2009 International Conference on Web Information Systems and Mining

Measuring Semantic Similarity between Words Using Wikipedia

Lu Zhiqiang
School of Computer Engineering and Science
Shanghai University
Shanghai, China
luzhiqiang@.shu.edu.cn

Shao Werimin
School of Computer Engineering and Science
Shanghai University
Shanghai, China
swm1956@vip.sina.com

Yu Zhenhua
Second Limited Liability Company of Rizhao Port
Shandong, China
nightwish62@163.com

***Abstract*—Semantic similarity measures play an important role in the extraction of semantic relations. Semantic similarity measures are widely used in Natural Language Processing (NLP) and information Retrieval (IR). This paper presents a new Web-based method for measuring the semantic similarity between words. Different from other methods which are based on taxonomy or Search engine in Internet, our method uses snippets from Wikipedia[1] to calculate the semantic similarity between words by using cosine similarity and TF-IDF. Also, the stemmer algorithm and stop words are used in preprocessing the snippets from Wikipedia. We set different threshold to evaluate our results in order to decrease the interference from noise and redundancy. Our method was empirically evaluated using Rubenstein-Goodenough benchmark dataset. It gives higher correlation value (with 0.615) than some existing methods. Evaluation results show that our method improves accuracy and more robust for measuring semantic similarity between words.**

Keywords-Text semantic similarity, wikipedia, TF-IDF, cosine similarity

The characteristics of polysemy and synonymy that exist in words of natural language have always been a challenge in the fields of Natural Language Processing (NLP) and Information Retrieval (IR). In many cases, humans have little difficulty in determining the intended meaning of an ambiguous word, while it is extremely difficult to replicate this process computationally. For example, people can easily find that orange is more semantic similar to apple than book. Unfortunately, that is not easy for computer. For many tasks in psycholinguistics and NLP, a job is often decomposed to the requirement of resolving the semantic relation between words or concepts.

One needs to come up with a consistent computational model to assess this type of relation. When a word level semantic relation requires exploration, there are many potential types of relations that can be considered: hierarchical, associative, equivalence (synonymy), etc. Among these, the hierarchical relation represents the major and most important type, and has been widely studied and applied as it maps well to the human cognitive view of classification (i.e. taxonomy). The IS-A relation, in particular, is a typical representative of the hierarchical relation. It has been suggested and employed to study a special case of semantic relations — semantic similarity or semantic distance [11]. In this study of semantic similarity, we will take this view, although it excludes some potential useful information that could be derived from other relations.

The study of words relationships can be viewed in terms of the information sources used. The least information used is knowledge-free approaches that rely exclusively on the corpus data themselves. Under the corpus-based approach, word relationships are often derived from their co-occurrence distribution in a corpus [4]. Unfortunately, taxonomy-based methods can not include some new entity in WordNet[2] which is mainly used by taxonomy-based methods. So, the information sources should be new emerging words or entities because language is live and active.

The Web is a heterogeneous document collection. Huge-scale and dynamic nature are characteristics of the Web. Regarding the Web as a live corpus becomes an active research topic recently. How to utilize the huge volume of Web data to measure association of information is an important issue. Besides, how to get the word counts and the word association counts from the web pages without scanning over the whole collections is indispensable. Directly managing the web pages is not an easy task when the Web grows very fast.

In this paper, a new way for measuring semantic similarity between words based on Wikipedia is described. The remainder of the paper is organized as follows. The Section 2 introduces some related works. The Section 3 introduces some related conceptions and detailed algorithm to measure semantic similarity using Wikipedia. The results are evaluated by Rubenstein-Goodenough Benchmark Datasets in Section 4.Some conclusion are presented in Section 5.

I. RELATED WORK

There many researchers who have done closely relate works to measure semantic similarity. Rada et al. and Lee et al.[11] derived semantic distance formulas using the edge counting principle, which were then used to support higher level result ranking in document retrieval. Sussna [16] defined a similarity measure that takes into account taxonomy structure information. Resnik's [12] information content

[1] http://www.wikipedia.org/

[2] http://wordnet.princeton.edu/

978-0-7695-3817-4/09 $26.00 © 2009 IEEE
DOI 10.1109/WISM.2009.59

measure is a typical representative of the node-based approach. Richardson and Smeaton [13, 14] worked on a combined approach that is very similar to ours. One of the many applications of semantic similarity models is for word sense disambiguation (WSD). Agirre and Rigau [1] proposed an interesting conceptual density concept for WSD. Given the WordNet as the structured hierarchical network, the conceptual density for a sense of a word is proportional to the number of contextual words that appear on a sub-hierarchy of the WordNet where that particular sense exists. The correct sense can be identified as the one that has the highest density value. Using an online dictionary, Niwa and Nitta [10] built a reference network of words where a word as a node in the network is connected to other words that are its definitional words. The network is used to measure the conceptual distance between words. A word vector is defined as the list of distances from a word to a certain set of selected words. These selected words are not necessarily its definitional words, but rather certain types of representational words called origins. Word similarity can then be computed by means of their distance vectors. They compared this proposed dictionary-based distance vector method with a corpus-based co-occurrence vector method for WSD and found the latter has a higher precision performance. However, in a test of leaning positive or negative meanings from example words, the former gave remarkable higher precision than the latter. Kozima and Furugori [8] also proposed a word similarity measure by spreading activation on a semantic net composed by the online dictionary LDOCE. In the area of IR using NLP, approaches have been pursued to take advantage of the statistical term association results [15]. Typically, the text is first parsed to generated syntactic constructs.

As Internet becoming popular, some methods based on Web to measuring semantic similarity between words are proposed. A method which defines a point-wise mutual information using the number of hits returned by Web search engine to recognize synonyms is proposed by Turney [5]. Chen [6] proposed a novel method named Co-Occurrence Double Check which uses Web as a live corpus.

II. MEASURING SEMANTIC SIMILARITY BETWEEN WORDS USING WIKIPEDIA

A. About Wikipedia

Wikipedia is the world's largest collaboratively edited source of encyclopedic knowledge, which provides important semantic information for us. So we can get external knowledge about words from Wikipedia to analyze semantic similarity between words.

Firstly, we must decide which part in Wikipedia for a word is useful for us. For example, if we search word "*car*" in Wikipedia, we can get much information about "*car*", such as car's history, its production and its safety, and so on. But we can't use all of them for not all snippets are useful for us to analyze semantic similarity.

Usually, Wikipedia return some top result for the word for which we search information in Wikipedia. These snippets use simply vocabulary to explain the word, or give simply definition or some description about the word. We select these for analytic target to measure semantic similarity between words.

B. Preprocessing the snippets from Wikipedia

We can't use the snippets downloaded from Wikipedia directly for there are a lot of semantic-unrelated words and word in different form will bring in Negative impact in our calculating. Therefore, we must deal with the snippets by following steps:

1) Delete stop words. Words like "a","the","of" and so on called stop words are meaningless for semantic analysis. Before we do some calculating on snippets from Wikipedia, we must delete them.

Fortunately, paper [3] gives us a stop words list for general text. We use it to filter the stop words in snippets from Wikipedia.

2) Because we will do some statistical work on the snippets from Wikipedia, words in different form will bring in disadvantage influence. We must use some algorithm to void it.

Stemmer algorithm[3] gives us critical help to deal the text. We use the algorithm to deal with every word in the snippets from Wikipedia.

C. About TF-IDF

The TF–IDF (term frequency–inverse document frequency) is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. Variations of the TF–IDF weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

Consider a document containing 100 words where in the word *cow* appears 3 times. The term frequency (TF) for cow is then 0.03 (3/100).

The TF-IDF weighting scheme is often used in the vector space model together with cosine similarity to determine the similarity between two documents. We will use TF-IDF and cosine similarity to analyze the text which is from Wikipedia after preprocessing in 3.2.

D. Calculate the Semantic Similarity by cosine similarity

In this section a method which integrates TF-IDF and cosine similarity is proposed in details to measure semantic similarity between words.

[3] http://tartarus.org/~martin/PorterStemmer/csharp2.txt

TABLE I. COMPARISON TO SOME OTHER METHODS

Word Pair	R-G Ratings	PMI	Our Result (a=0)	Our Result (a=0.1)	Our Result (a=0.11)
car-automobile	3.92	6.170	0.753	0.753	0.753
gem-jewel	3.84	7.233	0.085	0	0
journey-voyage	3.84	7.39	0.237	0.237	0.237
boy-lad	3.76	6.604	0.128	0.128	0.128
coast-shore	3.7	7.401	0.349	0.349	0.349
asylum-madhouse	3.61	8.025	0.402	0.402	0.402
magician-wizard	3.5	8.954	0.399	0.399	0.399
midday-noon	3.42	8.327	0.746	0.746	0.746
furnace-stove	3.11	8.548	0.363	0.363	0.363
food-fruit	0.89	6.377	0.164	0.164	0.164
bird-cock	3.05	6.484	0.029	0	0
bird-crane	2.97	6.802	0.013	0	0
tool-implement	2.95	6.756	0.120	0.120	0.120
brother-monk	2.82	6.821	0.091	0	0
crane-implement	1.68	6.786	0.046	0	0
lad-brother	1.66	6.653	0.012	0	0
journey-car	1.16	5.194	0.031	0	0
monk-oracle	1.1	5.521	0.096	0	0
cemetery-woodland	0.95	6.032	0.069	0	0
food-rooster	0.89	5.479	0.038	0	0
coast-hill	0.87	6.326	0.085	0	0
forest-graveyard	0.84	6.742	0.080	0	0
shore-woodland	0.63	6.012	0.165	0.165	0.165
monk-slave	0.55	7.318	0.026	0	0
coast-forest	0.42	6.785	0.100	0.100	0
lad-wizard	0.42	5.990	0.088	0	0
chord-smile	0.13	5.334	0.039	0	0
glass-magician	0.11	7.812	0.007	0	0
noon-string	0.08	5.553	0.049	0	0
rooster-voyage	0.08	5.561	0.007	0	0
Correlation	1	0.534	0.600	0.615	0.597

Cosine similarity is a measure of similarity between two vectors of n dimensions by finding the cosine of the angle between them, often used to compare documents in text mining. Given two vectors of attributes, A and B, the cosine similarity, θ, is represented using a dot product and magnitude as:

$$\text{Similarity}(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||}. \quad (1)$$

For text matching, the attribute vectors A and B are usually the TF vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

In our experiments, as for words "a" and "b", we define their vectors A and B as follow:

Support $W = W_a \cup W_b = \{w_1, w_2, \ldots, w_n\}$, where W_a and W_b is a word set of snippet from Wikipedia after preprocessing in Section 3.2 of word "a" and "b" respectively, Where $n=|W|$. Support TF_{ai} is the TF of the word w_i in the snippet

of word a, and TF_{bi} is the TF of the word w_i in the snippet of word b, we define:

$$A = \{TF_{a1}, TF_{a2}, \ldots, TF_{an}\}$$
$$B = \{TF_{b1}, TF_{b2}, \ldots, TF_{bn}\}$$

Then we can use cosine similarity to calculate the semantic similarity between them. The results of similarity range from -1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating independence, and in-between values indicating intermediate similarity or dissimilarity. In fact, because TF > 0, the $\cos(\theta)$ of word a and b is greater than or equal 0.

III. Experiments & Evaluations

A. Rubenstein-Goodenough Benchmark Datasets

Rubenstein-Goodenough [7] proposed a datasets of 65 word-pairs rate by a group of 51 human subjects. Most methods used 28 word-pairs of Rubenstein-Goodenough Benchmark Datasets only. Rubenstein and Goodenough rating are a reliable benchmark for evaluating semantic similarity measure. In static, the person correlation coefficient is a common measure of the correlation between two variables X and Y. The formula is given as follows:

$$\rho = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \mu_X}{\sigma_X}\right)\left(\frac{Y_i - \mu_Y}{\sigma_Y}\right), \qquad (2)$$

where $\frac{X_i - \mu_X}{\sigma_X}$, μ_X and σ_X are the standard score, population mean, and population standard deviation respectively. It is apparent that the higher the correlation coefficient against R-G ratings, the more accurate for the method measuring semantic similarity between words.

B. Impact of the threshold a

If the similarity of A and B which is calculated by method aforementioned in the Section 3.4 less than a, then the similarity of A and B will be zero. For example a=0.05 means that if similarity of A and B less than 0.05, we will consider that A and B is non-related. Table 1 presents that, the correlation coefficient p with Rubenstein and Goodenough rating will be maximum when a=0.1.

C. Comparison to other Web-based methods

Table 1 presents a comparison of proposed method to another measuring method (PMI) [5].PMI based method for calculating the semantic similarity between words, which uses Pointwise Mutual Information to sort list of important neighbor words of the two target words. The correlation coefficient of PMI against R-G rating is 0.534. The correlation coefficient of method which this paper proposes is 0.615, which is more accurate than PMI.

IV. Conclusions

In this paper, we have presented a new approach for measuring semantic similarity between words. It uses the definition or introduce about the words in Wikipedia. This paper makes some contributions to propose a more accurate method for measuring the semantic similarity between words. Snippets in Wikipedia are used to measure semantic similarity between words. The result demonstrates that the snippets in Wikipedia have a significantly influence on the accuracy of semantic similarity measure between words.

The major contributions of this paper consist of:

1). Measuring semantic similarity between words using Wikipedia is firstly proposed in this paper;

2). The stop words and the stemmer algorithm are used in preprocessing snippets from Wikipedia;

3). TF-IDF and cosine similarity are used in details to measure semantic similarity between words.

Acknowledgment

The authors are grateful for their colleague Z. Xu for his helpful discussions and valuable guidance. This paper was supported by Shanghai Leading Academic Discipline Project.

References

[1] Agirre, E. and G. Rigau, 1995, "A proposal for Word Sense Disambiguation Using Conceptual Distance", Proceedings of the First International Conference on Recent Advanced in NLP, Bulgaria.

[2] Akiko, Aizawa. An information-theoretic perspective of tf–idf measures. National Institute of Informatics, 2002.

[3] Christopher Fox, 1989, "A stop list for general text", ACM SIGIR Forum, Volume 24, Issue 1-2 (Fall 89/Winter 90) Pages: 19-21.

[4] Church, K.W. and P. Hanks, 1989, "Word Association Norms, Mutual Information, and Lexicography", Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, ACL27'89, 76-83.

[5] D.Turney. Mining the Web for synonyms: PMI-IR verus LAS on TOEFL. In Proceedings of the 12th European Conference on Machine Learning, pages 591-502.

[6] H. Chen, M. Lin et al. Novel association measures using web search with double checking. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of ACL, pages 1009-1016, 2006, Sydney.

[7] H. Rubenstein and J.B. Goodenough. 1965. Contextual Correlates of Synonymy. Communications of the ACM, 8(10): 627-633.

[8] Kozima, H. and T. Furugori, 1993, "Similarity Between Words Computed by Spreading Activations on an English Dictionary", Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics, EACL-93, 232-239.

[9] Nir Oren. Reexamining tf.idf based information retrieval with genetic programming, ACM International Conference Proceeding Series, pages 1009-1016,pages 224 - 234 ,2002.

[10] Niwa, Y. and Y. Nitta. 1994, "Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries", Proceedings of the 17th International Conference on computational Linguistics, COLING'94, 304-309.

[11] Rada, R., H. Mili, E. Bicknell, and M. Bletner, 1989, "Development and Application of a Metric on Semantic Nets", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 19, No. 1, 17-30.

[12] Resnik, P., 1995, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy", Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 1,448-453, Montreal, August 1995.

[13] Richardson, R. and A.F. Smeaton, 1995, “Using WordNet in a Knowledge-Based Approach to Information Retrieval”, Working Paper, CA-0395, School of Computer Applications, Dublin City University, Ireland.

[14] Smeaton, A.F. and I. Quigley, 1996, “Experiments on Using Semantic Distance Between Words in Image Caption Retrieval”, Working Paper, CA-0196, School of Computer Applications,Dublin City University, Ireland.

[15] Strzalkowski, T. and B. Vauthey, 1992, “Information Retrieval Using Robust Natural Language Processing”, Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics, ACL’92, 104-111.

[16] Sussna, M., 1993, “Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network”, Proceedings of the Second International Conference on Information and Knowledge Management.