Measuring Semantic Similarity between Words using Page-Count and Pattern Clustering Methods

Prathvi Kumari, Ravishankar K

Abstract*— Web mining involves activities such as document clustering, community mining etc. to be performed on web. Such tasks need measuring semantic similarity between words. This helps in performing web mining activities easily in many applications. Despite the usefulness of semantic similarity measures in these applications, accurately measuring semantic similarity between two words remains a challenging task. In this paper to find the semantic similarity between two words it makes use of information available on the web and uses methods that make use of page counts and snippets to measure semantic similarity between two words. Various word co-occurrence measures are defined using page counts and then integrate those with lexical patterns extracted from text snippets. To identify the numerous semantic relations that exist between two given words, a pattern extraction and clustering methods are used. The optimal combination of page counts-based co-occurrence measures and lexical pattern clusters is learned using support vector machine used to find semantic similarity between two words. Finally semantic similarity measure what is got is in the range [0, 1], is used to determine semantic similarity between two given words. If two given words are highly similar it is expected to be closer to 1, if two given words are not semantically similar then it is expected to be closer to 0.

***Index Terms*— Natural Language Processing, Semantic Similarity, Support Vector Machine, Text Snippets, Web Mining**.

I. INTRODUCTION

Information available on the web considered as vast, hidden network of classes of objects, are interconnected by various semantic relations. Semantics identify concepts which allow extraction of information from data. For a machine to be able to decide the semantic similarity, intelligence is needed. It should be able to understand the semantics or meaning of the words. But a computer being a syntactic machine, semantics associated with the words or terms is to be represented as syntax. The measurement of semantic similarity between words remains a challenging task in many Natural Language Processing tasks and information retrieval such as Query Expansion, Query suggestion, Word Sense Disambiguation etc.Semantic similarity measures are successfully employed in various natural language tasks such as word sense disambiguation, language modeling, and synonym extraction. The semantic similarity between entities changes over time and domain.

Manually compiled taxonomies such as WordNet [1] and text corpora have been used in previous work on semantic similarity [2, 3, 4, and 5]. However, Semantic Similarity between two words is a dynamic phenomenon that varies over time and across domains.

One major issue behind taxonomies and corpora oriented approaches is that they might not necessarily capture similarity between proper names such as named entities (e.g., personal names, location names, product names) and the new uses of existing words. For example, blackberry is frequently associated with phones on the Web. However, this sense of blackberry is not listed in most general-purpose thesauri or dictionaries. A user, who searches for blackberry on the Web, may be interested in this sense of blackberry and not blackberry as a fruit. New words are constantly being created as well as new senses are assigned to existing words. Manually maintaining thesauri to capture these new words and senses is costly if not impossible.

To overcome the drawbacks mentioned above, this paper proposes a method that finds semantic similarity between two words based on the page counts and text snippets retrieved from web search engines like Google. Because of the various documents and the wide growth rate of the Web, it is difficult to analyze each document separately. Web search engines provide a vital interface to this vast information. Page counts and snippets are two useful knowledge sources provided by most Web search engines. Page count of a query is the no. of pages which contain the query words. Snippet is some text extracted by web search engine based on the query term given. And then proposes methods such as lexical pattern extraction and pattern clustering to accurately measure semantic similarity between two words.

II. RELATED WORKS

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.

Words and phrases acquire meaning from the way they are used in society, from their relative semantics to other words and phrases. For computers, the equivalent of "society" is "database,"

Manuscript Received on July 2013.

Prathvi Kumari, (Mtech), Computer Science Engineering, SIT, Mangalore.

Prof. Ravishankar K, Associate Professor, Computer Science Engineering, SIT, Mangalore.

and the equivalent of "use" is "a way to search the database". A new theory of similarity between words and phrases is presented based on information distance and Kolmogorov complexity. To fix thoughts, the World Wide Web (WWW) is used as the database, and Google as the search engine. The method is also applicable to other search engines and databases. This theory is then applied to construct a method to automatically extract similarity, the Google similarity distance, of words and phrases from the WWW using Google page counts. The WWW is the largest database on earth, and the context information entered by millions of independent users averages out to provide automatic semantics of useful quality [6].

Multiple information sources explore the determination of semantic similarity by a number of information sources, which consist of structural semantic information from a lexical taxonomy and information content from a corpus. To investigate how information sources could be used effectively, a variety of strategies for using various possible information sources are implemented. The knowledge bases may be constructed in a hierarchy that is common place in the world. A new measure is then proposed which combines information sources nonlinearly [7].

III. PROPOSED METHOD

The proposed method that finds similarity between two words P, Q is supposed to return a value between 0.0 and 1.0. The proposed method makes use of page counts and text snippets retrieved by search engine like Google. If P and Q are highly similar, we expect semantic similarity value to be closer to 1, otherwise semantic similarity value to be closer to 0.

The proposed system first downloads few WebPages from Google and stores it in the database before the proposed system methods are applied to it. Semantic Similarity found here depends on the downloaded WebPages and the methods applied here.

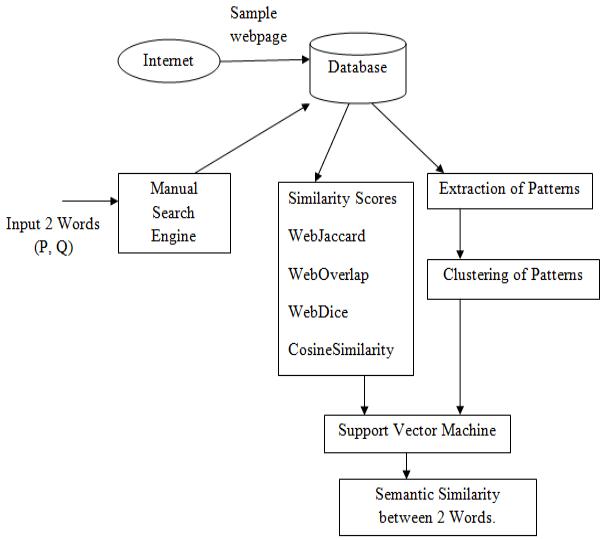

Fig. 1: Architecture of the Proposed System

Figure 1 illustrates an example of using the proposed method to compute the semantic similarity between two words for example consider “cricket” and “sport”. First, a web search engine is queried and retrieves page counts for the two words and for their conjunctive (i.e., “cricket,” “sport”, and “cricket AND sport”). Four popular similarity scores using page counts are used here. 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 cooccur.On the other hand, snippets returned by a search engine represent the local context in which two words co-occur on the web.

A .Page-count-based Similarity Scores

The notation H (P) is used to denote the page count for query P in a search engine. Traditional Jaccard, Overlap, Dice and Cosine Similarity measures are modified for the purpose of measuring similarity using page counts.

WebJaccard coefficient between words P and Q, WebJaccard (P, Q), is defined by,

$$\text{WebJaccard}(P,Q) = \begin{cases} 0, & \text{if } H(P \wedge Q) \le C \\ \dfrac{H(P \wedge Q)}{H(P)+H(Q)-H(P \wedge Q)} & \text{Otherwise} \end{cases} \quad (1)$$

Likewise, define WebOverlap coefficient, WebOverlap (P, Q), as,

$$\text{WebOverlap}(P,Q) = \begin{cases} 0 & \text{if } H(P \wedge Q \le C \\ \dfrac{H(P \wedge Q)}{\text{Min}(H(P), H(Q))} & \text{Otherwise} \end{cases} \quad (2)$$

Defines WebDice as a variant of Dice coefficient. WebDice (P, Q) is defined as,

$$\text{WebDice}(P,Q) = \begin{cases} 0, & \text{if } H(P \wedge Q) \le C \\ \dfrac{2H(P \wedge Q)}{H(P)+H(Q)} & \text{Otherwise} \end{cases} \quad (3)$$

Cosine Similarity is defined as CS (P, Q),

$$CS(P,Q) = \begin{cases} 0, & \text{if } H(P \wedge Q) \le C \\ \dfrac{H(P \wedge Q)}{\text{Sqrt}(H(P)) \times \text{Sqrt}(H(Q))} & \text{Otherwise} \end{cases} \quad (4)$$

Here, P∧Q denotes the conjunction query P AND Q. The notation H (Q) is used to denote the page count for query *Q* in a search engine.

Given the scale and noise in web data, it is a possible that two words may appear on some pages accidentally. In order to reduce the adverse effect due to random co-occurrences, The 4

co-occurrences are set to zero if the page counts for the query P ∩ Q are less than a threshold *c* (c is assumed to be 5).

B. Extracting Lexical Patterns from Snippets

Page counts based similarity measures do not consider the relative distance between *P* and *Q* in a page or the length of the page. Although *P* and *Q* may occur in a page they might not be related at all. Therefore, page counts based similarity measures are prone to noise and are not reliable when H(P∩Q) is low. On the other hand snippets capture the local context of query words. Snippet contains a window of text selected from a document that includes the queried words. For example, consider the snippet in Fig. 2. Here, the phrase is a 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. Thus such lexico-syntactic patterns extracted from snippets are a solution to the problems with page counts based similarity measures.

"**Cricket** is a **sport** played between two teams, each with eleven players"

Fig.2: Snippet retrieved for the query "Cricket and sport"

Given two words P and Q, we query a web search engine using the wildcard query "P***** Q" and download snippets. The "*" operator matches one word or none in a webpage. Therefore, our wildcard query retrieves snippets in which P and Q appear within a window of seven words. 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. It is a pattern extracted from snippet. The extracted lexical patterns are then clustered based on the similarity with respect to given cluster. 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.

Page counts along with the similarity scores such as WebJaccard, WebDice, WebOverlap and Cosine Similarity are calculated .Meanwhile the database also retrieves the snippets for the conjunctive query. Since the snippets contain the summary of contents and url of different web pages, the user can identify the relevant information for the requested query. Relevant patterns are identified for the requested query using a pattern extraction method. The patterns showing similar semantic relations are clustered together using pattern clustering method. Lexical patterns thus extracted are clustered together and given to SVM. The SVM acts up on both results of word co-occurrence measures and also pattern clusters in order to calculate semantic similarity between two given words. A machine learning approach, support vector machine is used to combine both page counts-based co-occurrence measures, and snippets-based lexical pattern clusters to construct an accurate semantic similarity measure.

To analyse this proposed system implemented here is an offline search engine with few WebPages are downloaded from the internet and kept in the database and with the downloaded WebPages and the methodology used here, the proposed system is tested. Table 1 shows the resultant 4 co-occurrence measures and semantic measure for the few word pair. These few word pair is taken from Miller Charles dataset [8] and the result for it are shown in the table below.

Table 1. Different measures for few words are shown

Word 1	Word 2	WebJaccard	WebOverlap	WebDice	CosineSimilarity	Semantic Measure
Gem	Jewel	0.14	0.27	0.24	0.24	0.15
Food	Fruit	0.08	0.17	0.14	0.14	0.09
Came	Came	0.03	0.08	0.06	0.06	1.00
Gem	Gem	0.09	0.07	0.16	0.07	1.00
Noon	String	0.03	0.10	0.07	0.03	0.04
Bird	Crane	0.01	0.05	0.03	0.03	0.02

Figure 3 shows the Graphical View of the Semantic Measure got from the table 1.

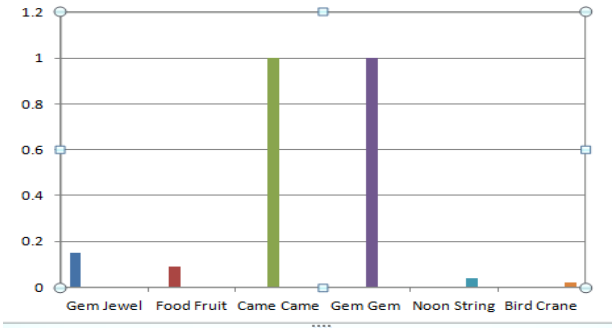

Fig. 3: Graphical View of the Semantic Measure

In general we can say that, the existing system when the same query is submitted by the different users, a search engine returns the same results, regardless of the submitted query. This may not be suitable for users to get their needed information. Semantic Search in general is more suitable to display the results based on the user interest. Figure 4 shows that the semantic search provides more results than the direct search. In the direct search the results are displayed based on the submitted query. But in the semantic search it considers the meaning for that submitted query and it displays the related results for that query. For example in figure 4 shows that the direct search finds out 10 results for the query apple. But the semantic search displays 12 queries for the same query. Thus the semantic search shows more results than the direct search [9].

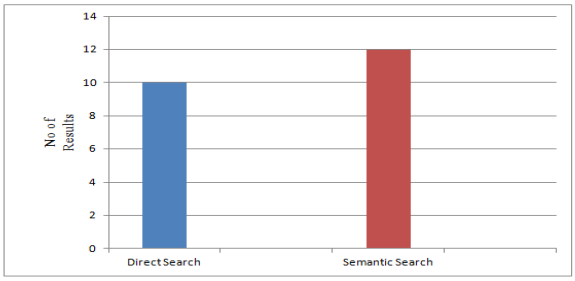

Fig. 4: Comparision of results between the Direct Search and the Semantic Search

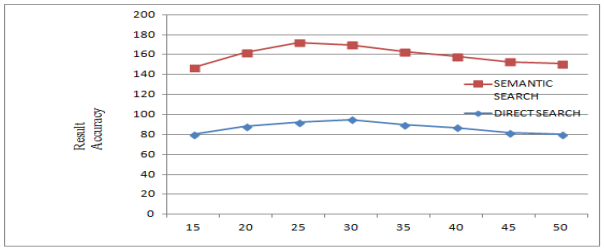

Fig. 5: Accuracy chart for the Direct Search and the Semantic Search

Figure 5 shows the accurate results for the direct search and the semantic search. In the semantic search it provides more accuracy than the direct search and it is represented in the chart. Thus the semantic search based on user personalization is giving efficient results based on user interest at minimum processing time with high accuracy of results.

IV. CONCLUSION

Sample WebPages are first downloaded from internet and proposes a semantic similarity measure which is based on the page counts and text snippets .The aim of this paper is to measure semantic similarity between two given words with utmost accuracy. In this paper, a measure is proposed that uses snippets to strongly calculate semantic similarity between two given words. To achieve these, techniques like pattern extraction from the snippets and pattern clustering are introduced and also consist of four page-count-based similarity scores. These methods help in finding various relationships between words. Then both page- count based co-occurrence measures and pattern clusters are integrated using support vector machines to define semantic score for a word pair.In future, the proposed system can make use of artificial intelligence with expert systems involved and the proposed similarity measure is used for automatic synonym extraction, query suggestion and name alias recognition. Also it is intended to apply the lexical patterns extracted by this proposed system to find synonyms from the web.

REFERENCES

1. George A. Miller , “WordNet: A Lexical Database for English”.
2. D. Lin. Automatic retreival and clustering of similar words.In Proc. of the 17th COLING, pages 768–774 1998.
3. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proc. of 14th Internation JoinT Conference on Aritificial Intelligence, 1995.
4. J.J. Jiang and D.W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In Proc. of the International Conference on Research in Computational Linguistics ROCLING1998.
5. D. Lin. An information-theoretic definition of similarity. In Proc. Of the 15th ICML, pages 296–304, 1998.
6. D. Mclean, Y. Li, and Z.A. Bandar, “An Approach for Measuring SemanticSimilarity between Words Using Multiple Information Sources,” July/Aug. 2003.
7. R. Cilibrasi and P. Vitanyi, “The Google Similarity Distance,” IEEE Trans. Knowledge and Data Eng., vol.19, no. 3, pp. 370-383, Mar.2007.
8. G. Miller and W. Charles, “Contextual Correlates of Semantic Similarity,” Language and Cognitive Processes, vol. 6, no. 1, pp. 1-28, 1998.
9. V.Hemalatha and Mrs .K. Sarojini, “semantic similarity approach using rsvm based on personalized search in web search engine”,vol 1,November 2012