LEXICAL PATTERN EXTRACTION USING RANDOM FOREST

SHIVA PRASAD YADAV,
Assistant Professor,
Department of Computer Science and Engineering,

RAMESH BHANOTHU,
Assistant Professor,
Department of Electronics and Communications Engineering,

Siddhartha Institute of Technology and Sciences,
Narapally, Hyderabad, Telangana – 500 088.

Abstract

Using page counts and text samples from two phrases, an internet search engine generated an empirical technique for measuring semantic similarity. It accomplishes this by using page counts to create numerous word co-occurrence metrics and combines them with lexical patterns extracted from text snippets. To discover the many semantic linkages that exist between two supplied words, a novel pattern extraction technique and a pattern clustering algorithm are created. Random forest are used to find the optimal combination of co-occurrence metrics based on page counts and lexical pattern clusters. The proposed technique improves accuracy significantly in a group mining job. In a community extraction task, the recommended semantic similarity measure is used to discover links between elements, especially persons. The proposed approach outperforms the baselines with statistically significant accuracy and recall values. The findings of the community mining challenge show that the proposed approach may be used to compare the semantic similarity of not just words, but also named things for which manually constructed lexical ontologies are either lacking or incomplete.

1. Introduction

The primary goal of data process is to discover information from such a data gathering in a human-readable manner, which comprises database and administration procedures in addition to the raw analysis stage. The genuine data mining task is to automatically or semi-automatically analyse massive volumes of data in order to detect previously found interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly identification), and correlations (association rule mining). This usually includes the use of database techniques like spatial indexes.

The use of data mining technologies to find trends on the Internet is known as web mining. Based on the analytical goals, web mining may be divided into three types: web content mining, web use mining, and web structure mining. This can be performed by creating topological similarities, using ontologies to define a distance between words, or using statistical approaches, such as a vector space model, to correlate words and textual contexts from a suitable text corpus.

To discover information on the World Wide Web and FTP servers, a web search engine is utilized. The search results are frequently shown as a list of results, which is known as a search engine results page (SERPs). The data may contain web pages, images, data, and other types of media. Some search engines collect information from databases or open directories as well. Search engines preserve genuine information by executing a program on a web crawler, as opposed to online directories, which are simply updated by individual editors.

The implementation determines the criterion for detecting similarity. Data clustering is a mechanism for physically storing conceptually related information. To increase database system efficiency, the number of disc accesses must be decreased. Objects with similar characteristics are clustered with each other in a single class, and a single disc access renders the overall class available.

2. Literature survey

In text analysis, there are various instances where we wish to determine how similar two small samples are. For example, there might be several ways to characterize a topic or person, 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 connotations of the terms "AI" and "Artificial Intelligence" are quite similar, even though they do not share any genuine vocabulary.

The first is about graph structures, meanwhile the second is about visual displays. As a consequence, while the cosine score between these two snippets would be 0.5 due to the same lexical word "graphical," the use of this identical phrase at a semantic level does not actually suggest similarity. To tackle this difficulty, they would require a mechanism for analyzing the similarity of such tiny text samples that captures more of the semantic context of the snippets rather than just calculating term-wise similarity.

A Web-based Kernel Function for Comparing Short Text Snippets, Traditional document similarity metrics, such as cosine, perform poorly when detecting the similarity of short text snippets, such as search queries, since there are typically few, if any, phrases in common between two short text snippets, according to the authors. They solved the problem by devising a novel method for measuring the similarity of short text snippets (even those with no overlapping words) that takes advantage of online search results to provide additional context for the short texts.

However, applying standard document similarity approaches to such tiny text segments, such as the regularly used cosine coefficient, typically produces inadequate results. Because each text pair has no common phrases, employing the cosine would result in a similarity of 0 in each of the situations above. Even though two snippets use the same word, they may utilize it in different contexts.

3. Methodology

In web mining, information retrieval, and natural language processing, accurately quantifying semantic similarity between words is a major challenge. Web mining applications that demand the capacity to

reliably evaluate the semantic similarity of ideas or things include community extraction, connection identification, and entity disambiguation. One of the most difficult challenges in information retrieval is retrieving a group of documents that are semantically relevant to a given user query.

Various natural language processing tasks, such as word sense disambiguation (WSD), textual entailment, and automatic text summarization, need accurate assessment of semantic similarity between terms. In manually generated general-purpose lexical ontologies like WordNet, semantically related terms of a certain word are listed. A synset is a collection of synonyms for a certain sense of a word in WordNet. Semantic similarity between things, on the other hand, varies across time and between domains.

- **Lexical Pattern Extraction**

The search pattern, which includes the first and last words, is typed here. The phrase is tested in the web pages so that the pattern is initial word, any number of words, and last word. The skip count number of words in the phrase identified in the web pages can be disregarded during pattern extraction.

- **Lexical Pattern Clustering**

The lexical pattern clustering technique may be used to group the patterns. The patterns are grouped, and then the count and co-occurrence of the term are examined. The word can be deduced from this. The cluster may be grouped depending on the threshold value entered in the textbox control, the words are clustered, and the results are displayed in the listbox control.

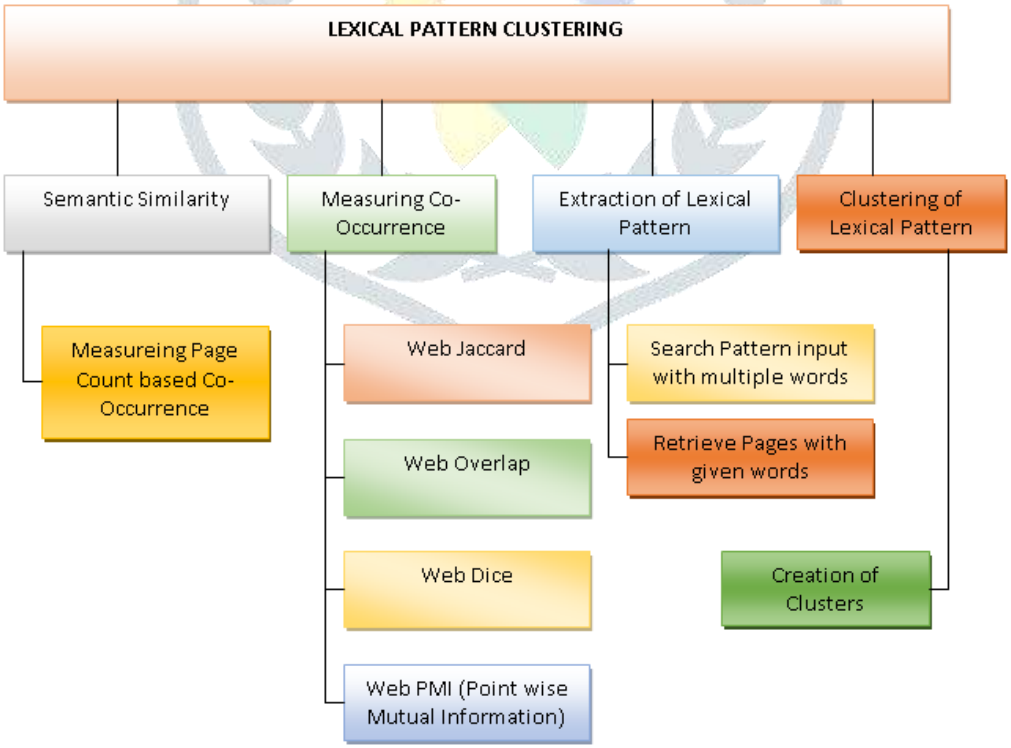

Fig.1 System Flow Diagram

The most crucial factor to consider when creating a database is how the data will be used. The major goals of database design are as follows:

- **Data Integration**

Information from many files is combined, accessed, and worked on as if it were in a single file in a database. Although the data is logically centralized, it may be physically dispersed among several devices that are linked by data connection infrastructure.

- **Data Integrity**

Data integrity refers to keeping all data in a single location and determining how each program may access it. This method produces more consistent data, as one update is enough to change the record status for all apps that utilize it. This results in less data redundancy; data items do not need to be duplicated; and direct access storage requirements are reduced.

4. Result and discussion

The lexical pattern clustering technique may be used to organize the patterns. After clustering the patterns, the word count and co-occurrence can be examined. The term may be deduced from this information. The words are clustered, and the results are shown in the listbox control based on the threshold value supplied in the textbox control. It displays the message "Please enter a value between 0 and 1" if the threshold value range exceeds 1.

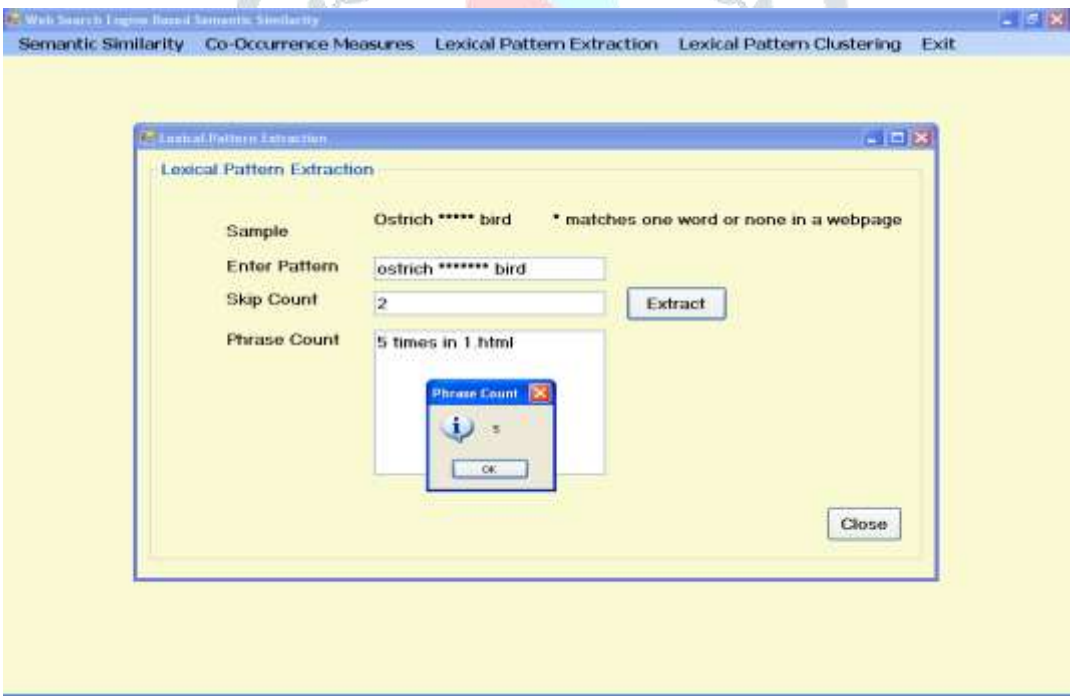

Fig 2. Lexical Pattern Extraction

This form shows the word's lexical pattern extracted from the webpage.

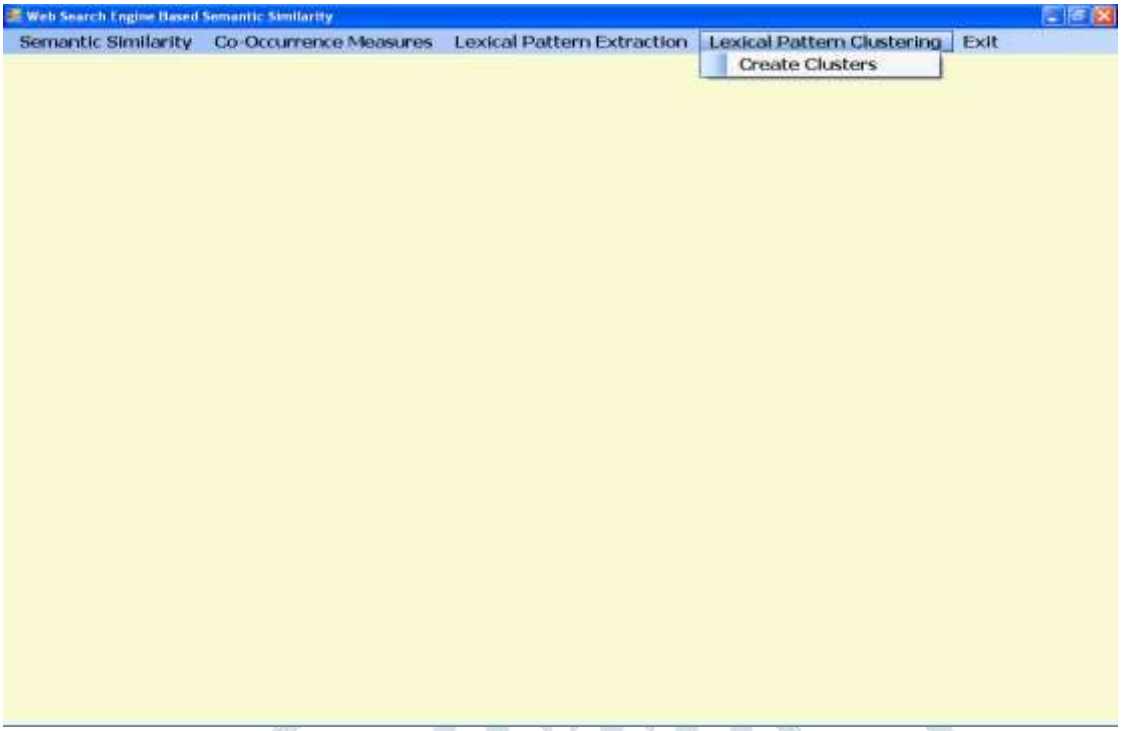

Fig.3 Lexical Pattern Clustering Menu

This form displays the lexical pattern clustering menu.

5. Conclusion

In order to find various lexical patterns that represent the same semantic relationship, a sequential pattern clustering technique was also designed. It combines lexical patterns extracted from text snippets with page counts to produce multiple word co-occurrence metrics. Both page counts-based co-occurrence metrics and lexical pattern clusters were used to determine the properties of a word pair. The properties acquired from WordNet synsets for synonymous and non-synonymous word pairs were utilized to train random forest. The study proposed a semantic similarity score based on page counts and snippets from an online search engine for two phrases. Four word co-occurrence metrics were derived using page counts. It established a method for obtaining a large number of semantic linkages between two words called lexical pattern extraction.

References

1. 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.
2. Jiang.J and Conrath.D,(1997), "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," Proc. Int'l Conf Research in Computational Linguistics.
3. 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.

4. 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.
5. 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.
6. 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).
7. Sahami.M and Heilman.T,(2006) "A Web-Based Kernel Function for Measuring the Similarity of Short Text Snippets," Proc. 15th Int'lWorld Wide Web Conf.
8. Strube.M and Ponzetto.S.P,(2006) "Wikirelate! Computing Semantic Relatedness Using Wikipedia," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), pp. 1419-1424.
9. Ted Pedersen, Amruta Purandare, and Anagha Kulkarni,(2005) 'Name discrimination by clustering similar contexts', in Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics.
10. Xin Li, Paul Morie, and Dan Roth,(2005) 'Semantic integration in text, from ambiguous names to identifiable entities', AI Magazine, American Association for Artificial Intelligence, Spring, 45–58.