AN EFFICIENT WEB PERSONALIZATION APPROACH TO DISCOVER USER INTERESTED DIRECTORIES

M. Robinson Joel[1], M.V. Srinath[2] and A. Adhiselvam[3]

[1]*Department of Computer Science and Engineering, Shree Motilal Kanhaiyalal Fomra Institute of Technology, India*
E-mail: joelnazareth@gmail.com

[2,3]*Department of Computer Applications, Sengamala Thayaar Educational Trust Women's College, India*
E-mail: [2]sri_induja@rediffmail.com, [3]adhiselvam@yahoo.com

Abstract

Web Usage Mining is the application of data mining technique used to retrieve the web usage from web proxy log file. Web Usage Mining consists of three major stages: preprocessing, clustering and pattern analysis. This paper explains each of these stages in detail. In this proposed approach, the web directories are discovered based on the user's interestingness. The web proxy log file undergoes a preprocessing phase to improve the quality of data. Fuzzy Clustering Algorithm is used to cluster the user and session into disjoint clusters. In this paper, an effective approach is presented for Web personalization based on an Advanced Apriori algorithm. It is used to select the user interested web directories. The proposed method is compared with the existing web personalization methods like Objective Probabilistic Directory Miner (OPDM), Objective Community Directory Miner (OCDM) and Objective Clustering and Probabilistic Directory Miner (OCPDM). The result shows that the proposed approach provides better results than the aforementioned existing approaches. At last, an application is developed with the user interested directories and web usage details.

Keywords:

Advanced Apriori Algorithm, Fuzzy Clustering Algorithm, Pattern Analysis, Web Directories and Web Personalization

1. INTRODUCTION

The Web site is a collection of a number of web pages grouped under the same domain name. Generally web pages may contain images, text, graphics, videos, advertisements and navigate between them via hyperlinks. Whenever the user accesses the website, a log file is recorded. It contains the information about the usage of the web user. Recently, web log files are growing at a faster rate due to the tremendous usage of web users. These data are usually stored in web log files by using different text-based formats like National Center for Supercomputing Applications (NCSA) Common log file format, Internet Information Services (IIS) log file format, World Wide Web Consortium (W3C) extended log file format. Moreover these log files are located in different locations like a web proxy server, web server and the client browser. To preprocessing the log file, it needs to follow the cleanup steps like data cleaning, user identification, session identification and data summarization. This process helps to remove the noisy and unnecessary data on the log files.

Web mining is the combination of information gathered by outdated data mining concepts and techniques with information collected over the World Wide Web. It is used to estimate the effectiveness of a particular web site, understand customer behavior and support the success of a marketing/business campaign. It also allows looking for patterns in data over content mining, structure mining and usage mining. Web content mining is the procedure of extracting information from the contents of the web documents. Content data corresponds to the group of facts a web page was considered to convey to the users. Problems addressed in text mining are topic discovery, extracting association patterns, grouping of web documents and organization of web pages. Web Structure mining is the procedure of finding and extracting useful information from semi structured data sets. This can be further divided into two types built on the kind of structure information used, namely hyperlinks and document structure. A hyperlink that connects to a diverse part of the similar page is called an Intra-Document Hyperlink. And a hyperlink that links two diverse pages is called an Inter-Document Hyperlink. In document structure, the content inside a web page can also be ordered in a tree-structured format established on the several Hyper Text Markup Language (HTML) and Extensible Markup Language (XML) tags within the page. Mining efforts focus on spontaneously mining Document Object Model (DOM) structures out of documents. Usage Mining is the application of data mining systems to large web data repositories in order to produce results. Usage data capture the uniqueness or origin of web consumers along with their surfing behavior on a web site.

Web site personalization is the procedure of modifying the content and structure of a web site to the precise requirements of each user taking benefit of the user's directional behavior. The phases of the web personalization comprises of: 1) the collection of web data, 2) the preprocessing phase of these data, 3) the analysis of the collected data and 4) the purpose of the actions that should be performed. In the proposed work, the log file is collected from the proxy server log. The gathered data are undergoing a preprocessing phase to remove the unwanted and noisy information. The web directories are discovered based on the user and session clustering. For grouping the user and session, the Fuzzy Clustering Approach (FCA) is applied. To sequence the patterns, Advanced Apriori (A-Apriori) algorithm is used.

The rest of the paper is organized as follows. Section 2 presents a description about the previous research which is relevant to web personalization. Section 3 involves the detailed description about the proposed method. Section 4 presents the performance analysis. The conclusions are discussed in section 5.

2. RELATED WORK

This section deals with the works related to the web usage mining concepts. Bouras et al proposed a methodology for active web personalization and document alignment infrastructure for meta portals. The web personalization mechanism was based on

dynamic creation and updation of user profiles according to the users preferences. This mechanism was used to built and maintain a user profile without disturbing the user. Also the real-time user-centered document grouping steps was implemented to support the web personalization system [1]. Carmona et al. presented an approach used in an e-commerce website of extra virgin olive oil sale. Some set of phases was carried out including data collection, data preprocessing, extraction and analysis of knowledge. The knowledge was extracted using supervised and unsupervised data mining algorithms through clustering, association and subgroup discovery [2]. Gholamzadeh et al. proposed a method based on data mining techniques to support and improve the web service discovery process and service-oriented applications. The proposed assistant discovery approach was based on automatic finding of semantic similarity between web services with the help of clustering methods. A fuzzy semantic clustering algorithm was proposed which assists web service consumers to discover the group of similar web services through unique query [3].

Guerbas et al. proposed an enhanced web log mining process and online navigational pattern prediction. A refined time-out based heuristic approach used for session identification and the usage of a specific density based algorithm for navigational pattern discovery was also presented [4]. Hurtado et al. proposed an unsupervised computational method that conducts the upholding of the directory by examines of user browsing data. This approach was built on the extraction and classification of user sessions into the categories of the directory. The directory maintenance method was considerably modified to find queries. It was used to discover the related properties permitting users to change from directory browsing to query preparation. This method attained detection of new pages in separate category and also related queries with high precision were suggested without the need for labeled data to conduct usual web page and query classification tasks [5]. Lee et al proposed a prediction model called Two-Level Prediction Model (TLPM). This method uses the normal hierarchical possessions from web log data. TLPM can reduce the extent of candidate set of web pages and escalate the speed of predicting with tolerable accuracy [6]. Lefever et al presented a fuzzy ant based clustering approach for multi-document person name disambiguation problem [7].

Li et al. proposed a Web User approach to discover unexpected usage in web access log. A brief-driven method was presented to extract the unexpected web usage sequences, where the brief system consists of a temporal relation and semantics constrained sequence rules acquired with respect to prior knowledge. The discovered rules of unexpected web usage were used for web content personalisation and recommendation, site structure optimisation and critical event prediction [8]. Mabroukeh et al. presented taxonomy of sequential pattern-mining techniques with web mining as an application. These algorithms were tested by introducing a taxonomy for classifying sequential pattern-mining algorithms based on key features supported by the technique. This classification aims at enhancing the sequential pattern mining problems [9]. Matthews et al. proposed a Genetic Algorithm (GA) based solution that consumes the plastic nature of the 2-tuple verbal illustration. It was used to determine rules that happens at the intersection of fuzzy set boundaries. The GA-based approach was enhanced by comprising a graph illustration and an improved fitness function [10].

Mishra et al proposed a Frequent Pattern (FP)-growth algorithm to extract the frequent access patterns from the web log data and valuable information about the user's interest. In the pattern analysis phase interesting knowledge was retrieved based on frequent patterns [11]. Moawad et al proposed a new multi-agent system based method for initialing the web search results. This model builds an user profile from initial and simple information, and sustain it over implicit user feedback to launch user profile. In the web search process, this model semantically optimizes the user query in two steps. One is query optimization using user profile preferences and the second one is query optimization using the WordNet ontology [12]. Paramasivam et al. presented a complete outline and discoveries in mining web usage patterns from web log files of a real web site. This work describes the discovering and tracking of user profiles. An objective validation strategy was also used to assess the quality of the mined profiles. This method allows discovery frequent behavioral patterns in real time, whatever the amount of connected users had been measured [13].

Pierrakos et al. presented a knowledge discovery outline for the structure of Community Web Directories. In this work, the Web directory was considered as a thematic hierarchy and personalization was realized by creating user community models based on the usage data. This method combines the users browsing behaviour with thematic information from the Web directories. Clustering and probabilistic approaches was combined and presented [14]. Sandhyarani et al. designed the structure of community web directories. This structure depends on the web usage data. User community models take the method of thematic hierarchies and they were constructed by employing clustering approach. This methodology was applied to the Open Directory Project (ODP) directory and also to a synthetic web directory. To discover the community models a new criterion was introduced. The criterion associate a priori thematic informativeness of the web directory groups with the level of interest observed in the usage data. The proposed clustering methodology was estimated both on a dedicated artificial and a community web directory, representing its value to its user of the web [15].

Sha et al proposed a method named Enterprise Proxy Log Cleaner (EPLogCleaner) that filter out plenty of irrelevant items based on the common prefix of their URLs. The proposed system make an evaluation of EPLogCleaner with a real network traffic trace captured from the enterprise proxy. EPLogCleaner improve the data quality of enterprise proxy logs [16]. Tao et al. proposed a knowledge-based model for web information assembly. This scheme uses a world knowledge base and user local instance sources for user profile achievement and the capture of user evidence model. This method gives to better designs of knowledge-based and identifies web information gathering system [17].

Tsay et al developed an algorithm for category mapping between hierarchical directories. The algorithm was based on the two techniques: consistency refinement and hierarchical substitution which were developed with extensive of hierarchical structures [18]. Vaishnavi et al. proposed a technique for the growth of the web personalization scheme using Modified Fuzzy Possiblistic C-Means (MFPCM) [19]. Varghese et al. proposed a cluster optimization methodology based on fuzzy logic. Web usage mining method was used for eliminating the redundancies

occurred in data after clustering. Fuzzy Cluster-Chase algorithm was presented for cluster optimization to personalize web page clusters of end users [20].

3. DISCOVERING USER INTERESTED DIRECTORIES

On the concept of community Web directory, a Web directory is formed based on the needs and interest of particular user communities. To create such directories, web usage data are used. User community representations take the system of thematic hierarchies and they are usually constructed using the clustering approach. In this work, a novel method is introduced that combine an Advanced Apriori informativeness of the web directory types with the level of user's interest observed in the usage data. This approach is tested using the logs driven from the representation servers of an Internet Service Provider (ISP), which results in the usability of community web directories. The collected raw data are needed to undergo a preprocessing stage in order to remove the noise and unwanted information.

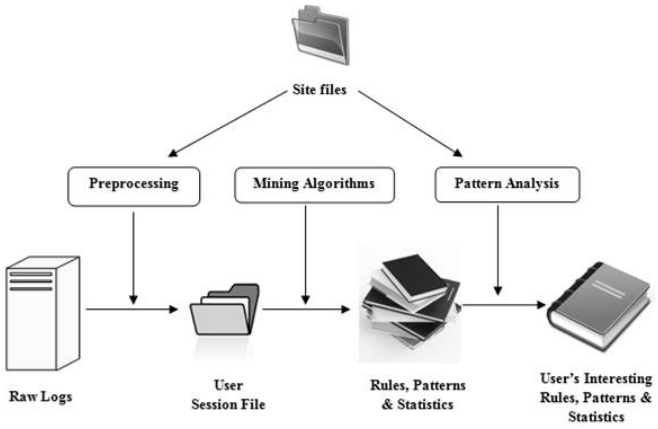

Fig.1. High Level Usage Mining Process

The overall web usage mining process is depicted in Fig.1. There are three main stages that are performed on the usage mining process. They are preprocessing the log files, mining algorithms to predict the rules and patterns, and finally the patterns are analyzed to categorize the user interested rules.

3.1 PROXY SERVER LOG DATA

Web plays a major role in extracting and searching useful information. There is a need for data log to monitor any transaction of the communications. Log file data provides an essential informative insight into the website usage. It categorizes the activity of different users over a long period of time. The proxy server log provides technical details about the server load, management activities, successful transactions and request and assisting in marketing and site development. The server log contains the following set of attributes: Date, Time, Client IP Address, User Authentication, Server IP Address, Server Post, Server Method, URL and Agent Log [11]. The complete flow of the proposed methodology is shown in Fig.2.

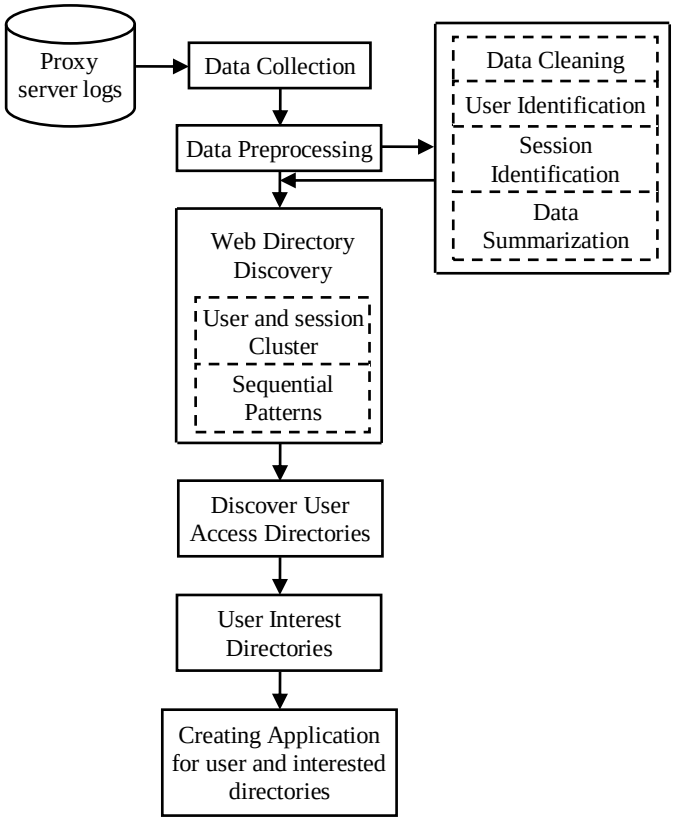

Fig.2. Flow of Proposed Work

3.2 PREPROCESSING OF PROXY SERVER LOGS

Preprocessing is an important step to mine the data, because the log file may contain noise and unwanted data which could be affecting the result of the mining process. So data preprocessing phase is necessary to filter and organize only appropriate details before applying any web mining algorithms. The advantage of data preprocessing technique is to improve the data quality and improve the mining accuracy. It consists of data cleaning, user identification, session identification and data summarization.

3.2.1 Data Cleaning:

This stage consists of removing the useless or unwanted requests from the log files. Normally, this stage removes requests like images and other multimedia files. It also finds the Web robots and removes their requests.

3.2.2 User Identification:

In many cases, the log file provides only the IP address and the user agent. Web sites require user registration and the log file contains the user login information. This information is used for user identification. This step finds the individual user by using their IP address. A new IP address is sensed, and then the proxy server assumes that there is a new server. If IP address is similar but browser version or operating system is diverse then it denotes as different user.

3.2.3 Session Identification:

Each user spends certain period of time for each web page. Session refers time duration spent for each web page. A referrer-based method is used for finding sessions. If IP address and operating system are similar, then the referrer data should be taken.

3.2.4 Data Summarization:

This stage transfers the file into relational databases. Later it applies the data generalization at the request level and then aggregated data computed for visits and user sessions.

3.3 WEB DIRECTORY DISCOVERY

The web directories are discovered based on the user and cluster session. This section involves the clustering of user and session.

3.3.1 Clustering the User and Session using Fuzzy Clustering Approach (FCA):

The user is identified by the IP address of the respective user. Both the user and session information are retrieved from the log file. The user and session details are clustered using the fuzzy clustering algorithm. The clustering strategy starts by partitioning the data set into a large number of small clusters. Let $A = \{a_j \mid j = 1,\ldots,n\}$ be the set of n vectors and let $C = (\beta_1,\ldots,\beta_c)$ represents a c tuple of prototypes each of which characterizes one of the c clusters. The FCA minimizes

$$S(C,V) = \sum_{i=1}^{c}\sum_{j=1}^{n}(v_{ij})^2 l_{ij}^2 - \alpha\sum_{i=1}^{c}\left[\sum_{j=1}^{n} v_{ij}\right]^2 \quad (1)$$

Subject to $$\sum_{i=1}^{c} v_{ij} = 1 \quad \forall j = \{1,\ldots,n\} \quad (2)$$

Here l_{ij}^2 denotes the space from feature vector a_j, to the prototype β_c, v_{ij} denotes the amount of membership of feature point a_j in cluster β_c, and $V = [v_{ij}]$ is a $c \times n$ constrained fuzzy c-partition matrix. The number of cluster c in Eq.(1) is dynamically updated in the algorithm. The membership v_m minimizes Eq.(1) with respect to V.

$$v_{sm} = v_{sm}^{FCA} + v_{sm}^{Bias} \quad (3)$$

Here, $$v_{sm}^{FCA} = \frac{P1/l_{sm}^2}{\sum_{k=1}^{c} 1/l_{km}^2} \quad (4)$$

$$v_{sm}^{Bias} = \frac{\alpha}{d_{sm}^2}(n_s - \bar{n}_m) \quad (5)$$

where, n_s is the cardinality of cluster s.

$$n_s = \sum_{j=1}^{n} v_{sj} \quad (6)$$

And $$\bar{n}_m = \frac{P\sum_{k=1}^{c} 1/l_{km}^2 \; n_k}{\sum_{k=1}^{c} 1/l_{km}^2} \quad (7)$$

The user and sessions are clustered into discrete groups based upon the Eq.(1) to Eq.(7).

3.3.2 Sequential Pattern Analysis: Advanced-Apriori (A-Apriori):

In this approach, Advanced Apriori algorithm is utilized to analyze the sequential patterns.

In the algorithm steps, PI denotes Probable Infrequent Itemset, PF denotes the Probable Frequent Itemset and DI denote the Definite Infrequent Itemset.

The interesting directory is ranked based on the support and confidence criteria. Confidence measures the percent of times that a rule fulfills when it is applicable. Support is the fraction of the transactions in the database containing the number of transactions the rule predicts correctly.

Advanced Apriori Algorithm

1) Initialize PF = PI = DI = ϕ
2) Set run = 0, A, minSupport
3) Increment run
4) Set ItemRead = 0
5) While ItemRead < total-transaction
6) Get PI and PF
7) For all i < A
8) i = i+1
9) For each itemset T∈ PI or PF
10) If T ∈transaction
11) T=T+1
12) End for
13) For each itemset T ∈ PI
14) If T ≥ minSupport
15) Move T from PI to PF
16) If (itemset I ⊃ T) ∈ PF
17) Add I in PI
18) End for
19) For each itemset T ∈ PF
20) If T.start = ItemRead and T.run= run-1
21) Move T into PF
22) End for
23) For each itemset T∈ PI
24) If T.start = ItemRead and T.run = run – 1
25) Move T into PI
26) End for
27) ItemRead = ItemRead + A
28) End

Consider a rule E that implies:

$$E: \text{In the context } (\Theta_1,\ldots,\Theta_p)$$

$$(m_1 \Lambda \ldots \Lambda m_s) \Rightarrow (n_1 \Lambda \ldots \Lambda n_r) \quad (8)$$

The support and confidence of the above rule are computed according to the following equations:

$$Support(E) = \frac{s(m_1,\ldots,m_s,n_1,\ldots,n_r,\Theta_1,\ldots,\Theta_p,\text{All},\ldots,\text{All})}{s(\text{All},\ldots,\text{All},\Theta,\ldots,\Theta,\text{All},\ldots,\text{All})} \quad (9)$$

$$Confidence(E) = \frac{s(m_1,..,m_s,n_1,...n_r,\Theta_1,..,\Theta_p,\text{All},..,\text{All})}{s(m_1,..,m_s,\text{All},...,\text{All},\Theta_1,..,\Theta_p,\text{All},..,\text{All})} \quad (10)$$

where, $s(m_1,..,m_s, n_1,..,n_r, \Theta_1,..,\Theta_p, \text{All},..,\text{All})$ defines the sum-based cumulative measure of the itemset. Based upon Eq.(9) and Eq.(10) the interesting directories are discovered.

3.4 WEB APPLICATION CREATION

Based upon the discovered interesting directories an application can be created. The application is composed of two sections. One is users view and the other one is directory view. The users view link contains all the interested users IP address. If the user view icon is clicked, it shows all the interested users

IP address. The directory view incorporates the list of interested directories that are retrieved based on the A-APRIORI. If one directory is selected, then it displays some list of contents that are accessed by web users.

4. PERFORMANCE ANALYSIS

The methodology introduced in this paper for discovering the web directories has been tested, which concentrated on the analysis of usage data from the web proxy server logs of ISP. The estimation of this approach employs two criteria: coverage and user gain.

4.1 COVERAGE

Coverage is defined as the number of target web pages that are covered by the session based directories.

Consider if e_t is the examined target, then the coverage of e_t is defined by the following equation:

$$Cov(e_t) = \begin{cases} 1, & \text{if } e_t \in \Theta_i \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

Suppose D is the total number of target web pages and D' is the number of target pages that are covered by the session based interested web directories, and then the coverage of the set of web directories is defined as follows:

$$Cov_{\Theta} = \frac{\|D'\|}{\|D\|} \quad (12)$$

4.2 USER GAIN

User gain is the estimated actual gain that a user follows the interested web directories instead of accessing the preliminary web directories to get the preferred web page. In order to estimate this measure, a criterion called navigation path i.e. navpath is introduced. This measure is used to calculate the depth of the navigation path. The navpath of a target page e_t is defined by the following equation:

$$navpath(e_t) = d(e_t) + \sum_{i=1}^{d(e_t)} s_i \quad (13)$$

Here $d(e_t)$ is the depth of the path to e_t and s_i is the branching factor at the ith step of the path from the root to e_t. The user gain for each target page is described as follows:

$$UG(e_t) = \frac{navpath_{WD}(e_t) - navpath_{IWD}(e_t)}{navpath_{WD}(e_t)} \quad (14)$$

Here $navpath_{WD}(e_t)$ represents the original web directory and $navpath_{IWD}(e_t)$ represents the interested web directories.

The average $UG_{D'}$ is defined as:

$$UG_{D'} = \frac{\sum_{e_t \in D'} UG(e_t)}{|D'|} \quad (15)$$

4.3 RESULTS

4.3.1 Support and Confidence:

The support and confidence are estimated among the discovered directories using the Eq.(9) and Eq.(10). The resulted values are shown in Fig.3.

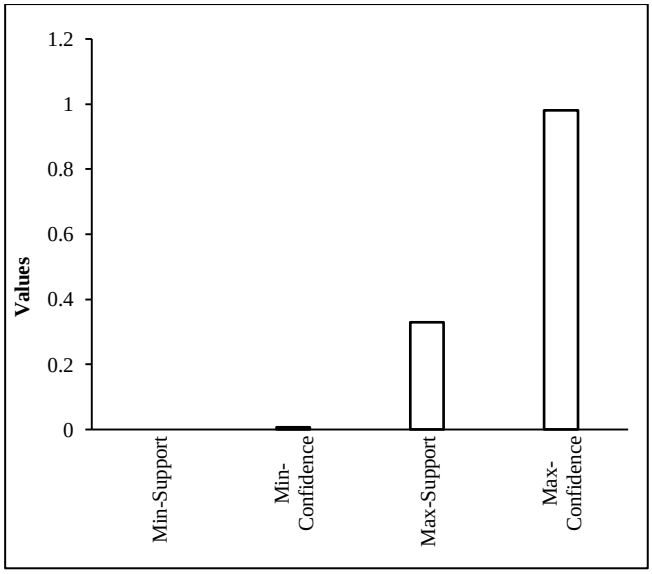

Fig.3. Support & Confidence values

4.3.2 User's Interested Web Directories:

The retrieved user interested web directories based on our proposed approach is shown in Fig.4. The 'total' block is the web directories are retrieved before applying the A-Apriori algorithm. The 'final' block is the user interested web directories, which are extracted after applying the A-Apriori algorithm.

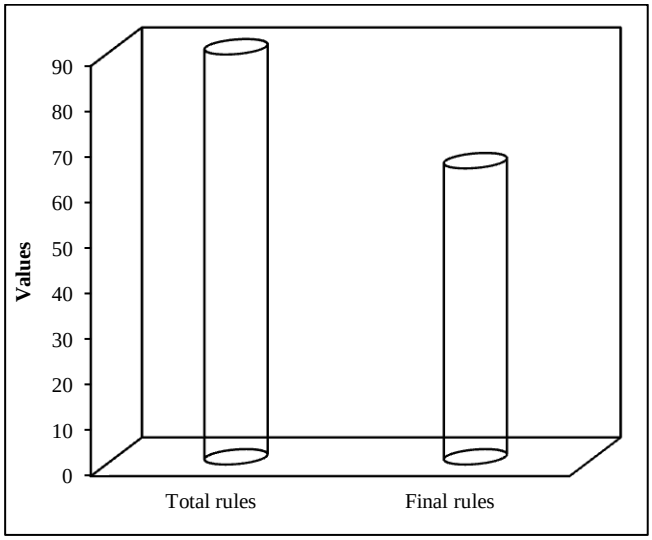

Fig.4. Total directories vs User Interested Web Directories

4.3.3 Coverage:

The coverage and user gain are calculated by using Eq.(11-15). The resulted coverage against user gain is shown in Fig.5. The proposed Discovered User Interested Web Directories (DUID) results better coverage than the existing approaches like OPDM, OCDM and OCPDM [14].

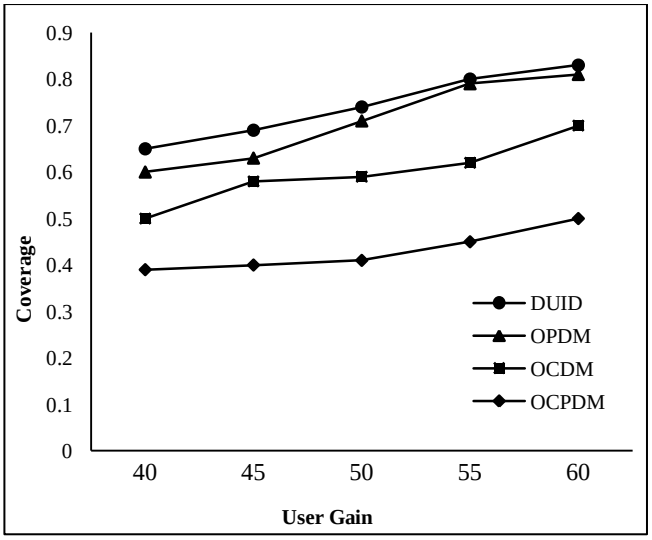

Fig.5. Coverage vs User Gain

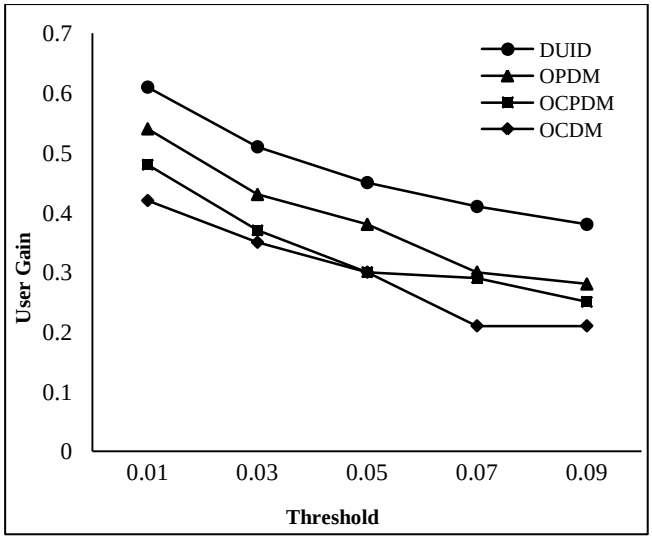

Fig.7. User Gain vs Threshold

4.3.4 Threshold:

The proposed DUID removes the directories which are all having the threshold value less than the fixed threshold. The proposed approach provides better coverage and user gain against threshold. The result shows that our proposed DUID is better than the existing techniques as shown in Fig.6 and Fig.7.

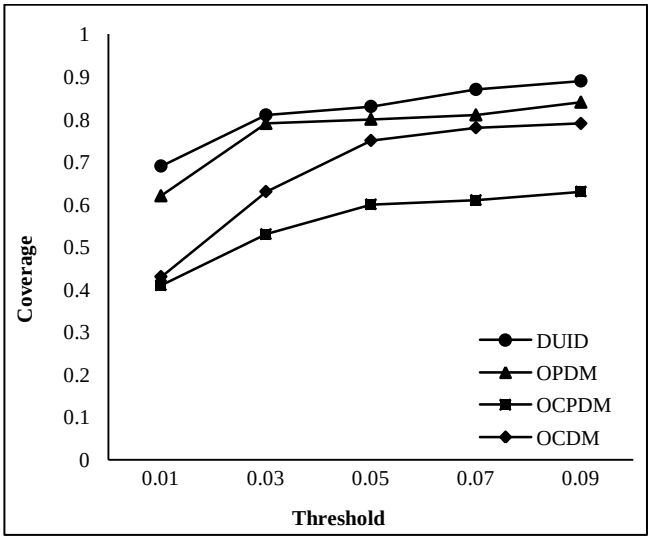

Fig.6. Coverage vs Threshold

4.4 APPLICATION CREATION

The user interested web directories are used to built a web directory application to visually show the details about the user and directory information's.

Fig.8. Developed Application

The Directory view link holds the IP addresses of all the interested web users shown in Fig.8.

Fig.9. Directory view

The IP address links holds the history about the web usage data. It is shown in Fig.9.

5. CONCLUSION AND FUTURE WORK

In this paper, a novel approach is presented for discovering the user interested web directories. In order to achieve a high quality web directory, the proxy server log files undergone a preprocessing phase. Moreover, the complete methodology is described for the construction of such web directories with the use of Fuzzy clustering and Advanced Apriori algorithms. The implementation proves that our proposed approach is well suited to retrieve the user interested directories. Also the result shows that the proposed approach provides good results when compared with the existing methods.

This method is implemented based on Fuzzy clustering to group the user sessions and Advanced Apriori algorithm to analyze the frequent pattern mining. The following idea can be implanted in the future to enhance the proposed structure. In future, these resulted web directories can be classified into different orderings like highly interested directories, medium interested directories and low interested directories based upon the user's web usage. To achieve this objective, various classification algorithms are applied on the resulted web directories.

REFERENCES

[1] C. Bouras and V. Poulopoulos, "Enhancing meta-portals using dynamic user context personalization techniques", *Journal of Network and Computer Applications*, Vol. 35, No. 5, pp. 1446-1453, 2012.

[2] C.J. Carmona, S. Ramírez-Gallego, F. Torres, E. Bernal, M.J. del Jesus and S. García, "Web usage mining to improve the design of an e-commerce website: OrOliveSur.com", *Expert Systems with Applications*, Vol. 39, No. 12, pp. 11243-11249, 2012.

[3] N. Gholamzadeh and F. Taghiyareh, "Ontology-based fuzzy web services clustering", *5th International Symposium on Telecommunications*, pp. 721-725, 2010.

[4] A. Guerbas, O. Addam, O. Zaarour, M. Nagi, A. Elhajj, M. Ridley and R. Alhaji, "Effective web log mining and online navigational pattern prediction", *Knowledge-Based Systems*, Vol. 49, pp. 50-62, 2013.

[5] C. Hurtado and M. Mendoza, "Automatic maintenance of web directories by mining web browsing data", *Journal of Web Engineering*, Vol. 10, No. 2, pp. 153-173, 2011.

[6] C.H. Lee, Y.-l. Lo and Y.-H. Fu, "A novel prediction model based on hierarchical characteristic of web site", *Expert Systems with Applications*, Vol. 38, No. 4, pp. 3422-3430, 2011.

[7] E. Lefever, T. Fayruzov, V. Hoste and M. De Cock, "Clustering web people search results using fuzzy ants", *Information Sciences*, Vol. 180, No. 17, pp. 3192-3209, 2010.

[8] D. Li, A. Laurent and P. Poncelet, "Web User: mining unexpected web usage", *International Journal of Business Intelligence and Data Mining*, Vol. 6, No. 1, pp. 90-111, 2011.

[9] N.R. Mabroukeh and C.I. Ezeife, "A taxonomy of sequential pattern mining algorithms", *ACM Computing Surveys*, Vol. 43, No. 1, Article No. 3, 2010.

[10] S.G. Matthews, M.A. Gongora, A.A. Hopgood and S. Ahmadi, "Web usage mining with evolutionary extraction of temporal fuzzy association rules", *Knowledge-Based Systems*, Vol. 54, pp. 66-72, 2013.

[11] M.R. Mishra and M.A. Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 9, pp. 311-318, 2012.

[12] I.F. Moawad, H. Talha, E. Hosny and M. Hashim, "Agent-based web search personalization approach using dynamic user profile", *Egyptian Informatics Journal*, Vol. 13, No. 3, pp. 191-198, 2012.

[13] S.P. Paramasivam and D.S. Srinivasan, "Perfections and psychiatring user profile in web sites using web usage mining and clustering sessions", *International Journal of Research in Computer and Communication Technology*, Vol. 2, No. 2, pp. 78-85, 2013.

[14] D. Pierrakos and G. Paliouras, "Personalizing Web Directories with the Aid of Web Usage Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 9, pp. 1331-1344, 2010.

[15] R. Sandhyarani, B. Rajkumar and J. Gyani, "Construction of Community Web Directories based on Web usage Data", *An International Journal on Advanced Computing*, Vol. 3, No. 2, pp. 41-48, 2012.

[16] H. Sha, T. Liu, P. Qin, Y. Sun and Q. Liu, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", *Procedia Computer Science*, Vol. 17, pp. 812-818, 2013.

[17] X. Tao, Y. Li and N. Zhong, "A knowledge-based model using ontologies for personalized web information gathering", *Journal on Web Intelligence and Agent Systems*, Vol. 8, No. 3, pp. 235-254, 2010.

[18] J.J. Tsay and C.-H. Lin, "Hierarchical directory mapping for category-constrained meta-search", *Journal of Intelligent Information Systems*, Vol. 42, No. 1, pp. 75-94, 2014.

[19] A. Vaishnavi, "Effective Web personalization system using Modified Fuzzy Possibilistic C Means", *Bonfring International Journal of Software Engineering and Soft Computing*, Vol. 1, pp. 01-07, 2011.

[20] N.M. Varghese and J. John, "Cluster optimization for enhanced web usage mining using fuzzy logic", *World Congress on Information and Communication Technologies*, pp. 948-952, 2012.