INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 3 ISSUE 4 (2017) PAGES 2266 – 2270
RECEIVED : 05.07.2017 PUBLISHED : 06.08.2017 Aug 06, 2017

WEB STRUCTURE MINING ASSOCIATE BETWEEN DSPACE LOG FILES

P.Nivetha[1] , D.B.Shanmugam[2], S.Munusamy[3]

[1]M.Phil, Research Scholar, Dr.M.G.R.Chockalingam Arts College, Arni.

[2]Associate Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni,

[3]Assistant Professor, Department of MCA, Sri Balaji Chockalingam Engineering College , Arni.

ABSTRACT

Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs. In this thesis, the DSpace log files have been preprocessed to convert the data stored in them into a structured format. Thereafter, the general procedures for Bot-removal and session-identification from a web log file, have been written down with certain modifications pertaining to the DSpace log files, in an algorithmic form. This algorithm is based on the artificial immune system model and uses this model to learn and extract information present in the web data server logs this algorithm has been duly modified according to Website structure.

Keywords: DSpace, Website structure, Log files.

1. INTRODUCTION

Web Usage Mining is a part of Web Mining, which, in turn, is a part of Data Mining. As Data Mining involves the concept of extraction meaningful and valuable information from large volume of data, Web Usage mining involves mining the usage characteristics of the users of Web Applications. This extracted information can then be used in a variety of ways such as, improvement of the application, checking of fraudulent elements etc. Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned.

The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information. In this project, we have taken up a small part of the Web Usage Mining process, which involves the Preprocessing, User Identification, Bot-removal and Analysis of the Web Server Logs. In the current era, we are witnessing a surge of Web Usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has what made Web Mining one of the hot topics in the field of Information Technology.

2. RELATED WORK

In Web Usage Mining, data can be collected in server logs, browser logs, proxy logs, or obtained from an organization's database. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation. These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. Information about the request, including client IP address, request date/time, page requested, HTTP code, bytes served, user Agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information. These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. A statistical analysis of the server log may be used to examine traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, adequate hosting resources and the fine tuning of sales efforts can be aided by analysis of the web server logs. Marketing departments of any organization that owns a website should be trained to understand these powerful tools. A message given to a Web browser by a Web server. The browser stores the message in a text file called cookie. The message is then sent back to the server each time the browser requests a page from the server. The main purpose of cookies is to identify users and possibly prepare customized Web pages for them. When you enter a Web site using cookies, you may be asked to fill out a form providing such information as your name and interests. This information is packaged into a cookie and sent to your Web browser which stores it for later use. The next time you go to the same Web site, your browser will send the cookie to the Web server.

3. LINK ALGORITHM

As the use of Web is increasing more day by day, the web users get easily lost in the web’s rich hyper structure. The main aim of the owner of the website is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach. In this thesis we discuss and compare the commonly used algorithms i.e. Page Rank, Weighted Page Rank and HITS. He World Wide Web (WWW) is rapidly growing on all aspects and is a massive, explosive, diverse, dynamic and mostly unstructured repository of data. Till now, WWW is the huge information repository referenced for Knowledge. Web mining techniques provides the additional information through hyperlinks where different documents are connected. We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph. There are number of algorithms proposed based on link analysis.

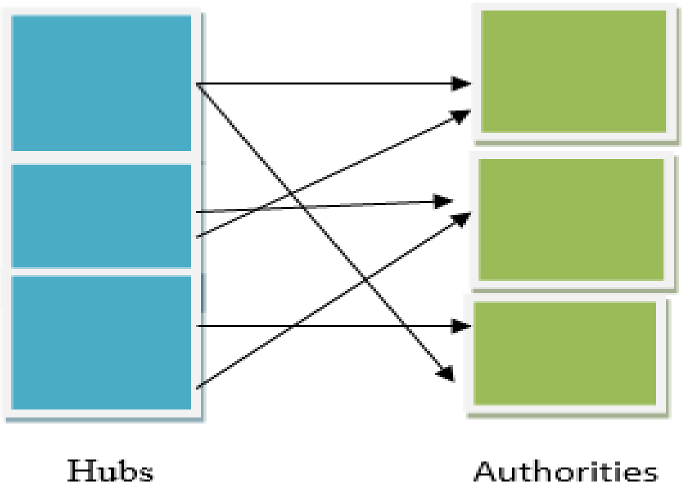

Fig.1. Hubs and Authorities

In the area of software, data mining technology has been considered as important mean for discovering patterns and trends of large amount of data. So, this approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. Data mining is a computational intelligence discipline which/has emerged as a effective tool for data analysis, new KDD and good decision making. The raw and unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the work of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar.

4. TECHNIQUES FOR CLUSTERING

In the area of software, data mining technology has been considered as important mean for discovering patterns and trends of large amount of data. So, this approach is basically used to extract the unknown pattern from the large set of data for business as well as real time applications. Data mining is a computational intelligence discipline which/has emerged as a effective tool for data analysis, new KDD and good decision making. The raw and unlabeled data from the large volume of dataset can be classified initially in an unsupervised fashion by using cluster analysis i.e. clustering the work of a set of observations into clusters so that observations in the same cluster may be in some sense be treated as similar. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Clustering is the main task of exploratory data mining, and basically common technique for statistical data analysis used in many fields, like machine learning, pattern discovery, information retrieval and biomedical data information. In a k-medoids methods a cluster is represented by one of its points. This is an easy solution because it covers any attribute type and medoids are insensitive to outliers because peripheral cluster points do not affect them. When medoids in this algorithm are selected, clusters are known as subsets of points or values near to respective medoids, and the objective function is defined

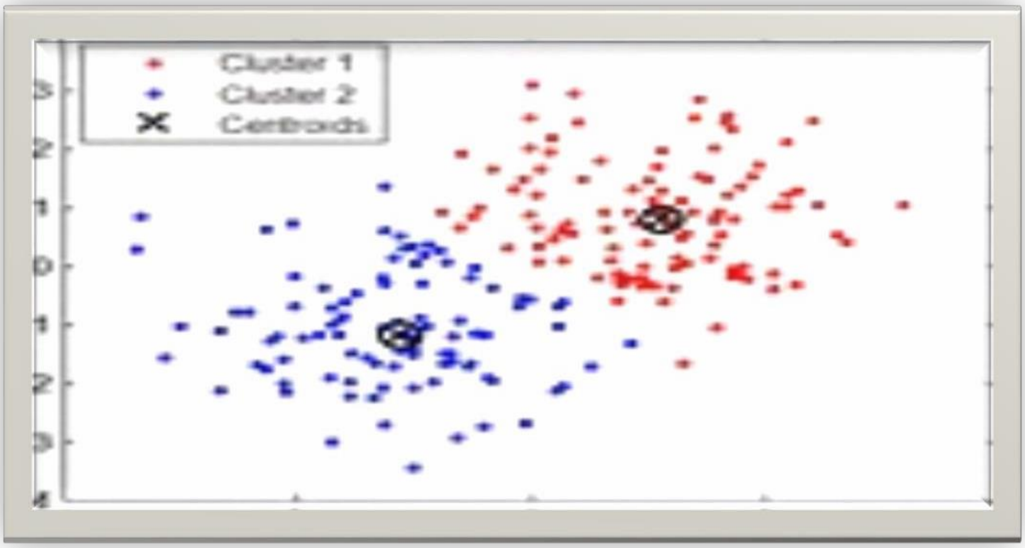

Fig.2. Groups Involved Clustering

as the averaged distance. PAM (Partitioning around Medoids) algorithm was one of the first *k*-medoids algorithms.

5. ANALYSIS

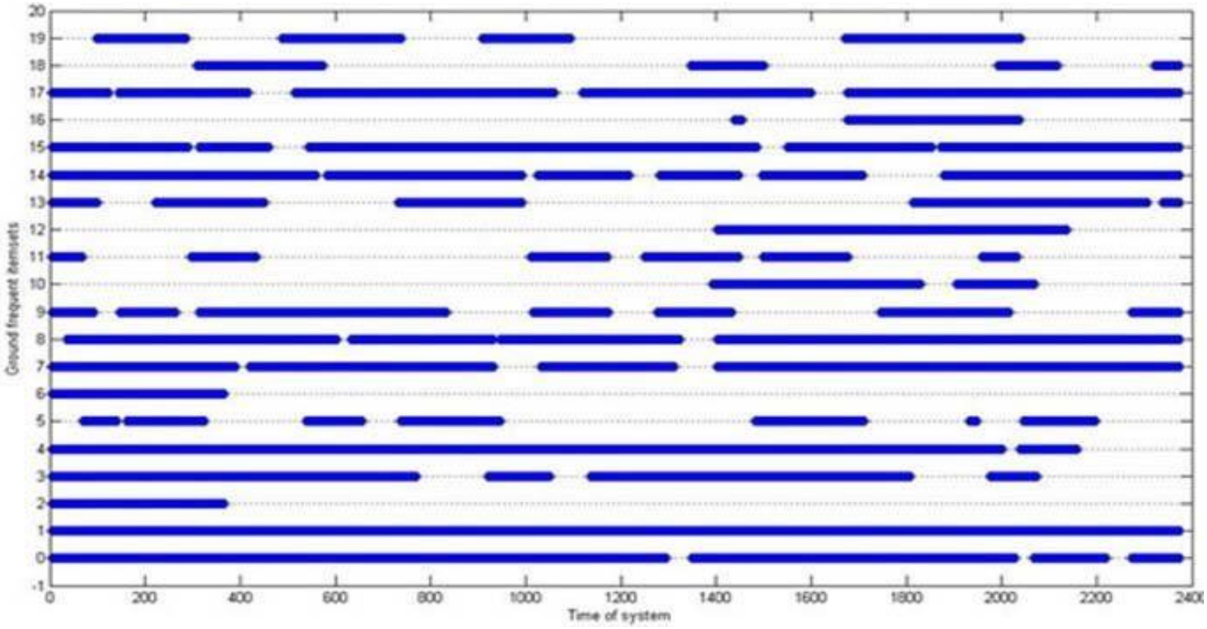

Fig.3. Output Profile

This chart produced the level of storage. Even those three algorithm are different platform, but the task are only same. Link and clustering are less slow and inefficient compare to EIN_WUM.This Thesis clearly produced that result from the previous table. EIN_WUM are advanced and comfort zone between the previous two from the web mining area. Now a day millions of people using web. Day by day increasing web browsing user.so we following EIN_WUM algorithms from the extraction and storage process, it may produce the good result. Only the aim of the thesis to explain to introduced smart on

CONCLUSION

In this Thesis a new robust algorithm with several novelties such as incorporating danger theory, new immune network model and directed mutation is presented. This algorithm is able to learn frequent patterns of Web usage data in single pass of input data. The main factor of the algorithm that has made it capable of learning the frequent patterns in single pass is its rich and manageable immune network The proposed methods were successfully tested on the log files for bot removal and user sessions identification. The results which were obtained after the analysis wares at is factory and contained valuable information about the Log Files. The subjective interpretation and efficient of the EIN-WUM algorithm produced results which depicted the usage patterns (frequently accessed contents) of the WEBSITE.

REFERENCES

[1] Adel T. Rahmani and B. Hoda Helmi, EIN-WUM an AIS-based Algorithm for Web

[2] A. Secker. Artificial Immune Systems for Web Content Mining: Focusing on the Discovery of Interesting Information. University of Kent in Canterbury, UK, 2006.

[3] A. Secker and A. A. Freitas and J. Timmis. A Danger Theory Inspired Approach to Web Mining. In Proc. 1st Int. Conf. on Artificial Immune Systems (ICARIS), Lecture Notes in Computer Science 2787, pages 156–167. Springer-Verlag, 2003.

[4] Bain Tony SQL Server 2000 Data Warehouse and Analysis Services. Beijing: China Electric Power Press, 2003, p443-470.

[5] B. H. Helmi and A. T. Rahmani. An AIS Algorithm for Web Usage Mining with Directed Mutation. In Proceedings of the World Congress on Computational Intelligence (WCCI'08) (To be published). 2008.

[6] C.Bin, Lin jie, de, Liu ming and xiang, Chen Data mining and OLAP Theory & Practice [M]. Beijing: Tsinghua University Press, 2003, p194- 244.

[7] Dunham., Margaret H., Data Mining Introductory and Advanced Topics. Beijing: Tsinghua University Press, 2003, p195-220.