ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

AN EFFECTIVE APPROACH FOR MEASURING SEMANTIC RESEMBLANCE USING WEB SEARCH ENGINE

K.V.Rama Murthy[1], M.Srinivasa Rao[2]

[1]MCA Student, Dept of MCA, DRK College of Engineering and Technology, Hyderabad, Andhra Pradesh, India

[2]Associate Professor, Dept of CSE, DRK College of Engineering and Technology, Hyderabad, Andhra Pradesh, India

ABSTRACT:

An important component in various tasks on the web is by measuring the semantic resemblance between words such as relation origin, community withdrawal, document gathering, and habitual metadata origin. Accurately measuring semantic similarity between two words (or entities) remains a challenging task even though the usefulness of semantic similarity measures in these applications. From a web search engine to estimate semantic similarity an empirical method using page counts and text snippets is retrieved for two words. Specifically, using page counts we define various word co-occurrence measures and incorporate those with lexical patterns extracted from text snippets. We propose a novel pattern extraction algorithm and a pattern clustering algorithm to identify the numerous semantic relations that exist between two given words. Using support vector machines the optimal combination of page counts-based conccurrence measures and lexical pattern clusters is learned. Previously proposed web-based semantic similarity measures on three benchmark data sets showing a high correlation with human ratings as the proposed method outperforms various baselines. Moreover, the accuracy in a community mining task is significantly improved by the proposed method.

Keywords: Explicit Data, Data Extraction, Text Analysis.

1. INTRODUCTION:

An important problem in web mining and data recovery is accurately measuring the semantic similarity between words. One of the main problems is to retrieve a set of documents in information retrieval is semantically related to a given user query [1] [2] [3]. For various natural languages processing tasks is efficient estimation of semantic similarity between words. General purpose lexical ontology's such as Word Net is semantically related words of a particular word are listed in manually created. In Word Net, for a particular sense of a word a synset contains a set of synonymous words. However, semantic resemblance between units changes over time and across domains [4]. To existing words, new words are constantly being created as well as new senses are assigned. To capture these new words and senses is costly if not impossible is maintained manually by ontologies. To estimate the semantic similarity between words or entities using web search engines an automatic method is proposed [6]. It is time consuming to analyze each document separately because of the vastly numerous documents and the high growth rate of the web [7] [8]. To this vast information web search engines provide an efficient interface which is shown in fig 1. Most web search

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

engines provide two useful information sources they are page counts and snippets [9] [10] [12].

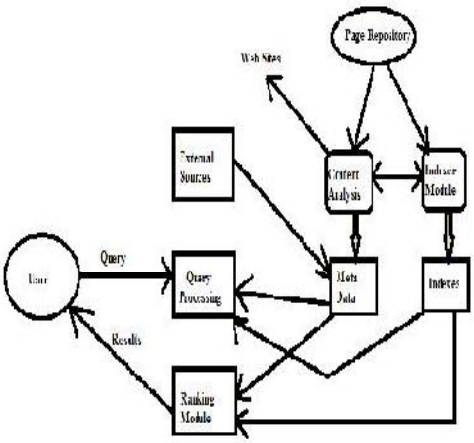

Fig 1:Architecture of web search engine

The number of pages that contain the query words is estimated by page count of a query [11].

2. PAGE COUNT-BASED COOCCURRENCE PROCEDURE:

In general, because of the queried word might appear many times on one page, the page count may not necessarily be equal to the word frequency. Page count for the query X AND Y can be considered as a global measure of co-occurrence of words X and Y. Using page counts alone as a measure of co-occurrence of two words presents several drawbacks in spite of its simplicity [13] [14]. First, the position of a word in a page is ignored by the analysis of page count. Therefore, they might not be actually related even though two words appear in a page. Second, a combination of all its senses might contain by page count of a polysemous word. For those reasons, measuring semantic similarity is unreliable by page counts. Snippets provide useful information regarding the local context of the query term by a brief window of text extracted by a search engine around the query term in a document [15] [16] [17]. Depending on the size of the pages it obviates the trouble of downloading web pages, which might be time consuming and processing snippets is also efficient. However, because of the huge scale of the web and the large number of documents in the result set a widely acknowledged drawback of using snippets is that a query can be processed efficiently with only those snippets for the top ranking results.

3. MEASURING SEMANTIC RESEMBLANCE:

Ranking of search is determined by a composite arrangement of a variety of factors distinctive to the underlying search engine. Therefore, in the top-ranking snippets exists no guarantee for all the information we need to measure semantic similarity between a given pair of words is contained [18] [19]. To overcome the above mentioned problems we experimentally show a method that considers both page counts and lexical syntactic patterns extracted from snippets is proposed. To calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy a straightforward method for a given taxonomy of words is proposed. The multiple paths might exist between the two words if a word is polysemous. In such cases, for calculating similarity only the shortest path between any two senses of the words is considered [20]. It relies on the notion that all links in the taxonomy represent a uniform distance a problem that is frequently acknowledged with this approach. A similarity measure using information content is proposed in this paper. He defined the similarity between two concepts A1 and A2 in the taxonomy as the maximum of the information content of all concepts S that subsume both A1 and A2. Then, the maximum of the similarity between any concepts that the words belong to the similarity between two words is defined. Information content is calculated using the Brown corpus as he used Word Net as the taxonomy. From a corpus in a nonlinear model combined structural semantic information from a lexical taxonomy and information content. A similarity measure that uses direct path duration, depth, and local concentration in taxonomy is proposed.

4. RESULTS:

A semantic similarity computation is estimated by the use of correlation among the similarity scores created for the word pairs in a benchmark data set and the human ratings. Both the coefficients such as Pearson correlation coefficient

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

and Spearman correlation coefficient have been used for the estimation measures on semantic resemblance. It is remarkable that Pearson correlation coefficient can get brutally affected by nonlinearities in ratings. Contrastingly, Spearman correlation coefficients primarily allocate ranks to every list of scores, and then calculate the correlation between the two lists of ranks. Consequently, Spearman correlation is more suitable for evaluating semantic resemblance measures, which may not be fundamentally linear.

5. CONCLUSION:

Using both page counts and snippets a semantic similarity measure is proposed and retrieved from a web search engine for two words. Using page counts four word cooccurrence measures were computed. To extract numerous semantic relations that exist between two words is proposed by a lexical pattern extraction algorithm. Moreover, to identify different lexical patterns that describe the same semantic relation a sequential pattern clustering algorithm was proposed. To define features for a word pair both page counts-based cooccurrence measures and lexical pattern clusters were used. For synonymous and non synonymous word pairs selected from Word Net synsets a two-class support vector machine (SVM) was trained using those features extracted. By achieving a high correlation with human ratings, the proposed method outperforms a variety of baselines as well as previously proposed web-based semantic resemblance measures.

REFERENCES:

[1] A. Kilgarriff, "Googleology Is Bad Science," Computational Linguistics, vol. 33, pp. 147-151, 2007.

[2] M. Sahami and T. Heilman, "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'l World Wide Web Conf., 2006.

[3] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Disambiguating Personal Names on the Web Using Automatically Extracted Key Phrases," Proc. 17th European Conf. Artificial Intelligence, pp. 553- 557, 2006.

[4] H. Chen, M. Lin, and Y. Wei, "Novel Association Measures Using Web Search with Double Checking," Proc. 21st Int'l Conf. Computational Linguistics and 44th Ann. Meeting of the Assoc. for Computational Linguistics (COLING/ACL '06), pp. 1009-1016, 2006.

[5] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. 14th Conf. Computational Linguistics (COLING), pp. 539-545, 1992.

[6] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain, "Organizing and Searching the World Wide Web of Facts - Step One: The One- Million Fact Extraction Challenge," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), 2006.

[7] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets," IEEE Trans. Systems, Man and Cybernetics, vol. 19, no. 1, pp. 17-30, Jan./Feb. 1989.

[8] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proc. 14th Int'l Joint Conf. Aritificial Intelligence, 1995.

[9] D. Mclean, Y. Li, and Z.A. Bandar, "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 871-882, July/Aug. 2003.

[10] G. Miller and W. Charles, "Contextual Correlates of Semantic Similarity," Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 1998.

[11] D. Lin, "An Information-Theoretic Definition of Similarity," Proc. 15th Int'l Conf. Machine Learning (ICML), pp. 296-304, 1998.

[12] R. Cilibrasi and P. Vitanyi, "The Google Similarity Distance," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383, Mar. 2007.

[13] M. Li, X. Chen, X. Li, B. Ma, and P. Vitanyi, "The Similarity Metric," IEEE Trans. Information Theory, vol. 50, no. 12, pp. 3250- 3264, Dec. 2004.

[14] P. Resnik, "Semantic Similarity in a Taxonomy: An Information Based Measure and Its Application to Problems of Ambiguity in Natural Language," J. Artificial Intelligence Research, vol. 11, pp. 95- 130, 1999.

[15] R. Rosenfield, "A Maximum Entropy Approach to Adaptive Statistical Modelling," Computer Speech and Language, vol. 10, pp. 187-228, 1996.

ISSN-2320-7884 (ONLINE)
ISSN-2321-0257 (PRINT)

[16] D. Lin, "Automatic Retrieval and Clustering of Similar Words," Proc. 17th Int'l Conf. Computational Linguistics (COLING), pp. 768- 774, 1998.
[17] J. Curran, "Ensemble Methods for Automatic Thesaurus Extrac- tion," Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002.
[18] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic Query Expansion Using Smart: Trec 3," Proc. Third Text REtreival Conf., pp. 69-80, 1994.
[19] V. Vapnik, Statistical Learning Theory. Wiley, 1998.
[20] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, pp. 22-29, 1991.

BIOGRAPHY:

K V Rama Murthy has completed B.SC (Computers) from Sri RamaKrishna Degree College, Ongole, Andhra Pradesh, and pursuing MCA in DRK College of Engineering andTechnology, JNTUH, Hyderabad. His main research interest includes Data Mining and Databases.

M.Srinivasa Rao is working as an Associate Professor in DRK College of Engineering and Technology, JNTUH, Hyderabad, Andhra Pradesh, India. He is pursuing Ph.D in Information Security. He has completed M.Tech (C.S.E) from JNTUH. His main research interest includes Information Security and Computer Ad-Hoc Networks.