Discovering Semantic Similarity between Words Using Web Document and Context Aware Semantic Association Ranking

P.Ilakiya

Abstract— **The growth of information in the web is too large, so search engine come to play a more critical role to find relation between input keywords. Semantic Similarity Measure is widely used in Information Retrieval (IR) and also it is important component in various tasks on the web such as relation extraction, community mining, document clustering, and automatic metadata extraction. An empirical method to estimate semantic similarity using page counts and text snippets retrieved from a web search engine for two words. Specifically, define various word co-occurrence measures using page counts and integrate those with lexical patterns extracted from text snippets. Pattern clustering is used to identify the numerous semantic relations that exist between two given words. The optimal combination of page counts-based co-occurrence measures and lexical pattern clusters is learned using support vector machines. The proposed method context Aware Semantic Association Ranking discovering complex and meaningful relationships, which we call Semantic Associations.**

***Index Terms*— Semantic Similarity, Pattern Extraction, Pattern Clustering, Page Count, Snippet, Semantic Association.**

I. INTRODUCTION

Highlight Due to the rapid development of technologies, we all receive much more information than we did years ago. Organizing and comparing information effectively has now become a real problem. The string comparison functions available in most high-level languages make it easy to compare the lexical similarity between text passages. But in most cases to avoid information overloading and to improve the quality of search result, we want to consider the semantics rather than the lexical of these text passages. The traditional lexical similarity measurements do not adapt well to the requirements of semantic contexts. While searching in the web, it normally looks for matching documents that contain the keywords specified by the user. Web mining is use of data mining technique which extracts the information from the web documents. Keywords are the input for the searching and retrieving the document in the web. For example, when we give "Tablet" and "PC" as keyword, the search engine displays the document that contains the words Tablet and PC but it does not relate the words semantically.

Manuscript received June, 2013.

Ilakiya P, Department Of Computer Science and Engineering, SNS College Of Technology, Coimbatore, India, 9677466553

So we get the unwanted information i.e. user in focused to get the information related to the Tablet products like Tablet PC, but in ordinary web it also gives the related information to Tablet and PC it will be avoided through the semantic web. This Unwanted information is said to be as Information Overload. [8]

Semantic similarity is the degree to which text passages have the same meaning. This is typically established by analyzing a set of documents or a list of terms and assigning a metric based on the likeness of their meaning or the concept they represent or express. To be semantically similar, two terms do not have to be synonyms one. Instead, two terms are given a number to indicate the semantic distance between them, i.e., how term A is relates to term B. Semantic similarity provides a common way to build ontology's, which in turn provides the knowledge base for the semantic web.[8]

Semantic similarity between words is achieved in web with the help of semantic web which is an extension of the current web that contains information with their well defined meaning. It describes the relationship between keywords and the properties of the keywords. Some of the examples for semantic web are swoogle, Hakia and Yebola.

A. Lexical based search

Searching key word in the semantic web is totally different from the Text based search engine Google. In this, while giving the keyword it uses the Cashing Built and Bit table. Cashing Built is like call log in the Mobile Phone and also it is called as the temporary buffer that stores recent search history. Bit Table is an algorithm Centric Database, which contains the huge collection of data and indexes of all the keywords in the document except the stopping words. Both the Cashing Built and the Bit table are searched simultaneously, and when the search engine founds the keyword it automatically disconnect the another connection and display the result.

The searching Result in the Google is mainly based on four criteria: Architecture, Site content, Credibility and Back link, Engagement. However it displays the result based on hits and Page Rank Algorithm. So there are more chances to extract irrelevant documents also. This is a major disadvantage of Google search. So it paved the way to the semantic web.

B. Semantic based search

Semantic web consists of three parts: RDF (Resource Description Framework), Ontology, and OWL (Web Ontology Language). In this, the keyword is taken as a snippet. Snippets are useful for search because most of the time a user can read the snippet and decide whether a particular search is relevant without opening the "url" [14]. First the snippet is processed into the RDF. It is like the XML Language, RDF divides the snippet as three parts Subject, Predicate, Object, and then constructs the combination between these that results are stored in the Database.

Second Ontology is applied to the Database, it helps to find entity in the database and the same entity are group, after this it finds the number of the relationship between the entities; finally OWL cuts the minimal number of Semantic Measure and display the result to user. These are the searching process behind the Semantic Web Search Engine.

II. RELATED WORKS

A. Relation Based Page Rank Algorithm

Relation based page rank algorithm to be used in conjunction with semantic web search engines that simply relies on information that could be extracted from user queries and on annotated resources [7]. Its help to give the result in ordered manner that is, best fit user query document is displayed first. Flow of query extraction and result is shown in figure1. First the user post there query in search engine, the web crawler download the document from the web page database. Then the initial Result set is constructed that contain query keyword and concept.

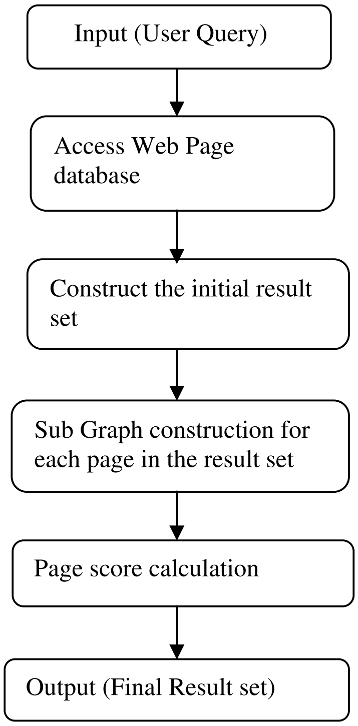
Input (User Query)
Access Web Page database
Construct the initial result set
Sub Graph construction for each page in the result set
Page score calculation
Output (Final Result set)

Figure 1: Flow chart for query extraction and presentation of results

For that resultant set search engine construct the query sub graph for each page then page score is calculated. Based on the page score final result set is constructed and given to the user. [11]

Advantage

i. Effectively manage the search space.

ii. Reduce the Complexity associated with the ranking task.

Disadvantage

i. Less scalable.

B. Unsupervised Semantic Similarity between terms

Page count and context based metrics are discussed in the Unsupervised semantic similarity computation. Page count metrics consider the hits returned by a search engine this can be done by following three sequential executions.
1) Jaccard and Dice Coefficient are used to find the similarity between set of word. Example W1 and W2 are the two words if the co occurrence is equal to one means; both the W1 and W2 are same. Co occurrence zero means W1 and W2 did not same.
2) Mutual Information is used to calculate the number of document that contain the word W1 and W2 that number is assigned to a random variable X, Y.[12] Then the mutual dependence between these two words are calculated it can be evaluated by following two case i) If X and Y independent means both X and Y does not share any information so in this case mutual information is 0. ii) If X and Y are equal means X share some value and the mutual information is 1.
3) Google based semantic relatedness calculates the similarity between two words and then compared with the distance between these two words. If the similarity is high means distance between these words will be low [5].Context based semantic similarity algorithm download the top ranked document based on user query and also compute the frequent occurrence of these words.

Advantage

i. Provides good performance.

ii. Fully automatic and required little computation power.

Disadvantage

i. Related document selection feature selection does not discuss.

C. Measure Word-Group Similarity Using Web Search Result

Distribution similarity measure is performed in Word-Group Similarity measure. [1] Measure the search counts within the small number of word pair. Based on Dataset, that contains some limited number of document. Apply the page count based measure in the particular dataset, results the similarity score always high, this is possible because of the limited number of document.

Advantage

i. Proposed method can be utilized to measure the similarity of entire sets of word in addition to word-pairs

Disadvantage

i. Optimal values are not obtained

D. A Study on similarity and relatedness Using Distributional and Word Net -based Approaches

Word Net is the Database for English words it contains the synonyms for most of the words repeatedly using by the users. Some versions of word net are MCR (Multilingual Central Repository) contains some other language like Spanish, Italian [10].

Cross Linguality Similarity method is applied to calculate the same word in different languages[6]. For example Car and the Coche having the same meaning, Car is from English and coche is from Spanish that can be analyses from MCR and graph is constructed to calculating the similarity and page rank.

Advantage

i. Effective way to compute cross-lingual similarity with minor losses.

E. Measuring Semantic Similarity between Words Using Web Document.

One of the approaches to measure the semantic similarity between words is snippet that is returned by the Wikipedia. First the snippet is extracted from Wikipedia based on the user query, then the snippet is preprocessed because some time semantically unrelated snippet are extracted by the search engine. Finally stop words and suffix are removed from the snippet [12]. If the system is not removing the Stop Key word and Suffixes means total number of keywords for a particular document will be high and also the time taken to find the similarity also is increases.

There is a five similarity measure of association that is simple similarity, Jaccard similarity, Dice similarity, Cosine similarity, Overlap Similarity. Jaccard similarity comparing the similarity and diversity of given sample set. Dice similarity also related to the jaccard measure. Over Lap method is used to find the overlapping between the two sets. The cosine similarity is a measure of similarity between two vectors of n dimensions by finding the angle between them [12].

F. A Web Search engine-Based Approach to Measure Semantic Similarity between Words

Web search engine based approach only combines the page count and snippets together and then the resultant set will be given to the SVM. While we are searching keywords in the search engine, it passes the keywords to page count and snippet that shown in the figure 2. Page Count represents the number of pages that contain the keyword. [13]It is the global co occurrence of the keyword. Page count for the separate word and also for both the words is calculated by using the four similarity score method or co occurrence measure that are Web Jaccard, Web Overlap, Web Dice, Web PMI. Page count not only sufficient to measure the semantic similarity, why means it does not gives the position of the word in the document, semantic of the word does not possible in this.

Text snippet is also help to find the similarity between the given keyword. For example the query words are cricket and sport. Cricket is a sport played between two teams is the snippet retrieved from the search engine then the user easily knows, it is related to the query or not without viewing the whole document. In the above example “is a” is the semantic relationship and some of the semantic relationship are also know as, part of etc.

Then the Lexical Pattern extraction and pattern Clustering is applied. Lexical Pattern extraction must contain subsequence with one occurrence of each keyword that is X and Y. Maximum length of the subsequence is L Words. Subsequence is allowed to skip some words [4]

Lexical Pattern Clustering group set of object into classes of similar object. And also it defines the features of the words. Although it defines the similarity vary from one cluster to another cluster model.

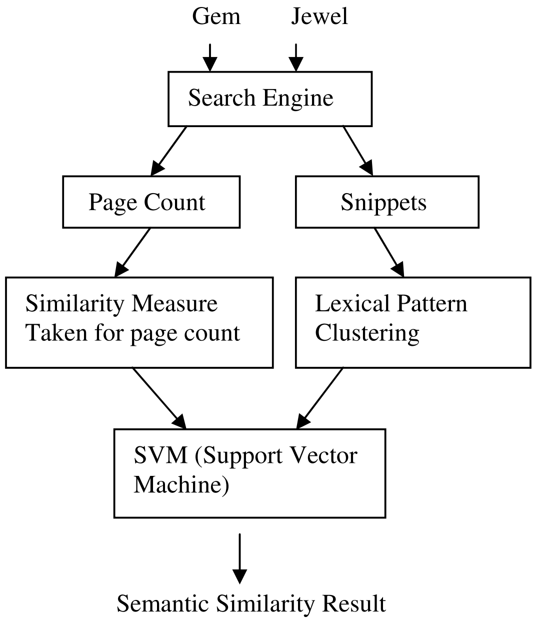
Gem
Jewel
Search Engine
Page Count
Snippets
Similarity Measure Taken for page count
Lexical Pattern Clustering
SVM (Support Vector Machine)
Semantic Similarity Result

Figure 2: Search Engine working Outline Diagram

Finally the pattern cluster result and the similarity score result is given to the SVM.SVM is the Support Vector Machine this will be the supervised learning model used for classification that analyze the data and recognize the patterns. It is the one of the classifier that helps to classify the synonymous and non synonymous data.

Advantage

i. Precision and recall is improved in this.

III. PROPOSED SEMANTIC SIMILARITY METHOD

A search engine is an interesting task. 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. So by this method most similar document are extracted and ranked.

A. Dataset Preprocessing

The dataset preprocessing can often have a significant impact on generalization performance of any data mining algorithm. The elimination of noise instances is one of the most difficult problems in such scenarios. Usually the removed instances have excessively deviating instances that have too many null feature values. These excessively deviating features are also referred to as outliers. In addition, a common approach to cope with the infeasibility of learning from very large data sets is to select a single sample from the large data set.

Missing data handling is another issue often dealt with in the data preparation steps. WordSimilarity353 (WS) dataset has been used throughout the project and the same has been preprocessed here.

B. Lexical Pattern Extraction

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. For example, consider the snippet cricket and sport. Here, the phrase is indicates a semantic relationship between cricket and sport. Many such phrases indicate semantic relationships. Lexical pattern Extraction gives page count of the single word and combination of two words using following four similarity measure, Jaccard, Overlap (Simpson), Dice, and Pointwise mutual information (PMI).

The Web Jaccard coefficient between words (or multiword phrases) or the similarity between two words P and Q.

$$WebJaccard(P,Q)=\begin{cases} 0, & if\, H(P\cap Q)\le c, \\ \dfrac{H(P\cap Q)}{H(P)+H(Q)-H(P\cap Q)}, & otherwise. \end{cases}$$

Web Overlap is a natural modification to the Overlap coefficient i.e., Overlap of two keywords.

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

The Web Dice coefficient as a variant of the Dice coefficient P and Q. It defines the dependence between two probabilistic events.

$$WebDice(P,Q)=\begin{cases} 0, & if\, H(P\cap Q)\le c, \\ \dfrac{2H(P\cap Q)}{H(P)\_H(Q)}, & otherwise. \end{cases}$$

Pointwise mutual information is a measure that is motivated by information theory; it is intended to reflect the dependence between two probabilistic events. The Web PMI define as

$$WebPMI(P,Q)=\begin{cases} 0, & if\, H(P\cap Q)\le c, \\ \log_2\left(\dfrac{\frac{H(P\cap Q)}{N}}{\frac{H(P)}{N}\frac{H(Q)}{N}}\right), & otherwise. \end{cases}$$

C. 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.

D. Measuring Semantic Similarity

Semantic similarity defined four co-occurrence measures using page counts and shows how to extract clusters of lexical patterns from snippets to represent numerous semantic relations that exist between two words. In this 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.

E. Context Aware Semantic Association Ranking

Context-Aware Semantic Association Ranking is applied to enhance the results with ranking. Discovering complex and meaningful relationships, which we call Semantic Associations, is an important challenge.[2] Just as ranking of documents is a critical component of today's search results, ranking of relationships will be essential in tomorrow's semantic search results that would support discovery and mining of the Semantic Web. Building upon the above proposed work, a framework is proposed where ranking techniques can be used to identify more interesting and more relevant Semantic Associations. These techniques utilize alternative ways of specifying the context using ontology. This enables capturing users' interests more precisely and better quality results in relevance ranking.

IV. CONCLUSION

The World Wide Web is huge, widely distributed; global information service centre in this 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, and also efficient estimation of semantic similarity between words is critical for various natural language processing tasks such as word sense disambiguation, textual entailment, and automatic text summarization. In information retrieval, one of the main problems is to retrieve a set of documents that is semantically related to a given user query. For example, the word "apple" consists of two meaning one indicates the fruit apple and the other is the apple company. Retrieving accurate information to users to such kind of similar words is challenging. Some existing system proposed an architecture and method to measure semantic similarity between words, which consists of snippets, page-counts and two class support vector machine. Proposed approaches to compute the semantic similarity between words or entities using text snippets is good. Context-Aware Semantic Association Ranking is applied to enhance the results with ranking.

ACKNOWLEDGMENT

The authors would like to thank the Editor-in-Chief, the Associate Editor and anonymous Referees for their comments.

REFERENCES

[1] Ann Gledson and John Keane," Using Web-Search Results to Measure Word-Group Similarity" Proceedings of the 22nd International Conference on Computational Linguistics, pages 281–288 Manchester, August 2008.

[2] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth "Context-Aware Semantic Association Ranking", Semantic Web and Databases Workshop Proceedings. Belin, September 7,8 2003.

[3] Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring Semantic Similarity between Words Using Web Search Engines," Proc. Int'l Conf. World Wide Web (WWW '07), pp. 757-766, 2007.

[4] DanushkaBollegala, Yutaka Matsuo, and Mitsuru Ishizuka," A Web Search Engine-Based Approach to Measure Semantic Similarity between Words" IEEE transactions on knowledge and data engineering, vol. 23, no. 7, July 2011

[5] Elias Iosif and Alexandros Potamianos, "Unsupervised Semantic Similarity Computation between Terms Using Web Documents" IEEE transactions on knowledge and data engineering, vol. 22, no.11, November 2010.

[6] Eneko Agirre,Enrique Alfonseca,Keith Hall,Jana Kravalova, Marius Pasca,Aitor Soroa," A Study on Similarity and Relatedness Using Distributional and Word Net-based Approaches" The 2009 Annual Conference of the North American Chapter of the ACL, pages 19–27,Boulder, Colorado, June 2009.

[7] Fabrizio Lamberti,Andrea Sanna, and Claudio Demartini,"A Relation-Based Page Rank Algorithm for Semantic Web Search Engines" IEEE transactions on knowledge and data engineering, vol. 21, no. 1, January 2009.

[8] Gabriela Polcicova and Pavol Navrat "Semantic Similarity in Content-Based Filtering" ADBIS 2002, LNCS 2435, pp. 80–85, 2002. Springer-Verlag Berlin Heidelberg 2002.

[9] Sheetal A. Takale and Sushma S. Nandgaonkar," Measuring Semantic Similarity between Words Using Web Documents" (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 1, No.4 October, 2010

[10] Gledson and J. Keane, "Using Web-Search Results to Measure Word-Group Similarity," Proc. Int'l Conf. Computational Linguistics (COLING '08), pp. 281-288, 2008.

[11] Jiang and D. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," Proc. Int'l Conf. Research in Computational Linguistics (ROCLING X), 1997.

[12] Strube and S.P. Ponzetto, "Wikirelate! Computing Semantic Relatedness Using Wikipedia," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), pp. 1419-1424, 2006.

[13] Wu and M. Palmer, "Verb Semantics and Lexical Selection," Proc. Ann. Meeting on Assoc. for Computational Linguistics (ACL '94), pp. 133-138, 1994.

Ms P.Ilakiya currently pursuing her Master of Engineering in Software Engineering at SNS College of Technology, affiliated to Anna University Chennai, Tamilnadu, India. She received her B.E degree from Arunai Engineering College, affiliated to Anna University Chennai. Her research interests include data mining and networking.