Volume 6, Issue 2, February 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# Innovative Pre-Processing Technique and Efficient Unique User Identification Algorithm for Web Usage Mining

[1]Ranjena Sriram, [2]Dr. R. Mallika

[1] Research Scholar, Computer Science, Karpagam University, Eicheneri Post, Coimbatore, Tamil Nadu, India

[2] Assistant Professor, Department of Computer Science, CBM College, Kovaiputhur, Coimbatore, Tamil Nadu, India

***Abstract - This current study focuses on proposing a new pre-processing and unique user identification algorithms for Web Usage Mining to discover and analyse user access pattern through mining of log files or log databases and associated data from a particular websites. Pre-processing technique to clean the data and user identification process to identify unique users. Since number of users interacting with web sites around the world are increasing day by day, the amount of data generated and information gathered could help the organisations to improve their business according to the Customers' needs and behaviour. This work proposes an efficient pre-processing technique and an innovative Hashing technique - (a Hash table and a Hash function have been proposed) to identify Distinct User for Web Usage Mining. The proposed pre-processing technique has been evaluated by comparing with existing pre-processing techniques to prove its accuracy and efficiency. Similarly the Hashing techniques have been compared with existing searching methodologies and it has been proved that the proposed technique is quick in searching according to Big O notation.***

***Key Words: Web Usage Mining, Hashing Techniques, UUI(Unique User Identification).***

## I. INTRODUCTION

Millions of users access several websites for day to day transactions. Web servers play an important role in mining these transactions by various ways. Each web server maintains a server log which holds and maintains many information like user information (Users who access the site), Server behaviour, potential benefits of new technical developments etc. Many institutions have not been able to perform an effective use of Web Server log files for enhancing and improving server performance and design we need to identify the way of user accessing the web pages in particular session time.

Huge volume of data is stored in the server log depending on the capacity of server storage. When data has to be mined for identifying unique pattern it took more time due to the data stored in the server. To overcome these problems this study focuses on two major areas.

i. An efficient algorithm to clean web log data

ii. A unique algorithm using hashing technique to identify unique users.

The study proposes a fast active distinct user identification algorithm which uses a Hashing technique blended with an IP address and a finite user's inactive time to identify different users in the web log file. Though man works have been proposed and implemented, none produced a better quality in the results produced when the data size increases in the Web Server. Experimental results have proved that the algorithms proposed in this work shows better results for Web Servers with huge data size. The results have also proved the generalized behaviour of the algorithms for different Web Serves with different data and attribute information.

## II. WEB USAGE MINING

Web usage mining is the application of Data Mining Techniques to discover interesting usage patterns from web data, in order to understand and better serve the needs of web-based applications. It tries to make sense of the data generated by the web surfer's Sessions behaviours. While the web content and structure mining utilize the primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web.Web usage mining analyses results of user interactions with a web server, including weblogs, click streams, and database transactions at a web site of a group of related sites.

Web Usage Mining is a three phase process consisting of

a) Pre-processing / Data Preparation

b) Pattern Discovery

c) Pattern Analysis

**Pre-processing / Data Preparation :** The important task in any data mining application is the creation or selection of proper dataset to which suitable data mining or statistical algorithms can be applied to generate classes for further processes. This is particularly important in Web Usage Mining due to the characteristics of click stream data and its relationship with other related data collected from multiple sources across multiple channels. The data preparation step is the most tome consuming and computationally intense step in Web Usage Mining process, and often requires special algorithms and heuristics not common to other domains. This process is critical to extract useful information for mining. The process may involve pre-processing the original data, integrating data from different sources, and transforming the integrated data into a form suitable for input into specific data mining operations. Collectively we refer to the process as pre-processing preparation [1].

**Pattern Discovery**

Statistical methods, Data mining methods, Associate rule, Sequential methods and cluster techniques are used to identify unique patterns.

**Pattern Analysis Phase**

Discovered patterns are analyzed using OLAP tools, knowledge query management system and intelligent agents to remove uninteresting rules/patterns.

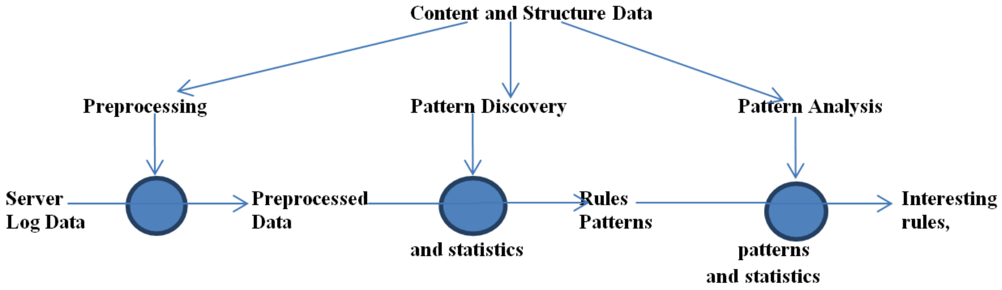

Figure 1: Process of Web Usage Mining.

## III. SOURCES AND TYPES OF DATA

The primary data source used in Web usage mining is the server log files, which include Web server access logs and application server logs. Additional data sources that are essential for both data preparation and pattern discovery include the site files and meta-data, operational databases, application templates, and domain knowledge. In some cases and for some users, additional data may be available due to client side or proxy level (Internet Service Provider) data collection , as well as from external clickstream or demographic data sources such as those provided by data aggregation services . The data obtained from various sources can be categorized into four primary groups[2].

### *A. Usage Data*

The log data collected automatically by the Web ad application server's represents the fine-gained navigational behaviour of visitors.It is the primary source of data in Web usage mining. Each hit against the server, corresponding to an HTTP request, generates a single entry in the server access logs. Each log entry (depending on the log format) may contain fields identifying the time and date of the request, the IP address of the client, the resource requested, possible parameters used in invoking a Web application, status of the request, HTTP method used, the user agent (browser and operating system type and version), the referring Web resource, and, if available, client-side cookies which uniquely identify a repeat visitor.

Depending on the goals of the analysis, this data needs to be transformed and aggregated at different levels of abstraction. In Web Usage Mining, the most basic level of abstraction is that of a page view. A page view is an aggregate representation of a collection of Web objects contributing to the display on a user's browser resulting froma single user action. Conceptually, each page view can be viewed as a collection of Web objects or resources representing a specific user event.

### *B. Content Data*

Content Data in a site is the collection of objects and relationships that is conveyed to the user. For the most part, this data is comprised of combinations of textual materials and images. The data sources used to deliver or generate this data include static HTML/XML pages, multimedia files, dynamically generated page segments from scripts, and collections of records from the operational databases. The site content data also includes semantic or structural meta-data embed within the site or individual pages, such as descriptive keywords, document attributes, semantic tags, or HTTP variables. The underlying domain ontology for the site is also considered part of the content data. Domain ontologies may include conceptual hierarchies over page contents, such a product categories, explicit representation of semantic content and relationships **via** an ontology language such as RDF, or a database schema over the data contained in the operational database.

### *C. Structured Data*

The structured data represents the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. The structure data also includes the intra-page structure of the content within a page. For example, both HTML and XML documents can be represented as tree structures over the space of tags in the page. The hyperlink structure for a site is normally captured by an automatically generated site map. A site mapping tool must have the capability to captire and represent the inter and intra page view relationships. For dynamically generated pages, the site mapping tools must either incorporate intrinsic knowledge o the underlying applications and scripts that generate HTML content, or must have the ability to generate content segments using a sampling of parameters passed to such applications or scripts.

### *D. User Data*

The operation database(s) for the site may include additional user profile information. Such data may include demographic information about registered users, user ratings on various objects such as product or movies, past purchases or visit histories of users, as well as other explicit or implicit representations of user's interests. Some of this data can be captured anonymously as long as it is possible to distinguish among different users. For example, anonymous information contained in client-side cookies can be considered a part of the users profile information, and used to identify repeat visitors to a site. Many personalization applications require the storage of prior profile information.

## IV. RELATED WORK

The study focuses on data cleaning and user identification steps in pre-processing technique for Web Usage Mining.

An efficient algorithm for data cleaning and a unique and versatile algorithm for identifying distinct users have been proposed in this work. The data preparation process is often the most time consuming and computationally intensive step in the Web usage mining process. The process may involve

- Preprocessing the original data
- Integrating data from multiple sources
- Transforming the integrated data into a specific format useful for mining process.(This process is called as Data Preparation).

The input for Web Usage Mining is a user session file which consists of the following information

- Exact account of who accessed the Web site,
- What pages and in what order,
- How long each page was viewed.

Data Preprocessing consists of following steps

1. Data Cleaning
2. User Identification
3. Session Identification
4. Path Completion.

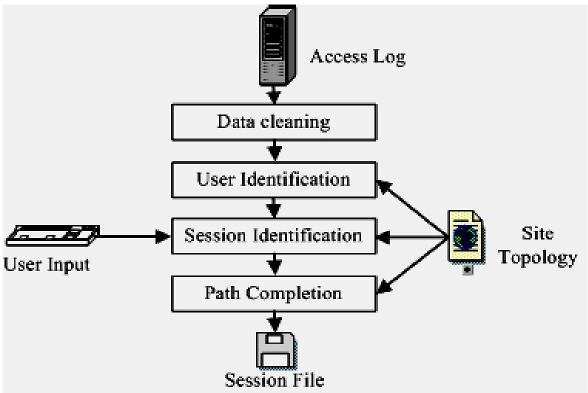

Figure 2: Complete Pre-processing Technique

Figure 2 illustrates the sequence of steps involved in Pre-processing technique. The role of each process and the implementation of the proposed algorithms are discussed below.

### *A. Data Cleaning*

The principle of data cleaning is to reduce extraneous items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. More emphasis has to be made on the following issues discussed below to make a perfect cleaning process to extract clear data for further mining.Extraneous records in web access file will be eliminated. The following records have to be eliminated. The records containing videos, graphics and containing file extensions of GIF, JPEG, and CSS, which can be found in URL, field of every record.The records with the failed HTTP status code. By grouping the Status field of every record in the web ac-cess log, the records with status codes over 299 or un-der 200 are removed. Different from most other researches, records having value of POST or HEAD in the Method field are re-served in present study for acquiring more accurate referrer information [3].

Though several algorithms have been proposed none produced accurate results, keeping these facts the proposed algorithm pays more emphasis on how to eliminate extraneous records and also the key factors which would decrease the accuracy of data to be mined. The algorithm carefully checks each record and eliminates then if it contains any record which matches to the above points discussed . References taken from the work done by ShuyanBai ; Vocational Coll. of

Yantai, Yantai ; Qingtian Han ; Qiming Liu ; XiaoyanGao in their work "**Research of an Algorithm Based on Web Usage Mining",** published in IEEE gave a brief idea on how to proceed in designing and developing the algorithm.[4].

### *B. Proposed Data Cleaning Algorithm*

```
The proposed algorithm for data cleaning is given below:
Input: Web server Log File
Output: Log Database
        Step1: Read LogRecord from Web Server Log File
        Step2: If (LogRecord .url-stem (gif.jpegjpg.cssjs))
                AND(LogRecord.method='GET') AND
                LogRecord.Sc-status<>(301,404,500)AND
                (LogRecord.Useragent<>Crawler.Spider.Robot))
                Then insert LogRecord in to LogDatabase.
        End of If condition.
        Step3: Repeat the above two steps until eof
        (Web Server LogFile)
        Step4: Stop the process.
```

Figure 3. Proposed Data Cleaning Algorithm

### Execution of the Algorithm

The organizations Web server log file is fed as input and each record is taken in sequence to check for the following prevailing conditions. If the log record containsany image files extensions gif, jpeg, css or js, if the record contains Get method , if the record returns 301, 404, 500, if the record contains Crawler, Spider or Robot agents, the record is eliminated preventing from insertion into the Log database. Successful records dissatisfying these conditions are inserted into the Log database for further mining to generate unique classes. The steps are repeated until end of file condition is satisfied.

### *C. User Identification*

User's identification is, to categorize who access web site and which pages are accessed. Different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study.

The rules adopted to distinguish user sessions can be described as follows:

- Each IP address represents one user;
- For more logs, if the IP address is the same, but the agent log shows a change in browser software or operating system, an IP address represents a different user
- Using the access log in conjunction with the referrer logs and site topology to construct browsing paths for each user. If a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, there is another user with the same IP address.

### *D. Session Identification*

A user session means a delimited set of user clicks (click stream) across one or more Web servers.

The following is the rules we use to identify user session in our experiment

- If there is a new user, there is a new session.
- In one user session, if the refer page is null, there is a new session.
- If the time between page requests exceeds a certain limit (30 minutes), it is assumed that the user is starting a new session.

### *E. Path Completion*

As the reality of local cache and proxy server, there are many important accesses that are not recorded in the access log. The task of path completion is to fill in this missing page similar to those used for user identification can be used for path completion. Methods similar to those used for user identification can be used for path completion.

### *F. Proposed Unique User Identification Algorithm Using Hashing Technique*

Unique user identification is important process next to data cleaning. Unique users are identified based on the rules suggested in User Identification section. Though many efficient algorithms are there, many fail in accuracy and efficiency (time taken to identify users) when the size of the Log Database increases. Today's modern web servers are capable of handling terabytes of data conventional algorithms are obsolete in handling these scenarios. Considering the above facts this study has proposed an efficient Unique User Identification algorithm that uses modern Hashing techniques to identify unique user quickly inspire the huge size of the database. A new hashing key has been prosed and successfully implemented in the algorithm to locate the user [7].

Hashing Techniques

For a huge database structure, it can be almost next to impossible to search all the index values through all its level and then reach the destination data block to retrieve the desired data. Hashing is an effective technique to calculate the direct location of a data record on the disk without using index structure.

Hashing uses hash functions with search keys as parameters to generate the address of a data record[5].

Hash Organization

- **Bucket** − A hash file stores data in bucket format. Bucket is considered a unit of storage. A bucket typically stores one complete disk block, which in turn can store one or more records.
- **Hash Function** − A hash function, **h,** is a mapping function that maps all the set of search-keys **K** to the address where actual records are placed. It is a function from search keys to bucket addresses.

Dynamic Hashing

The problem with static hashing is that it does not expand or shrink dynamically as the size of the database grows or shrinks. Dynamic hashing provides a mechanism in which data buckets are added and removed dynamically and on-demand. Dynamic hashing is also known as **extended hashing [6]**.

Hash function, in dynamic hashing, is made to produce a large number of values and only a few are used initially.

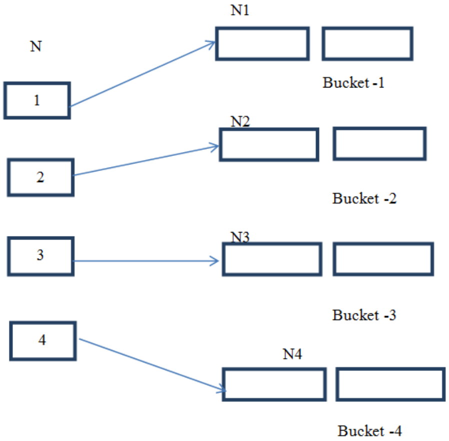

Figure 4. Dynamic Hashing Technique

Considering this actuality, we presented a new algorithm called "UUI (UNIQUE USER IDENTIFICATION)". It analyses more factors, such as user's IP address, Web site's topology, browser's edition, operating system and referrer page. This algorithm possesses preferable precision and expansibility. It can not only identify users but also identify session. Session identification will be discussed in next section. Proposed method shows comparison not only based on User_IP some-where same User_IP may generate the different web users, based on path which chosen by any user and access time with referrer page we find out the distinct web user [8].

When huge databases were taken for consideration the time taken to locate the records was much, hence appropriate methodology has to be incorporated to make the process faster. Taking these prevailing conditions, the study has proposed a new Hashing formulation, to minimize the searching time for large datasets. The formation of the Hashing technique is discussed below.

Proposed Hash Function

$$\mathbf{Nmod_2 * K} \qquad \mathbf{(1.1)}$$

Where N refers the record number indirectly pointing the data (an IP address or an Operating system or a browser) and K refers to the virtual address of the bucket. The multiplied factor gives the original location of the data.

Substitute Nmod$_2$ with parameter H equation (1.1) becomes

$$\mathbf{H(K)} \qquad \mathbf{(1.2)}$$

This proposed function is substituted in the algorithm to identify the unique user in a quick manner.

```
Unique User Identification (UUI)
Definition: given a clean and filtered web log file and record set web log file
Records  R= {r1,r2,r3……r.n}
where n>0
Step1:    input Log database RUser of N records
Step2:    Distinct User identification base
Step3:    RUser=P<url, ip_addr, agent, method, operating system,status,session id,time_stamp>
Step4:    RUser=<r1,r2,r3…rn> where n!=0,i=0
Step5:    while (i<n)
Step 6:   while (Log database<>eof)
Step7:    read Log database RUser
Step8 :   Substitute the proposed Hash Function $R(i)mod_2 * K(i)$.userip not part of Distinct user
          identification basethen it treated as new user and copy userip in distinct user Identification base.
Step9:    end if
Step 10   end loop(Log database)
        Step11:    i=i+1;
Step12:   end loop (Web log file)
Setp13:   end
```

Figure 5. Unique User Identification (UUI) Algorithm

## V. RESULTS AND DISCUSSIONS

To validate the effectiveness and efficiency of the algorithms proposed, an experiment with the web server logs of MurdocUniversity, Dubai and Emirates College of Management and Information Technology, Dubai was made. The initial data source of our experiment is from JAN 1, 2014 to Aug 3, 2015, with data size of $10^{12}$ records. Our experiments were per-formed on a 2.8GHz IntelICeleronI, CPU, 2.00 GB of main memory, Windows 2000 professional, SQL Server 2000 and MATLAB (7.9.0.529). MATLAB tool was used to develop applications to evaluate the performance of the proposed algorithms. The table listed below illustrates the overall performance of UUI algorithm.

Table I. Reults of Datacleaning Process For Murdoc University And Ecmit College

| DATA SOURCES | MURDOC UNIVERSITY | EMIRATES COLLEGE OF MANAGEMENT AND INFORMATION TECHNOLOGY |
|---|---|---|
| Entries in raw web log | 100000279900 (records) | 100450279900 (records) |
| Entries after data cleaning | 100000002783 (records) | 100270002783 (records) |
| Number of users | 567502876 | 606920287 |
| Number of Unique users | 436675422 | 445275422 |
| Execution time of UUI(Algorithm) | 3.257(s) | 4.437(s) |
| Number of sessions | 546744372 | 586744372 |

From Table I it is clearly evident that the algorithm works fine for a large dataset. Results prove that the proposed UUI algorithm consumes relatively less time to find whether the user record already exists or a new one. The following sections describe the overall performance of the UUI Experiments using MORDOC University server log data to evaluate its performance.

Table II. Overall Performance of Proposed UUI Algorithm

| Month | Unique Visitors | Number of Visits | Pages | Hits | Bandwidth |
|---|---|---|---|---|---|
| Jan 2014 | 747792371 | 747592371 | 947592371 | 1947592371 | 1.9 GB |
| Feb 2014 | 726527342 | 736527342 | 836527342 | 2836527342 | 1.85 GB |
| Mar 2014 | 718945720 | 718945720 | 718945720 | 1718945720 | 1.65 GB |
| Apr 2014 | 727654381 | 717654381 | 817654381 | 1817654381 | 1.7 GB |
| May 2014 | 625678990 | 655678990 | 755678990 | 1755678990 | 1.54 GB |
| June 2014 | 543298760 | 443298760 | 643298760 | 1643298760 | 1.22 GB |
| July 2014 | 456789321 | 556789321 | 656789321 | 1656789321 | 1.02 GB |
| Aug 2014 | 326789900 | 326789900 | 526789900 | 1526789900 | 1.00 GB |
| Sep 2014 | 0 | 0 | 0 | 0 | 0 |
| Oct 2014 | 0 | 0 | 0 | 0 | 0 |
| Nov 2014 | 0 | 0 | 0 | 0 | 0 |
| Dec 2014 | 0 | 0 | 0 | 0 | 0 |
| **Total** | **4873476785** | **4903276785** | **5903276785** | **14903276785** | **11.88** |

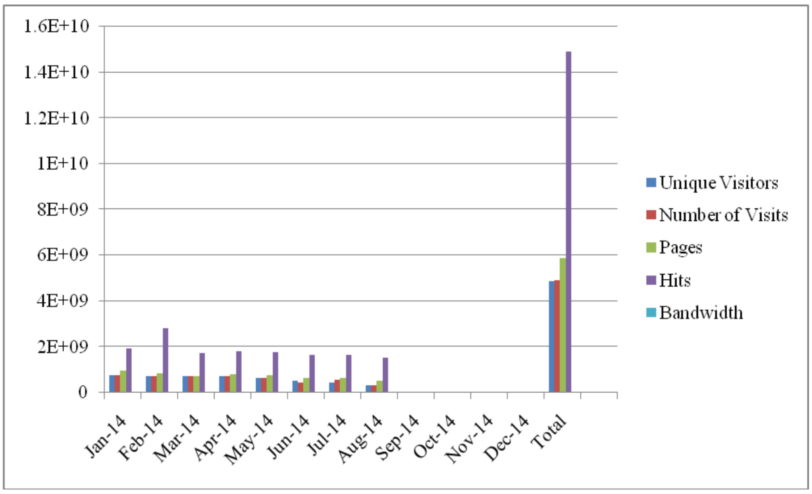

Figure 6. Graphical results of (UUI) Algorithm

## VI. CONCLUSION

In this Research a Unique user identification technique which enhancement of pre-processing steps of web log usage data in data mining has been proposed. The study uses two pre-processing technique combine within one pre-processing step time of user identification we find out distinct user based on their attended session time. Here introduced one proposed algorithm for advanced pre-processing DUI algorithm is very efficient as compare to other identification techniques. Future work needs to be done to combine whole process of Web Usage Mining. A complete methodology covering such as pattern discovery and pattern analysis will be more useful in identification method.

## REFERENCES

[1] Web UsageMining by *BamshadMobasher*

[2] "An effective data pre-processing method for Web Usage Mining",IEEE, ISBN978-1-4673-5786-9 ,Page(s):7-10.

[3] "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data" JaideepSrivastava * t, Robert Cooley:l: , MukundDeshpande, Pang-Ning Tan Department of Computer Science and EngineeringUniversity of Minnesota 200 Union St SE Minneapolis, MN 55455

[4] "Research of an Algorithm Based on Web Usage Mining",ShuyanBai ; Vocational Coll. of Yantai, Yantai; Qingtian Han ; Qiming Liu ; XiaoyanGao IEEE, ISBN: 978-1-4244-3893-8, Page(s): 1-4.

[5] "Review and Analysis of Hashing Techniques", International Journal of Advanced Research in Computer Science and Software Engineering", Volume 4, Issue 5, May 2014, Page(s) : 296-297.

[6] "Comparative study of Hashing Algorithm Using Cryptographic and Steganography Using Audio Files", SangeetaRaheja, International Journal of Advanced Research in Computer Science and Software Engineering", Volume 4, Issue 5, May 2014, Page(s):292-294.

[7] "Advanced Preprocessing using Distinct User Identification in web log usage data", Sheetal A. Raiyan, International Journal of Advanced Research in Computer and Communication Engineering Vol. 1, Issue 6, August 2012 Copyright to IJARCCE www.ijarcce.com 418 , Page(s) : 418-420.

[8] "An Exclusive Survey on Web Usage Mining For User Identification", Satpal Singh," International Journal of Innovative Research in Computer and Communication Engineering", Vol. 2, Issue 11, November 2014 Page(s):6582-6586.