RESEARCH ARTICLE OPEN ACCESS

Web Page Suggestion in Web Mining

Aparna Gupta, Prof. Aarti Karndikar,
M.tech student, Dept.of Computer Science, Nagpur University, India.
Professor, Dept. of Computer Science, Nagpur University, India.
guptaaparna111@gmail.com , aartikarandikar@gmail.com

Abstract—
The World Wide Web continuously growing repository of web pages and links at an exponential rate which makes exploiting all useful information a standing challenge. It has recently a wide range of applications in E-commerce web site and E-services such as building interactive marketing strategies, Web recommendation and Web personalization.

The paper concerns Web server log file analysis to discover knowledge and by applying Clustering and optimization technique to get user interest which is helpful or useful for giving suggestion about specific user's interest.

Keywords—. Web Mining, Pre-processing, FCM Clustering , Cluster Chase optimization Algorithm, Suffix Tree

I. INTRODUCTION

The World Wide Web is an interlinked collection of billions of documents formatted using HTML. The size of this collection has become difficult for information retrieval. The user has to shift through scores of pages to come upon the information customer desires. Internet has became an habitual part of our lives now a days so the techniques which are helpful in extracting data present on the web is an interesting area of research. These techniques helps us to extract knowledge from large Web log data which is used in the mining process . According to survey , web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. With the explosive growth of information available on the World Wide Web and the highly increasing pace of adoption to Internet commerce, the Internet that dynamically generates information that is beneficial to E-businesses or web marketing [1].

Web Usage mining is the application of data mining techniques to discover usage patterns from web data. Log Data is usually collected from user's interaction with the web, e.g. web/proxy server logs, user queries, registration data. Usage mining tools discover and predict user behavior, using this tool which help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service. For decision management, the result of web usage mining can be used for advertisement, improving web design, improving satisfaction of customer, and marketing analysis etc [3][8].

II. RELATED WORK

Web Mining is technique in data mining to extract knowledge from web data, including web documents, hyperlinks between documents, us-age logs of web sites, etc. There are three types of web mining process: *Web Usage Mining*-Web usage mining is the process of extracting useful information from server logs i.e. users history, *Web Structure Mining* - structure mining is to extract previously unknown relationships between Web pages, *Web Content Mining*-Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents[2].

Clustering analysis aims to group similar web usage sessions into identical clusters. The process can not be performed unless WUM data is passed through sophisticated pre-processing steps. We clustered the pre-processed WUM data using a swarm intelligence based optimization, PSO based clustering algorithm. In this paper, showed the performance of the Particle Swarm Optimization (PSO) algorithm is better than K-means clustering .The result of clustering of server log data based on these parameters: (a) time and request per 30 minutes distribution (b)page viewed and number of user distribution (c) session-number of request distribution (d) session-time distribution [5].

In [5], a cluster optimization technique is proposed to improve web usage mining using ant nestmate approach. As the size of the cluster increases, it will become an inevitable need to optimize the clusters. Cluster optimization methodology is based on ant nestmate recognition ability and is used to eliminate the data redundancies. For clustering ART1-nueral network based approach is used. The accuracy and completeness of the user profiles increases by cluster optimization.

Time aware web users clustering [6] emphasize the help us to discover similarities in usage patterns with respect to the time locality of their visit. Two clustering methods are used for tuning and binding the page and time visiting criteria. The clusters developed by this method presents similar behaviour at the same time period, by varying the priority given to page or time visits.

III. ARCHITECTURE

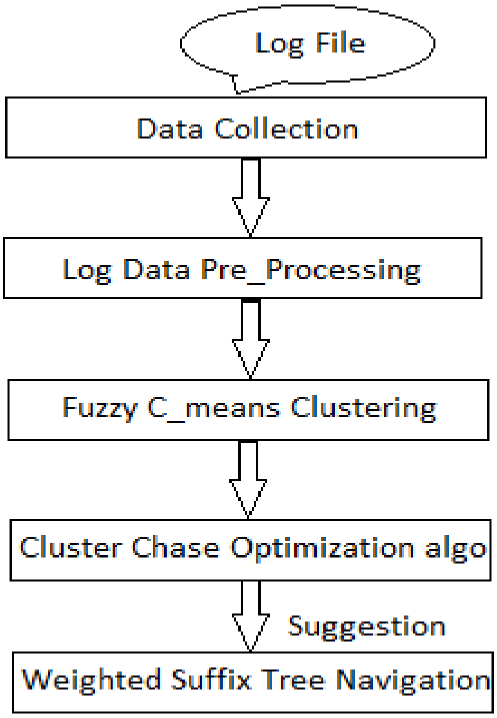
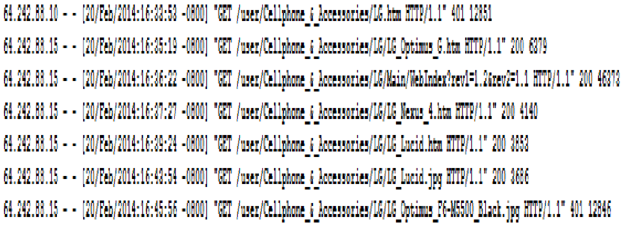


Fig.1 Proposed Architecture

A. Data Collection

In our model ,web log data is colleced from web server which is navigation history of web site maintained in Web access sequence. Web access sequences are sequence of web pages which express a session of the user (Clicked sequence)and that can be collected by the log file of the web site. It is a huge repository of web pages and links, accesses web sites are recorded in web logs file. In thisapproach, Web access log(W3C Extended log file format) is used as data which is generated.



```
64.242.88.10 - - [20/Feb/2014:16:33:53 -0800] "GET /user/Cellphone_&_Accessories/LG.htm HTTP/1.1" 401 12851
64.242.88.15 - - [20/Feb/2014:16:35:19 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Optimus_G.htm HTTP/1.1" 200 6379
64.242.88.15 - - [20/Feb/2014:16:36:22 -0800] "GET /user/Cellphone_&_Accessories/LG/Main/WebIndex?rev1=1.2&rev2=1.1 HTTP/1.1" 200 46373
64.242.88.15 - - [20/Feb/2014:16:37:27 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Nexus_4.htm HTTP/1.1" 200 4140
64.242.88.15 - - [20/Feb/2014:16:39:24 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Lucid.htm HTTP/1.1" 200 3853
64.242.88.15 - - [20/Feb/2014:16:43:54 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Lucid.jpg HTTP/1.1" 200 3686
64.242.88.15 - - [20/Feb/2014:16:45:56 -0800] "GET /user/Cellphone_&_Accessories/LG/LG_Optimus_F6-M5500_Black.jpg HTTP/1.1" 401 12846
```

B. Web Log Pre-processing

Preprocessing is necessary, because Log file contain noisy & ambiguous data which may affect result of mining process [7]. The input of the proposed system is a web log file.First, raw data were read from Web server log files and for each HTTP request the following data were distinguished: the IP address of the Web client, the identifier of the Web client, the user identifier, the timestamp, the HTTP method, the URI of the resource requested, the version of HTTP protocol, the HTTP status code, the size of the object sent to the client. Examlpe: 204.31.113.138 -[03/Jul/1996:06:56:12 -0800] "GET PowerBuilder/Compny3.htm HTTP/1.0" 200 5593

The data preprocessing step has data cleaning, user identification and session identification.

Data Cleaning- First stage of data cleaning is connected with elimination of useless data. Data cleaning is related to site specific, and involves extraneous references to embedded objects that may or may not be important for purpose of analysis, including references to style files, graphics or sound files. Therefore some of entries are useless for analysis process that is cleaned from the log files. By Data cleaning, errors and inconsistencies will be detected and removed to improve the quality of data[8]. Since our analysis concerns the behavior of users and involves a click-stream analysis, the following requests have been excluded from analysis: hits for embedded objects (e.g. images), automatically generated by Web client browsers, requests generated by Web bots (e.g. Web crawlers). Thus, data cleaning includes the elimination of irrelevant entries like:

- Removes requests concerning non-analyzed resources such as images,multimedia files, and page style files.
- Entries with unsuccessful HTTP status codes; HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299.

- Entries with request methods except GET and POST

User Identification- Identifying the individual uses by observing their IP address means user Identification. For Identifying the Unique User ,proposed some Rules:

- If there is new IP address , then there is a new user, a reasonable assumption is that each different agent type for an IP address represents a different user.

Session Identification- After user identification, the pages accessed by each user must be divided into individual session, which is known as session identification . The goal of session identification is to find each user's access pattern and frequently accessed path.[4] Mechanism used in the paper for time out a)defines a time limit for the access of a particular page and this limit is 25 minutes
divided into more than one session. A session refers user's navigation behaviours in a Website, b) to identify the access time of the user for a respective web page.

C. *Fuzzy C_Means Clustering*

Clustering is the process of collecting similar object one another.In this paper,we are using object as user session as time generated by pre_processing stage .In clustering , grouping performed based on users having similar access sequences. The data objects are represented by the feature vector. Given a set of data objects S = {X1 , X2 ,…Xn }, where
$X = (Xi1, Xi2, \ldots. Xil)^P \in R^1$ is a feature vector and the similarity is calculated by the distance function D defined as D :S x S → R such that for distinct X i , X j ε S. The distance between two data object is calculated as follow Fuzzy C-means (FCM) clustering is of overlapped clustering which allows one data

$$D(x_i,x_j) = \sum_{k=1}^{i} (x_{ki} - x_{kj})^2 \quad \forall i,j = 1{:}n \text{ and } i \neq j$$

object to belong to two or more clusters. It is based on minimization of the following objective function:

$$j_m = \sum_{i=1}^{N} \sum_{k=1}^{c} U_{ij}^m \|X_{ki} - X_{kj}\|^2 , \quad 1 <= m < infinity$$

Where m- is a real number greater than 1,

U ij- is the degreeof membership of X i in the cluster j,
C is the total number of clusters,
N- is the total number of user sessions,
X i is the feature vector,
C j is the center of the cluster, and
||*|| is the any norm that expresses the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with update of membership U ij and the centers C j by:

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

$$C_j = \frac{\sum_{K=1}^{C} U_{ij}^m x_{ki}}{\sum_{K=1}^{C} U_{ij}^m} \qquad (3)$$

Step 4: Repeat step (2) and (3) until the termination criterion is satisfied.
Step 5: Stop
The fuzzy c-means procedure stars until the termination criterion is satisfied. Termination criteria can be that the difference between updated and previous objective function value J, is less than a predefined minimum threshold.

The following steps explain the working of FCM:
Input: The feature vector X i that represent the navigational sequence of each user and the number of clusters.
Output: The clusters having users with similar access sequence.
Step 1: Start
Step 2: Initialize or update the fuzzy partition matrix Uij with equation (2)
Step3: Calculate the center vectors Cj using equation (3)
Step 4: Repeat step (2) and (3) until the termination criterion is satisfied.
Step 5: Stop
The fuzzy c-means procedure stars until the termination criterion is satisfied. Termination criteria can be that the difference between updated and previous objective function value J, is less than a predefined minimum threshold.

D. *Cluster Chase optimization Algorithm*

The objective of this approach is to reduce the inter cluster similarity and increase the intra cluster similarity along with scalability. The clustering routine optimizes number of clusters as well as cluster assignment, and cluster prototypes .

This paper proposes a Fuzzy Cluster-chase algorithm which takes the input from fuzzy clustering approach FCM that check the similarity by analysing the fuzziness measure. The fuzzy dispersion of clusters i and j is given as:

$$FD(Pi) = \{\frac{1}{ni(\sum_{xk \in Pi} uik[xk - ci]^{\frac{1}{2}})}\}$$

The separation or dissimilarity between the clusters is measured as: DSMC(Pi,Pj)= Ci – Cj (5)

Similarity of two clusters are:
SM(Pi,Pj)=FD(Pi)+FD(Pj)/DSMC(Pi,Pj) (6)

The following steps explain the Fuzzy Cluster-chase algorithm:

Input: N clusters which gives the representation of the URLs most frequently accessed by all members of that clusters.

Output: M clusters which minimizes intra cluster distance and maximizes the inter cluster distance.

```
Step 1: Start
Step 2: Initialize the value of i as 1
Step 3: Repeat the following steps until i is equal to N
Step 4: For each cluster i to N
Step 5: Check the similarity between two clusters Pi and Pi+1 by the equation (6)
Step6:If the similarity> $\partial$
Step 7: Check whether same user exist in both clusters
Step 8:If yes then check the membership value of the user in both clusters and delete the user form the cluster having low membership value and remains in the cluster having highmembership value.
Step 9: Stop
```

In this proposal, the value of ∂ = 0.65

Once all the iterations are finished we get M clusters which is less than the initial N clusters (M<N). Some clusters will have higher densities and some of them will be vanished.

E. SUGGESTION MODEL: WEIGHTED SUFFIX TREE

When cluster optimization procedure of access sequence is over then output of optimization algorithm that is each optimized cluster is expressed as a weighted sequences. In this approach web access sequence combined using their multiple sequence alignment. Finally when multiple sequence alignment of cluster's sequence has been calculated then all weighted sequences has been produced, their corresponding generalized weighted suffix tree is constructed. The generalizes weighted Suffix tree will represents all possible weighted sub_wards of all weighted sequences. The weighted suffix tree[10,11] can implicitly capture most important navigation experience. It works like prediction Model.

CONCLUSION

In this paper we proposed pre-processing of Web log data, apply clustering and optimization methods to get similar interest of particular user and finally to provide user related suggestion using prediction model. There are lots of challenges in web Mining and we need to solve them by apply different techniques on big log data .Our future work includes the prediction on Big Data.

REFERENCES

[1] Mr. P. Mittal, M. Yadav "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science and Software Engineering Z, Volume 3, Issue 3, March 2013.

[2] Cyrus Shahabi • Farnoush Banaei-Kashani, "Efficient and Anonymous Web-UsageMining for Web Personalization", INFORMS Journal on Computing © 2003 INFORMSVol. 15, No. 2, Spring 2003, pp. 123–147.

[3] V. Losarwar, Dr. M. Joshi " Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapor.

[4]] S. Alam, G. Dobbie, P. Riddle, "Particle Swarm Optimization Based Clustering Of Web Usage Data",IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 2008.

[5] A. Alphy, S.Prabakaran, "Cluster Optimization for Improved web Usage Mining using Ant-Nestmate Approach", IEEE-International Conference on Recent Trends in Information Technology, June 3-5, 2011.

[6] Petridou, S.G.; Koutsonikola, V.A.; Vakali, A.I.; Papadimitriou, G.I.,"time aware web users clustering",Knowledge and Data Engineering, IEEE Transactions on Volume:20,Issue:5,Page(s): 653 - 667 ,2008.

[7] P. Patil , and U.Patil , "Preprocessing of web server log file for web mining", World Journal of Science and Technology 2012, 2(3):14-18.

[8] Theint Theint Aye, " Web Log Cleaning for Mining Of Web Usage Patterns", IEEE 2011.

[9] R. Khanchana and Dr. M. Punithavalli, "A Web Usage Mining Approach Based On New Technique In Web Path Recommendation Systems", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 1, January- 2013.

[10] C.Dimopoulos, C. Makris, Y. Panagis, E. Theodoridis ,A. Tsakalidis,"A web page usage prediction scheme using sequence indexing and clustering techniques", ScienceDirect2010.

[11] Costas S. Iliopoulos, Christos Makris, Yannis Panagis, "The Weighted Suffix Tree: An Efficient Data Structure for HandlingMolecular Weighted Sequences and its Applications", Fundamenta Informaticae 71 (2006) 259–277 259,IOS Press.