International Innovative Research Journal of Engineering and Technology
ISSN NO: 2456-1983

International Conference on Emerging Innovation in Engineering and Technology

ICEIET-2017

Nearest Keyword Set Search In Multidimensional Dataset

Mrs.D.Maladhy[1],M.Risvaanabegum[2],G.Sivapriya[3],V.Suryadharshini[4]

[1] Assistant Professor, Department of Information Technolgy, RGCET
[2,3,4] Department of InformationTechnology RGCET

[1]maladhymtech@gmail.com,[2]risvaanariz@gmail.com,[3]priyagunasekaram@gmail.com,[4]suryadharshini15@gmail.com

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 word counts and text snippets retrieved from a web search engine for two words. Specifically, we define various word co-occurrence measures using word 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 word 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 query benchmark data sets showing a high correlation with human ratings. Moreover, the proposed method significantly improves the accuracy in a community mining task.

Keywords-Clustering,DataMining,Semantic web,Data Warehousing

I . INTRODUCTION

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users. Commercial databases are growing at unprecedented rates. A recent META Group survey of data warehouse projects found that 19% of respondents are beyond the 50 gigabyte level, while 59% expect to be there by second quarter of 1996. In some industries, such as retail, these numbers can be much larger. The accompanying need for improved computational engines can now be met in a cost-effective manner with parallel multiprocessor computer technology. Data mining algorithms embody techniques that have existed for at least 10 years, but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older statistical methods. In the evolution from business data to

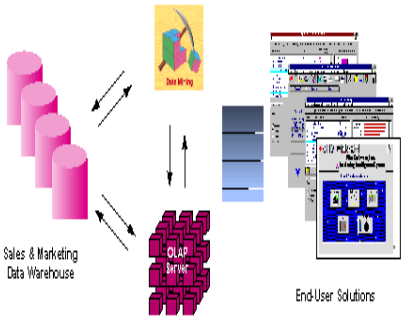
business information, each new step has built upon the previous one. For example, dynamic data access is critical for drill-through in data navigation applications, and the ability to store large databases is critical to data mining. From the user's point of view, the four steps listed in Table 1 were revolutionary because they allowed new business questions to be answered accurately and quickly.

*.a)**.T**he Scope of Data Mining*

Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities

b)How Data Mining works

How exactly is data mining able to tell you important things that you didn't know or what is going to happen next? The technique that is used to perform these feats in data mining is called modeling. Modeling is simply the act of building a model in one situation where you know the answer and then applying it to another situation that you don't. For instance, if you were looking for a sunken Spanish galleon on the high seas the first thing you might do is to research the times when Spanish treasure had been found by others in the past. You might note that these ships often tend to be found off the coast of Bermuda and that there are certain characteristics to the ocean currents, and certain routes that have likely been taken by the ship's captains in that era. You note these similarities and build a model that includes the characteristics that are common to the locations of these sunken treasures. With these models in hand you sail off looking for treasure where your model indicates it most likely might be given a similar situation in the past. Hopefully, if you've got a good model, you find your treasure. This act of model building is thus something that people have been doing for a long time, certainly before the advent of computers or data mining technology. What happens on computers, however, is not much different than the way people build models. Computers are loaded up with lots of information about a variety of situations where an answer is known and then the data mining software on the computer must run through that data and distill the characteristics of the data that should go into the model. Once the model is built it can then be used in similar situations where you don't know the answer. For example, say that you are the director of marketing for a telecommunications company and you'd like to acquire some new long distance phone customers. You could just randomly go out and mail coupons to the general population - just as you could randomly sail the seas looking for sunken treasure. In neither case would you achieve the results you desired and of course you have the opportunity to do much better than random - you could use your business experience stored in your database to build a model. As the marketing director you have access to a lot of information about all of your customers: their age, sex, credit history and long distance calling usage. The good news is that you also have a lot of information about your prospective customers: their age, sex, credit history etc. Your problem is that you don't know the long distance calling usage of these prospects (since they are most likely now customers of your competition). You'd like to concentrate on those prospects who have large amounts of long distance usage. You can accomplish this by building a model. Table 2 illustrates the data used for building a model for new customer prospecting in a data warehouse An Architecture for Data Mining-To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysisinalargedatawarehouse.



II. RELATED WORK

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 entities .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. For example, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries.

DRAWBACKS OF EXISTING SYSTEM

-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, apple is frequently associated with computers on the web. However, this sense of apple is not listed in most general-purpose thesauri or dictionaries In information retrieval, the delay is the one of the main problem to retrieve the document in sequence manner. In web application, the information retrieval is very very difficulties to manage the collusion on the retrieve information. In existing process, the wordcounts of function are not efficient.

III. PROPOSED WORK

To 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. Word counts and snippets are two useful information sources provided by most web search engines. Word count of a query is an estimate of the number of words that contain the query words. In general, word count may not necessarily be equal to the word frequency because the queried word might appear many times on one word.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.

ADVANTAGES OF PROPOSED SYSTEM

- We proposed a semantic similarity measure using both word counts and snippets retrieved from a web search engine for two words. Four word co-occurrence measures were empirical results show that PROMISH is faster than based state of the art tree technique having performance improvements of multiple order of magnitude computed using word 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 word counts-based co-occurrence measures and lexical pattern clusters were used to define features for a word pair.

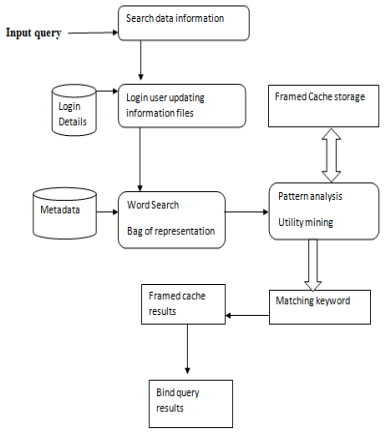


Figure 2.Architecture of propose

IV. CONCLUSION-

In this paper we propose a problem of top k nearest keyword search in multidimensional datasets .We develop an exact (PROMISH-E)and an approximate (PROMISH-A) method. We designed a noel index based an random projections hashing index is used to find subset of points containing the true results .we also proposed an efficient solution to query results from subset of data points .our.

REFERENCES

1. W.Li and c,x.Chen."Efficient data modeling and querying system for multi-dimensional spatial data,"in GIS,2008,pp.58:1-58:4.
2. D.Zhang,B.C,Ooi,andA,K,H,Tung,"Locatin g mapped resources in web 2.0," in ICDE,2010,pp.521-532.
3. V.Singh,S.Venkatesha,andA,K.Singh,"Geo-clustering of images with missing geotags," in GRC,2010,pp.420-425.

Authors

Mrs.D.Maladhy Assistant Professor, Department of Information Technolgy.Rajiv Gandhi College Of Engineering & Technology

Ms.RisvaanaBegum Final Year Student , Department of

Ms.SivaPriya Final Year Student , Department of Information Technolgy.Rajiv Gandhi College Of Engineering & Technology

Ms.SuryaDharshini Final Year Student , Department of Information Technolgy.Rajiv Gandhi College Of Engineering & Technology