An Effective Web Usage Analysis using Fuzzy Clustering

[1] P.Nithya, [2] P.Sumathi

[1] Doctoral student in Computer Science, Manonmanaiam Sundaranar University, Tirunelveli
[2] Assistant Professor, PG & Research Department of Computer Science, Govt.Arts College, Coimbatore.

[1] nithi.selva@gmail.com

ABSTRACT

Nowadays, internet is a useful source of information in everyone's daily activity. Hence, this made a huge development of World Wide Web in its quantity of interchange and its size and difficulty of websites. Web Usage Mining (WUM) is one of the main applications of data mining, artificial intelligence and so on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. Since WUM directly involves in large range of applications, such as, e-commerce, e-learning, Web analytics, information retrieval etc. Web log data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. There are varieties of problems related with the existing web usage mining approaches. Existing web usage mining algorithms suffer from difficulty of practical applicability. So, a novel research is very much necessary for the accurate prediction of future performance of web users with rapid execution time. WUM consists of preprocessing, pattern discovery and pattern analysis. Log data is characteristically noisy and unclear, so preprocessing is an essential process for effective mining process. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots. Fuzzy algorithm is a distinctive clustering algorithm available to cluster unlabeled data that produces both membership and typicality values during clustering process. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique.

Keywords: *Preprocessing, Data Cleaning, Path Completion, Travel Path set, Content Path Set*

1. INTRODUCTION

As we have millions of data on the internet, the WWW has turned into a network of data without any proper structural organization. Due to the heterogeneous nature of data it is very difficult to search for a particular data. This particular nature of data makes the users to feel to be overloaded with data. In this internet era e-business and web marketing are the most fascinating areas; here finding out the customer requirements plays a vital role to improve the business. In this regard web user identification or personalization is very essential thing. Personalization involves using technology to accommodate the differences between individuals.

In general the web related data can be categorized into three types, they are i) its structure (how the data is organized in the web page), ii) its content (the data in the web page), iii) usage (how the users used the content) and the data mining methods that are used to handle these data are web structure mining, web content mining, web usage mining respectively.

Web usage mining [16, 17] consists of three main steps. Weblog file is usually given as input.

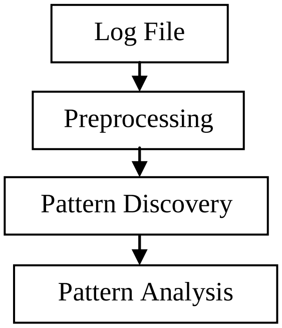


Fig 1.1: Steps in Web Usage Mining

Preprocessing is a significant step as it takes maximum time in mining process. Initially the web data goes for preprocessing in which the following tasks will be carried out, data cleaning, identifying the users, identifying the session, path completion and transaction construction. Data cleaning is removing the unnecessary data for mining. It includes,

- Exclusion of local and global noise.
- Eliminating format information and videos.
- Amputation of records with the failed HTTP status code.
- Cleaning the Robots.

User identification is to relate the page references with similar IP address with diverse users. Splitting of user’s page references into sessions is referred as session identification. Neglected page references in the session can be loaded using path completion techniques. To better understand the user’s interest classification technology is being used. The next step in web usage mining is extracting the knowledge through various data mining techniques like classification, clustering, association rule mining in the preprocessed data. The last step in mining process is to convert the information into knowledge by analyzing the patterns obtained from the previous step using different tools. The conventional method used for pattern analysis is SQL Queries. But in the proposed method path completion technique is used to fill the missing pages to construct the transactions in the preprocessing step.

2. RELATED WORKS

The discovery of the users' navigational patterns using SOM is proposed by Etminani et al., [1]. Jianxi et al., [2] presented a Web usage mining technique based on fuzzy clustering in Identifying Target Group. Nina et al., [3] suggests a complete idea for the pattern discovery of Web usage mining. Wu et al., [4] given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. The author discovers the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases. Aghabozorgi et al., [5] proposed the usage of incremental fuzzy clustering to Web Usage Mining. Rough set based feature selection for web usage mining is proposed by Inbarani et al., [7]. Jalali et al., [8] put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. For providing the online prediction effectively, Shinde et al., [9] provides a architecture for online recommendation for predicting in Web Usage Mining System.

Baraglia et al., [15] proposed a Web usage mining (WUM) system, called SUGGEST, which continuously creates the suggested connections to Web pages of probable importance for a user. Lee et al., [19] put forth a Web Usage Mining technique based on clustering of browsing characteristics.

3. METHODOLOGY

Web log data preprocessing is a complex process and takes 80% of total mining process. Log data is pretreated to get reliable data. The aim of data preprocessing is to select essential features clean data by removing irrelevant records and finally transform raw data into sessions.

3.1 The Fuzzy Clustering Algorithm

The basic theory about fuzzy clustering analysis based o fuzzy equivalent relation is that because fuzzy corresponding R is a subset of U×U (universe of fuzzy set), when we are using λ -threshold, the subset of U×U will be just an ordinary equivalent relation and the objects in the U will be classified. When λ is reduced from 1 to 0, the classification will become gradually merging, forming a dynamic clustering dendritic diagram therefore. Thus, the establishment of fuzzy relation R is a key step of fuzzy analysis method. Many initial objects are expressed by variables (also called attribution), for example, an employee could be expressed by age, gender, wage, position and so on. This relation can be transformed into a relation table or n×p matrix.

Fuzzy Clustering is an iterative algorithm. The aim of Fuzzy Cluster is to find cluster centers (centroid) that minimize a dissimilarity function. This algorithm works by assigning membership to each data point corresponding to each cluster center on the basis of distance between the cluster center and the data point [20].

3.2 Data Cleaning

The process of data cleaning is removal of outliers or irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files. Thus removal process in the experiment includes

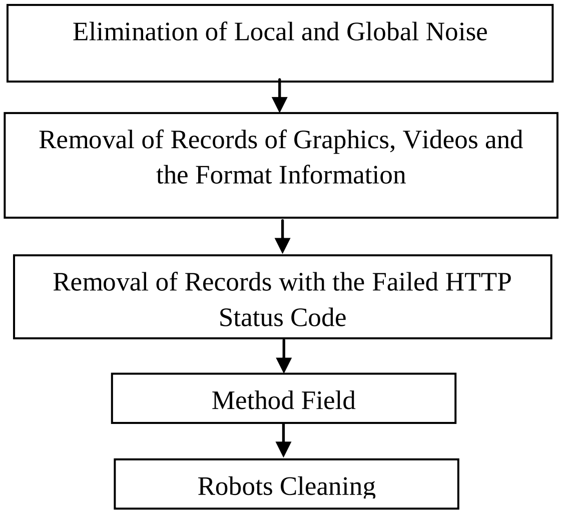

Elimination of Local and Global Noise
Removal of Records of Graphics, Videos and the Format Information
Removal of Records with the Failed HTTP Status Code
Method Field
Robots Cleaning

Fig 2.1: Steps in Data Cleaning

3.2.1 Exclusion of Local and Global Noise

There are two categories of web noise, i) Global noise and ii) Local noise based on their granularities which are not less than a page.

Global noise: These are the unwanted objects with large granularities which are not less than page. This may includes mirror sites, duplicated WebPages, previous versioned WebPages.

Local noise: These are the unnecessary items in the web page. This may include the decoration pictures, navigational guides, advertisements.

3.2.2 The Records of Graphics, Videos and the Format Information

The records have filename extension of GIF, JPEG, CSS, and so on, which can be found in the URI field of the every record, can be removed. This extension files are not actually the user interested web page, rather it is just the documents embedded in the web page. So it is not necessary to include in identifying the user interested web pages. This cleaning process helps in discarding unnecessary evaluation and also helps in fast identification of user interested patterns.

3.2.3 The Records with the Failed HTTP Status Code

The HTTP status code is then considered in the next process for cleaning. By examining the status field of every record in the web access log, the records with status codes over 299 or fewer than 200 are removed. This cleaning process will further reduce the evaluation time for determining the used interested patterns.

3.2.4 Method Field

It should be pointed out that different from most other researches, records having value of POST or HEAD in Method field are reserved in present study for acquiring more accurate referrer information.

3.2.5 Robots Cleaning

Web robot (WR) (also called spider or bot) is a software tool that periodically scans a web site to extract its content. Web robots automatically follow all the hyperlinks from a web page. Search engines, such as Google, periodically use WRs to gather all the pages from a web site in order to update their search indexes. The number of requests from one WR may be equal to the number of the web site's URIs. If the web site does not attract many visitors, the number of requests coming from all the WRs that have visited the site might exceed that of human-generated requests.

Eliminating the Web Robots (WR) generated log entries will simplify the mining process and also eliminates uninterested sessions from the log file. Generally WR will produce a large number of requests on a website. WR's sessions are really not needed for the analysis because web usage mining focuses only on user interested patterns.

Most of the Web robots identify themselves by using the user agent field from the log file. Several databases referencing the known robots are maintained [Kos, ABC]. However, these databases are not exhaustive as each day new WRs appear or are being renamed, making the WR identification task more difficult.

To identify web robots' requests, the data cleaning module implements two different techniques.

- In the first technique, all records containing the name “robots.txt” in the requested resource name (URL) are identified and straightly removed.
- The next technique is based on the fact that the crawlers retrieve pages in an automatic and exhaustive manner, so they are distinguished by a very high browsing speed. Therefore, for each different IP address, the browsing speed is calculated and all requests with this value more than a threshold are regarded as made by robots and are consequently removed. The value of the threshold is set up by analyzing the browser behavior arising from the considered log files.

This helps in accurate detection of user interested patterns by providing only the relevant web logs. Only the patterns that are much interested by the user will be resulted in the final phase of identification if this cleaning process is performed before start identifying the user interested patterns.

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed preprocessing phase with robots cleaning, experiments were carried out using UCI Machine Learning Repository (University of California, Irvine). This repository contains 211 datasets.

For the purpose of evaluating the proposed robot cleaning preprocessing phase, it is evaluated against,

- Initial log file and
- Preprocessed log file without removing robots.

Three standard datasets from the UCI Machine Learning Repository datasets and a real dataset is collected from reputed college were selected for the evaluation purpose. Following is the data sets used for evaluating the proposed preprocessing phase with robots cleaning.

- Anonymous Microsoft Web Dataset [21],

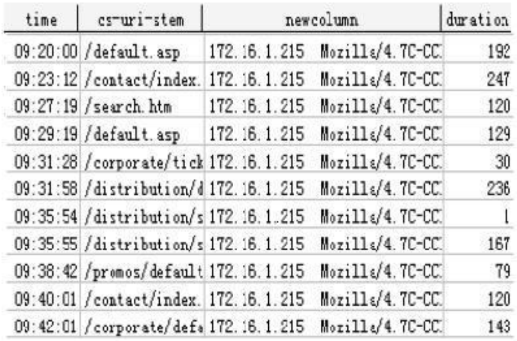
4.1 Anonymous Microsoft Web Dataset

This dataset consists of 37711 records in the log file. Then the data cleaning process is carried out. Initially, after removing records with local and global noise, graphics and videos format such gif, JPEG, etc., 29862 records are obtained. Then by checking the status code and method field, the total of 26854 records is resulted. Finally, 18452 records are resulted after applying robot cleaning process and it is shown in table 4.1.

Table 4.1: Number of Records Resulted After Three Data Cleaning Phases in Anonymous Microsoft Web Dataset

Data Cleaning Phase	**Number of Records**
Initial Log	37711
After removing local and global noise, graphics and videos format records	29862
After checking status code and method field	26854
After robot cleaning process	18452

Table 4.2: Session Identified List



time	cs-uri-stem	newcolumn	duration
09:20:00	/default.asp	172.16.1.215 Mozilla/4.7C-CC	192
09:23:12	/contact/index.	172.16.1.215 Mozilla/4.7C-CC	247
09:27:19	/search.htm	172.16.1.215 Mozilla/4.7C-CC	120
09:29:19	/default.asp	172.16.1.215 Mozilla/4.7C-CC	129
09:31:28	/corporate/tick	172.16.1.215 Mozilla/4.7C-CC	30
09:31:58	/distribution/d	172.16.1.215 Mozilla/4.7C-CC	236
09:35:54	/distribution/s	172.16.1.215 Mozilla/4.7C-CC	1
09:35:55	/distribution/s	172.16.1.215 Mozilla/4.7C-CC	167
09:38:42	/promos/default	172.16.1.215 Mozilla/4.7C-CC	79
09:40:01	/contact/index.	172.16.1.215 Mozilla/4.7C-CC	120
09:42:01	/corporate/defa	172.16.1.215 Mozilla/4.7C-CC	143

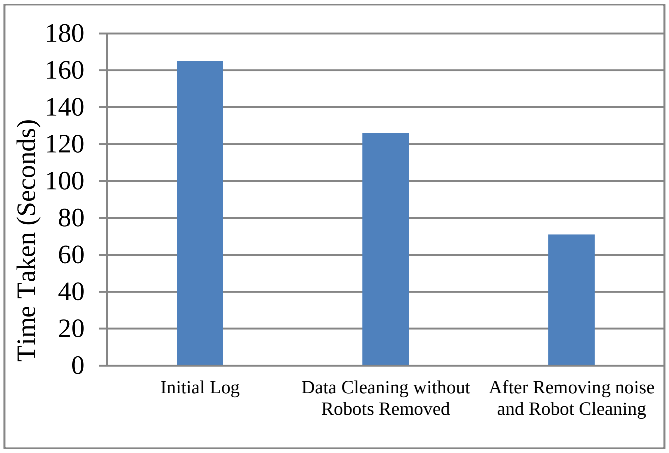
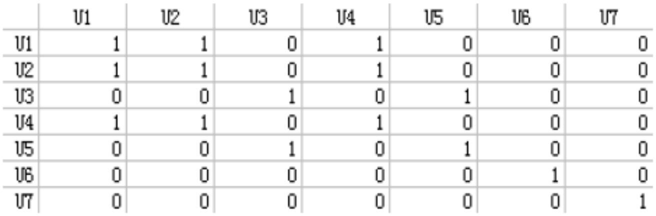


Fig 4.1: Time Taken for User Interested Pattern Prediction in Anonymous Microsoft Web Dataset

Table 4.3: A Cutting Matrix With λ =0.7



	U1	U2	U3	U4	U5	U6	U7
U1	1	1	0	1	0	0	0
U2	1	1	0	1	0	0	0
U3	0	0	1	0	1	0	0
U4	1	1	0	1	0	0	0
U5	0	0	1	0	1	0	0
U6	0	0	0	0	0	1	0
U7	0	0	0	0	0	0	1

When λ =0.7, the users can be divided into five categories: {U1, U4}, {U2}, {U3, U5}, {U6}, {U7}. When λ =0.6, the users can be divided into four categories :{ U1, U2, U4}, {U3, U5}, {U6}, {U7}. Table 4.3 shows the result with λ =0.6. Experiments showed that when taking λ = 0.6 for the optimal threshold, the result of the algorithm was the most ideal and the algorithm is faster and need less storage space.

Figure 4.1 shows that the time required for the prediction of user interested pattern using initial log is 165 seconds, whereas, 126 seconds after cleaning by gif status removal and it takes only 71 seconds.

5. CONCLUSION

Data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. Data cleaning phase includes the removal of records of graphics, videos and the format information, the records with the failed HTTP status code and finally robots cleaning. Different from other implementations records are cleaned effectively by removing local and global noise and robot entries. This preprocessing step is used to give a reliable input for data mining tasks. Accurate input can be found if the byte rate of each and every record is found. The data cleaning phase implemented in this paper will helps in determining only the relevant logs that the user is interested in. Anonymous Microsoft Web Dataset. Anonymous Web Dataset are used for evaluating the proposed preprocessing technique and it reveals that number of records

REFERENCES

[1] Etminani, K., Delui, A.R., Yanehsari, N.R. and Rouhani, M., "Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM", First International Conference on Networked Digital Technologies, Pp.224-249, 2009.

[2] Jianxi Zhang, Peiying Zhao, Lin Shang and Lunsheng Wang, "Web Usage Mining Based On Fuzzy Clustering in Identifying Target Group", International Colloquium on Computing, Communication, Control, and Management, Vol. 4, Pp. 209-212, 2009.

[3] Nina, S.P., Rahman, M., Bhuiyan, K.I. and Ahmed, K., "Pattern Discovery of Web Usage Mining", International Conference on Computer Technology and Development, Vol. 1, Pp.499-503, 2009.

[4] Chih-Hung Wu, Yen-Liang Wu, Yuan-Ming Chang and Ming-Hung Hung, "Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment", International Conference on Machine Learning and Cybernetics (ICMLC), Vol. 6, Pp. 2909-2914, 2010.

[5] Aghabozorgi, S.R. and Wah, T.Y., "Using Incremental Fuzzy Clustering to Web Usage Mining", International Conference of Soft Computing and Pattern Recognition, Pp. 653-658, 2009.

[6] Maratea, A. and Petrosino, A., "An Heuristic Approach to Page Recommendation in Web Usage Mining", Ninth International Conference on Intelligent Systems Design and Applications, Pp. 1043-1048, 2009.

[7] Inbarani, H.H., Thangavel, K. and Pethalakshmi, A., "Rough Set Based Feature Selection for Web Usage Mining", International Conference on Conference on Computational Intelligence and Multimedia Applications, Vol. 1, Pp. 33-38, 2007.

[8] Jalali, M., Mustapha, N., Sulaiman, N.B. and Mamat, A., "A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems", 12th International Conference Information Visualization, Pp. 302-307, 2008.

[9] Shinde, S.K. and Kulkarni, U.V., "A New Approach for on Line Recommender System in Web Usage Mining", International Conference on Advanced Computer Theory and Engineering, Pp. 973- 977, 2008.

[10] Zhang Huiying and Liang Wei, "An intelligent algorithm of data pre-processing in Web usage mining", Fifth World Congress on Intelligent Control and Automation, Vol. 4, 3119- 3123, 2004.

[11] Nasraoui, O., Soliman, M., Saka, E., Badia, A. and Germain, R., "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 2, Pp. 202-215, 2008.

[12] Hogo, M., Snorek, M. and Lingras, P., "Temporal Web usage mining", International Conference on Web Intelligence, Pp. 450-453, 2003.

[13] DeMin Dong, "Exploration on Web Usage Mining and its Application", International Workshop on Intelligent Systems and Applications, Pp. 1-4, 2009.

[14] Yan Li, Boqin Feng and Qinjiao Mao, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, Vol. 1, Pp. 554-559, 2008.

[15] Baraglia, R. and Palmerini, P., "SUGGEST: a Web usage mining system", International Conference on Information Technology: Coding and Computing, Pp. 282-287, 2002.

[16] Jian Chen, Jian Yin, Tung, A.K.H. and Bin Liu, "Discovering Web usage patterns by mining cross-transaction association rules", International Conference on Machine Learning and Cybernetics, Vol. 5, Pp. 2655-2660, 2004.

[17] Wu, K.L., Yu, P. S. and Ballman, A., "Speed Tracer: A Web usage mining and analysis tool", IBM Systems Journal, Vol. 37, No. 1, Pp. 89-105, 1998.

[18] Labroche, N., Lesot, M.J. and Yaffi, L., "A New Web Usage Mining and Visualization Tool", 19th IEEE International Conference on Tools with Artificial Intelligence, Vol. 1, Pp. 321-328, 2007.

[19] Chu-Hui Lee and Yu-Hsiang Fu, "Web Usage Mining Based on Clustering of Browsing Features", Eighth International Conference on Intelligent Systems Design and Applications, Vol. 1, Pp. 281-286, 2008.

[20] J. Bezdek, “Pattern Recognition with Fuzzy Objective Function Algorithms”, New York: Plenum Press, 1981.

[21] http://archive.ics.uci.edu/ml/datasets/Anonymous+Microsoft+Web+Data

[22] http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data