IOSR Journal of Engineering (IOSRJEN)
ISSN: 2250-3021 ISBN: 2878-8719 PP 52-54
National Symposium on engineering and Research

Measuring Semantic Similarity between Words Using Web Search Engines:A Survey

Shweta A.Koparde
Matoshri College Of Engineering,Nashik

Abstract: - *Semantic similarity measures play important part in information retrieval and Natural Language Processing. Earlier work in semantic web-related applications such as community mining, relation extraction, automatic metadata extraction has used various different semantic similarity measures. In spite of the usefulness of semantic similarity measures in these applications, measuring semantic similarity in two words (or entities) remains a challenging task. This paper performs a survey on semantic similarity measure and proposes a strong semantic similarity measure that uses the information available on the web to measure similarity between two words or entities. The proposed method exploits text snippets returned by a web search engine. This paper proposes new approach to compute semantic similarity using automatically extracted Lexico-syntactic patterns from text. These different similarity scores are included using support vector machines, to leverage a strong semantic similarity measure.*

Keywords*: - Semantic similarity, Web mining*

I. INTRODUCTION

The study of semantic similarity between words has long been an essential part of information retrieval and natural language processing. Semantic similarity is a central concept that finds great importance in various fields such as artificial intelligence, natural language processing, cognitive science and psychology. Accurate measurement of semantic similarity between words is essential for various tasks such as, document clustering, information retrieval, and synonym extraction. For a machine to be able to decide the semantic similarity, intelligence is needed. It should be able to understand the semantics or meaning of the words. Semantic similarity between entities can change over time and all over the world in different domains. For example, *apple* is commonly linked with *computers* on the Web. Though, apple is not listed in most general-purpose thesauri or dictionaries. A user, who searches for *apple* on the Web, may be interested in this sense of apple as a company and not apple as a fruit. New words are constantly being created as well as new senses are given to existing words. Physically maintaining thesauri to capture these new words and senses is costly if not impossible.

This paper proposes an automatic method to measure semantic similarity between words or entities using web search engines. Because of the various documents and the wide growth rate of the Web, it is difficult to analyze each document separately. Web search engines provide a vital interface to this vast information. Page counts and snippets are two useful knowledge sources provided by most Web search engines. Page count of a query is the no. of pages which contain the query words. 1. Page count for the query X *AND Y* can be considered as a global measure of co-occurrence of words X and Y. For example, the page count for the query *\apple"* AND *\computer"* in *Google* 273*;* 000*;* 000, whereas the same for *\mango"* AND *\computer"* is only 3*; 220;* 000. The more than 80 times more several page counts for *\apple"* AND *\computer"* shows that *apple* is more semantically similar to *computer* than is *mango*. In spite of its simplicity, using page counts alone as a measure of co-occurrence of two words presents numerous drawbacks. One of the drawback is page count analysis ignores the position of a word in a page. That is why even if two words appear in a page, they might not be related. The other drawback is page count of a word with multiple senses may contain a mixture of all its senses. For an example, page counts for *apple* contain page counts for *apple* as a fruit and *apple* as a company. In addition, given the range and noise in the Web,some words might occur randomly. For those reasons, page counts alone are volatile when measuring semantic similarity.

A snippet is a small window of text extracted by a search engine around the query in a document, which gives important information about the local context of the query term. Semantic similarity measures defined over text in snippets have been used in query expansion [1], personal name disambiguation [2] and community mining [3]. Processing snippets is also important as it obviates the trouble of downloading web pages, which may be time consuming depending on the size of the pages. However, a widely known drawback of using snippets is that, because of the enormous scale of the web and the large number of documents in the result set, only those text snippets

with the top-ranking results for a query can be processed efficiently. Ranking of search results or the snippets is determined by a complicated combination of numerous factors unique to the underlying search engine. Therefore, no guarantee that exists which will provide information we need to measure semantic similarity between a two words is enclosed in the top-ranking snippets.

This paper does a survey on semantic similarity measure and a brief into of the system which include both page count as well as snippets [9].

II. RELATED WORK

Semantic similarity measures are important in various Web related tasks. In query expansion [4] a user query is changed using synonymous words to advance the relevancy of the search. One method to find correct words to include in a query is to check and compare the previous user's queries using semantic similarity measures. If there exist a previous query which is semantically related to the current query, then it can be suggested either to the user or internally used by the search engine to change the original query. Semantic similarity measures have been used in Semantic Web related applications such as automatic annotation of Web pages, community mining [5], and keyword extraction for inter-entity relation representation [1].Semantic similarity measures are important for different applications in natural language processing such as word-sense disambiguation [2], language modeling [4], and synonym extraction [5], and automatic thesauri extraction [3]. Manually compiled taxonomies such as WordNet and large text corpora is used in previous works on semantic similarity [5]. About the Web as a live corpus has become a dynamic research topic recently. Unsupervised models apparently perform better when *n*-gram counts are obtained from the Web rather than from a large corpus [6, 7]. Resnik and Smith [8] extracted bilingual sentences from the Web to create parallel corpora for machine translation. Turney [12] defined a point-wise mutual information (PMI-IR) measure using the number of hits returned by a Web search engine to recognize synonyms. Matsuo et. al, [13] used a similar approach to measure the similarity between words and apply methods in a graph based word clustering algorithm.

Given taxonomy of concepts, a simple method to calculate similarity between two words (concepts) is to find the length of the shortest path connecting the two words in the taxonomy [9]. If a word is polysemous then multiple paths might exist between the two words. In such cases, only the shortest path between any two senses of the words is used for calculating similarity. A problem that is considered with this approach is that it depends on the idea that all links in the taxonomy represent a uniform distance. Resnik [8] proposed a similarity measure using information content. He defined the similarity between two concepts $C1$ and $C2$ in the taxonomy as the maximum of the information content of all concepts C that include both $C1$ and $C2$. Then the similarity between two words is defined as the maximum of the similarity between any concepts that the words belong to. He used WordNet as the taxonomy; information content is measured using the Brown corpus.

Li et al., [14] combined structural semantic information from a lexical taxonomy and information content from a corpus in a nonlinear model. They projected a similarity measure that uses shortest path length, depth and local density in taxonomy. Their experiments reported a Pearson co-relation coefficient of 0:8914 on the Miller and Charles [10] benchmark dataset. They did not evaluate their method on the condition of similarities among named entities. Lin [11] defined the similarity between two concepts as the information that is in common to both concepts and the information included in each individual concept.

III. METHOD

We propose a method which includes snippets to measure semantic similarity between a given pair of words. We define four similarity scores using page counts. We then describe an automatic lexico-syntactic pattern extraction algorithm. We grade the patterns extracted by our algorithm according to their ability to state semantic similarity. We use two class support vector machines (SVMs) to find the best combination of page counts-based similarity scores and top ranking patterns. The SVM is trained to categorize synonymous word-pairs and non-synonymous word-pairs. We select synonymous word-pairs from WordNet. Non-synonymous word-pairs are automatically created using a random shuffling technique. We convert the output of SVM into a posterior probability. We describe the semantic similarity between two words as the posterior probability that they belong to the synonymous-words (positive) class.

3.1 Extracting Lexico Syntactic Patterns from Snippets

Text snippets are returned by search engines along with the search results. They give important information regarding the local context of a word. We extract lexico-syntactic patterns that indicate different aspects of semantic similarity. Consider the following text snippet returned by Google for the query *\cricket" AND \sport"*.

"*Cricket is a sport played between two teams, each with eleven players."*

Figure 1: Pattern Extraction Example

Here, the phrase *is a* indicates a semantic relationship between cricket and sport. Many such phrases indicate semantic relationships. For example, *also known as, is a, part of, is an example of* all indicate semantic relations of different types. In the example given above, words indicating the semantic relation between *cricket* and *sport* appear between the query words. Replacing the query words by wildcards *X* and *Y* we can form the pattern *X is a Y* from the example given above. Though, in some cases the words that indicate the semantic relationship do not fall between the query words. For example, consider the following example.

"Honda and Toyota are two major Japanese car manufacturers."

Figure 2: Pattern Extraction Example

Here, the relationship between *Honda* and T*oyota* is that they are both *car manufacturers*. Identifying the exact set of words that express the semantic relationship between two entities is a difficult problem which requires more semantic analysis. However, such an analysis is not possible considering the different ill-formed sentences we need to process on the Web. This paper, propose a one dimensional pattern extraction method to capture the semantic relationship between words in text snippets.

Pattern extraction algorithm is as follows: Algorithm 3.1: ExtractPatterns(*S*)

comment: Given a set *S* of word-pairs,extract patterns.

for each word-pair (*A;B*) € *S* do *D* ← GetSnippets("*A B*")

N ← *null*

for each snippet *d* € *D*

do *N* ← *N* + GetNgrams(*d; A;B*)
Pats ← CountFreq(*N*)

return (*Pats*)

Given a set of S synonymous word pairs, GetSnippet is a function which returns text snippets from the query A and B. For each snippet, we can replace two words in wildcards.

IV. CONCLUSION

In this paper, we proposed a measure that uses snippets to strongly calculate semantic similarity between two given words or named entities. The method consists of four page-count-based similarity scores and automatically extracted lexico-syntactic patterns. We can include page-counts-based similarity scores with lexico syntactic patterns using support vector machines. Training data were automatically created using WordNet synsets. Proposed method outperformed all the baselines including previously proposed Web-based semantic similarity measures on a benchmark dataset.

REFERENCES

[1] M. Sahami and T. Heilman. A web-based kernel functions for measuring the similarity of short text snippets. In Proc. of 15th International World Wide Web Conference, 2006.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Disambiguating personal names on the web using automatically extracted key phrases. In Proc. of the 17th European Conference on Artificial Intelligence, pages 553{557, 2006.

[3] H. Chen, M. Lin, and Y. Wei. Novel association measures using web search with double checking. In Proc. of the COLING/ACL 2006, pages 1009{1016,2006.

[4] C. Buckley, G. Salton, J. Allan, and A. Singhal.Automatic query expansion using smart: Trec 3. In Proc. of 3rd Text REtreival Conference, pages 69{80,1994.

[5] P. Mika. Ontologies are us: A unifed model of social networks and semantics. In Proc. of ISWC2005, 2005.

[6] F. Keller and M. Lapata. Using the web to obtain frequencies for unseen bigrams. Computational Linguistics, 29(3):459{484, 2003.

[7] M. Lapata and F. Keller. Web-based models of natural language processing. ACM Transactions on Speech and Language Processing, 2(1):1{31, 2005.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proc. of 14th International Joint Conference on Artificial Intelligence, 1995.

[9] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, IEEE,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 7, JULY 2011 on "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words"

[10] G. Miller and W. Charles. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1):1{28, 1998.

[11] D. Lin. An information-theoretic definition of similarity. In Proc. of the 15th ICML, pages 296{304,1998.

[12] P. D. Turney. Mining the web for synonyms: Pmi-ir versus lsa on toe°. In Proc. of ECML-2001, pages 491{502, 2001.

[13] Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka. Graph-based word clustering using web search engine. In Proc. of EMNLP 2006, 2006

[14] D. M. Y. Li, Zuhair A. Bandar. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4):871{882, 2003.