IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.
www.ijiset.com
ISSN 2348 – 7968

Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern

Wasvand Chandrama, Prof. P.R.Devale, Prof. Ravindra Murumkar

Department of Information technology, Research scholar of Bharati Vidyapeeth University College of Engineering, Pune, Maharashtra 411046, India.,

Department of Information technology, Bharati Vidyapeeth University college of Engineering, Pune, Maharashtra411046, India,

Department of Information technology. Pune University, Pune Institute of Computer Technology Pune, Maharashtra411046, India.

Abstract

With the explosive growth of the World Wide Web and the rapidly increasing speed of adoption to Internet commerce, the Internet has evolved that contains or dynamically generates information that is beneficial to E-businesses. Web mining is used to extract the content of web site where web usage mining helps to define user pattern. This paper works on data preprocessing of web log file. The model enables the administrator to access the web log file and perform data preprocessing on it i.e. data cleaning on web log file and identifying user navigational pattern. By using classification algorithm we identify user interested web site.

Keyword: *Web usage mining, data preprocessing, classification, pattern discovery, clustering.*

1.Introduction

The World Wide Web is a repository of web pages that provides the lot of information to the internet users. For internet users the information available on web has become a vital source. Because of these reasons, there is increasing growth and complexity of websites available on internet, the size of web is large. A web site is the link the customer to company. The companies can study visitor's activities through web analysis, and find the patterns.

Web mining is broadly defined as discovering and analysis of useful information from the World Wide Web. Web mining divided into three parts: Web Contents Mining, Web structure mining and Web Usage Mining. Web Contents Mining can be as the automatic search and extraction of information and resources available from number of sites and on-line databases though search engines or web spiders. Web Usage Mining can be as the discovery and analysis of access patterns of user, through the mining of log files. The output of the WUM can be used in web personalization, improving the system performance, site modification, usage characterization etc.

Web log file is a server log file which is a basic data sources in Web usage mining, in which it contain - access logs of the web server.. The important task in the WUM is Data Preprocessing phase. It consists of data cleaning, user identification, session identification, path completion. Data preprocessing is used to clean the irrelevant data from log file so it can be provide to the pattern discovery to identify the user pattern.

ISSN 2348 – 7968

2. Related Work

User navigational patterns using SOM [1] these papers identify the user navigational pattern that useful for e-commerce site and web site owners to attract new customers, tracking online customer. For preprocessing the log file they used programming logic. After the preprocessing, for clustering the similar types of web site used Kohonen's SOM i.e. Self organizing Map and extract the most frequent patterns.

SOM is like it not needs any complex mathematical relation and it just extracting the pattern of any length and different patterns can be extracted as per the occurrence of page group in cluster. But the drawback is that the patterns that have a least access in hours of day do not exist.

Behavior of female students by analyzing log files [2] here they to get detail about the internet usage and the behavior of female student. It defining the pattern of different department teacher's and student, to improve the teaching methods with respect to internet usage and gender differentiation. By using time spent on Internet cannot explain in detail the difference between the excellent student and Weak student. Need some more attribute with respect to useful visited website related to academic. It uses classification to categorize student.

Customer behavior using web usage mining in E-commerce.[3] E-commerce is commonly used for online shopping so need to understand the customer requirement, what they want to buy. The input is enterprise server log file, including the field like client registration and some extra information left by client through browser and internet. For analyzing data used some term like correlation regulation discovery for the next page predicting for the client want to visit. Apriori algorithm is used for frequent pages mining that help site designer to get better site structure. To observe the maximum forward path and frequently visited path by client they used directed tree model. Clustering used K-mean algorithm.

Improving methods of preprocessing in Web Log Mining.[4] it provide improved filtration method of frame page, the filtering algorithm is judging the each page is Frame or SubFrame and if it is SubFrame, deletes it one by one. Filtration it gets recovery in path supplement. Decision tree is created by using ID3 algorithm and divide and rule method that improves the efficiency and filtration method.

Web log cleaning for mining of Web Usage Patterns [5] focus is on data preprocessing by providing field Extraction and Data Cleaning algorithm. Field Extraction is used to separating fields from log file that removes unnecessary data, Data cleaning algorithm is also used.

3. System Architecture

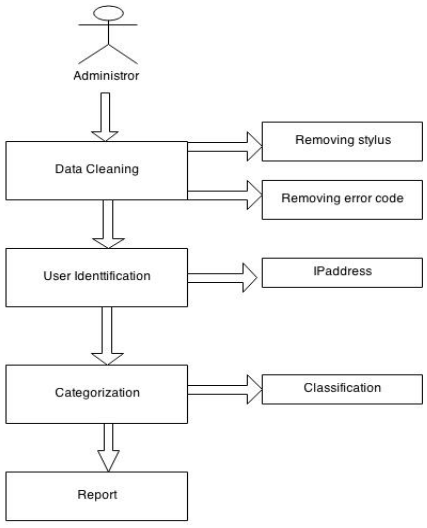

Fig. 1: System architecture

ISSN 2348 – 7968

Here proxy server log file is taken as a data source which is collage proxy server log file. Our application is implementing for collage purpose to know the top most visited site and user navigational pattern. Admin is main user of our application. Firstly data preprocessing is done on log file and then categorized the file for generating user navigational pattern.

Proposed Work

For identifying the user navigational pattern we proposed work for an efficient method of data preprocessing for data cleaning and identifying user navigational pattern has consist of different approaches and methods. To get better results of user category we used classification in this model. The model has consisted of functional blocks such as:

3.1 Data Preprocessing

The data cleaning process is used here to removing the unnecessary and duplicate data from log file. The objective of WUM is to have a clear picture of the web user requirement or behavior; hence it required the removing of files having the suffixes such as, jpeg. jpg, gif, css, cgi, etc, error codes like 401,404 are not relevant and not useful for mining.

3.2 Defining User by using IPaddress

In user identification, IPadress helps to represent unique user. From this we identified which IPaddress uses which web site or web pages are visited.

3.3 Categorizing visited URL

After preprocessing web log file the text file generated, which is used for categorization. Classification is used for categorization. Categorization implements the objects that are grouped into categories for some specific purpose. Using Naïve Bayes classifier we defined the category of visited site.

Algorithm for text categorization.

- Select the IPaddress for which categorization of web site is perform. Select words that have frequency of more than once occurrence in the web page.
- View the frequent words as in word sets.
- Search for matching the word sets or its subset containing items more than one in the list of word sets collected from training data with that of subset of frequent word set of new document.
- Collect the probability value matching with word set for each target class.
- Calculate the probability values for each target class.

3.4 Generating report:

After categorization of each URL, generating report which contains the graphical representation of visited URL. Top most visited site like educational site, finance site, social media site are plotted in graph. Report is generated hourly, daily, weekly, monthly basis

ISSN 2348 – 7968

4. Implementation result

After the data cleaning process is done on web log file, it shows how much memory space is will be utilize and maintain quality of it. After data preprocessing the text file uses the classification to categorize the visited URL of web site. And generating result as in graphical report format.

Removing irrelevant data:

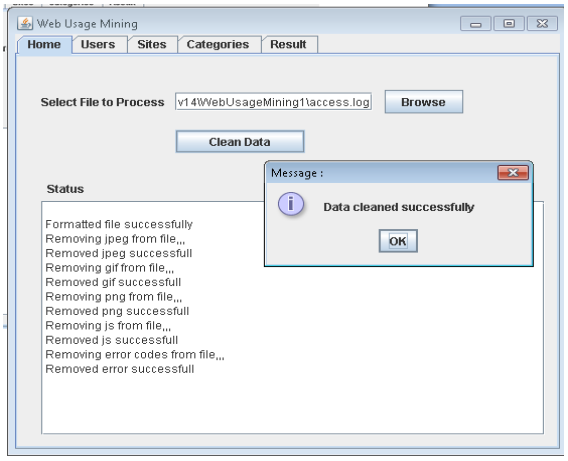

Fig 2: Data cleaning

Unique user with IPaddress:

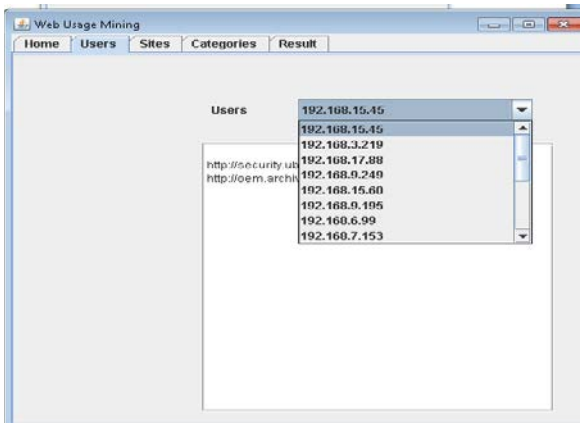

Fig 3: User identification

Categorizing shown:

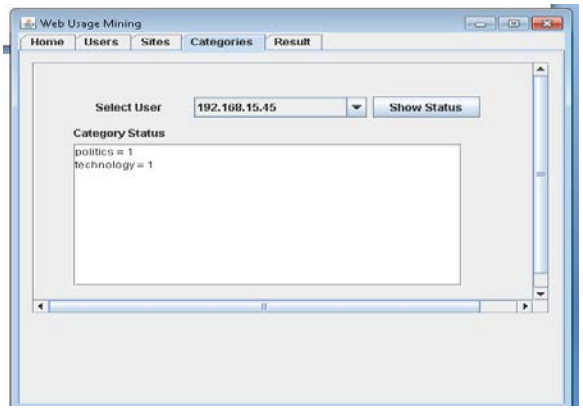

Fig 4: Categorization

Result:

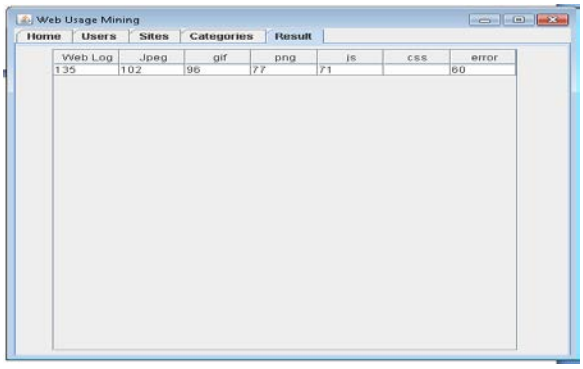

Fig 5: Removing stylus

5. Conclusion:

To get user interested web site or user behavior data preprocessing for data cleaning and defining the user pattern is most useful. This model helps the for data cleaning and categorizing the web site so understand the user interested web site and per user need or request. By using Navie Bayes algorithm of classification it helps easily to categorize users visited web site and provide better efficiency and performance as compare to other algorithm.

ISSN 2348 – 7968

6. References

[1] Kobra Etminani, Amin Rezaeian Delui,Noorali Raeeji Yanehsart, Modjtaba Rouhani.

"Web Usage Mining: Discovery Of The User's Navigational Patterns Using Som'', IEEE 2009

[2] Rozita Oskouei "Behavior Mining Of Female Students By Analyzing Log Files". IEEE 2010

[3] Mahendra Pratap Yadav, Mhd Feeroz, Vinod Kumar Yadav,"Mining The Customer Behavior Using Web Usage Mining In E-Commerce" IEEE 2012

[4] Huaqiang Zhou, Hongxia Gao, Han Xiao, "Research On Improving Methods Of Preprocessing In Web Log Mining", IEEE 2010.

[5] Theint Theint Aye, "Web Log Cleaning For Mining Of Web Usage Patterns", IEEE 2011

[6] K Sudheer Reddy, "Preprocessing of web server logs- an illustrative approach for effective usage mining", ACM 2012

[7] Ida Mele, "WUM for enhancing search-result delivery and Helping user to find interesting web content" ACM 2013

[8] "An efficient algorithm for data cleaning of Log file using file extensions"

[9] Bina kotiyal,ankit kumar. Bhaskar pant, R.H. goudar, shivaji chauhan and sonam junee, "User behavior analysis in web log through comparative study of Eclat and Apriori" IEEE 2012

[10] K.Santhisree, Dr A.Damodaram, S.Appaji, D.Nagerjunadevi, "Web usage data clustering using Dbscan algorithm and set Similarities" IEEE 2010

[11] tasawar hussain, dr. sohail asghar," A hierarchical Cluster based preprocessing methodology for web usage mining" IEEE.

[12] jose Roberto, Geraldo xexeco; "An architecture for web usage mining"

[13] K.sudheer reddy, m. kantha reddy,v,sitaramula,"An effective data preprocessing method for web usage mining" IEEE conference 2013

7. Biography:

Wasvand chandrama s. pursuing M.tech in Information Technology from Bharati Vidyapeeth Deemed University collage of enginering, received the B.Tech degree in computer engineering in 2011.

P.R.Devale pursuing PH.D in natural language processing and received M.E. in Information Technology also working as a Professor in Bharti Vidhyapeeth Deemed University Collage of Engineering.

Ravi Murumkar received M.E. in Information Technology and working as Assistant Professor in Pune Insititute of Conputer Technology.