IJARES/July 2014/Volume-2/Issue-7/1250-1255 **ISSN 2347- 9337**

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND SCIENCE

AN EXPOSURE TOWARDS COMPUTATION OF SIMILARITY SEARCH ON WEB SERVICES

G.Swathi[1], A.Satchidanandam[2]

[1]M.Tech Student, Dept of CSE, RRS College of Engineering & Technology, Muthangi (V), Patancheru (M), Hyderabad, T.S, India

[2]Assistant Professor, Dept of CSE, RRS College of Engineering & Technology, Muthangi (V), Patancheru (M), Hyderabad, T.S, India

ABSTRACT:

Measuring semantic resemblance among named entities is very important in numerous applications for instance query expansion, as well as community mining. Resourceful assessment of semantic similarity connecting words is crucial for a variety of natural language processing tasks for instance word sense disambiguation, as well as automatic text summarization. Semantic similarity methods were used in a variety of applications in natural language processing for instance word sense disambiguation, language modelling, as well as automatic thesauri extraction. We recommend an automatic means to approximate semantic similarity among words or else entities by means of web search engines. Semantic similarity process defined over snippets, were employed in query expansion, personal name disambiguation, as well as community mining. The projected semantic similarity assess is appealing for applications since it does not necessitate precompiled taxonomies. Ranking of search results, consequently snippets, is determined by means of complex grouping of a variety of factors exceptional to basic search engine.

Keywords: Similarity, Ranking, Natural language processing, Automatic text, Web search engines.

1. INTRODUCTION:

Web search engines make available a competent interface to this enormous information. Page counts as well as snippets are two constructive information sources offered by the majority of web search engines. Page count might not essentially be

equivalent to word frequency as queried word might come into view numerous times on one page [1][2]. In information recovery, one of most important problems is to recover a set of documents that is semantically connected to a specified user query. Resourceful assessment of semantic similarity connecting words is crucial for a variety of natural language processing tasks for instance word sense disambiguation, as well as automatic text summarization. Semantic similarity methods were used in a variety of applications in natural language processing for instance word sense disambiguation, language modelling, as well as automatic thesauri extraction. Semantic similarity measures are significant in numerous web connected tasks. In query expansion, a user query is amended by means of synonymous words to get better relevancy of search. One scheme to discover proper words to comprise in a query is to evaluate the preceding user queries by means of semantic resemblance measures. If there exist a preceding query that is semantically connected towards current query, then it is moreover suggested to user, or else internally used by search engine to change the innovative query [4][5]. No assurance exists that the entire information we require to compute semantic resemblance among a specified pair of words is controlled in top-ranking snippets. Page counts-based resemblance scores believe comprehensive co-occurrences of two words on web as shown in fig1. On the other hand, they do not believe the local context in which two words co-occur. Page counts-based co-occurrence measures as well as lexical pattern clusters were employed to describe features in support of a word pair [6][7]. Conversely snippets returned by means of search engine correspond to local context in which two words co-occur on web. We discover the frequency of several lexical syntactic models in snippets returned for conjunctive query of two words. Clustering algorithms based on pair wise assessment between the entire patterns are prohibitively time intense when the patterns are abundant.

2. METHODOLOGY:

Web mining applications for instance community mining, relation discovery, and entity disambiguation; necessitate capability to precisely compute semantic similarity among concepts or else entities. We recommend an automatic means to approximate semantic similarity among

words or else entities by means of web search engines. Because of abundant documents as well as high growth rate of web, it is time intense to analyze every document independently. Snippets, a concise window of text taken out by search engine around query term in a document, make available constructive information concerning local context of query term. Semantic similarity process defined over snippets, were employed in query expansion, personal name disambiguation, as well as community mining. Processing snippets is moreover competent as it obviates problem of downloading web pages, which valour be time intense depending on size of pages. In query expansion, a user query is amended by means of synonymous words to get better relevancy of search [8][9]. An extensively recognized problem of using snippets is that, because of huge extent of web and huge number of documents in consequence set, merely those snippets for top ranking results in support of a query is processed resourcefully. Ranking of search results, consequently snippets, is determined by means of complex grouping of a variety of factors exceptional to basic search engine. In query expansion, a user query is amended by means of synonymous words to get better relevancy of search [10]. No assurance exists that the entire information we require to compute semantic resemblance among a specified pair of words is controlled in top-ranking snippets.

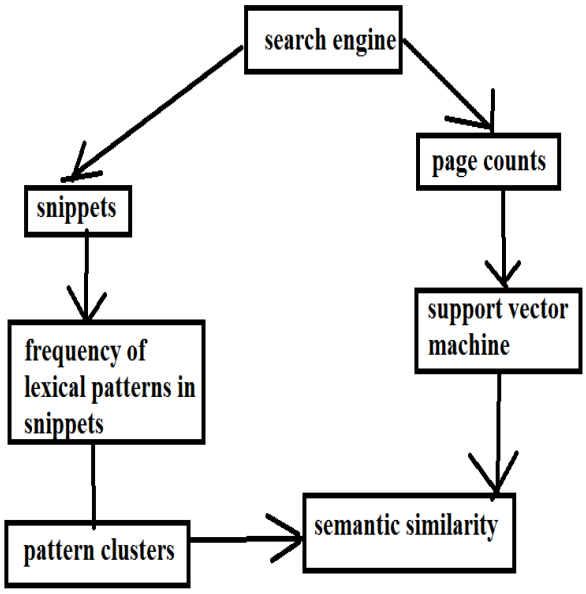

Fig1: An overview of proposed system.

3. AN OVERVIEW OF MEASURING SEMANTIC RESEMBLANCE:

Measuring semantic resemblance among named entities is very important in numerous applications for instance query expansion, as well as community mining. Resourceful assessment of semantic similarity connecting words is crucial for a variety of natural language processing tasks. Semantic similarity methods were used in a variety of applications in natural language

processing. As most named entities are not covered by means of Word Net, resemblance measures that are based on Word Net cannot be employed directly in these responsibilities. Dissimilar from common English words, named entities are being produced continually. Maintaining an advanced taxonomy of named entities is expensive, if not impracticable. The projected semantic similarity assess is appealing for applications since it does not necessitate precompiled taxonomies. We projected a semantic comparison measure by means of both page counts as well as snippets recovered from web search engine in support of two words. Four word co-occurrence processes were worked out by means of page counts. We projected a lexical pattern removal algorithm to take out several semantic relations that exist among two words. We discover the frequency of several lexical syntactic models in snippets returned for conjunctive query of two words. A sequential pattern clustering algorithm was projected to recognize dissimilar lexical patterns that explain similar semantic relation. Page counts-based co-occurrence measures as well as lexical pattern clusters were employed to describe features in support of a word pair. Page counts as well as snippets are two constructive information sources offered by the majority of web search engines. A two-class SVM was trained by means of those characteristics extracted for synonymous as well as non-synonymous word pair's chosen from Word Net synsets. Outcomes on three standard data sets showed that projected means outperform a variety of baselines and earlier projected web-based semantic similarity measures, attain a high association by means of human rating.

4. CONCLUSION:

Semantic similarity measures are significant in numerous web connected tasks. In information recovery, one of most important problems is to recover a set of documents that is semantically connected to a specified user query. Because of abundant documents as well as high growth rate of web, it is time intense to analyze every document independently. We projected a semantic comparison measure by means of both page counts as well as snippets recovered from web search engine in support of two words. Clustering algorithms based on pair wise assessment between the entire patterns are prohibitively time intense when the patterns are abundant. Web mining applications for

instance community mining, relation discovery, and entity disambiguation; necessitate capability to precisely compute semantic similarity among concepts or else entities. Snippets, a concise window of text taken out by search engine around query term in a document, make available constructive information concerning local context of query term. One scheme to discover proper words to comprise in a query is to evaluate the preceding user queries by means of semantic resemblance measures. Processing snippets is moreover competent as it obviates problem of downloading web pages, which valour be time intense depending on size of pages. A sequential pattern clustering algorithm was projected to recognize dissimilar lexical patterns that explain similar semantic relation. An extensively recognized problem of using snippets is that, because of huge extent of web and huge number of documents in consequence set, merely those snippets for top ranking results in support of a query is processed resourcefully.

REFERENCES

[1] 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.

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

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

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

[5] 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.

[6] 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.

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

[8] D. Lin, "Automatic Retrieval and Clustering of Similar Words," Proc. 17th Int'l Conf. Computational Linguistics (COLING), pp. 768-
774, 1998.

[9] J. Curran, "Ensemble Methods for Automatic Thesaurus Extraction," Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2002.

[10] 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.

[11] K. Church and P. Hanks, "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, pp. 22-29, 1991.

[12] Z. Bar-Yossef and M. Gurevich, "Random Sampling from a Search Engine's Index," Proc. 15th Int'l World Wide Web Conf., 2006.

[13] F. Keller and M. Lapata, "Using the Web to Obtain Frequencies for Unseen Bigrams," Computational Linguistics, vol. 29, no. 3, pp. 459-484, 2003.