Volume 3, Issue 4, April 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Web Search Engine Using Semantic Approach

Ms. N.Subathra M.E.,[1]
CSE, PSNACET, India.

Mr. M.S.Thanabal M.E., (PhD) [2]
CSE, PSNACET, 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 remains a challenging task. We propose an 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. Identifying semantic relations that lies between two given words, we propose a 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 method that is proposed outperforms previously proposed web-based semantic similarity measures on three benchmark data sets shows a high correlation.

Keywords -- community mining, pattern clustering, text snippets, semantic, metadata.

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 WordNet; 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 co-occurrence 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. For those reasons, page counts alone are unreliable when measuring semantic similarity. Snippets, a brief window of text extracted by a search engine around the query term in a document, provide useful information regarding the local context of the query term. Semantic similarity measures defined over snippets have been used in query expansion, personal name disambiguation, and community mining. Processing snippets is also efficient because it obviates the trouble of downloading web pages, which might be time consuming depending on the size of the pages. However, a widely acknowledged drawback of using snippets is that, because of the huge scale of the web and the large number of documents in

the result set, only those snippets for the to pranking results for a query can be processed efficiently. Ranking of search results, hence snippets is determined by a complex combination of various factors unique to the underlying search engine. Therefore, no guarantee exists that all the information we need to measure semantic similarity between a given pair of words is contained in the top-ranking snippets. We propose a method that considers both page counts and lexical syntactic patterns extracted from snippets that we show experimentally to overcome the above mentioned problems. The phrase is the largest indicates a hypernymic relationship between Jaguar and cat. Phrases such as also known as, is a, part of, is an example of all indicate various semantic relations. Such indicative phrases have been applied to numerous tasks with good results, such as hypernym extraction and fact extraction. From the previous example, we form the pattern X is the largest Y, where we replace the two words Jaguar and cat by two variables X and Y. Our contributions are summarized as follows:

- We present an automatically extracted approach lexical syntactic patterns that computes the semantic similarity of two words or entities using text snippets retrieved from a web search engine. We propose a lexical pattern extraction algorithm that considers word subsequences in text snippets. Moreover, the extracted sets of patterns are clustered to identify the different patterns that describe the same semantic relation.
- We integrate different web-based similarity measures using a machine learning approach. We extract synonymous word pairs from Word Net synsets as positive training instances and automatically generate negative training instances. We then train a two-class support vector machine (SVM) to classify synonymous and non synonymous word pairs. The integrated measure outperforms all existing web based semantic similarity measures on a benchmark data set.
- We apply the proposed semantic similarity measure to identify relations between entities, in particular people, in a community extraction task. In this experiment, the proposed method outperforms the baselines with statistically significant precision and recall values. The results of the community mining task show the ability of the proposed method to measure the semantic similarity between not only words, but also between named entities, for which manually created lexical ontologies do not exist or incomplete.

II. Related Work

Given taxonomy of words, a straightforward method to calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy. If a word is polysemous, then multiple paths might exist between the two words. In such cases, only the shortest path between any two senses of the words is considered for calculating similarity. A problem that is frequently acknowledged with this approach is that it relies on the notion that all links in the taxonomy represent a uniform distance. Resnik proposed a similarity measure using information content. He defined the similarity between two concepts C1 and C2 in the taxonomy as the maximum of the information content of all concepts C that subsume both C1 and C2. Then, the similarity between two words is defined as the maximum of the similarity between any concepts that the words belong to. He used WordNet as the taxonomy; information content is calculated using the Brown corpus. Semantic similarity measures have been used in various applications in natural language processing such as word sense disambiguation, language modeling, synonym extraction, and automatic thesauri extraction. Semantic similarity measures are important in many web related tasks. In query expansion, a user query is modified using synonymous words to improve the relevancy of the Search. One method to find appropriate words to include in a query is to compare the previous user queries using semantic similarity measures. If there exists a previous query that is semantically related to the current query, then it can be either suggested to the user, or internally used by the search engine to modify the original query.

III. Method Outline

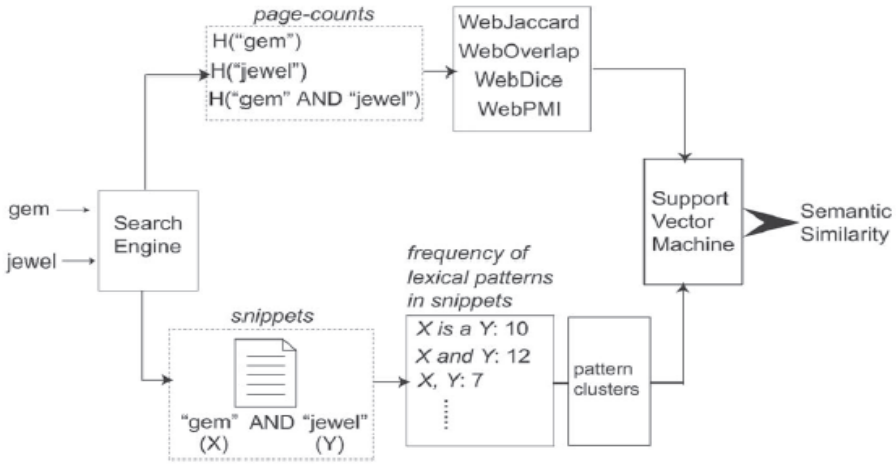

FIG 1 : OUTLINE OF METHOD

Fig illustrates an example of using the proposed method to compute the semantic similarity between two words, gem and jewel. First, we query a web search engine and retrieve page counts for the two words and for their conjunctive (i.e., "gem," "jewel," and "gem AND jewel"). we define four similarity scores using page counts. Page counts-based similarity scores consider the global co-occurrences of two words on the web. However, they do not consider the local context in which two words co occur. On the other hand, snippets returned by a search engine represent the local context in which two words co-occur on the web. Consequently, we find the frequency of numerous lexical syntactic patterns in snippets returned for the conjunctive query of the two words. It is noteworthy that a semantic relation can be expressed using more than one lexical pattern. Grouping the different lexical patterns that convey the same semantic relation, enables us to represent a semantic relation between two words accurately. For this purpose, we propose a sequential pattern clustering algorithm in Section 3.4. Both page counts-based similarity scores and lexical pattern clusters are used to define various features that represent the relation between two words.

IV. PAGECOUNT-BASEDCO-CURRENCE MEASURES

Page counts for the query P AND Q can be considered as an approximation of co-occurrence of two words (or multiword phrases) P and Q on the web. However, page counts for the query P AND Q alone do not accurately express semantic similarity. For example, Google returns 11,300,000 as the page count for "car" AND automobile," whereas the same is 49,000,000 for "car" AND "apple Outline of the proposed method. Automobile is more semantically similar to car than apple is, page counts for the query "car" AND "apple" is more than four times greater than those for the query "car" AND "automobile." One must consider the page counts not just for the query P AND Q, but also for the individual words P and Q to assess semantic similarity between P and Q. We compute four popular co-occurrence measures; Jaccard, Overlap (Simpson), Dice, and Point wise mutual information (PMI), to compute semantic similarity using page counts.

V. LEXICAL PATTERN EXTRACTION

Page counts-based co-occurrence measures described do not consider the local context in which those words co-occur. 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. The phrase indicates a semantic relationship between cricket and sport. Many such phrases indicate semantic relationships. For example, also known as, is a, part of, is an example of all indicate semantic relations of different types. In the example given above, words indicating the semantic relation between cricket and sport appear between the query words. Replacing the query words by variables X and Y , we can form the pattern X is a Y from the example given above. Despite the efficiency of using snippets, they pose two main challenges: first, a snippet can be a fragmented sentence; second, a search engine might produce a snippet by selecting multiple text fragments from different portions in a document. Because most syntactic or dependency parsers assume complete sentences as the input, deep parsing of snippets produces incorrect results. Consequently, we propose a shallow lexical pattern extraction algorithm using web snippets, to recognize the semantic relations that exist between two words. Lexical syntactic patterns have been used in various natural language processing tasks such as extracting hypernyms, or meronyms, question answering, and paraphrase extraction. Although a search engine might produce a snippet by selecting multiple text fragments from different portions in a document, a predefined delimiter is used to separate the different fragments. For example, in Google, the delimiter "..." is used to separate different fragments

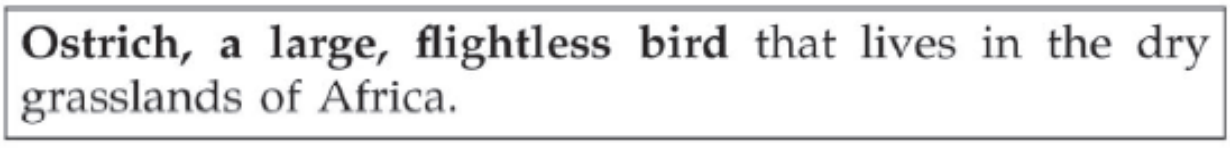

FIG 2: A snippet retrieved for the query "ostrich……bird"

VI. 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 an 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. We represent a pattern a by a vector a of word-pair frequencies.

VII. 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 section, 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.

VIII. CONCLUSION

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 WordNet 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 task, thereby underlining its usefulness in real-world tasks that include named entities not adequately covered by manually created resources.

References

[1] A. Kilgarriff, "Googleology Is Bad Science," Computational Linguistics, vol. 33, pp. 147-151, 2010.
[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., 2009.
[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, 2009.
[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, 2009.
[5] M. Hearst, "Automatic Acquisition of Hyponyms from Large Text Corpora," Proc. 14th Conf. Computational Linguistics (COLING), pp. 539-545, 2010.
[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), 2009.
[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. 2008.
[8] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proc. 14th Int'l Joint Conf. Aritificial Intelligence, 2007.
[9] 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. 2008.
[10] G. Miller and W. Charles, "Contextual Correlates of Semantic Similarity," Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 2008

