International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
383 – 389

Measuring Semantic Similarity among Text Snippets and Page Counts in Data Mining

[1] V. Sobana
M. Phil Scholar, Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

[2]Mr. T. Muthusamy
MCA.,M.Phil,Assistant Professor,
Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

[3] Mrs. K. K. Kavitha
M.C.A., M.Phil., SET., (Ph.D).,
Vice Principal, Head of the Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

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 (or entities) remains a challenging task. We propose an empirical 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. To identify the numerous semantic relations that exist between two given words, we propose a novel 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 proposed method outperforms various baselines and previously proposed web-based semantic similarity measures on three benchmark data sets showing a high correlation with human ratings. Moreover, the proposed method significantly improves the accuracy in a community mining task.

*****

I. INRODUCTION

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.

Semantic similarity plays an iconic role in the field of data processing, artificial intelligence and data mining. It is also useful in information management, especially in the context of environment such as semantic web where data may originate from different sources and has to be integrated in flexible way. Semantic similarity is basically a measure used to compute the extent of similarity between two concepts based on the likeliness of their meaning. The concepts can be sentences, words or paragraphs. It finds the distance between different concepts in semantic space in such a way that lesser the distance, greater the similarity. In other words, semantic similarity identifies the common characteristics. Concepts can be similar in two ways that is either lexically or semantically. If words are having a similar character sequence then they are lexically similar, while semantics is concerned with the meaning of the words.

Application

The data mining methodologies are defined and implemented in various field and they are used in the real time and this searching technique has been introduced in some more efficient algorithms are some felt where mining has been used are listedbelow:

Future Health Care

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use different data mining approaches like multidimensional databases, robotics, soft computing, data resolving and statistics.

Mining can be used to predict the volume of patients in every category. Processes are been developed to make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

WEB MINING CONSISTS OF THE FOLLOWING TASKS

Resource Finding

The task of retrieving intended Web documents. Information selection and pre-processing: automatically

 ISSN: 2454-4248
383 – 389

selecting and pre-processing specific information from retrieved Web resources.

Generalization

Automatically discovers general patterns at individual Web sites as well as across multiple sites.

Analysis

Validation and/or interpretation of the mined patterns. There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM).

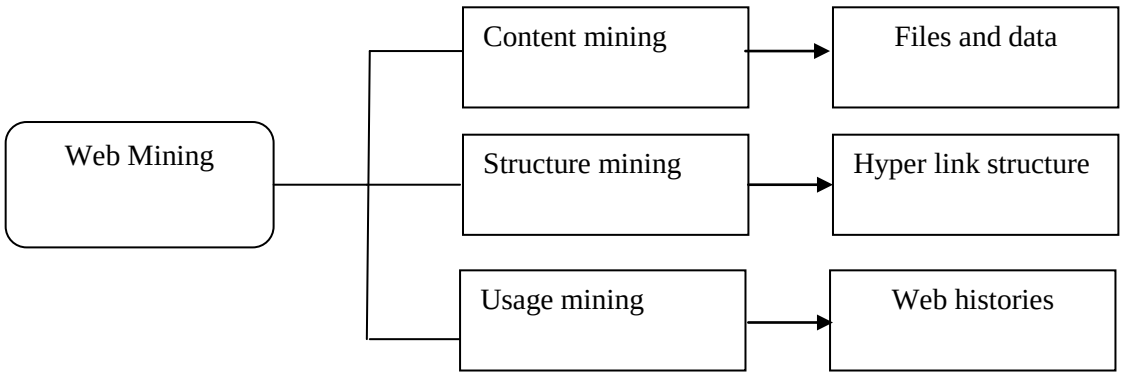

Figure 1.1 Data mining process

Web content usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks which helps to investigate the node and connection structure of web sites.

PROCESS OF WEB MINING

The complete process of extracting knowledge from Web data is shown in figure1.

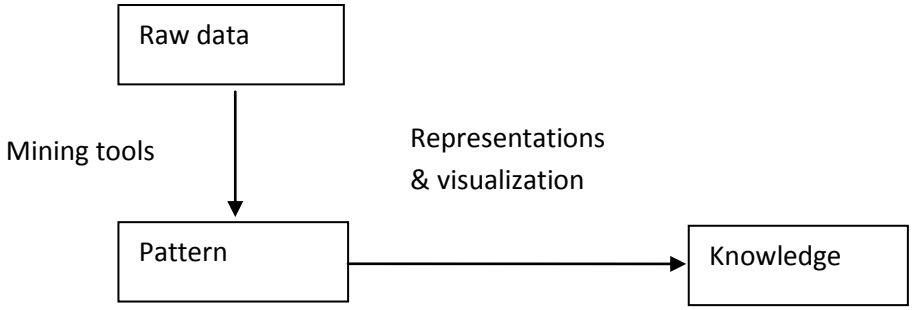

Figure 1.2 Process of Web Mining

The various steps are explained as follows.

Resource finding

It is the task of retrieving intended web documents.

Information selection and pre-processing

Automatically selecting and pre- processing specific from information retrieved Web resources.

Generalization

Automatically discovers general patterns at individual Web site as well as multiple sites.

Analysis

Validation and interpretation of the mined patterns.

SEARCH ENGINE ARCHITECTURE

Search engines are the key to finding specific information on the vast expanse of the World Wide Web. We use the term search engine in relation to the Web. These usually refer to the actual search forms, which searches through databases of the HTML documents.

There are at least three elements which contain important: information for a search engine. Discovery & the database, the user search, presentation and ranking of

 ISSN: 2454-4248
383 – 389

results.Here we represent the elements of search engine architecture.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration r taken into account for developing the proposed system.

Web Mining Information and Pattern Discovery onthe World Wide Web

In this paper we provide an overview of tools techniques and problems associated with both dimensions We present a taxonomy of Web mining and place various aspects of Web mining in their proper contextThere are several important issues unique to the Web paradigm that come into play if sophisticated types of analyses are to be done on server side data collections. These include integrating various data sources such as server access logs referrer logs user registration or prole informationresolving difficulties in the identification of users due to missing unique key attributes in collected dataand the importance of identifying user sessions or transactions from usage data site topologies and models of user behaviorWe devote the main part of this paper to the discussion of issues and problems that characterize Web usage miningFurthermore we survey some of the emerging tools and techniques and identify several future research directions.

Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely pre-processing, pattern discovery, and pattern analysis. This paper describes each of these phases in detail. Given its application potential, Web usage mining has seen a rapid in- crease in interest, from both the research and practice com- munities.

BACKGROUND OF THE STUDY

Semantic similarity approaches have been applied in semantic internet associated applications such as regular annotation of internet pages, social mining and root term extraction for inter-object association representation. There is an broad novel on calculating the similarity between paragraphs or documents, but there is less work connected to the computation of similarity between sentences or small texts. Related research can roughly be classified into four major categories: word co-occurrence/vector-based document model methods, corpus-based methods, hybrid methods, and descriptive feature information-based methods. The information on theoretical similarity measure has been analyzed using the ontology structure based on hierarchical and non-hierarchical ways.

PROBLEM DEFINTION

Given two words A and B, we model the problem of measuring the semantic similarity between A and B, as a one of constructing a function semanticsim (A, B) that returns a value in the range of 0 and 1. If A and B are highly similar (e.g. synonyms), we expect semantic similarity value to be closer to 1, otherwise semantic similarity value to be closer to 0. We define numerous features that express the similarity between A and B using page counts and snippets retrieved from a web search engine for the two words. Using this feature representation of words, we train a two-class Support Vector Machine (SVM) to classify synonymous and non-synonymous word pairs. Our objective is to find the semantic similarity between two words and improves the correlation value.

STATEMENT OF PROBLEM

- 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.
- For example, the word “apple” consists of two meaning one indicates the fruit apple and the other is the apple company. So retrieving accurate information to users to such kind of similar words is challenging.
- In the base paper, the authors proposed an architecture and method to measure semantic similarity between words. Which consists of snippets, page-counts and support vector machine.
- The authors proposed an approach to compute the semantic similarity between words or entities using text snippets. But in this project we are going to implement and compute the semantic similarity between words in Search engine without using Snippets or Support Vector Machines.
- Because using Snippets or Support Vector Machines makes the job of finding similarity easier. So we are going to implement the same concept without using snippets or support Vector machines.

 ISSN: 2454-4248
383 – 389

III. RESEARCH METHODOLOGY

INTRODUCTION

A novel method to provide better Webpage recommendation based on Web usage and domain knowledge. Supported by three new knowledge representation models and a set of Web-page recommendation strategies. The first model is an ontology-based model that represents the domain knowledge of a website. The construction of this model is semi-automated so that the development efforts from developers can be reduced. The second model is a semantic network that represents domain knowledge, whose construction can be fully automated. This model can be easily incorporated into a Web-page recommendation process because of this fully automated feature. The third model is a conceptual prediction model, which is a navigation network of domain terms based on the frequently. Web-pages and represents the integrated Web usage and domain knowledge for supporting Web-page prediction. The construction of this model can be fully automated

EXISTING METHODOLOGY

It has been reported that the approaches based on tree structures and probabilistic models can efficiently represent Web access sequences (WAS) in the Web usage data. These approaches learn from the training datasets to build the transition links between Web-pages.

The performance of existing approaches depends on the sizes of training datasets. The bigger the training dataset size is, the higher the prediction accuracy is. However, these approaches make Web-page recommendations solely based on the Web access sequences learnt from the Web usage data. Therefore, the predicted pages are limited within the discovered Web access sequences “New-page problem”.

PROPOSED METHODOLOGY

Proposed system presents a novel method to provide better Webpage recommendation based on Web usage and domain knowledge, which is supported by three new knowledge representation models and a set of Web-page recommendation strategies. The first model is an ontology-based model that represents the domain knowledge of a website. The construction of this model is semi-automated so that the development efforts from developers This model can be easily incorporated into a Web-page recommendation process because of this fully automated feature. can be reduced. The second model is a semantic network that represents domain knowledge, whose construction can be fully automated.

Major advantages of the system are as follows

- Simplicity,
- More reliable and efficiency,
- Time consuming,
- Most related data to the given words,
- Ranking based results,
- Semantic similarity based on words co occurrences, and
- Automatic metadata extraction.

The proposals havebe divided into four modules:

1. Lexical Pattern Extraction
2. Lexical Pattern Clustering
3. Measuring Semantic Similarity
4. Ranking search results

PATTERN EXTRACTION

The input texts are then split for page count calculation, this will helps to get the related items of those inputs as individual and also in combine. After that the text snippets, which are used to retrieve the related item from the online will get and make an extraction from that too.

Snippets

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.

Pre-processing

Pre-processing is an important step for mining tasks, whereby, the features of a data set are modified so as to make information extraction reliable and convenient. Preprocessing is necessitated due to one or more of the following reasons. • Presence of noise in data: noise may disturb the information extraction process by making the data less than ideal. • sparsity of data: this results in a lack of information regarding certain portions of the data space, and consequently, inference cannot be generalized easily to unseen examples

- Text based Pre-processing
- Link based Pre-processing
- Query terms
- URL
- Rank pre-processing these items for getting clean data. Semantic Web Search The aim of this paper is to show how to make use of relations in Semantic Web page annotations with the aim of generating an ordered result set, where pages that best fit, the user query is displayed first.

 ISSN: 2454-4248
383 – 389

The ideas of exploiting ontology-based annotations for information retrieval are considered to play a key role in the Semantic Web.

PATTERN EXTRACTION ALGORITHM

- The input texts are then split for page count calculation, this will helps to get the related items of those inputs as individual and also in combine.\
- After that the text snippets, which are used to retrieve the related item from the online will get and make an extraction from that too.
- For each entry, the definition is processed looking for words that are connected with the entry in Wikipedia by means of a hyperlink.
- If there is a relation in Word Net between the entry and any of those words, the context is analysed and a pattern is extracted for that relation.

PATTERN CLUSTERING ALGORITHM

1. The extracted lexical patterns are clustered based on the similarity with respect to given cluster.
2. Each cluster contains patterns that express similar semantic relations.
3. The sorted clusters in ascending order do mean that the most useful clusters are at the top.
4. A two- class SVM is trained with both synonymous and non-synonymous word pairs generated from Word Net.
5. For 3000 words the word pairs are extracted.
6. The total number of words in the training data is 6000.
7. Then lexical patterns are extracted subject to specified threshold.
8. Lexical patterns thus extracted are clustered and given to SVM.

Extracting Lexico-Syntactic Patterns from Snippets

Text snippets are returned by search engines alongside with the search results. They provide valuable information regarding the local context of a word. We extract lexicosyntactic patterns that indicate various aspects of semantic similarity. For example, consider the following text snippet returned by Google for the query “cricket” AND “sport”. Here, the phrase is a indicates a semantic relationship between cricket and sport.

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

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.

TESTING AND IMPLEMENTATION

DATA SETS

The word pairs are identified by using the WordNet and thefeature 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” areextracted and stored in local database.
- Wr (P, Q) is found out and it is applied in SWD. This is oneof the feature vector obtained from SWD.
- The pattern clusters are formed from the patterns that areidentified 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 pairsare obtained through SWD and snippet, and the SVM istrained using these feature vectors.

TECHNIQUES AND ALGORITHM USED:

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and it’s constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

- We propose an automatic method to estimate the semantic similarity between words or entities using web search engines.
- 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.
- We present an automatically extracted lexical syntactic patterns-based approach to compute the semantic similarity between words or entities using text snippets retrieved from a web search engine.

EXPERIMENTS

We conduct two sets of experiments to evaluate the proposed semantic similarity measure. First we compare the

 ISSN: 2454-4248
383 – 389

similarity scores produced by the proposed measure against Miller-Charles benchmark dataset. We analyze the behavior of the proposed measure with the number of patterns used as features, the number of snippets used to extract the patterns, and the size of the training dataset. Secondly, we apply the proposed measure in two real-world applications: community mining and entity disambiguation.

The Benchmark Dataset We evaluate the proposed method against Miller-Charles [24] dataset, a dataset of 30 word-pairs6 rated by a group of 38 human subjects. The word pairs are rated on a scale from 0 (no similarity) to 4 (perfect synonymy). Miller-Charles' data set is a subset of Rubenstein-Goodenough's [35] original data set of 65 word pairs. Although Miller-Charles experiment was carried out 25 years later than RubensteinGoodenough's, two sets of ratings are highly correlated (pearson correlation coefficient=0.97). Therefore, Miller-Charles ratings can be considered as a reliable benchmark for evaluating semantic similarity measures.

PAGE COUNT-BASED COOCCURRENCE PROCEDURE

In general, because of the queried word might appear many times on one page, the page count may not necessarily be equal to the word frequency. Page count for the query X AND Y can be considered as a global measure of co-occurrence of words X and Y. Using page counts alone as a measure of co-occurrence of two words presents several drawbacks in spite of its simplicity. First, the position of a word in a page is ignored by the analysis of page count. Therefore, they might not be actually related even though two words appear in a page. Second, a combination of all its senses might contain by page count of a polysemous word. For those reasons, measuring semantic similarity is unreliable by page counts. Snippets provide useful information regarding the local context of the query term by a brief window of text extracted by a search engine around the query term in a document.

MEASURING SEMANTIC RESEMBLANCE

Ranking of search is determined by a composite arrangement of a variety of factors distinctive to the underlying search engine. Therefore, in the top-ranking snippets exists no guarantee for all the information we need to measure semantic similarity between a given pair of words is contained. To overcome the above mentioned problems we experimentally show a method that considers both page counts and lexical syntactic patterns extracted from snippets is proposed. To calculate similarity between two words is to find the length of the shortest path connecting the two words in the taxonomy a straightforward method for a given taxonomy of words is proposed. The multiple paths might exist between the two words if a word is polysemous. In such cases, for calculating similarity only the shortest path between any two senses of the words is considered.

IV. CONCLUSION

This thesis has presented a new method to offer better Web-page recommendations through semantic enhancement by three new knowledge representation models. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology-based model which can be semi-automatically constructed and the other is a semantic network of Webpages, which can be automatically constructed. A conceptual prediction model is also proposed to integrate the Web usage and domain knowledge to form a weighted semantic network.

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 nonsynonymous word pairs selected from WordNetsynsets. 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

FUTURE ENHANCEMENTS

Feature development 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 nonsynonymous word pairs selected from WordNetsynsets.

BIBLIOGRAPHY

[1] Liu, B. Mobasher, and O. Nasraoui, "Web Usage Mining," in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, B. Liu, Ed.: Springer-Verlag Berlin Heidelberg, 2011, pp. 527-603.

[2] B. Mobasher, "Data Mining for Web Personalization," in The Adaptive Web. vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds.: Springer-Verlag Berlin, Heidelberg, 2007, pp. 90-135.

 ISSN: 2454-4248
383 – 389

[3] G. Stumme, A. Hotho, and B. Berendt, "Usage Mining for and on the Semantic Web," AAAI/MIT Press, 2004, pp. 461-480.

[4] H. Dai and B. Mobasher, "Integrating Semantic Knowledge with Web Usage Mining for Personalization," in Web Mining: Applications and Techniques, A. Scime, Ed. Hershey, PA, USA: IGI Global, 2005, pp. 276 - 306.

[5] S. A. Rios and J. D. Velasquez, "Semantic Web Usage Mining by a Concept-Based Approach for Off-line Web Site Enhancements," in Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT '08. IEEE/WIC/ACM International Conference on, 2008.

[6] S. Salin and P. Senkul, "Using Semantic Information for Web Usage Mining based Recommendation," in 24th International Symposium on Computer and Information Sciences, 2009., 2009, pp. 236-241.

[7] A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, "Incorporating Concept Hierarchies into Usage Mining Based Recommendations," in Proceedings of the 8th Knowledge discovery on the web international conference on Advances in web mining and web usage analysis Philadelphia, PA, USA: Springer- Verlag, 2007, pp. 110-126.

[8] N. R. Mabroukeh and C. I. Ezeife, "Semantic-Rich Markov Models for Web Prefetching," in 2009 IEEE International Conference on Data Mining Workshops Miami, Florida, USA, 2009, pp. 465-470.