Semantic Similarity Measurement Between Words using Swd & Snippets

R.Menaha
Asst Prof, IT Department
Dr.MCET, Pollachi, Coimbatore

G.Anupriya
Asst Prof (SG), CSE Department
Dr.MCET, Pollachi, Coimbatore

ABSTRACT

Semantic similarity plays a significant role in the areas of Web mining, Information Retrieval, NLP and Text mining. Even though it is exploited in various applications accurately measuring semantic similarity still remains a challenging task. In this paper a method is proposed to measure semantic similarity between words using web as information source and by combining two existing approaches to measure semantic similarity they are: Semantic Word Distance (SWD) and Snippets. The SWD measure finds the semantic similarity by determining the frequency of occurrences of the words in web pages (corpus).The semantic relation between words are also obtained through lexical patterns which are extracted from text snippets. A robust method is used to integrate these similarity scores using support vector machine. For the experimental purpose 100 word pairs are used to train the support vector machine and it classifies the word pair as either synonymous or non synonymous with higher accuracy.

Keywords

Page count measures, semantic similarity, semantic word distance, snippet.

1. INTRODUCTION

Measuring semantic similarity between words is an important task in the fields of web mining, information retrieval, text mining, semantic web and natural language processing. In the field of web mining, semantic similarity is exploited in applications like relation detection, community extraction, and automatic data extraction. In the area of Information Retrieval, semantic similarity is used in query expansion, and query suggestion system. Some of the natural language applications like word sense disambiguation, and language modelling require finding semantic similarity between words accurately. In Text Mining the semantic similarity can be used in text summarization. Despite it’s usefulness in various applications the task of accurately measuring semantic similarity remains a challenging task.

Semantic relation between two terms means how much two terms are related even if they are dissimilar in meaning [1]. For example the words “gem” and “jewel” are semantically related than “car” and “jewel”. Semantically related words of a particular word are listed in manually created general purpose lexical ontologies called WordNet. For example, Google is frequently associated with search on the web. However, this is not listed in most general-purpose WordNet or dictionaries. 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 a difficult process [2].

An empirical method is suggested to estimate the semantic similarity between the words using web as information source. The two approaches used for measuring semantic similarity are SWD and Snippets. SWD measures the frequency of the words in each document and normalizes it over all documents. The page count measure can also be used to find semantic similarity but it does not indicate the number of times a word has occurred in each of this page. A word may appear many times in a document and once in another document, but page count measure will ignore this. So the page count measure is not sufficient to measure the semantic relation between two words [1].

SWD considers only the global context of given words in web pages and it doesn’t give importance to the semantic relationship that exists between the word pairs. Therefore snippets are used for finding semantic similarity in local context [2]. Snippet is a brief window of text extracted by a search engine around the query term in a document. A snippet for the query car AND automobile is given in fig 1. Here “**is a”** indicates semantic relationship between the car and automobile. Many such phrases indicate semantic relationships. For example, ‘also known as’, ‘is a part of’, ‘is an example of’ indicate various semantic relations of different types.

> *“A**n*** **automobile** ***or*** **car** ***is a wheeled vehicle that carries its own motor and transports passengers”.***

Fig 1. Snippet for the query car AND automobile.

Processing snippets is also efficient because it ignores the trouble of downloading WebPages, which might be time consuming depending on the size of the pages [2]. 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 top ranking results for a query can be processed efficiently.

In the proposed system, a method is used, that integrates both SWD and Snippets to determine semantic relation between two words using the web as information source. The optimal combination of SWD score and Snippets score is learned through support vector machine. And the proposed system significantly improves the accuracy than existing one.

2. RELATED WORK

Different approaches have been followed in the past decade to measure the semantic similarity and they are broadly categorized into two ways: Distance based approach & Corpus based approach.

Distance based approaches measure the semantic similarity between two words using the distance defined in lexicon or knowledge base. Some instances of this kind of approach are published by Rada, Leacock and Chodorrow, Yang and Powers. A distance based approach maps the semantic similarity between two words by a formula defined as follows [4]:

$$\text{Sim}\,(w1, w2) = \Phi\,(\text{dist}\,(w1, w2)) \qquad \text{- (1)}$$

Where dist (.) returns distance between w1 and w2 and Φ (.) is function that transforms distance to similarity defined with various considerations.

Cilibrasi and Vitnayi have proposed a distance metric (distance based approach) between words using only page counts retrieved from a web search engine [3]. The proposed metric is named as Normalized Google Distance (NGD) and is given by

$$\text{NGD}\,(P, Q) = \frac{\max\{\log H(P), \log H(Q)\} - \log H(P,Q}{\log N - \min\{\log H(P), \log H(Q)\}} \qquad \text{- (2)}$$

The corpus auxiliary approaches measure word similarity by considering not only lexical information and also auxiliary information such as word co-occurrence. According to the published results, corpus auxiliary approaches outperform distance based approaches in some degree while are more complex [4].

Takale, S.A. and Nandgaonkar, S.A (2010) have used five different semantic similarity measures [9]. This method understands the semantics associated with the word by making use of snippets returned by Wikipedia for the given word pair.Liu, B., Dai, L. Xia, Y. and Wu, S. (2008) measured semantic similarity between words using HowNet as an information source [5]. Similarly Pedersen, T., Patwardhan, S. and Michelizzi, J. (2004) measured the semantic relatedness between concepts using WordNet as an information source [7].

Basu & Murthy [2008] have proposed a new measure called SWD for measuring the semantic similarity between words [1] and it is a corpus based approach. They have used WebKb Dataset as an information source. Bollegala, D., Matsuo, Y. and Ishizuka, M [2011], proposed a method to measure the semantic similarity between words using web as information resource. This method integrates the page count & snippets to measure the semantic similarity [2].

The proposed method is a corpus based approach and it uses extracted web pages (corpus) from search engine to find the semantic similarity between words.

3. METHOD

3.1 Outline

The system is designed to measure semantic similarity between words using web search engine. The similarity between the words P and Q is expressed through N+ 1 features which are extracted from Semantic Word Distance measure and N Pattern clusters which are obtained through the snippets retrieved from the web search engine. Using this feature representation of words, a two class SVM is trained and it returns the value 0 for the synonymous word pairs or 1 for the non synonymous word pairs.

Fig .2 illustrates the system design for any two word pairs which are denoted as "P" and "Q". First the search engine is queried for the web pages having "P","Q", and conjunctive query "P" and "Q". Then the system finds the word frequency of the given word pairs in all the extracted web pages (i.e., Wr (P, Q) and Wr (Q, P)) after which it finds the semantic similarity using SWD measure. For comparing the performance with the existing approach, the page counts for "P", "Q" and "P" and "Q" are also extracted from the web pages and the extracted page counts were used in four page count measures namely Web jaccard, Web dice, Web Overlap, Web PMI to measure similarity of the words.

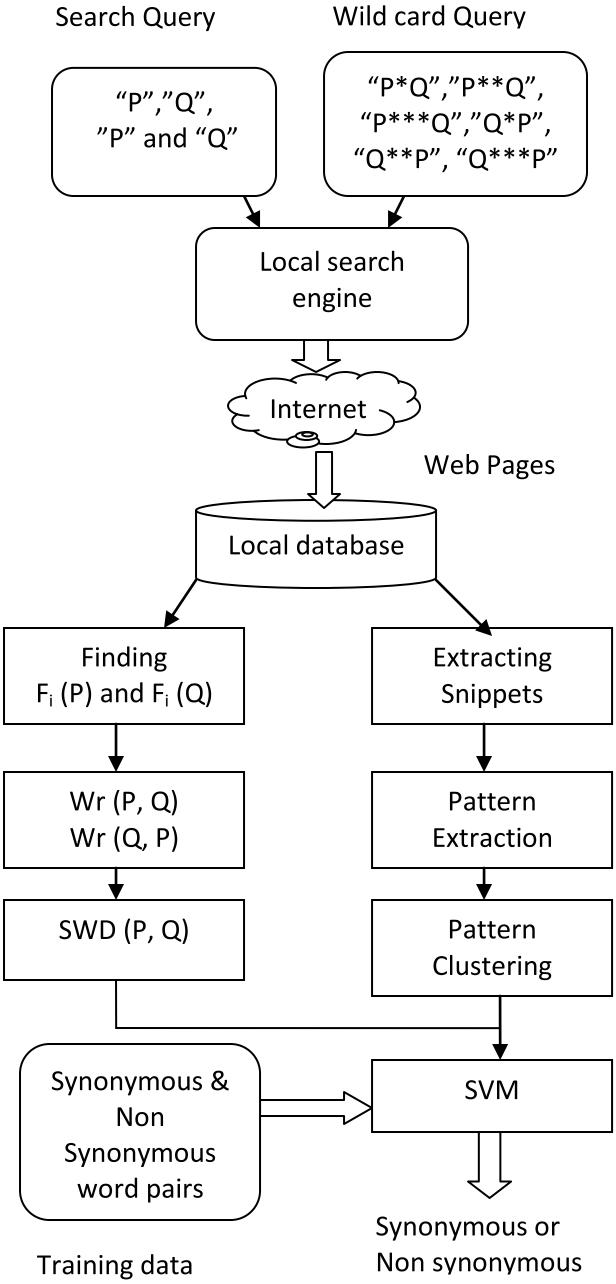

Fig.2. Overall System Block Diagram

Another part of finding similarity, is extracting snippets for word pairs by issuing wildcard query to the search engine. The wildcard query for the given word pairs is built by (P*Q, P**Q, P***Q, Q*P, Q**P, Q***P) and the snippets are extracted from the web pages returned by the search engine. The system extracts the patterns from the snippets by using pattern extraction algorithm, and then it finds the frequency of numerous lexical patterns in the snippets. The pattern which has same semantic relation is formed as cluster using the procedure which is given in section 3.3.2.From the N clusters N Features are Obtained.

Both the SWD Measure and Snippets express N+1 features that represent word pairs, by using these features a two class SVM is trained. Finally the SVM classifies whether a given word pair is synonymous or non synonymous and also the performance of existing method and proposed method is compared in terms of accuracy.

3.2 SWD Measure

The page count measures are not sufficient to measure semantic relation between two words because it provides the number of pages in which a word occurs [1]. It does not indicate the number of times the word has occurred in web pages. A word may appear many times in a document or once in a document but page count simply ignores this. Given two words P and Q, the semantic relation between them must be found on the basis of a corpus (web pages).Therefore the SWD measure is suggested here to measure semantic similarity between words because it takes into account the frequency of occurrences of the word in web pages [1].

SWD(x, y) is given by Eqn - (3)

$$= \begin{cases} -1 & if \quad fi(x) = fi(y) = 0 \ \forall i \in M \\ \left| \dfrac{wr(x,y) + wr(y,x)}{2} - 1 \right| & Otherwise \end{cases}$$

Where M is the total number of pages from which the relation between the words will be found.

Word ratio is defined as

$$\text{Wr (x, y)} = \frac{1}{M} \sum_{i=1}^{M} \max\ (I_i(x), I_i(y)) \quad \frac{f_i(x)+1}{f_i(y)+1} \qquad \text{- (4)}$$

Where Ii(X) is given by

$$\text{I}_\text{i}\ \text{(x)} = \begin{Bmatrix} 0 & if \ fi(x) = 0 \\ 1 & otherwise \end{Bmatrix} \qquad \text{- (5)}$$

- $f_i(x)$ is the number of times the word x occurs in i^{th} page.
- If $f_i(x)$ and $f_i(y)$ =0 then that two words have strongest semantic relation.
- Two words are dissimilar when the SWD value is greater than 2.

The lower limit of SWD is -1 which indicates, that the frequency of the two words are none in web pages. If the SWD value is 0 then it indicates that the semantic relation between the words is strongest. The relationship between the word pairs decreases as long as the value grows from 0[1].

3.3 EXTRACTION OF SNIPPETS & PATTERN CLUSTERING

Snippets are a brief window of text extracted by a search engine around the query term in a document. They 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 [2].

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. They provide valuable information regarding the local context of a word. Lexico syntactic patterns that indicate various aspects of semantic similarity can be extracted. For example, consider the following text snippet returned by Google for the query *"apple" AND "fruit"*.

"Apple is a Pome **fruit".**

Fig. 3. Snippet for the Query Apple and Fruit.

Here, the phrase '***is a'*** indicates a semantic relationship between apple and fruit. 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 *apple* and *fruit* appear between the query words. Replacing the query words by wildcards *X* and *Y* the pattern *X is a Pome Y* can be formed from the example given above. However, in some cases the words that indicate the semantic relationship do not fall between the query words.

3.3.1 Pattern Extraction

For the given word pairs P and Q, wildcard query is framed like P*Q, P**Q, P***Q, Q*P, Q**P, Q***P and these queries are searched in web search engine. The "*" operator matches one word or none in a webpage. Therefore, wildcard query retrieves snippets in which P and Q appear within a window size of maximum of seven words. To extract the patterns from the snippets the following procedure is followed [2].

A. For each snippet retrieved for wild card queries from search engine,

- First the words P and Q are replaced by the Variables X and Y
- All the numeric values are replaced as D.

B. Sub sequences are generated from the snippet by satisfying the following conditions [2].

- A subsequence must contain exactly one occurrence each of X and Y.
- The maximum length of a subsequence is L words.
- A subsequence is allowed to skip one or more words. However, the system does not skip more than g number of words consecutively.
- The system expands all negation contractions in a context. For example, can't is expanded to can not. While skipping the word the system should not skip the word 'not' when generating sub sequences
- Remove the duplicate snippets.

Finally, the system counts the frequency of all generated sub sequences and only uses sub sequences that occur more than 2 times as lexical patterns in experiments.

3.3.2 Pattern Clustering

A semantic Relation can be expressed by using more than one pattern [2]. For example, by considering the following two patterns, X is a Y, and X is a small Y, both the patterns indicates that there exists an 'is a' relation between X and Y. Identifying the different patterns that express the same relation enables to represent the relation between two words accurately. The patterns which express same semantic relation are grouped in a same cluster based on their frequency.

For the given a set of patterns, the clustering of patterns is done by using the following procedure [2].

i) Sort all the patterns based on their frequency. The frequency of a pattern is measured by using Eqn (6). After sorting, the rare patterns moved to the end.

$$\mu(a) = \sum f(P_i, Q_{i,} a) \qquad - (6)$$

Where $\mu(a)$ is the total frequency of pattern **a** with the word pairs P, Q.

ii) The patterns that express same semantic relation are identified and formed as cluster.

By sorting the patterns in descending order based on their frequency and clustering those first results in clusters with more common relation. This enables the rare patterns to move to an end. Here hard clustering is performed, so a pattern can belong to only one cluster.

3.3.3 Training SVM

Support Vector Machine is used as a classifier to classify the given words as synonymous or non synonymous word pairs. LIBSVM is currently one of the most widely used SVM software for support vector classification, regression and distribution estimation and it also supports multi-class classification. It has been used here for classifying the word pairs.

The Word pairs (50 Synonymous and 50 Non synonymous) are identified from WordNet. Once the SVM is trained with synonymous and non synonymous word pairs, SVM will be used to classify new word pairs either as synonymous or non synonymous. And also accuracy is measured to compare the performance of Existing and Developed system.

3.3.4 Measuring Semantic Similarity

A pair of words is represented through (N+1) feature vector. Here N represents number of cluster features and 1 represents the SWD feature. The cluster feature is computed as follows:

$$w_{ij} = \frac{\mu(a_i)}{\sum_{t \in c_j} \mu(t)} \qquad - (7)$$

Here $\mu(a_i)$ is the total frequency of pattern **a** in all the word pairs , C_j is the cluster j and W_{ij} is the weighted sum of all patterns in cluster C_j. Eventually the feature of j^{th} cluster is computed as follows.

$$\sum_{a_i \in c_j} w_{ij}\, f(P, Q, a_i) \qquad - (8)$$

The j^{th} feature value is given in Eqn (8) expresses the significance of the semantic relation represented by cluster j for word pair (P, Q).

N+1 features are extracted for each word pair selected from the dataset, and by using these features SVM is trained to classify whether a given word pair is synonymous or non synonymous word pair. The SVM classification process was carried out by using both existing system features consisting of (N+4) features and the proposed system consisting of (N+1) features and the accuracy values are compared.

Table 1: Semantic Similarity Scores of Selected Word pairs

Word pair	**Web jaccard**	**Web Dice**	**Web Overlap**	**Web PMI**	**SWD**	**Cluster Score**
Car-automobile	0.021	0.041	0.771	0.456	0.512	0.6
Gem-jewel	0.137	0.241	0.247	0.352	0.815	0.5
Journey-voyage	0.077	0.143	0.282	0.557	1.0	0.4
Tool-implement	0.028	0.056	0.379	0.488	0.253	0.7
Midday-noon	0.024	0.047	0.164	0.585	0.975	0.1
Food-fruit	0.095	0.173	0.660	0.352	0.975	0.3
Lad-brother	0.561	0.719	0.632	0.101	0.871	0.5
Shore-woodland	0.040	0.077	0.192	0.511	0.916	0.4
Magician-wizard	0.026	0.051	0.108	0.603	1.0	0.5
Food-rooster	0004	0.009	0.305	0.675	0.975	0.4

4. RESULTS

4.1 Data sets

The word pairs are identified by using the WordNet and the feature vector for each word pair is computed to train SVM. The feature vector is computed by using the following procedure.

For each word pair, named as (P, Q)

- The web pages for the query "P","Q", and "P and Q" are extracted and stored in local database.
- Wr (P, Q) is found out and it is applied in SWD. This is one of the feature vector obtained from SWD.
- The pattern clusters are formed from the patterns that are identified from the snippets. The cluster feature is obtained by using the Eqn (8) for each cluster.
- By using the above method the features for 100 word pairs are obtained through SWD and snippet, and the SVM is trained using these feature vectors.

4.2 Evaluation

Here Google is used as a search engine to extract the web pages for given word pairs. Table 1 illustrates the semantic similarity scores of SWD with other four measures namely web jaccard, Web dice, Web PMI, and Web Overlap. The cluster scores of the word pairs are measured and the SVM is trained to classify either the given word pair as synonyms or nonsynonyms word pair. Eventually the accuracy of the developed system and existing system is compared and illustrated in Table 2.

Table 2: Performance Evaluation

Method	Accuracy
SWD & Snippets	95 %
Page count measures & Snippets [Existing System-ref[2]]	92%

Table 2 show the performance when 80 word pairs were used for training and 20 word pairs were used for testing purpose. From Table.2 it can be inferred that the accuracy of the proposed system is better.

5. CONCLUSION

A new method for measuring semantic similarity between words is experimented in this paper. The method integrates the SWD and Snippet for measuring semantic similarity and uses SVM as a classifier to classify the given word pairs. From the experimental results, it can be seen that the accuracy of the proposed system has improved. In future the developed system can be applied for query expansion application by means of suggesting list of semantically related words for a given word to get more accurate results from web search engine.

6. REFERENCES

[1] Basu, T. and Murthy, C.A. (2009) 'Semantic Relation between words with the web as Information source', PReMI- Proc of the 3rd International Conference on Pattern Recognition and Machine Intelligence, LNCS 5909, pp.267-272.

[2] Bollegala,D., Matsuo,Y., and Ishizuka, M. (2011), 'A Web Search Engine Based Approach to Measure Semantic Similarity between Words', IEEE Transactions on Knowledge and Data Engineering, Vol 23, NO 7, pp.977-990.

[3] Cilibrasi, R. and Vitanyi, P. (2007) 'The google Similarity distance', IEEE Tansactions on Knowledge and Data Engineering, pp.370-383.

[4] Jinwu HU., Liuling DAI, Bin LIU, (2008) ' Measure Semantic Similarity between english words', ICYCS- The 9th International Conference for Young Computer Scientists, pp. 1689-1694.

[5] Liu, B., Dai, L. Xia, Y. and Wu, S. (2008) ' Measuring semantic similarity between words using How net', ICCSI - International conference on Computer Science and Information Technology, pp.601-605.

[6] Mehmet Ali Salahli (2009) ' An approach for measuring semantic relatedness between words via related terms', Mathematical and Computational Applications, Vol.14, No.1, pp.55-63.

[7] Pedersen, T., Patwardhan, S. and Michelizzi, J. (2004) 'WordNet::Similarity - Measuring the Relatedness of Concepts', HLT-NAACL Demonstration papers, pp: 38-41.

[8] Sahami, M. and Heilman, T. (2006) 'A Web-based Kernal Function for Measuring the Similarity of Short Text Snippets', Proc of 15_{th} Int'l World Wide Web Conf.pp.377-386.

[9] Takale, S.A. and Nandgaonkar, S.A (2010) 'Measuring semantic similarity between words using web documents', IJASCA- International Journal of Advanced Computer Science and Applications, Vol 1, No.4,pp.78-82.

[10] Zhiqiang, L., Werimin, S., and Zhenhua, Y. (2009), 'Measuring Semantic Similarity between Words Using Wikipedia', WISM - International Conference on Web Information Systems and Mining, pp: 251-255.