A WEB SEARCH ENGINE-BASED APPROACH TO MEASURE SEMANTIC SIMILARITY BETWEEN WORDS

KURRELA CHANDRAM,

Assistant Professor,

Department of Computer Science and Engineering,

KODI RAJESH,

Associate Professor,

Department of Electronics and Communications Engineering,

Siddhartha Institute of Technology and Sciences,

Narapally, Hyderabad, Telangana – 500 088.

Abstract

An online search engine yielded an empirical approach for estimating semantic similarity using page counts and text samples for two terms. It does so by utilizing page counts to establish multiple word co-occurrence metrics and combining them with lexical patterns taken from text excerpts. A unique pattern extraction approach and a pattern clustering algorithm are developed to detect the multiple semantic links that exist between two provided words. Support vector machines are used to discover the best mix of page counts-based co-occurrence metrics and lexical pattern clusters. In a group mining task, the suggested strategy considerably enhances accuracy. In a community extraction task, it uses the suggested semantic similarity metric to find relationships between items, namely individuals. With statistically substantial accuracy and recall values, the suggested technique outperforms the baselines. The results of the community mining challenge demonstrate that the suggested technique may be used to compare the semantic similarity of not just words, but also named entities for which manually produced lexical ontologies are either missing or incomplete.

1. Introduction

A web search engine is used to find information on the World Wide Web and FTP servers. The search results are often shown as a list of results, which is referred to as a search engine results page (SERPs). Web pages, photos, data, and other sorts of files may be included in the data. Some search engines also harvest data from databases or open directories. Unlike online directories, which are solely updated by individual editors, search engines keep actual information by executing a program on a web crawler.

The main purpose of the data mining process is to extract information from a data collection in a human-understandable structure, which includes database and administration steps in addition to the raw analysis stage. The true data mining job is to analyse vast amounts of data automatically or semi-automatically in order to identify previously discovered interesting patterns such as groupings of data records (cluster analysis), atypical records (anomaly identification), and relationships (association rule mining). Typically, this entails the use of database techniques such as spatial indexes.

The criterion for determining similarity depends on the implementation. Data clustering is a strategy for physically storing information that is conceptually comparable. The amount of disc accesses must be reduced in order to improve database system efficiency. Objects with comparable features are grouped together in one class of objects, and a single disc access makes the entire class available.

Web mining is the use of data mining tools to uncover trends on the Internet. Web mining may be classified into three forms based on the analytical objectives: web content mining, web usage mining, and web structure mining. This can be accomplished, for example, by establishing a topological similarity, utilizing ontologies to define a distance between words, or employing statistical methods, such as a vector space model, to connect words and textual contexts from a relevant text corpus.

2. Literature survey

A Web based Kernel Function for Measuring the Similarity of Short Text Snippets," the authors stated that traditional document similarity measures (e.g., cosine) perform poorly when determining the similarity of short text snippets, such as search queries, because there are often few, if any, terms in common between two short text snippets. They solved this issue by developing a unique approach for calculating the similarity of short text snippets (including those with no overlapping phrases) by exploiting online search results to offer more context for the brief texts.

There are numerous cases in text analysis when we want to know how similar two brief snippets are. For example, there may be several ways to define a topic or people, such as "United Nations Secretary-General" and "Kofi Annan," and they want to know whether there is a high level of semantic similarity between these two text samples. Similarly, the meanings of the words "AI" and "Artificial Intelligence" are quite similar, even if they may not share any real terminology.

However, employing typical document similarity techniques, such as the commonly used cosine coefficient, to such brief text fragments frequently yields insufficient results. Indeed, in each of the above cases, using the cosine would result in a similarity of 0 because each text pair has no common phrases. Even though two snippets share terminology, they may use the phrase in distinct contexts.

The first refers to graph structures, whereas the second refers to visual presentations. As a result, while the cosine score between these two snippets would be 0.5 owing to the common lexical term "graphical," the use of this same phrase at a semantic level does not genuinely indicate similarity between the snippets. They would need a technique for assessing the similarity of such brief text snippets that captures more of the semantic context of the snippets rather than only calculating term-wise similarity to solve this challenge.

3. Methodology

Using web search engines, the project provides an algorithmic approach for estimating semantic similarity between words or concepts. It takes a long time to study each document individually due to the large number of papers and the rapid development pace of the internet. Web search engines give a quick and easy way to access all of this data. Most web search engines give two useful information sources: page counts and snippets.

A query's page count is an estimate of how many pages include the query terms. Because the same word may appear many times on a single page, page count may not always be equal to word frequency. As a result, the research provides an approach that takes into account both page counts and lexical syntactic patterns retrieved from snippets, and shows that it can solve the challenges outlined above.

The novel system uses text samples collected from a web search engine to compute semantic similarity between words or entities using an automatically extracted lexical syntactic patterns-based method. The novel method presents a lexical pattern extraction approach that takes into account text snippets' word subsequences. Furthermore, the extracted collection of patterns is grouped to locate patterns that express the same semantic association.

- **Page Count Based Co-Occurrence Measures**

Words 1 and 2 are typed in here. The words are joined into a word pair and presented. HTML pages are stored in the 'web document' folder in the application's root folder. These terms are used to search the pages. There are three list boxes available. The first listbox is filled with page names that contain the term 'word1'.

The second listbox is filled with page titles that contain the term 'word2'. The third listbox is filled with page titles that include both terms. Label controls additionally show the counts of word1 pages, word2 pages, and both words. The settings are saved in the class 'GlobalClass' and utilized in subsequent modules.

- **Web Jaccard**

The H (P) page count with word1, the H (Q) page count with word2, and the H(PQ) page count with word pair are presented in label controls, and the Web Jaccard Value is computed and displayed in another label control.

- **Web Overlap**

In this case, the H (P) page count with word1, the H (Q) page count with word2, the H(PQ) page count with word pair, the minimum of H(P) and H(Q), and the Web Overlap Value are all presented in label controls.

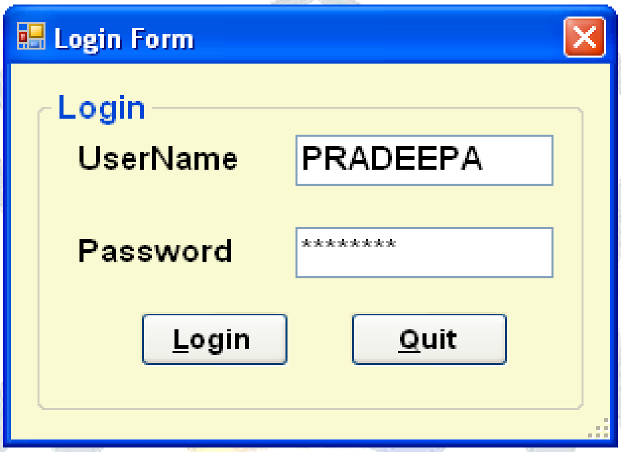
- **Web Dice**

The H(P) page count with word1, the H(Q) page count with word2, the H(PQ) page count with word pair, and the 2 * H(PQ) page count with word pair are presented in label controls, and the Web Dice Value is computed and displayed in a label control.

- **Web PMI (Point Wise Mutual Information)**

H (P) page count with word1, H (Q) page count with word2, H (PQ) page count with word pair, H(P)/N, H(Q)/N, H(PQ)/N, and Web PMI Value are computed and presented in label controls. 'N' is considered to be 10 in the sample values. The 'N' will be 10 to the power of 10 or greater in real time.

4. Result and discussion



Login Form

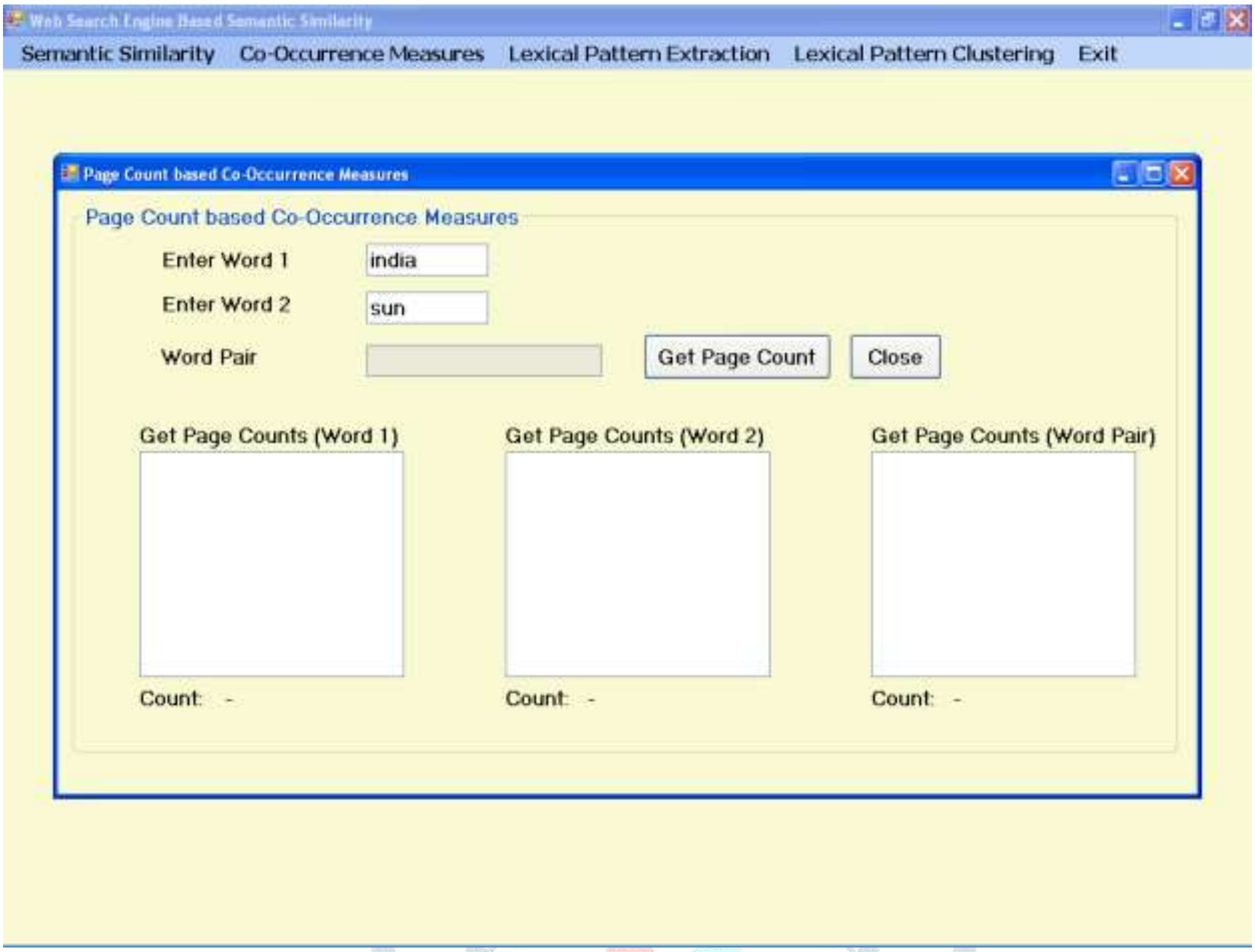
Login

UserName PRADEEPA

Password *********

Fig.1 Login Form

This form is for security purposes only. If the login or password is incorrect, an error message will appear. It will login to the application if this is not the case.


Web Search Engine Based Semantic Similarity
Semantic Similarity
Co-Occurrence Measures
Lexical Pattern Extraction
Lexical Pattern Clustering
Exit
Page Count based Co-Occurrence Measures
Get Page Count
Close
Get Page Counts (Word 1)
Get Page Counts (Word 2)
Get Page Counts (Word Pair)

Page Count based Co-Occurrence Measures

Enter Word 1: india

Enter Word 2: sun

Word Pair:

Count: -

Count: -

Count: -

Fig.2 Page Count Based Co- Occurrence Measures

This form displays the page count in the listbox control.

5. Conclusion

For two terms, the research suggested a semantic similarity metric that used both page counts and excerpts from a web search engine. Using page counts, four word co-occurrence metrics were calculated. It introduced a lexical pattern extraction approach for extracting a large number of semantic links between two words. A sequential pattern clustering approach was also devised in order to detect distinct lexical patterns that reflect the same semantic connection. It uses page counts to establish several word co-occurrence metrics and combines them with lexical patterns taken from text excerpts. The characteristics for a word pair were defined using both page counts-based co-occurrence measures and lexical pattern clusters. Those characteristics retrieved for synonymous and non-synonymous word pairings picked from WordNet synsets were used to train a two-class Support Vector Machine (SVM).

References

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

2. Church.K and Hanks.P,(1991) "Word Association Norms, Mutual Information and Lexicography," Computational Linguistics, vol. 16, pp. 22-29.

3. Cilibrasi.R and Vitanyi.P,(2007) “The Google Similarity Distance,” IEEE Trans. Knowledge and Data Eng., vol. 19, no. 3, pp. 370-383.

4. Cristianini.N and Shawe-Taylor.J.(2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.

5. Gabrilovich.E and Markovitch.S,(2007), “Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis,” Proc.Int’l Joint Conf. Artificial Intelligence (IJCAI ’07), pp. 1606- 1611.

6. Mclean.D, Li.Y, and Bandar.Z.A.,(2003), “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.

7. Jiang.J and Conrath.D,(1997), “Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,” Proc. Int’l Conf Research in Computational Linguistics.

8. Lin.D,(2007) “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.

9. Lapata.M and Keller.F,(2005), “Web-Based Models for Natural Language Processing,” ACM Trans. Speech and Language Processing, vol. 2, no. 1, pp. 1-31.

10. Pasca.M, Lin.D, Bigham.J, Lifchits.A, and Jain.A,(2006) “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).