Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 4, April 2013, pg.160 – 163

SURVEY ARTICLE

A Survey on Web Usage Mining with Fuzzy c-Means Clustering Algorithm

Chhaman Lakheyan[1], Usvir kaur[2]

[1]Student of masters of technology Computer Science, Department of Computer Science Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

[1] lakheyanchhaman@yahoo.in; [2] Usvirkaur@gmail.com

Abstract— Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. Based on the different emphasis and different ways to obtain information, web mining can be divided into two major parts: Web Contents mining and Web Usage Mining. Web Contents mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. In this paper we have we conduct survey on web usage mining along with its functionalities and Fuzzy c means algorithm for the retrieval of data for the search engine.

Key Terms: - Web usage mining; User/Session identification; Web Recommender; Web log; Fuzzy c algorithm

I. INTRODUCTION [9]

Web Usage Mining: -- Pattern Discovery and its applications

With the explosive growth of information sources available on the World Wide Web and the rapidly increasing pace of adoption to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses. A web site is the most direct link a company has to its current and potential customers. The companies can study visitor's activities through web analysis, and find the patterns in the visitor's behavior. These rich results yielded by web analysis, when coupled with company data warehouses, offer great opportunities for the near future.

Why Web Usage Mining-In this paper, we will emphasize on Web usage mining. Reasons are very simple: With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business. The growth of some E-businesses is astonishing, considering how E-commerce has made Amazon.com become the so-called "on-line Wal-Mart". Unfortunately, to most companies, web is nothing more than a place where transactions take place. They did not realize that as millions of visitors interact daily with Web sites around the world, massive amounts of data are being generated. And they also did not realize that this information could be very precious to the company in the fields of understanding customer behavior, improving customer services and relationship, launching target marketing campaigns, measuring the success of marketing efforts, and so on.

How to perform Web Usage Mining?

Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data (as discussed below). Web server log files were used initially by the webmasters and

system administrators for the purposes of "how much traffic they are getting, how many requests fail, and what kind of errors are being generated", etc. However, Web server log files can also record and trace the visitors' on-line behaviors. For example, after some basic traffic analysis, the log files can help us answer questions such as "from what search engine are visitors coming? What pages are the most and least popular? Which browsers and operating systems are most commonly used by visitors?"

Web log file is one way to collect Web traffic data. The other way is to "sniff" TCP/IP packets as they cross the network, and to "plug in" to each Web server.

After the Web traffic data is obtained, it may be combined with other relational databases, over which the data mining techniques are implemented. Through some data mining techniques such as association rules, path analysis, sequential analysis, clustering and classification, visitors' behavior patterns are found and interpreted.

The above is the brief explanation of how Web usage is done. Most sophisticated systems and techniques for discovery and analysis of patterns can be placed into two main categories, Pattern Analysis Tools and Pattern Discovery Tools, as discussed with the Diagram.

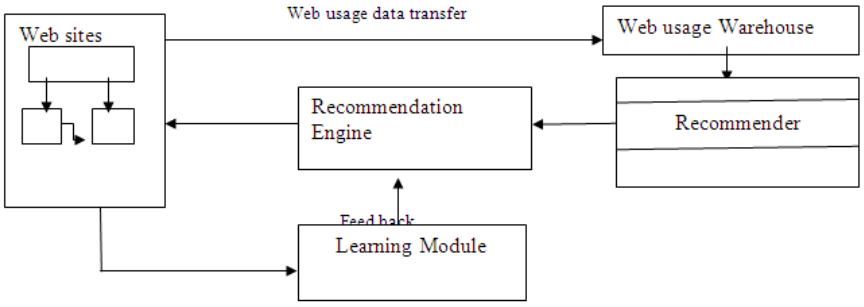

Fig [1] represents the basic functionality of web usage mining

II. CLASSIFICATION OF WEB MINING

Web Structure Mining

Web structure mining aims to discover useful knowledge from hyperlinks, which represent the structure of the Web. Hyperlink is a link that exists in a web page and refers to another region in the same web page or another web page [2]. The most popular application of web structure mining is to calculate the importance of web pages. This kind of application is used in Google search engine to order its search results. A web structure mining algorithm, Page Rank, is invented by Google founders: Larry Page and Sergey Brin. Web structure mining can also be applied to cluster or classify web pages (Gomes and Gong, 2005).

Web Content Mining

Web content mining extracts or mines useful information or knowledge from web page contents. There are two categories of web content mining: structured data extraction and text mining. The idea of structured data extraction is that many web site display important information retrieved from their database using some fixed templates. [2]

Web Usage Mining

Web usage mining aims to capture and model behavioral patterns and profiles of users who interact with a web site. Such patterns can be used to better understand the behaviors of different user segments, to improve the organization and structure of the site, and to create personalized experiences for users by providing dynamic recommendations of products and services. Unlike two previous web mining tasks, the primary data source for web usage mining is web server access log, not the web pages.

III. TEXT MINING

Text mining is defined as the automatic discovery of new, previously unknown, information from unstructured textual data. This process is done in three steps: information retrieval, information extraction and data mining. A primary reason for using data mining for biomedical text is to assist in the analysis of collections of the available biomedical text. Biomedical data is vulnerable to co linearity because of unknown interrelations. [4]

Before data mining algorithms can be used, a target data set will be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns. Pre-process is essential to analyze the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data. [4]

The biomedical data available with us is first put into a data warehouse. Before putting the data in the data warehouse the keyword extraction algorithm is used to find out the keywords from the full text. This keyword extraction uses partial parser to extract entity names (gene, protein names etc). This parser uses linguistic rules and statistical disambiguate to achieve greater precision. [2]

The data is then organized into clusters. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The clusters will be created based on the keywords extracted from our biomedical text. These clusters will be created using fuzzy C mean algorithm. The fuzzy c-means algorithm is one of the most widely used soft clustering algorithms. It is a variant of standard k-means algorithm that uses a soft membership function. Fuzzy C-Means (FCM) clustering algorithm is one of the most popular fuzzy clustering.[4]We have different algorithms but we are constricting mainly on Fuzzy C means algo and its issue which can solved.

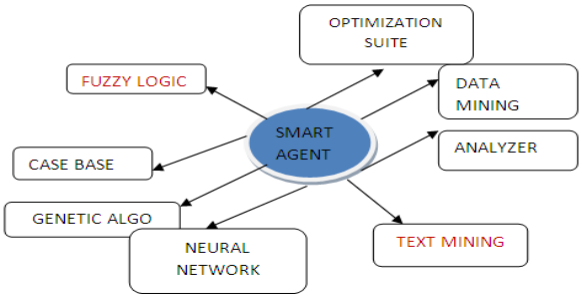

Fig [2] .How Fuzzy logic are interlinked with various disciplines

IV. BASIC FUZZY C-MEAN (FCM) ALGORITHM LOGIC [4]

Text mining is defined as the automatic discovery of new, previously unknown, information from unstructured textual data. This process is done in three steps: information retrieval, information extraction and data mining. A primary reason for using data mining for biomedical text is to assist in the analysis of collections of the available biomedical text. Biomedical data is vulnerable to co linearity because of unknown interrelations. The analysis in this paper will be augmented by using experiment-based approach.

Before data mining algorithms can be used, a target data set will be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns. Pre-process is essential to analyze the multivariate datasets before clustering or data mining. The target set is then cleaned. Cleaning removes the observations with noise and missing data.

The biomedical data available with us is first put into a data warehouse. Before putting the data in the data warehouse the keyword extraction algorithm is used to find out the keywords from the full text. This keyword extraction uses partial parser to extract entity names (gene, protein names etc). This parser uses linguistic rules and statistical disambiguity to achieve greater precision.

The data is then organized into clusters. Clustering is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. The clusters will be created based on the keywords extracted from our biomedical text. These clusters will be created using fuzzy C mean algorithm. The fuzzy c-means algorithm is one of the most widely used soft clustering algorithms. It is a variant of standard k-means algorithm that uses a soft membership function. Fuzzy C-Means (FCM) clustering algorithm is one of the most popular fuzzy clustering algorithms. FCM is based on minimization of the objective function Fm(u, c):

$$F_m(u,c) = \sum_{k-1}^{n} \sum_{i-1}^{c} (u_{ik})^m d^2(x_k, c_i)$$

FCM computes the membership *uij* and the cluster centers *cj* by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m}$$

where *m*, *the fuzzification factor* which is a weighting exponent on each fuzzy membership, is any real number greater than 1, *uij* is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *cj* is the dimension center of the cluster, *d2(xk,ci)*is a distance measure between object xk and cluster center ci, and ||*|| is any norm expressing the similarity between any measured data and the center.

The FCM algorithm involves the following steps:

1. Set values for c and *m*
2. Initial membership matrix U= [*uij*], which is U(0) (|i| = number of members, |j| = number of clusters)

3. At *k-step:* calculate the centroids for each cluster through equation (2) if $k \neq 0$. (If k=0, initial centroids location by random)

4. For each member, calculate membership degree by equation (1) and store the information in U(k)

5. If the difference between U(k) and U(k+1) less than a certain threshold, then STOP; otherwise, return to step 3.

V. PROPOSED FUZZY C MEANS ALGORITHM [4]

The proposed algorithm will take a complete list of all the biomedical articles and the output will be the XML files containing the clusters created using fuzzy c mean algorithm on keywords.

Input: List of full text biomedical articles.

Output: XML files containing the created clusters.

Algorithm

1. Read the next article in the list of biomedical text
2. Read the full text article
3. Extract the keywords from the article using KEA algorithm
4. Refer to the biomedical lexicon and discard the irrelevant keywords
5. Put the data in following relation so that the full text can be retrieved later using keywords only

Article UID	Article Name	Keywords	Full text	Source

6. Go to step 1 and repeat till all the articles in the list of biomedical articles are processed.
7. Use the fuzzy c-means algorithm to create clusters on keywords.
8. Save the article clusters in form of an XML file.

VI. CONCLUSION AND FUTURE WORK

Web mining can provide companies managerial insight into visitor profiles, which help top management take strategic actions accordingly. Also, the company can obtain some subjective measurements through Web Mining on the effectiveness of their marketing. Campaigning or marketing research, which will help the business to improve and align their marketing strategies timely.

For example, the company may have a list of goals as following:

- Increase average page views per session;
- Increase average profit per checkout;
- Decrease products returned;
- Increase number of referred customers;
- Increase brand awareness;
- Increase conversion rate (checkouts per visit).

These issues can be addressed with Fuzzy c means algorithm for text mining but still a lot of work can be done in this field to improve the efficiency of the search engine.

REFERENCES

[1] Raymond Kosala and Hendrik Blockeel. “Web Mining Research: A Survey”, A book.feb. 2005.

[2] Tamanna Bhatia, “Link Analysis Algorithms For Web Mining” ISSN :2229 - 423 (Print) |ISSN : 0976 - 8491 (Online) IJCST Vol. 2, Issue 2, June 2011.

[3] J.M. Kleinberg.” Authoritative sources in a hyperlinked environment” Journal of the ACM(JACM), 46(5):604{632, 1999.

[4] Sumit Vashishta, Dr. Yogendra Kumar Jain” Efficient Retrieval of Text for Biomedical Domain using Data Mining Algorithm”(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 4, 2011

[5] Kalyani M Raval “ International Journal of Advanced Research in Computer Science and Software Engineering”Volume 2, Issue 10, October 2012

[6] Hong-Jie Dai1,2, Yen-Ching Chang1 et al “New Challenges for Biological Text-Mining in the Next Decade” journal of computer science and technology 25(1): 169–inside back cover Jan. 2010.

[7] M.Lavanya & Dr.M.Usha Rani “Vision based deep web data extraction for web document clustering” global journal of computer science and technology volume 12 issue 5 version 1.0 March 2012.

[8] Steve Russell ,Intelligence for Business at E-Speed,” International journal of advanced computing” 1999. http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=1978

[9] Dr. Larry R. Harris Database Access Over the Web: Extending the Wire ,” journal of computer science and technology “25(1): 169–inside back cover Jan. 2009. http://www.dmreview.com/master.cfm?NavID=29

[10] Mary Garvin Data Mining and the Web: What They Can Do Together”, Journal of the ACM(JACM”2007. http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=420

