WEB USAGE MINING: ANALYSIS DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE ALGORITHM

K.Dharmarajan[1], Dr.M.A.Dorairangaswamy [2]

[1]*Scholar Research and Development Centre Bharathiar University Coimbatore – 641046, (India)*
[2]*Senior Professor & HOD - CSE&IT, AVIT , Chennai (India)*

ABSTRACT

The web usage mining techniques provides them with discovered ability to extracting useful information from the web log, but often lack of noisy data noisy and irrelevant. This indicates, in a data mining process, the large number of patterns discovered can easily exceed the capabilities of a user to identify succeed interesting domains. Therefore identify these issues web log data utility measures have been used to reduce the patterns prior to presenting them to the user. This paper proposes an improved system first uses Density based spatial clustering of application with noise algorithms can identify clusters in behavior of the users page visits, order of sequential of visits. Therefore, analyzing users' Web log data and extracting users' potential interested domains and identify frequent sequential web access patterns.

Keywords: DBSCAN, Preprocessing. Web Usage Mining, Web Log

I. INTRODUCTION

Internet is an enormous repository of web pages and links. Web pages provides huge amount of information for Internet users. Today, data mining techniques are used by many companies to focus the customer retention. Financial, artificial intelligence, communication and marketing organization are the companies using the data mining techniques. Web usage mining is one of main application of mining techniques in logs. There is tremendous growth and growth in internet. Users' accesses are documented in web logs. So on to the web data and forecast the user's visiting behaviors and obtains their interests by investigating the samples. The log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information on his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. Weblog mining is a special case of usage mining, which mines Weblog entries to discover user traversal patterns of Web pages Thus, Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Application of Web Usage Mining:
Personalization: Restructure website based on user’s profile and usage behavior.
System Improvement: Provide key to understanding web traffic behavior. Advanced load balancing, data distribution or policies for web caching are benefits of such improvements.

Modification of Website: Understanding visitors' behavior in a web site provides hints for adequate design and update decision.

Business intelligence covers the application of intelligent techniques in order to improve certain businesses, mainly in marketing.

Characterization of use: is based e.g. on models that determine the pages a visitor might visit on a given site.

Web Usage Mining involves determining the frequency of the page access by the clients and then finding the common traversal paths of the users. Long and convoluted user access paths along with low use of a web page indicate that the web site is not laid out in an intuitive manner. With the help of this analysis, one can re-structure the web site with the navigation results. Some of the most used algorithms in this mining process include association rule generation, sequential pattern generation and clustering.

II. WEB USAGE MINING PROCESS

The main aim of the innovation system is to find web user clusters from web server log files [14]. These discovered clusters show the characteristics of the underlying data distribution. Clustering is useful in characterizing user groups based on patterns, categorizing web documents that have similar functionalities.

This method allows for the collected works of Web log information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server

Web Usage Mining is a four-step process. The first step is data collection, the second step is data pre-processing, the third step is pattern discovery and the last step is pattern analysis.

Preprocessing: The pre-processing stage involves cleaning of the click stream data and the data is partitioned into a set of user transactions with their respective visits to the web site. "consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery". This step can break into at least four sub steps.

- Data Cleansing.
- User Identification.
- User Session Identification.
- Formatting

Data Cleaning is the process of removing irrelevant items such as jpeg, gif, sound files and references due to spider navigation to improve the quality of analysis. User Identification is the process of identifying users by using IPaddress and user agent fields of log entries. A user session is considered to be all of the page accesses that occur during a single visit to a Web site.

2.1 Pattern Discovery

Draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Several methods and techniques have already been developed for this step as summarized below:

Statistical Analysis such as frequency analysis, mean, median, etc.

- Clustering of users help to discover groups of users with similar navigation patterns (provide personalized Web content).
- Classification is the technique to map a data item into one of several predefined classes.
- Association Rules discover correlations among pages accessed together by a client.
- Sequential Patterns extract frequently occurring inter-session patterns such that the presence of a set of items s followed by another item in time order.
- Dependency Modeling determines if there are any significant dependencies among the variables in the Web.

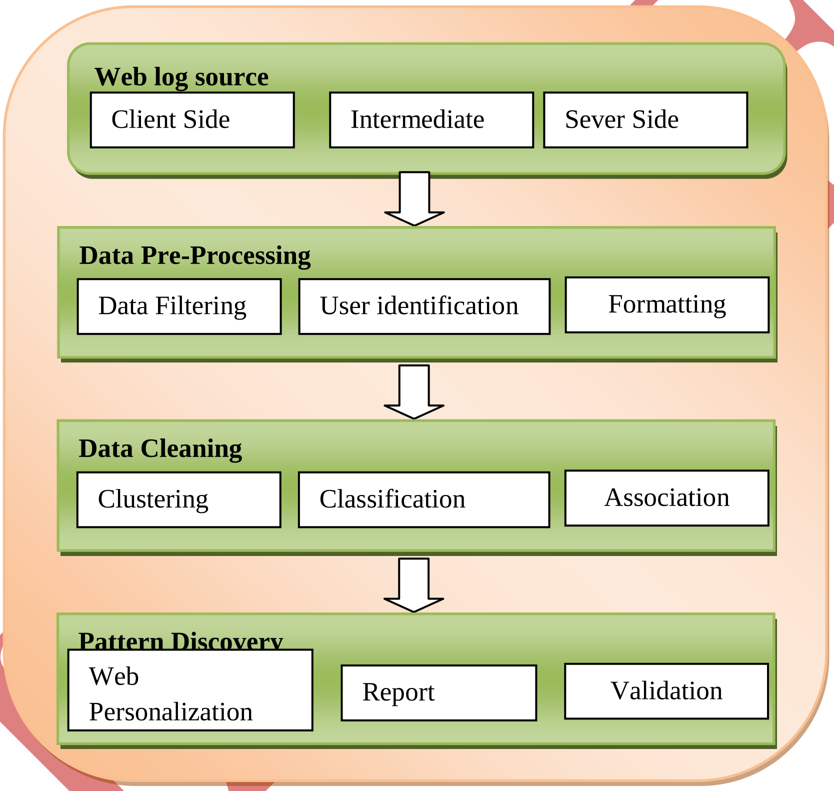

Fig. 1 Shows the Web Usage Mining Process

We choose clustering to discover users' navigational patterns. The pattern discovery stage is applying data mining techniques like path analysis, association rule mining, clustering, classification etc., on preprocessed log data. Here clustering technique is considered for pattern discovery. There are two types of clusters to be discovered: usage clusters and page clusters. Clustering usage data is to find visitor groups with common properties, interest or behavior. The aim of clustering web pages is to divide the dataset into groups of pages which have similar content. This study deals with clustering log data.

2.2 Pattern Analysis

Pattern Analysis is the final stage of WUM (Web Usage Mining), which involves the validation and interpretation of the mined pattern.

- Validation: to eliminate the irrelevant rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process.
- Interpretation: the output of mining algorithms is mainly in mathematic form and not suitable for direct human interpretations

The pattern analysis stage is to analyze the patterns found during the pattern discovery step. For analyzing multidimensional data OLAP cube or any visualization tool is used. Knowledge Query management or Intelligent Agents are also used for Pattern Analysis

III. WEB LOG FILES

Web Log Files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site, information of his request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. Log file range 1KB to 100MB.

3.1 Location of weblog file

Web log file is located in three different locations.

Web server logs: Web log files provide most accurate and complete usage of data to web server. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed.

Web proxy server: Web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server.

The two disadvantages are: Proxy server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction. The request interception is limited.

Client browser: Log file can reside in client's browser window itself. HTTP cookies used for client browser. These HTTP cookies are pieces of information generated by a web server and stored in user's computer, ready for future access.

3.2 Type of web log file

There are four types of server logs.

Access log file: Data of all incoming request and information about client of server. Access log records all requests that are processed by server.

Error log file: list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log .Access and error logs are mostly used, but agent and referrer log may or may not enable at server.

Agent log file: Information about user's browser, browser version.

Referrer log file: This file provides information about link and redirects visitor to site.

3.3 Web log file format: Web log file is a simple plain text file which record information about each user. Display of log files data in three different format.

- W3C Extended log file format
- NCSA common log file format
- IIS log file format

NCSA and IIS log file format the data logged for each request is fixed.W3C format allows user to choose properties, user want to log for each request. Normally weblog file contains data such as

- remotehost - domain name or IP address
- rfc931 - the remote logname of the user
- Authuser - user identification used in a successful SSL request
- [date] - the date and time of a request (e.g. day, month, year, hour, minute, second, zone)
- "request" - the request line exactly as it came from the client
- status - three-digit HTTP status code returned to the client (such as 404 for
- Page not found, or 200 for Request fulfilled)
- bytes - number of bytes returned to the client browser for the requested object ECFL has two additional elements:
- referrer - URL of the referring server and the requested file from a site
- agent - Browser and operating system name and version

IV. DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE ALGORITHM

A clustering algorithm that can do everything that DBSCAN can do is not yet available various new clustering algorithms appear occasionally. DBSCAN has been modified to great extent recently and used to derive a new procedure to calculate EPS (threshold distance) which are most important parameters [4]. The density-based clustering algorithm presented is different from the classical Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and has the following advantages: first, Greedy algorithm substitutes for R*-tree in DBSCAN(Density based spatial clustering of application with noise) to index the clustering space so that the clustering time cost is decreased to great extent and I/O memory load is reduced as well; second, the merging condition to approach to arbitrary- shaped clusters from [1] is designed carefully so that a single threshold can distinguish correctly all clusters in a large spatial dataset though some density-skewed clusters live in it from the outliers in [3]. Finally, authors investigate a robotic navigation and test two artificial datasets by the proposed algorithm to verify its effectiveness and efficiency.

V EXPERIMENTAL RESULTS

A clustering algorithm DBSCAN has been modified to great extent recently and used to derive a new procedure to calculate EPS (threshold distance) which are most important parameters from the section [4]. The log data used in this paper is extracted from a research website at www.mysite.com.

Algorithm: Proposed new Dbscan clustering:

Step1:

Construct the similarity matrix using S3M measure (Definition 1).

Step2:

```
Select all points from D that satisfy the Eps and Minpts
C = 0
for each unvisited point P in dataset D
mark P as visited
N = get Neighbors (P, eps)
if sizeof(N) < MinPts
mark P as NOISE
else
begin
C = next cluster
mark P as visited
end
add P to cluster C
for each point P' in N
if P' is not visited
mark P' as visited
N' = getNeighbors(P', eps)
if sizeof(N') >= MinPts
N = N joined with N'
if P' is not yet member of any cluster
add P' to cluster C
```

Step 3:

Return C

Step 4:

For all Ti є U Compute Si= R(Ti)

Using definition 2 for given threshold d.

Step 5:

Next compute the constrained-similarity upper Approximations Sj for relative similarity r using definition 3

if Si= SJ

end if

Step 6:

Repeat step 3 until U≠ Ø;

Return D.

End

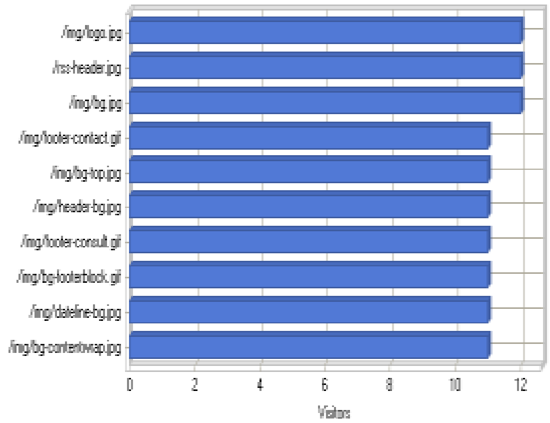

Fig. 2 Shows the DBSCAN Clustering Efficiency

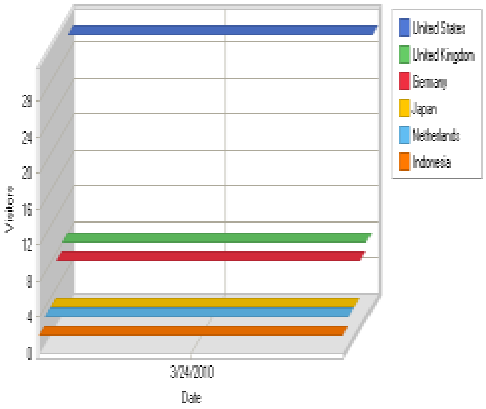

Fig. 3 Shows the DBSCAN algorithm

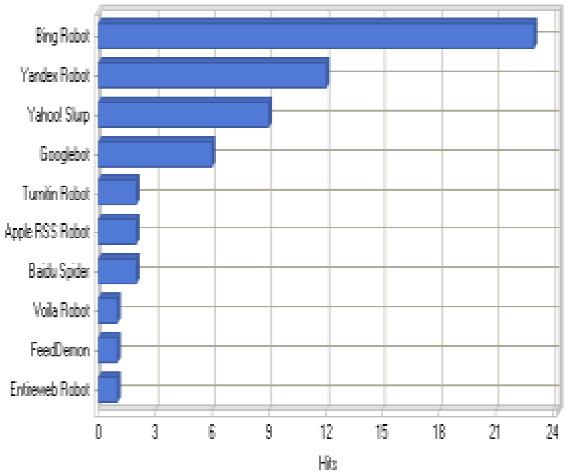

Fig. 4 shows the user access patterns

DBSCAN algorithm calculates the web site visits each point of the page, possibly multiple times view the page have calculated. For practical considerations, however, the time complexity is mostly run by the number of region Query invocations. DBSCAN executes exactly one such query for each point, and if an indexing structure is used from [6] that view such a neighborhood query in Ologn, an overall web page complexity of O(n.logn) is obtained . Without the use of an increase speed index structure, the page loading time complexity is O(n2). Often the distance matrix of size (n2 – n)/2 is materialized to avoid distance repetition.

The results on www.mysite.com which is useful in identify the user access patterns and behavior of visits of the hyperlinks of the each user and the inter cluster similarity among the clusters**.**

VI. CONCLUSION

This paper DBSCAN algorithm and presented experimental results for discovers frequent web accessing sequences from Weblog databases. It is useful in finding the user access patterns and the order of visits of the web page of the each user and the inter cluster similarity among the clusters. It is used for improving efficiency of website design, and also able to lead to excellent marketing decisions which introduce the concept of "utility" into web log file analysis. As utility measures the "interesting" or "usefulness" of a webpage, thus satisfies the DBSCAN in quantifying the user preferences of ease in web data transactions and also will gives better results compared to the rough set agglomerative clustering algorithms.

REFERENCES

[1] G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi, "Web Users Session Analysis Using DBSCAN and Two Phase Utility Mining Algorithms," IJSCE, Volume-1, Issue-6, January 2012

[2] Akiladevi R, Naveen Sundar, "A Survey On Web Log Mining Using Dbscan," IJRCAR , Vol.1 Issue.9, Pg: 144-149

[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". KDD-96.

[4] Priyamvada Paliwal and Meghna Sharma, "Enhanced DBSCAN Outlier Detection," IJARCSSE. Vol3, Issue 3, March 2013.

[5] S.Vijayalaksmi and M Punithavalli, "A Fast Approach to Clustering Datasets using DBSCAN and Pruning Algorithms," IJCA, Volume 60– No.14, December 2012.

[6] Bhaiyalal Birla, and Sachin Patel, "An Implementation on Web Log Mining" IJARCS. Volume 4, Issue 2, February 2014.

[7] Rekha Awasthi1, Anil K Tiwari, Seema Pathak3, "Analysis of Mass Based and Density Based Clustering Techniques on Numerical Datasets" IISTE. Vol.3, No.4, 2013.