Measuring Exegetic Semblance between Words

M.Sandhya
M.Tech in CSE Dept,
Balaji Institute of Technology and Science
Narsampet, Warangal, Andhra Pradesh,India

T.Upender
Assoc.prof
Balaji Institute of Technology and Science
Narsampet, Warangal, Andhra Pradesh,India

Abstract - Similarity between words that was concerned with the syntactic similarity of two strings. Semantic similarity is a confidence score that reflects the semantic relation between the meanings of two sentences. It is difficult to gain a high accuracy score because the exact semantic meanings are completely understood only in a particular context. The goals of the paper are to present you some dictionary-based algorithms to capture the semantic similarity between two sentences, which is heavily based on the semantic dictionary. A web search engine is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. To identify the numerous semantic relations that exist between two given words, we propose a novel pattern extraction algorithm and a pattern clustering algorithm. The optimal combination of page counts-based co-occurrence measures and lexical pattern clusters is learned using support vector machines.

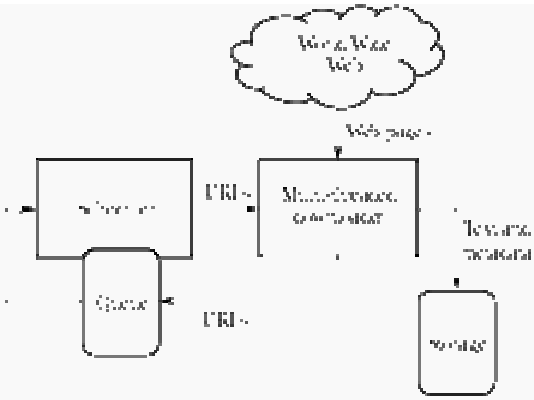
Key Words: - Web mining, information extraction, web text analysis.

I.INTRODUCTION

Web search engines work by storing information about many web pages, which they retrieve from the HTML itself. These pages are retrieved by a Web crawler (sometimes also known as a spider) — an automated Web browser which follows every link on the site. Exclusions can be made by the use of robots.txt. The contents of each page are then analyzed to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The purpose of an index is to allow information to be found as quickly as possible. Some search engines, such as Google, store all or part of the source page (referred to as a cache) as well as information about the web pages, whereas others, such as AltaVista, store every word of every page they find. This cached page always holds the actual search text since it is the one that was actually indexed, so it can be very useful when the content of the current page has been updated and the search terms are no longer in it. This problem might be considered to be a mild form of linkrot, and Google's handling of it increases usability by satisfying user expectations that the search terms will be on the returned webpage. This satisfies the principle of least astonishment since the user normally expects the search terms to be on the returned pages. Increased search relevance makes these cached pages very useful, even beyond the fact that they may contain data that may no longer be available elsewhere.

When a user enters a query into a search engine (typically by using keywords), the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. The index is built from the information stored with the data and the method by which the information is indexed. As early as 2007 the Google.com search engine has allowed one to search by date by clicking 'Show search tools' in the leftmost column of the initial search results page, and then selecting the desired date range. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or

phrases you search for. As well, natural language queries allow the user to type a question in the same form one would ask it to a human. A site like this would be ask.com.



High-level architecture of a standard Web crawler

The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another.[10] The methods also change over time as Internet usage changes and new techniques evolve. There are two main types of search engine that have evolved: one is a system of predefined and hierarchically ordered keywords that humans have programmed extensively. The other is a system that generates an "inverted index" by analyzing texts it locates. This first form relies much more heavily on the computer itself to do the bulk of the work.

Most Web search engines are commercial ventures supported by advertising revenue and, as a result, some employ, the practice of allowing advertisers to pay money to have their listings ranked higher in search results. Those search engines which do not accept money for their search engine results make money by running search related ads alongside the regular search engine results. The search engines make money every time someone clicks on one of these ads.

II.RELATED WORK

WordNet is a lexical database which is available online, and provides a large repository of English lexical items. There is a multilingual WordNet for European languages which is structured in the same way as the English language WordNet.

WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses = structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents. For example, the words night, night time, and dark constitute a single synset that has the following gloss: the time after sunset and before sunrise while it is dark outside. Synsets are connected to one another through explicit semantic relations. Some of these relations (hypernym, hyponym for nouns, and hypernym and troponym for verbs) constitute is-a-kind-of (holonymy) and is-a-part-of (meronymy for nouns) hierarchies.

For example, tree is a kind of plant, tree is a hyponym of plant, and plant is a hypernym of tree. Analogously, trunk is a part of a tree, and we have trunk as a meronym of tree, and tree is a holonym of trunk. For one word and one

type of POS, if there is more than one sense, WordNet organizes them in the order of the most frequently used to the least frequently used (Semcor).

III.PROPOSED WORK

Normalized Google distance is a semantic similarity measure derived from the number of hits returned by the Google search engine for a given set of keywords. Keywords with the same or similar meanings in a natural language sense tend to be "close" in units of Google distance, while words with dissimilar meanings tend to be farther apart.

Specifically, the Normalized Google Distance between two search terms x and y is

$$\mathrm{NGD}(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

where M is the total number of web pages searched by Google; f(x) and f(y) are the number of hits for search terms x and y, respectively; and f(x, y) is the number of web pages on which both xand y occur.

If the two search terms x and y never occur together on the same web page, but do occur separately, the normalized Google distance between them is infinite. If both terms always occur together, their NGD is zero, or equivalent to the coefficient between x squared and y squared.

The Normalized Google Distance is derived from the earlier Normalized Compression Distance (Cilibrasi & Vitanyi 2003).

IV. PROBLEM DEFINITION AND ASSUMPTIONS

In this section we will describe the various similarity features we use in our model. We utilize page counts and snippets returned by the Google search engine for simple text queries to define various similarity scores.

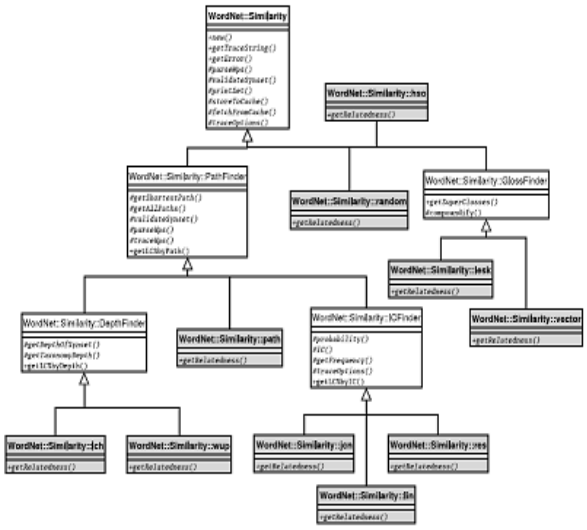
4.1 Semantic similarity between words

Given two words, the measurement determines how similar the meaning of two sentences is. The higher the score, the more similar the meaning of the two words.
Here are the steps for computing semantic similarity between two words:

- First, each word is partitioned into a list of tokens.
- Part-of-speech disambiguation (or tagging).
- Stemming words.
- Find the most appropriate sense for every word (Word Sense Disambiguation).
- Finally, compute the similarity of the words based on the similarity of the pairs of words.

***4.2** Semantic similarity algorithms*

These algorithms usually perform a tree walk over the relationships in WordNet to come up with a real-valued score of how related two terms are. These will be limited by how well WordNet models the concepts that you are interested in WordNet Similarity (written in Perl) is pretty good.



4.3 OpenCyc as a knowledge base

OpenCyc is a open-source version of Cyc, a very large knowledge base of 'real-world' facts. It should have a much richer set of sematic relationships than WordNet does. However, I have never used OpenCyc so I can't speak to how complete it is, or how easy it is to use.

4.4 n-gram frequency analysis

As mentioned by Jeff Moser. A data-driven approach that can 'discover' relationships from large amounts of data, but can often produce noisy results.

***4.5** Latent Semantic Analysis*

A data-driven approach similar to n-gram frequency analysis that finds sets of semantically related words.

V. Experiment and Result

The implementation environment has software such as Visual Studio(.Net) running in Windows XP operating system which is in a PC with 2GB RAM and 2.x MHz processing power. The system uses ASP.NET to build user interface. Enter the Word to be searched in the search box provided. The search box is auto complete which will populate the matched words automatically. After entering the Search item there are three radio button options All Words, Any Words, Phrase. Select any one option from the radio buttons.

After selecting the radio button click on the search button.

Fig.3 Search for the word.

The results will be displayed after clicking on the search button.

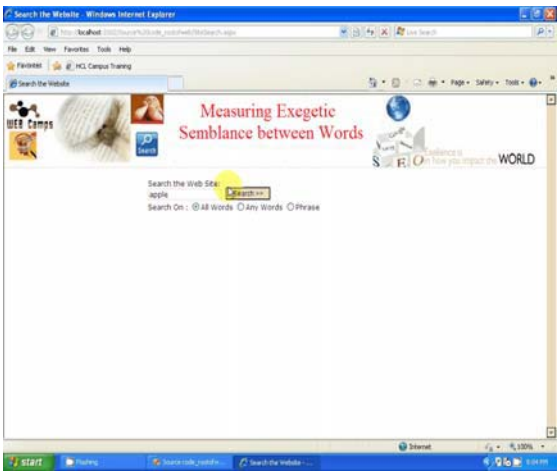


Fig .4 Enter the search and select any one option

The search word is searched in the websites based on the word match and the results with the search word will be displayed.

Fig.5 Links displayed based on the search

The number documents which are searched are also displayed.

Based on the click on link it will be redirected to the site where the word has been found. The complete details of the word are also displayed.

Fig.7 Link based on the search result.

VI.CONCLUSION

Wireless Sensor Networks have resource constraints. They need to have security systems such as intrusion detection system. However, an IDS running in wireless sensor node consumes more energy which leads to early demise of network. This paper presents an energy efficient intrusion detection mechanism that improves life of WSN. The effectiveness of the simulation model is tested with simulations using a custom-built simulator developed in Java programming language. The results revealed that the proposed analytical model is effective and can be used in real world applications.

REFERENCES

[1] A. Bagga and B. Baldwin. 1998. Entity-based cross document coreferencing using the vector space model. In *Proc. of 36th COLING-ACL*, pages 79–85.

[2] Z. Bar-Yossef and M. Gurevich. 2006. Random sampling from a search engine's index. In *Proceedings of 15th International World Wide Web Conference*. H. Chen, M. Lin, and Y. Wei. 2006. Novel association measures using web search with double checking. In *Proc. of the COLING/ACL 2006*, pages 1009–1016.

[3] J. Curran. 2002. Ensemble menthods for automatic thesaurus extraction. In *Proc. of EMNLP*.M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proc. of 14th COLING*, pages 539–545.

[4] J.J. Jiang and D.W. Conrath. 1998. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the International Conference on Research in Computational Linguistics ROCLING X*.

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

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

[7] D. Lin. 1998a. Automatic retreival and clustering of similar words. In *Proc. of the 17th COLING*, pages 768–774.

[8] D. Lin. 1998b. An information-theoretic definition of similarity. In *Proc. of the 15th ICML*, pages 296–304.C. D. Manning and H. Sch¨utze. 2002. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

[9] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. 2006a. Polyphonet: An advanced social network extraction system. In *Proc. of 15th International World Wide Web Conference*.

[10] Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka. 2006b. Graph-based word clustering using web search engine. In *Proc. of EMNLP 2006*.G. Miller and W. Charles. 1998. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

[11] M. Pasca, D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006. Organizing and searching the world wide web of facts - step one: the one-million fact extraction challenge.In *Proc. of AAAI-2006*.

[12] J. Platt. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74. R. Rada, H. Mili, E. Bichnell, and M. Blettner. 1989.

[13] Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 9(1):17–30.

[14] P. Resnik and N. A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380. P. Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. Of 14th .International Joint Conference on Aritificial Intelligence*.P. Resnik. 1999.

[15] Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Aritificial Intelligence Research*, 11:95–130.
[16] R. Rosenfield. 1996. A maximum entropy approach to adaptive statistical modelling. *Computer Speech and Language*, 10:187–228. H. Rubenstein and J.B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8:627–633.
[17] M. Sahami and T. Heilman. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *Proc. of 15th International World Wide Web Conference*. E. Terra and C.L.A. Clarke. 2003. Frequency estimates for statistical word similarity measures. In *Proc. of the NAACL/HLT*, pages 165–172.
[18] P. D. Turney. 2001. Minning the web for synonyms: Pmi-ir versus lsa on toefl. In *Proc. of ECML-2001*, pages 491–502.
[19] V. Vapnik. 1998. *Statistical Learning Theory*. Wiley, Chichester, GB.
[20] D. McLean Y. Li, Zuhair A. Bandar. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

.