2012 International Conference on Image, Vision and Computing (ICIVC 2012)
IPCSIT vol. 50 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V50.38

Measuring of Semantic Relatedness between Words based on Wikipedia Links

Rui-Qin WANG[+]

College of Physics & Electronic Information Engineering of Wenzhou University,
Wenzhou, 325035, China

Abstract. A novel technique of semantic relatedness measurement between words based on link structure of Wikipedia was provided. Only Wikipedia's link information was used in this method, which avoid researchers from burdensome text processing. During the process of relatedness computation, the positive effects of two-directional Wikipedia's links and four link types are taken into account. Using a widely used test set of manual defined measures of semantic relatedness as bench-mark, our method resulted in substantial improvement in correlation of computed relatedness score with human judgments, with a smaller amount of calculation comparing with other present popular methods in computing semantic relatedness between words.

Keywords: Semantic Relatedness, Wikipedia, Link Structure

1. Introduction

Semantic relatedness is the correlation degree of two concepts or terms in a classification system through a variety of semantic relations. Determiing the semantic relatedness between two words has a wide range of applications in natural language processing, such as text summarization, word sense disambiguation, semantic information retrieval, information extraction, etc.

Wikipedia [1] is the leading open encyclopedia that has evolved into a comprehensive resource with very good coverage on diverse topics, important entities, events, etc. Comparing with other repositories, Wikipedia has wider range, more comprehensive knowledge and faster update speed, which makes it become an ideal resource in semantic management. The English Wikipedia currently contains over 4 million articles (including redirection articles). Furthermore, Wikipedia contains quite a bit of structured information: it has a rich category structure, separate pages for ambiguous terms, and structured data for certain types of articles. Finally, it contains over 90 million links between articles, most of these links signify that some semantic relationship holds between the source and target concepts. Wikipedia links can be used to compute a measure of relatedness that typically outperforms traditional text similarity measures.

Wikipedia has grown into a high quality up-to-date knowledge base and can enable many intelligent systems that rely on semantic information. One of the most general and quite powerful semantic tools is measuring of semantic relatedness between words. Strube and Ponzetto [2] 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 lexical database and modified them to take advantage of the data found within Wikipedia. For example, their path-based measures were adapted to make use of Wikipedia's

[+] Corresponding author.
E-mail address: angelwrq@163.com.

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 [3] achieve extremely accurate results with a technique that is somewhat reminiscent of the vector space model (VSM) widely used in information retrieval. Their approach, known as ESA, 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. David Milne [4] proposed a technique called Wikipedia Link Vector Model or WLVM, which obtained the semantic relatedness between words using only the link structure of Wikipedia rather than its full textual content. The performance of WLVM is between the WikiRelate! and ESA.

This paper describes a new technique, which calculates semantic relatedness between words 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. That is similar to David's method, but we take consideration of different impacts of two directional links and all link types on semantic relatedness calculation, and we take a completely different measurement algorithm.

The rest of the paper is organized as follows. In Section 2 we present our measure of semantic relatedness based on Wikipedia link structure. In Section 3 we evaluate the performance of our method 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 in Section 4.

2. Computing Semantic relatedness with Wikipedia link information

As web pages, each article in Wikipedia contains many links point to other articles related to the current article. Wikipedia dump service extracts these links from time to time, storing them in form of database table files and making it available for download. In our work we have used the links between Wikipedia articles to measure the semantic relatedness of words. The specific process is as follows:

1) Finding the corresponding Wikipedia articles of each word for the word pair to be compared.
2) Calculating the semantic relatedness between the Wikipedia articles identified for each word according to the links they share.
3) Identifying the semantic relatedness of each word pair by selecting the two articles with the highest semantic relatedness.

2.1. Article retrieval and pruning

Our method extracts semantic relatedness for term pairs from the link structure of Wikipedia. To do so, it must first identify the articles that might discuss the terms of interest. For the reason that each article in the Wikipedia is a detailed description of a specific subject or an event, so intuitive idea is to compare the article title with the term. But due to the ambiguous property of word itself and the rules of writing of article's title, simple title match is not enough. The process of finding relevant articles must be implemented step by step, follow these steps:

1) **Direct-match articles**: Articles are used which title is fully matched with the term.
2) **Indirect-match articles**: Articles are used which title is partly matched with the term.
3) **Redirect articles:** Redirect links are followed so that their corresponding articles are used.
4) **Articles disambiguation page pointing to:** Disambiguation pages are processed so that every article that they link to is used.

Assuming that "plane" is the term to be compared in semantic relatedness measurement, table 1 below shows the distribution of the corresponding articles in Wikipedia this term respond to. From the table we can see that the corresponding articles found following the above process may be overlap, the mapping process from word to Wikipedia articles should eliminate the duplication to avoid double counting.

Table 1. Distribution of the corresponding articles of term “plane” in Wikipedia

Match category	Related articles
Direct (fully) match articles	plane
Indirect (partly) match articles	plane_(Dungeons_&_Dragons): *including redirect link* plane_(Magic:_The_Gathering) plane_(mathematics) : *including redirect link* plane_(metaphysics) : *including redirect link* plane_(tool)
Redirect articles	plane_(cosmology) : *from Plane_(metaphysics)* plane_(physics) : *from Plane_(mathematics)* plane_(geometry) : *from Plane_(mathematics)* plane_(D&D) : *from plane_(Dungeons_&_Dragons)* plane_(d&d) : *from plane_(Dungeons_&_Dragons)* plane_(D_&_D) : *from plane_(Dungeons_&_Dragons)* plane_(Dungeons_and_Dragons) : *from plane_(Dungeons_&_Dragons)*
Articles the disambiguation page “plane” points to	plain plane_(Dungeons_&_Dragons) plane_(tool) office_of_Film_and_Literature_Classification_(Australia) plane_(Magic:_The_Gathering) wiki_dictionary rafael_Ximeno_y_Planes plane_(metaphysics) planing_(sailing) planing plane_of_immanence fixed-wing_aircraft planes_Mistaken_for_Stars platanus plane_(mathematics)

2.2. Computing link relatedness between Wikipedia articles

Link information in the Wikipedia articles can be used to measure the link relatedness between two articles. Intuitively, if two articles are attached by a link, then the two articles must have a strong correlation. Similarly, if two articles share a link which has the same source or target, then the two articles must have a certain correlation.

In our work, we analyze a measure based on Dice’s measure that is commonly used in IR to measure the semantic relatedness between two Wikipedia articles. Dice’s measure has a very intuitive meaning: in case of Wikipedia articles two pages will be related, if the fraction of the links they have in common to the total number of links of both pages is high. More formally,

$$Dice(A,B) = \frac{2 \times |n(A) \cap n(B)|}{|n(A)| + |n(B)|}. \qquad (1)$$

where n(A) is the set of articles linking (considering both incoming and outgoing links) to article A, and n(B) is the set of articles linking to B.

While exploring the structure of Wikipedia, we have noticed that some types of links are extremely relevant to semantic relatedness, while some other types lead to wrong results. Hence we have added a weighting scheme to the basic measure, based on the following link types:

- **Redirect links:** Most Wikipedia articles have a See Also section that lists related articles. These links explicitly signify that a linked page is semantically related, which we called redirect page. Therefore, redirect links are very important for semantic relatedness and we assign the highest weight to the links of this type. Inverse redirect links (incoming links) are also quite important and they receive a relative high weight also.
- **Double links:** Articles that link to each other directly by regular links in most cases turn out to be quite related, hence these types of links come next in our weighting scheme.
- **Common category links**: Wikipedia has a rich category structure, and articles belonging to the same category are related. Therefore, we identify articles that have both a link between them and share the same category as the next most relevant type of link.
- **Regular links:** The rest of the links are regular links, which carry the least semantic information and

receive the lowest weights in our scheme. In our experiments we have used the type weighting scheme shown in Table 2, which is similar to the allocation scheme in reference [5] but not the same.

Table 2. Weights for various link types

Link type	Type weight
Redirect links	5
Inverse redirect links	3
Double links	2
Common category links	1.5
Regular links	1
Inverse regular links	0.5

It should be noted that due to the different popularity of each link, even the same type of links, the importance of degree in semantic relatedness calculation process will be different, so they should be treated differently. Reference to the TF-IDF weighting method used in traditional vector space model, we propose a novel method to calculate the popularity weight of two directional links as follows:

- Weight of the incoming links

$$w(a \to P) = |a \to P| \times \log\left(\frac{T}{\sum_{x=1}^{T} |a \to x|}\right) \quad (2)$$

where $|a \to P|$ is the links number from article a to current article P (usually 0 or 1), $\sum|a \to x|$ is the total number of links from article a, T is the total number of articles within Wikipedia. Thus links are considered less significant for judging the relatedness between articles if the same source points to many other articles.

- Weight of the outgoing links

$$w(P \to a) = |P \to a| \times \log\left(\frac{T}{\sum_{x=1}^{T} |x \to a|}\right) \quad (3)$$

where $|P \to a|$ is the links number from current article P to article a (usually 0 or 1), $\sum|x \to a|$ is the total number of links point to article a. Thus links are considered less significant for judging the relatedness between articles if many other articles also link to the same target.

Considering the type weight and the popularity weight of article links comprehensively, we construct the following formula to measure the link relatedness between Wikipedia articles:

$$r(A,B) = \frac{2 \times \sum_{i \in link(A) \cap link(B)} (CW_i \times PW_i)}{\sum_{i \in link(A)} (CW_i \times PW_i) + \sum_{j \in link(B)} (CW_j \times PW_j)} \quad (4)$$

where CW_i and PW_i are the type weight and the popularity weight of link i respectively, $link(A) \cap link(B)$ is the shared links of Wikipedia article A and B.

2.3. Semantic relatedness measurement between words

When the relatedness of the representative Wikipedia articles have been made, the semantic relatedness between the word pair is defined as the maximum relatedness value of their corresponding articles:

$$R(w_1, w_2) = \max_{\substack{\forall A \in D(w_1) \\ \forall B \in D(w_2)}} \{r(A,B)\} \quad (5)$$

where $D(w_1)$and $D(w_2)$ are the corresponding Wikipedia article set associated with word w_1 *and* w_2 respectively.

3. Experiments and Results

Experimental part of this paper tests and analyses the performance of the semantic relatedness between word pair based on Wikipedia link structure information. Meanwhile, in the same experimental environment, a number of existing popular semantic relatedness measures has been compared to highlight the advantages of our method.

3.1. Experimental data

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 Feb. 6, 2007. Our technique only requires the link structure and basic statistics for articles, which can be obtained separately as about 1.9 GB (compressed), while the whole Wikipedia contains approximately 10GB of compressed data. To further enhance the computing performance, we did a series of optimized operations to the original data, including indexing, removing part of the unrelated fields with the experiment, field conversion, etc. The final database size is about 1.1 GB, with such a relative smaller amount of data, the computation overload is perfectly acceptable.

The most direct method for evaluating the semantic relatedness measures is to compare them with judgments made manually. The largest, most widely used test set for this purpose is the WordSimilarity-353 collection [6]. This contains 353 word pairs for which at least 13 manual judgments of similarity (on a scale of 0-10) are specified. The average correlation between an individual participant's judgments and those of the whole group was 0.79 according to spearman rank-order correlation [7]. We use this collection to evaluate our method and compare with other mainstream techniques.

3.2. Experimental results and analysis

There are many indicators for measuring the correlation degree, the most commonly used are Spearman, Pearson and Kendall's tau-b correlation coefficient. Both the characteristics and computation processes of these three categories are different. Considering the usage and the evaluation capabilities of them, this study decides to use Spearman rank correlation coefficient as a measure of correlation between the manual judgment and the judgments of various semantic relatedness measures, so as to compare the performance of those methods. The formula of Spearman rank correlation coefficient is as follows:

$$r = 1 - \frac{6 \times \sum d^2}{n \times (n^2 - 1)} \qquad (6)$$

where d is the difference of two rank (grade), n is the sample size.

The semantic relatedness measurement method proposed in this study and the other three popular measures based on Wikipedia are compared in this experiment, the result Spearman correlation coefficients are shown in table 3. We can see from the table that ESA achieve extremely accurate results, which approaches the accuracy of manual judgments. But ESA approach must deal with all documents in the Wikipedia, so the computational complexity of this method can not be underestimated. WLVM got the performance between *WikiRelate!* and ESA, its main advantage lies in the simple in principle. WLVM does not require the process of heavy text processing and it is a practical application. Our method considers the link type and link popularity comprehensively during the process of calculating the semantic relatedness between word pair. At the cost of a small calculation, our method obtains a better performance than WLVM.

Table 3. Performance of semantic relatedness measures

Measures	Correlation with manual judgments	Reference
WikiRelate!	0.19-0.48	[2]
ESA	0.750	[3]
WLVM	0.52-0.59	[4]
OUR	0.63	-

In order to evaluate the impact of different link types on the performance of relatedness measurement, we made another experiment using different types of links each time, and deriving the test results as table 4. As

can be seen from the table, using only one or several types of links achieve relative poor results, this demonstrates that two-directional links and four types of links all have positive impacts on the performance of relatedness measurement. Additionally, we can see from the table that the incoming links have better performance on the relatedness measurement than the outgoing links, while the traditional measures like WLVM always ignored this type of links.

Table 4: Impact of different link types on the performance

Link type	Correlation with manual judgments
Incoming links + popular weight	0.46
Outgoing links + popular weight	0.41
Bi-directional links + popular weight	0.49
Four types of link + type weight	0.54
OUR	0.63

4. 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 WLVM, which also exploits the link structure of Wikipedia rather than its entire content for this purpose. The central point of difference is that our technique considers not only the link type but also the link population in the process of calculating the semantic relatedness. Another significant difference with WLVM method is that our method takes into account both the incoming links and the outgoing links to measure Wikipedia article's link relatedness, while WLVM only considers the traditional outgoing links, but ignoring the incoming links which proved to have more information.

ESA is a relative comprehensive method of calculating semantic relatedness, which uses the Wikipedia text and category information to get an accurate measurement results. However, in ESA method, in order to form a semantic vector space all Wikipedia documents need to be processed. At present Wikipedia contains more than 1200 million documents, text processing is a very time-consuming work, so the computational complexity of ESA is too high. The proposed method in this paper uses only the link structure of Wikipedia, the text content of Wikipedia is not involved, which avoid the heavy text processing work.

In the process of semantic relatedness calculation, our method takes into account two-directional links (incoming links, outgoing links) and four types of links (redirect links, double links, common category links, general link). In the experimental part of this paper, several mainstream semantic relatedness measures were tested and compared with our method. Experimental results show that our method achieved excellent results at the cost of a small calculation overload.

There is room for improvement, however. We have identified only four types of links in our measure, the vast number of other links found in Wikipedia that our measure does not yet consider is a particularly promising research direction. In addition, the category structure of Wikipedia articles have not been considered completely in our measure, including this information there may be a significant measure result. With such possibilities left to be explored, it seems likely that this comparatively accurate measure of semantic relatedness can be further improved, while bypassing the need to process Wikipedia's extensive textual content. We firmly believe that, with the increasingly growing rich and perfect of Wikipedia encyclopedia, it is bound to provide strong support to natural language processing.

5. Acknowledgements

This paper is supported by the Wenzhou city Science and Technology Plan Project of China under Grant No.S20090014 and No.H20090052.

6. References

[1] Wikipedia, the free encyclopedia. http://www.wikipedia.org/.

[2] Strube M.and Ponzetto S.P. *WikiRelate!* Computing Semantic Relatedness Using Wikipedia. Proc. of the AAAI, Boston: IEEE, 2006:1419-1424.

[3] Gabrilovich E. and Markovitch S., Computing Semantic Relatedness of Words and Texts in Wikipedia-derived Semantic Space, Proc. of the IJCAI, 2007: 1606-1611.

[4] David M. Computing Semantic Relatedness using Wikipedia Link Structure. Department of Computer Science, University of Waikato, Hamilton, 2007.

[5] Maxim G., Dmitry L., Denis T., etc. Efficient Ranking and Computation of Semantic Relatedness and its Application to Word Sense Disambiguation. Institute for System Programming, Russian Academy of Sciences, 2008.

[6] Finkelstein L., Gabrilovich Y.M., Rivlin E., etc. Placing search in context: The concept revisited. ACM Transactions on Information Systems, 2002, 20:116-131.