International Journal of Engineering Research in Computer Science and Engineering Engineering (IJERCSE) Vol 2, Issue 3, March 2015

Semantic Search Engine

[1]Swapna Dhondge,[2]Snehal Gaikwad,[3]Richa Sharma,[4]Abhaya Vagare
[1][2][3][4]K.K.Wagh College of Engineering
[1]swapna.dhondge@gmail.com,[2]sgaikwad069@gmail.com,[3]richas2807@gmail.com,[4]abhayavagare93@gmail.com

Abstract **The ability to accurately retrieve the required results from a huge database present over the internet is important. Web search engine based on semantics gives precise results. An empirical method is proposed to provide a semantic wise search that uses in one hand, a technical English dictionary and on the other hand, a page count based metric and a text snippet based metric retrieved from an existing web search engine. To identify the numerous semantic relations between the words, a novel pattern extraction algorithm and a pattern clustering algorithm is proposed. The page counts based co-occurrence measures and lexical pattern clusters extracted from snippets is learned using support vector machines. Integrate the page count, text snippet and dictionary based metric to accurately measure the semantic similarity search compared to normal search**

Key words- **Web search engine, Pattern extraction, Pattern clustering , natural language processing, co-occurrence measures, snippet, Technical dictionary.**

I. INTRODUCTION

A web search engine is a software system that is designed to search for information on the World Wide Web. Information on the web is vast with lot of hidden information, which are interconnected by various semantic relations. WordNet can be used for the analysis of semantic similarity but Since the semantic similarity between the words changes over time and domain, WordNet is not efficient. If semantic similarity between terms is not accurate the search results will not be precise. Using an automatic method to estimate the semantic similarity by an existing search engines is more efficient [1]. Page counts, dictionary based metric and snippets are some types of useful information provided by a search engine. Page count for a query is the number of web pages returned by the search engine. Page counts for two words provide the global co-occurrences of the two words on the web. If two words have more page count then they are likely to be more similar. But page counts alone as a measure of co-occurrence of two words presents lot of drawbacks.
1. Analysis of page count ignores the word position in a page.

2. A page count for a word with multiple senses contains a combination of all its senses.
for these reasons page count analysis alone is not reliable for measuring the semantic similarity. Snippet is a brief window of text extracted by a search engine around the query term in a document It is useful to get convenient summary about the search results. So it avoids the need to download the en- tire source document from the web. Consider a snippet from Google for the query Cricket AND sport

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

A considerable disadvantage of using snippets is that, due to the presence of large number of documents in the web result set, only the top-ranking snippet result set for a query is processed efficiently. In this paper, a method that considers technical dictionary based metric, page counts and lexical syntactic patterns extracted from snippets is proposed experimentally to overcome all the mentioned problems.

1.1 Related Work

The dataset of the words is been provided for the calculation of similarity between two words to find the length of the shortest path which connects two words[2]. Evaluating similarity considering distance now we consider other way which has been provided by Resnik et al[3] in which similarity is been calculated on the information content. Miller and Charles [4] calculated their experiments in which they reported high Pearson correlation coeffient of 0.8914. Normalized Google Distance which is used in Cilibrasi and Vitangi[4] where they considered the page counts which are retrieved from a web search engine. Semantic similarity between two queries using snippets which is used for those queries by the search engine has been proposed by Sahami[5].Calculation of semantic similarity between words using web search engine that is page count and snippets which included query

expansion and word sense disambiguation was proposed by Bollengala et al[1].The percentage where merged with support vector machine and then the measurement of similarity was done.

1.2 Outline

In the proposed model, we calculate semantic similarity between two words given as input to our system. Semantic similarity is represented by decimals in the range [0, 1]. If two words are semantically more similar then output will be nearer to 1. If two words are semantically not similar then output will be nearer to 0. We make use of search engine and technical dictionary to measure the semantic similarity between two input words.

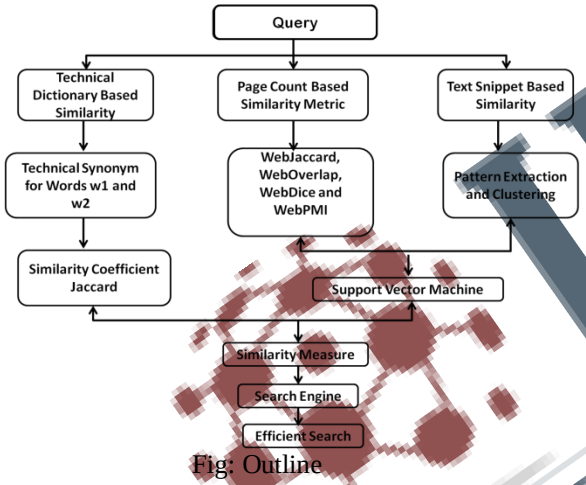

Fig: Outline

We fire a query to our model. We use stop word removing algorithm to remove the unwanted results in the query. Then the similarity between remaining words is calculated. Let us consider that after removing stop word, we get two word say w1 and w2. We then query two input words say w1 and w2 to the technical dictionary. It returns the technical synonyms for queried words from which similarity coefficient Jaccard is calculated. Two input word are also queried to the search engine. Search engine returns page count and text snippets related to our input words. Using page count, we calculate co-occurrence measures. We retrieve pattern from the text snippets and cluster them. Both page count based co-occurrence measures and lexical pattern clusters are used to define a feature vector that represents semantic similarity between words approximately. A two class SVM is trained [6] to find semantic similarity between words using feature vector representation. Final semantic similarity is calculated using feature vector and technical dictionary based similarity coefficient Jaccard. Using this similarity
count, the result page showing more similarity between words is arranged at the top. Likewise all result pages are arranged in decreasing order of semantic similarity count.

1.2.1 Co-occurrence measures

Page counts for the individual words P AND Q has been considered and semantic similarity has been found out. The cooccurence measures [1] has been calculated by using four formulas:- .

1.WebJaccard:-When the two words appear on the same page conjunction query is been used here and coeffient is set to zero.

$$WebJaccard(P,Q) = \begin{cases} 0 & \text{if } H(P \wedge Q) <= c \\ \dfrac{H(P \wedge Q)}{H(P)+H(Q)-H(P \wedge Q)} & \text{otherwise} \end{cases}$$

2. Webdice:-Coeffient is variant. The WebJaccard coefficient between words P and Q is defined as :

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

3. Weboverlap:-Natural modification coeffient. WebOverlap (P, Q) is defined as

$$WebOverlap(P,Q) = \begin{cases} 0 & \text{if } H(P \wedge Q) \le c \\ \dfrac{H(P \wedge Q)}{\min(H(P),H(Q))} & \text{otherwise} \end{cases}$$

4. PMI:-Number of documents indexed by the search engine it is set as N = 10Λ10 The WebPMI as a variant of point- wise mutual information using page counts is defined as

$$WebPMI(P,Q) = \begin{cases} 0 & \text{if } H(P \wedge Q) \le c \\ \log 2\left[\dfrac{H(P \wedge Q)/N}{H(P)/N, H(Q)/N}\right] & \text{otherwise} \end{cases}$$

Where N is the number of documents indexed by the search engine.

1.2.2 Lexical Pattern Extraction

Information extraction is defined as the detection and extraction of particular events of interest from text. Due to the draw- back of use of snippets, we propose an algorithm called as lexical pattern extraction algorithm.

The lexical pattern extraction algorithm is based on text snippets. We use this algorithm for finding the semantic relationship between the given two words. We are using this technique by various Natural Language Processing (NLP). Snippets are useful for search because a user can read the snippet and decide that a search result is directed related or not, without even opening the URL. Snippets are useful because it is unnecessary to download the documents from the web. It is efficient and time consuming if a document is large. Example of snippet :

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

Lexical patterns are the patterns that satisfy the following criteria. 1. A subsequence must exactly contain one occurrence of each A and B 2. The max length of subsequence is L words 3. In a subsequence one or more words can be skipped. How- ever, consequently it should be less than g.

4. Only negation contractions in a context are expanded. [1]

1.2.3 Lexical Pattern Clustering

New words are constantly created and existing words are as- signed with new senses on the Web. To decide that two words are semantically similar or not, it is important to know the semantic relations that hold between the words. Consider an example Horse and Cow or Horse and car. The words horse and cow can be considered semantically similar because both horses and cows are useful animals in agriculture. Similarly, a horse and a car can be considered semantically similar because cars, and historically horses, are used for transportation. It is note that all lexical patterns are not independent multiple lexical patterns can express the same semantic relation. Clustering means to group of something or bunch of items that are close to each other. By clustering the semantically related patterns into groups, it can both overcome the data sparseness problem. We are using lexical pattern clustering algorithm [1] for these. Steps for pattern clustering- Input: patterns = a1, . . . , an, threshold Output:

clusters C
SORT(Λ)
C ← {}
For pattern ai ∈ Λ
do Max ← −∞
c∗ ← null
for cluster cj ∈ C
do sim ← cosine(ai,cj)
if sim > max
then max ← sim
c∗ ← cj ⊕ai
else C ← C ∪ {ai}
End if
End for
Return C

1.2.4 Dictionary based similarity measure

The similarity of words is generally calculated through the word distance, word distance is a real number. It works as a similarity calculation method of English words based on WordNet, by extracting synonym sets, generic information and meaning interpretation of specific words from the dictionary WordNet. It provides a similar word set for information retrieval. Word- Net is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. The wordnet (ASCII) database format is well documented. The API documentation is available online and is distributed with the main WordNet packages. The calculation of the similarity between two words based on the technical dictionary(TD). In this, technical synonyms for each word are extracted from the technical English dictionary. The technical dictionary is used to diff erentiate the technical meaning of the words from normal meaning. The TD is used to extract sets of technical synonyms for each word. Once the two sets of synonyms for each word are collected, the degree of similarity, S (w1, w2) is calculated using the Jaccard coefficient:

S (w1,w2) = mc mw1+mw2−mc

Where mc: The number of common words between the two synonyms set mw1: The number of words contained in the w1 synonym set mw2: The number of words contained in the w2 synonym set If the group of synonyms for the words w1 explicitly contains the word w2 or vice versa, assign directly the value 1 to S (w1, w2).[7]

1.3 Measuring semantic similarity

As discussed earlier page counts alone is not reliable for measuring semantic similarity. Therefore, a machine learning approach is used for combining both page counts co-occurrence measures and snippet based measures. Given N clusters of lexical pattern, a pair of words (P,Q) is represented by(N+4) dimensional feature vector f_{PQ}. Co-occurrence measure based on page-counts are used as four distinct features in f_{PQ}. Then a feature from each of the N cluster is computed as follows [1]:

W_{ij} =

$$\frac{\mu(ai)}{\sum_{t \in cj} \mu(t)}$$

W_{ij} is a weight assigned to a pattern ai which is present in cj. μ(a) is total frequency of a pattern a in all word-pairs. Then finally the value of j^{th} feature in a vector for a word pair (P, Q) is computed as follows [1]:

$\sum_{ai \in cj} W_{ij} f(P,Q,ai)$

It is weighted sum of all patterns in cluster cj that co-occur with words P and Q. Using these features a two class SVM is trained to detect synonyms and non-synonyms word pair. Training dataset is automatically generated from wordnet synsets. This trained SVM is used to measure semantic similarity between words. After the computation of semantic similarity between words, it is used to design a semantic search engine which in turn returns the semantically related results for user query.

CONCLUSION

We proposed a semantic similarity measure using both page counts and snippets retrieved from a web search engine and technical dictionary for two words. Word co-occurrence measures were computed using page counts returned by search engine. Lexical pattern extraction algorithm was used to ex- tract semantic relations between two words from the snippets returned by search engine. Sequential pattern clustering algorithm was used to identify different lexical patterns that describe the same semantic relation. 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. Experimental results on benchmark data set showed that the proposed method outperforms various baselines as well as previously proposed web- based semantic similarity measures. Pro- posed method can be used to improve the accuracy of the search based on semantic relation between two words.

REFERENCES

[1] Bollegala D, Matsuo Y, and Ishizuka M (2011), "Measuring semantic similarity between words using web search engines ", IEEE Transactions on Knowledge and Data Engineering, vol.23, Issue 7, pp.977-990.

[2] R.Rana ,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.1pp. 17- 30,jan /feb 1989.

[3] P.Resnik, "Using Information Content to Evaluate se- mantic similarity in a taxonomy",Proc.14th Intl Joint conf.Artificial Intelligence (AAAI 06),2006.

[4] G.Miller and W.Charles , "Contextual correlates of se- mantic similarity" ,Language and cognitive processes vol.6,no.1,pp.1-28,1998.

[5] M.Sahami and T.Heilman , "A Web Based Kernel Function for Measuring the Similarity of short text snippets" ,Proc 15th Intl World Wide Web Conf,2006.

[6] V. Vapnik, "Statistical Learning Theory." Wiley, 1998

[7] Mrs.M.Karthiga,Mrs.PCD.Kalaivaanim, Mr.S.Sankarananth , "A Semantic Similarity Approach Based on Web Resources ", 3rd ed. Harlow, England: Addison-Wesley, 2013.