International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-4)

Research Article | **April 2015**

A Web Search Engine-Based Approach to Measure Semantic Similarity between Words

R. Karthikeyan
PG Student
Prist University, Tamil Nadu, India

V. Udhayakumar
MCA, MTech, HOD Department of CSE
Prist University, Tamil Nadu, India

Abstract:

Measuring the semantic similarity between words is an important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. Despite the usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words (or entities) remains a challenging task. We propose an empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically, we define various word co-occurrence measures using page counts and integrate those with lexical patterns extracted from text snippets. 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. The proposed method outperforms various baselines and previously proposed web-based semantic similarity measures on three benchmark data sets showing a high correlation with human ratings. Moreover, the proposed method significantly improves the accuracy in a community mining task.

Keywords: Web mining, Information extraction, Web text analysis

I. INTRODUCTION

ACCURATELY measuring the semantic similarity between words is an important problem in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation, require the ability to accurately measure the semantic similarity between concepts or entities. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. Efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation (WSD), textual entailment, and automatic text summarization. Semantically related words of a particular word are listed in manually created general-purpose lexical Ontologies such as WordNet.1 In Word Net, a synset contains a set of synonymous words for a particular sense of a word. However, semantic similarity between entities changes over time and across domains. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries. A user who searches for apple on the web, might be interested in this sense of apple and not apple as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining ontologies to capture these new words and senses is costly if not impossible. We propose an automatic method to estimate the semantic similarity between words or entities using web search engines. Because of the vastly numerous documents and the high growth rate of the web, it is time consuming to analyze each document separately. Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines. Page count of a query is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the queried word might appear many times on one page. Page count for the query P AND Q can be considered as a global measure of concurrence of words P and Q. For example, the page count of the query “apple” AND “computer” in Google is 288,000,000, whereas the same for “banana” AND “computer” is only 3,590,000. The more than 80 times more numerous page counts for “apple” AND “computer” indicate that apple is more semantically similar to computer than is banana. Despite its simplicity, using page counts alone as a measure of co-occurrence of two words presents several drawbacks. First, page count analysis ignores the position of a word in a page. Therefore, even though two words appear in a page, they might not be actually related. Second, page count of a polysemous word (a word with multiple senses) might contain a combination of all its senses. For example, page counts for apple contain page counts for apple as a fruit and apple as a company. Moreover, given the scale and noise on the web, some words might co-occur on some pages without being actually related [1]. For those reasons, page counts alone are unreliable when measuring semantic similarity.

II. PROBLEM DEFINITION

Retrieving accurate information for users in Search Engine faces a lot of problems. This is due to accurately measuring the semantic similarity between words is an important problem.

For example, the word "apple" consists of two meaning one indicates the fruit apple and the other is the apple company. So retrieving accurate information to users to such kind of similar words is challenging.

In the base paper, the authors proposed an architecture and method to measure semantic similarity between words. Which consists of snippets, page-counts and support vector machine.

The authors proposed an approach to compute the semantic similarity between words or entities using text snippets. But in this project we are going to implement and compute the semantic similarity between words in Search engine without using Snippets or Support Vector Machines. Because using Snippets or Support Vector Machines makes the job of finding similarity easier. So we are going to implement the same concept without using snippets or support Vector machines.

III. SYSTEM ANALYSIS

Existing Work:

- Accurately measuring the semantic similarity between words is an important problem in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation, require the ability to accurately measure the semantic similarity between concepts or entities.
- In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query.
- Efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation (WSD), textual entailment, and automatic text summarization.
- For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries.

Proposed Work:

- We propose an automatic method to estimate the semantic similarity between words or entities using web search engines.
- Web search engines provide an efficient interface to this vast information. Page counts and snippets are two useful information sources provided by most web search engines.
- Page count of a query is an estimate of the number of pages that contain the query words. In general, page count may not necessarily be equal to the word frequency because the queried word might appear many times on one page.
- We present an automatically extracted lexical syntactic patterns-based approach to compute the semantic similarity between words or entities using text snippets retrieved from a web search engine.

IV. FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ECONOMICAL FEASIBILITY
- TECHNICAL FEASIBILITY
- SOCIAL FEASIBILITY

Economical Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a

necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

V. MODULE DESCRIPTION

Lexical Pattern Extraction

In this module, Words in Page are extracted. It uses counts-based co-occurrence measures. Lexical Pattern Clustering. This can be problematic if one or both words are polysemous, or when page counts are unreliable. On the other hand, the snippets returned by a search engine for the conjunctive query of two words provide useful clues related to the semantic relations that exist between two words. A snippet contains a window of text selected from a document that includes the queried words. Snippets are useful for search because, most of the time, a user can read the snippet and decide whether a particular search result is relevant, without even opening the url. Using snippets as contexts is also computationally efficient because it obviates the need to download the source documents from the web, which can be time consuming if a document is large.

Lexical Pattern Clustering

Typically, a semantic relation can be expressed using more than one pattern. For example, consider the two distinct patterns, X is a Y, and X is a large Y. Both these patterns indicate that there exists and is-a relation between X and Y. Identifying the different patterns that express the same semantic relation enables us to represent the relation between two words accurately. According to the distributional hypothesis, words that occur in the same context have similar meanings. The distributional hypothesis has been used in various related tasks, such as identifying related words, and extracting paraphrases. If we consider the word pairs that satisfy (i.e., co-occur with) a particular lexical pattern as the context of that lexical pair, then from the distributional hypothesis, it follows that the lexical patterns which are similarly distributed over word pairs must be semantically similar.

Measuring Semantic Similarity

We defined four co-occurrence measures using page counts. We showed how to extract clusters of lexical patterns from snippets to represent numerous semantic relations that exist between two words. In this module, we describe a machine learning approach to combine both page counts-based co-occurrence measures, and snippets-based lexical pattern clusters to construct a robust semantic similarity measure.

Ranking search results

In this module, an automatic method to estimate the semantic similarity between words or entities using web search engines with ranking the search results occur. Accurately measuring the semantic similarity between words is an important problem in web mining, information retrieval, and natural language processing. Web mining applications such as, community extraction, relation detection, and entity disambiguation, require the ability to accurately measure the semantic similarity between concepts or entities. Based on the similarity between the user given search keyword, the ranking takes place.

VI. CONCLUSION AND FUTURE ENHANCEMENT

We proposed a semantic similarity measure using both page counts and snippets retrieved from a web search engine for two words. Four word co-occurrence measures were computed using page counts. We proposed a lexical pattern extraction algorithm to extract numerous semantic relations that exist between two words. Moreover, a sequential pattern clustering algorithm was proposed to identify different lexical patterns that describe the same semantic relation. Both page counts-based co-occurrence measures and lexical pattern clusters were used to define features for a word pair. A two-class SVM was trained using those features extracted for synonymous and non synonymous word pairs selected from Word Net synsets. Experimental results on three benchmark data sets showed that the proposed method outperforms various baselines as well as previously proposed web-based semantic similarity measures, achieving a high correlation with human ratings. Moreover, the proposed method improved the F-score in a community mining

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.

