. Unlike similar techniques, it is able to provide relatively accurate measures using only the link structure and titles of articles, rather than their textual content. Before delving into the details of this approach, we first describe its context in terms of the larger work in which it exists, and the other systems to which it can be compared. This is followed by a description of the algorithm, and its evaluation using a well known data set of manual judgments of semantic relatedness. The paper concludes with a discussion of the strengths and weaknesses of the approach, and directions for possible improvement.

## 2 Context of the research

The research described in this paper is motivated by the larger goal of improving the way in which we locate information. Whenever we seek out new knowledge—whenever we turn to the ubiquitous search engines—there is a fundamental paradox that must be grappled with: how can one describe the unknown? This is because a query is not simply a statement of intent as is commonly thought. Instead it is an excerpt, a few words or phrases, from within a relevant document. To form an effective query, one must predict not only what information this relevant document contains, but also the terms by which this is expressed. In short, one must already know a great deal of what is being sought, in order to find it.

What knowledge seekers need—at least those who are not clairvoyant—is a bridge between what they can describe and the information they seek; between their query terms and the topics and terminology of the documents available. One possible bridge is a thesaurus; a map of semantic relations between words and phrases. Knowledge seekers who cannot identify the effective terms for their query could use a thesaurus that covers the terminology of both documents and potential queries, and describes relations to bridge between them. Seekers who cannot form a specific query at all could use a sensibly organized thesaurus that exposes the topics available and allows them to be explored.

Current use of thesauri for retrieval is limited. They are extremely expensive to produce, and thus are only available for a small portion of document sets. The research which forms the context of this paper aims to address this by developing a framework by which thesauri can be produced cheaply for any document collection. Ideally, these thesauri should be as accurate, relevant, and browsable as their expensive, manually defined counterparts. This is an ambitious goal. Existing techniques for building thesauri automatically are attractive for their low cost and their ability to match the content of documents exactly, but are woefully inferior in terms of accuracy and conciseness. There are many other techniques available, such as query log mining and web link analysis, but their use is typically limited to behind-the-scenes processing due to their inaccuracy.

The most recent—and perhaps the most promising—development for this research is the emergence of internet communities such as Wikipedia and del.icio.us which can be exploited directly for the task of organizing information. These offer terms and

relations defined by human intelligence (as opposed to statistical or lexical approximations), constant maintenance, coverage of swiftly changing domains, and reflection of contemporary language and interests. All this is achieved through the exploitation of existing public efforts, without the cost associated with traditional thesauri.

Identifying the semantic relatedness of terms and concepts within Wikipedia is an important step in extracting sensible, brows-able thesauri from it. Previous work on this problem has shown that most of the relations described by Wikipedia’s structure—the hierarchical relations between categories and the interlinking of articles—cannot be directly mapped to traditional thesauri [1]. Many direct links are irrelevant and need to be discarded, while other additional relations need to be inferred from chains of links. An accurate measure of semantic relatedness would be a valuable guide in either case, and would be extremely useful for disambiguating raw text or entries in other thesauri to concepts in Wikipedia.

### 3 Related Work

Measures of semantic relatedness are used extensively in natural language processing, in such applications as word sense disambiguation, text summarization, and information extraction. The most direct method for evaluating these measures is to compare them with judgments made manually. The largest, most widely used test set for this purpose is the WordSimilarity-353 collection [2]. This contains 353 word pairs for which at least 13 manual judgments of similarity (on a scale of 0-10) are specified. Despite the subjective nature of such judgments, agreement between them was relatively high; the average correlation between an individual participant’s judgments and those of the whole group was 0.79 according to spearman rank-order correlation [3].

**Table 1:** Performance of semantic relatedness measures

Measure	Correlation with manual judgments	Reference
WordNet	0.33–0.35	[6]
Roget’s Thesaurus	0.55	[5]
LSA	0.56	[4]
WikiRelate!	0.19 – 0.48	[7]
ESA	0.75	[8]

As described previously, the central point of difference between the various techniques for obtaining semantic relatedness measures is their source of background knowledge. Corpus based approaches obtain this from performing statistical analysis of large untagged document collections. The most successful of these is Latent Semantic Analysis (LSA) whose measures, as shown in Table 1, have a 0.56 correlation with manual judgments [4]. Other techniques make use of structured

thesauri such as Roget [5] and Wordnet [6], but suffer reduced accuracy and can only provide judgments for a limited number of terms.

Strube and Ponzetto [7] were the first to compute measures of semantic relatedness using Wikipedia. Their approach, known as WikiRelate!, took familiar techniques that had previously been applied to WordNet and modified them to take advantage of the data found within Wikipedia. For example, their path-based measure was adapted to make use of Wikipedia's structure of categories rather than WordNet's relations between synsets, and their text overlap-based measures were based on the text found in Wikipedia's articles, rather than WordNet's glosses. These combined measures provide a level of accuracy that is comparable to those derived from WordNet.

Gabrilovich and Markovitch [8] achieve extremely accurate results with 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 a particular term or portion of text. The weights of these articles—the strength of their association with the input text—are calculated using a centroid based document classifier. The result is a measure that approaches the accuracy of manual judgments. As well as offering much improved accuracy over WikiRelate!, ESA offers the ability to provide relatedness measures for any length of text; there is no restriction that the input be matched to an article title.

#### 4 The Wikipedia Link Vector Model

The Wikipedia Link Vector Model (WLVM) extracts semantic relatedness measures for term pairs from the hyperlink structure of Wikipedia. To do so it must first identify the articles that might discuss the terms of interest. The most direct method for doing so is to obtain the article whose title matches the term directly, but this is complicated by the tendency for terms to have multiple meanings. Figure 1 illustrates the example of *plane*, an ambiguous term that might refer to a theoretical surface of infinite area and zero depth, or a tool for flattening wooden surfaces. In such cases the

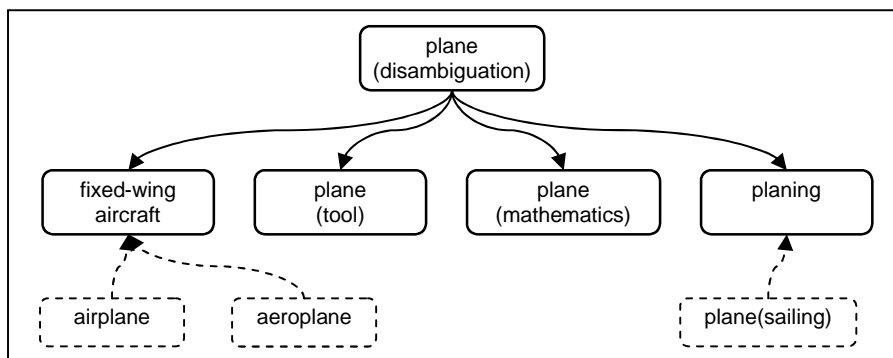


Figure 1: Candidate Wikipedia articles for the term *plane*

conventional solution for Wikipedia authors is to encase additional disambiguation information within parenthesis: the articles become *plane(mathematics)* and *plane(tool)* respectively. These brief scope notes can simply be stripped out when looking for relevant articles so that multiple items are returned for ambiguous terms. Wikipedia's contributors do not always follow this convention, however; *plane* can also refer to *planing*; the act of skimming over the surface of water, or to a *fixed-wing aircraft*. Sometimes this can be resolved by inspecting redirects; there exists a pseudo-article named *plane(sailing)* which redirects to the article *planing*. Other times one must consult disambiguation pages which list the various articles that a term might refer to; *fixed-wing aircraft's* only redirects are *airplane* and *aeroplane*, but it is listed under the disambiguation page for *plane* as the most common sense of the term. Thus the full procedure used to obtain all Wikipedia articles relating to a term is to first list all pages whose titles (sans scope notes) match the term, and then process this so that:

- Articles are used directly.
- Redirects are followed so that their corresponding articles are used.
- Disambiguation pages are processed so that every article that they link to is used.

The next step in obtaining a similarity measure between two terms is to judge the similarity between their representative articles identified in the previous step. In this approach, the semantic similarity of two Wikipedia articles is defined by the angle between the vectors of the links found within them. This is similar to the vector space model used extensively within information retrieval to judge the similarity between documents and queries [9]. Rather than constructing vectors of term counts weighted by their probability of the term occurring (traditionally given by *tf-idf* measures), we build them using link counts weighted by the probability of each link occurring. This probability is defined by the total number of links to the target article over the total number of articles. Thus if  $t$  is the total number of articles within Wikipedia, then the weighted value  $w$  for the link  $a \rightarrow b$  is:

$$w(a \rightarrow b) = |a \rightarrow b| \times \log \left( \sum_{x=1}^t \frac{t}{|x \rightarrow b|} \right) \quad (1)$$

In other words, the weight of a link within a source document is the number of times the source document contains that link (generally 0 or 1), multiplied by the inverse probability of any link to the target document. 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 *science* is much less significant than if they both link to a specific topic such as *atmospheric thermodynamics*. With these weights defined for all  $n$  links  $\{l_i \mid i=1..n\}$  found within a pair of articles  $x$  and  $y$ , the vector for each article is given by:

$$\begin{aligned} x &= (w(x \rightarrow l_1), w(x \rightarrow l_2), \dots, w(x \rightarrow l_n)) \\ y &= (w(y \rightarrow l_1), w(y \rightarrow l_2), \dots, w(y \rightarrow l_n)) \end{aligned} \quad (2)$$

Our similarity measure for the articles is then given by the angle between their vectors. This ranges from  $0^\circ$  if the articles contain identical lists of links, to  $90^\circ$  if there is no overlap between them. With these angles calculated for all possible mappings between the relevant articles for our two terms, the actual similarity between the terms is judged to be the lowest angle found between any pair of relevant articles. This is where articles are finally disambiguated, so that only the two articles that are most closely related to each other are used to form the final measure of similarity. Thus, the actual articles used for final calculation will differ if one is judging between *plane* and *jet*, or *plane* and *bezier curve*.

## 5 Evaluation

Our natural ability as humans to disambiguate topics and judge their relatedness can be considered the gold standard against which our technique should be compared. The WordSimilarity-353 dataset described in Section 3 was used for this purpose, and allows direct comparison with similar techniques. To the best of our knowledge, this is also the largest publicly available test set of its kind.

As an open source project, the entire content of Wikipedia is easily obtainable for studies such as this. It is made available in the form of database dumps that are released sporadically, from several days to several weeks apart. The version used in this evaluation was released on June 3, 2006. At this point Wikipedia contained approximately 2GB of compressed plain text, which explodes to 40 GB of compressed data if its full revision history is considered. Our technique only requires the link structure and basic statistics for articles, which can be obtained separately as about 500 MB (compressed). From this we identified over a million articles, which constitute the various concepts for which semantic relatedness judgments are available. These are highly inter-linked; each article links to an average of 26 others. A further million redirects doubles the available terminology, but this shrinks when scope notes are discarded from titles, due to the ambiguity problem described in the previous section. Thus the final vocabulary for which semantic relatedness judgments were available is a little under two million distinct terms.

Access to this data is facilitated by the Wikipedia Miner Toolkit,<sup>1</sup> which was developed by the author to allow rapid exploration of Wikipedia's link structure. It wraps a database representation of the encyclopedia with convenience classes for querying and browsing its structure. When mined according to the process described in Section 4, the 437 distinct terms contained within our test data set related to almost 5,000 Wikipedia concepts. This once again highlights the high degree of ambiguity involved; each term relates to an average of 11 articles, with a maximum of 118 articles for the highly ambiguous term *Jackson*.

The next step for this technique is to judge the similarity between the Wikipedia articles identified for each term. In order to evaluate the accuracy of this separately, we manually identified the correct articles for each term pair. The automatically generated similarity measures between these manually selected articles correlate

---

<sup>1</sup> In the final version of the paper (sorry!), this footnote will be a reference to a SourceForge page for the Wikipedia Miner toolkit. It has been submitted but is currently under review.

highly with the manual judgments of the terms they represent (according to spearman rank-order correlation coefficient). The correlation of 0.72 shown in Table 2 can be compared directly to the results described in Table 1; it is only slightly lower than the most accurate method ESA, and is a significant improvement over the remainder. Thus the angle between vectors of normalized link counts is a good measure of the semantic relatedness of Wikipedia articles.

The final step is to identify the semantic relatedness of each term pair by selecting the two articles with the highest semantic relatedness (or lowest vector angle). Unfortunately, accuracy degrades significantly when the algorithm is asked to disambiguate articles automatically in this way. As shown in Table 2, correlation with manual judgments drops to 0.45. This is understandable; disambiguation is inherently difficult when terms in each pair can only be disambiguated against each other. In Wikipedia there is the vast number of correct senses for many of the terms, which complicates matters. For example, *Arafat* and *Jackson* are almost completely divergent according to manual judgments, but are made moderately related when *Jackson* is disambiguated to *Jesse Jackson*, an American politician criticized for his anti-Semitic remarks.

**Table 2:** Performance of the *WLVM* measure for semantic relatedness

article selection (disambiguation)	correlation with manual judgements
Manual	0.72
automatic	0.45

There is room to improvement, however. Under *WLVM*, dissimilar links degrade weightings as much as shared links increase them. Large articles are typically multifaceted and have a greater chance of containing dissimilar links, and are consequently unfairly penalized. If anything, this bias should be reversed; when disambiguating manually, one tends towards the more common, easily recognized sense of a term—which typically corresponds to the larger article.

To evaluate the effect of this bias, we tested another measure: the sum of shared link weights. This is identical to *WLVM* except that the weights of shared links are added together and those of dissimilar links are discarded. Table 3 shows the effect of this simple modification: the accuracy of the measure between individual articles degrades significantly, and yet disambiguation improves. Thus the sum of shared link weights is a better measure of the semantic relatedness of terms, but *WLVM* is a much more accurate measure of the semantic relatedness between articles.

**Table 3:** Performance of the *sum of shared link weights* measure

article selection (disambiguation)	correlation with manual judgements
Manual	0.59
Automatic	0.52

## 6 Discussion and Conclusions

In this paper we proposed and evaluated a novel approach to computing semantic relatedness of terms with the aid of Wikipedia. Our approach is most similar to Explicit Semantic Analysis (ESA) and WikiRelate!, which also exploit the content of Wikipedia for this purpose. The central point of difference is that our technique uses only the skeleton structure of Wikipedia rather than its entire content. To obtain measures from either of the other techniques, one must obtain and pre-process a vast amount of textual data. By comparison this technique merely requires one to download a database of pages and links that is far smaller and already indexed. Unfortunately this comes at the cost of accuracy; our approach falls well behind ESA and only outperforms some of the measures provided by WikiRelate!.

There is room for improvement, however. We have identified a distinct bias in WLVM towards smaller, more obscure articles, which greatly degrades its ability to disambiguate terms. One simple modification resolves this to improve accuracy beyond all measures provided by WikiRelate!. Unfortunately it degrades the measure significantly when one considers the relatedness of articles rather than terms. Future work will be centered on resolving WLVM's disambiguation bias while maintaining or improving it as measure of relatedness between Wikipedia articles.

There are many possibilities to explore regarding this. Particularly promising is the vast number of other links found in Wikipedia that our measure does not yet consider. For example, we observed that the categories that articles descend from and the articles that link to them correlate quite highly with semantic relatedness even when un-weighted. With such possibilities left to be explored, it seems likely that this comparatively accurate measure of semantic relatedness can be further improved until it reaches the same level as ESA, while bypassing the need to process Wikipedia's extensive textual content.

## References

1. Milne, D., Medelyan, O. and Witten, I. H. Mining Domain-Specific Thesauri from Wikipedia: A case study. Proc. of WI 2006
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. Spearman, C. The Proof and Measurement of Association between Two Things. The American Journal of Psychology 100(3/4), 1987
4. Deerwester, S., Dumais, S. Furnas, G. Landauer, T. and Harshman, R. Indexing by latent semantic analysis. JASIS 41(6), 1990.
5. Jarmasz, M. (2003) Roget's thesaurus as a lexical resource for natural language processing. Unpublished Master's thesis, University of Ottawa, 2003.
6. Budanitsky, A. and Hirst, G. Evaluating wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1), 2006
7. Strube, M. and Ponzetto, S.P. WikiRelate! Computing Semantic Relatedness Using Wikipedia. Proc.of AAAI 2006
8. Gabrilovich, E. and Markovitch, S. Computing semantic relatedness of words and texts in Wikipedia-derived semantic space. Accepted for IJCAI 2007
9. Salton, G. and Wong, A. and Yang, C.S. A vector space model for automatic indexing. Communications of the ACM 18(11), 1975