“A new approach for user identification in web usage mining preprocessing”

1. Arvind kumar Dangi, 2. Asst. Prof. Sunita Sangwan
1. (Software Engineering, PDM College of engineering, India) akdangi@gmail.com
2. (Computer Science & Engineering, PDM College of engineering, India) sunita2009@gmail.com

Abstract: Web usage mining is a subset of data mining. In order to huge amount of data but the data is less appropriate “quantity and quality” of the web data is opposite to each other this is the main problem. Web data usage Mining is a variation of this field is untapped source of richly offered free textual information. The importance of web data usage mining is mounting along with the immense volumes of data generated in web habitual existence data always arrives in a various, continuous, rapid and time varying flow. Web data usage mining taking out procedures are important in extracting useful streaming on-line sources. As throughout the globe no. of web users are continuously and rapidly growing, it is necessary for the web usage miners to utilize efficient tools in order to discover, extort, clean and assess the desired information. The data pre-processing stage is the most important phase in the preprocessing for investigation of the web user & his usage behavior. To fulfill this requirement the navigations are recorded in web log file as well as the IP address of the website, session of usage & visited web link. In order to improve the performance & quality of data preprocessing in order to identify unique users and user sessions. We propose a new method for web data preprocessing in which it has three phases. “In the first phase some websites are selected and by different locations access these website & by applying the (java) tools & methods then find out the IP address of that websites, session usage time & navigations, in the final phase combine them i.e. (web link navigation + IP address of website + session of usage ). This framework helps to investigate the web user usage behavior efficiently.

Keywords: - webslink, navigation, threading, networking, diminution, web logs, session, Data clearout, Data makeover.

I. INTRODUCTION

Data pre-processing is an important and required phase in Web usage mining. Different methods & techniques are employed in web usage mining to remove irrelevant items and identify IP address, sessions of usage and navigation along with the extraction information. There is no of steps define for preprocessing as Data clear out also known as the routines work to “clean” the data by satisfying in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data is noisy, they are improbable to trust the results of any data mining that has been applied to it. Also, dirty data can cause uncertainty for the mining procedure, resulting in unreliable output. But, they are not always robust. a new method for web data preprocessing in which it has three phases. “In the first phase some websites are selected and by different locations access these website & by applying the (java) tools & methods then find out the IP address of that websites, session usage time & navigations, in the final phase combine them i.e. (web link navigation + IP address of website + session of usage ). This framework helps to investigate the web user usage behavior efficiently.
II. INDENTATIONS AND EQUATIONS

2.1 Data clear out (cleaning)

Data clear out also known as the routines work to “clean” the data by satisfying in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

2.2 IP address identification & Web link visited identification

Using the following function & code we identify the IP address & web link visited by the end user using the java language.

```java
try {
    Statement st = con.createStatement();
    int a = st.executeUpdate("INSERT INTO cdass.tracing (userid, 'IP', visited_page) 
" + "VALUES
("+userid+", "+ip+", "+page.toString()+");\n" + ");
} catch (Exception e) {
    System.out.println("Error During tracing..............." + e);
} finally {
    try {
        con.close();
    } catch (SQLException ex) {
        Logger.getLogger(track.class.getName()).log(Level.SEVERE, null, ex);
    }
}
```

2.3 Session of usage identification

Identify the session of usage by the user at particular web site or the web page.

```java
long d = session.getCreationTime();
long dd = session.getLastAccessedTime();
out.print(new Date(dd).getSeconds() - new Date(d).getSeconds());
session.removeAttribute("userid");
```

Session Tracing

```jsp
<%=
    tracing.track.setTracking((String) session.getAttribute("userid"), request.getRemoteAddr(), request.getRequestURL());
<%>

<%=
    long d = session.getCreationTime();
    long dd = session.getLastAccessedTime();
    // out.print(new Date(dd).getSeconds());
    out.print(new Date(dd).getSeconds() - new Date(d).getSeconds());
    // out.println((dd-d)/60);
    // System.out.println(session.getLastAccessedTime());
    session.removeAttribute("userid");
    // response.sendRedirect("login_form.jsp");
<%>
```

Packages used for achieving the above tasks

import java.sql.Connection;

The JDBC (Java Database Connectivity) API defines interfaces and classes for writing database applications in Java by making database connections. Using JDBC you can send SQL, PL/SQL statements to almost any relational database. JDBC is a Java API for executing SQL statements and supports basic SQL functionality. It provides RDBMS access by allowing you to embed SQL inside Java code. Because Java can run on a thin client, applets embedded in Web pages can contain downloadable JDBC code to enable remote database access.

import java.sql.Statement;

To execute a SQL statement on your table, you set up a Statement object. So add this import line to the top of your code:

import java.sql.Statement;
In the try part of the try ... catch block add the following line (add it just below your Connection line):

```java
Statement stmt = con.createStatement();
```

Here, we're creating a Statement object called `stmt`. The Statement object needs a Connection object, with the `createStatement` method.

We also need a SQL Statement for the Statement object to execute. So add this line to your code:

```java
String SQL = "SELECT * FROM Workers";
```

The above statement selects all the records from the database table called `Workers`.

We can pass this SQL query to a method of the Statement object called `executeQuery`. The Statement object will then go to work gathering all the records that match our query.

```java
import java.util.logging.Level;

The java.util.logging package provides the logging capabilities via the Logger class.

To create a logger in your Java coding you can use the following snippet.

```java
Private final static logger  logger = logger.getLogger(myclass.class.getName());
```

The Logger you create a actually a hierarchy of Loggers, and a . (dot) in the hierarchy indicates a level in the hierarchy. So if you get a Logger for the `com.example` key this Logger is a child of the `com` Logger and the `com` Logger is child of the Logger for the empty String. You can configure the main logger and this affects all its children.

<table>
<thead>
<tr>
<th>SIGN IN TIME</th>
<th>SIGN OUT TIME</th>
<th>BROWSER AGENT</th>
<th>URL</th>
<th>IP</th>
</tr>
</thead>
</table>

### 2.4 Data integration
Integration we mean that a dataset can be retrieved and added to other datasets to create greater, more robust and useful datasets. The discipline of data integration comprises the practices, architectural techniques and tools for achieving the consistent access and delivery of data across the spectrum of data subject areas and data structure types in the enterprise to meet the data consumption requirements of all applications and business processes.

### 2.5 Data makeover
Data makeover process applies on the data which is collected from different sources of the data. In the make over process it organized the data minimizes redundancy and dependency of the data.

### 2.6 Data diminution
These are the technique that can be applied to get a reduced representation of the data set that is much smaller in volume maintain the data integrity of the original data.

### 2.7 Usage mining
Web Usage Mining process is a user session file that gives an exact account about the user accessed the Web site, what pages were visited and in what order, and how much time each page was visited. A user session is the set of the page accesses that occur during a single visit to a Web site. the information contained in a raw Web server log represent a user session file before data preprocessing.

#### 3.1 Framework of activities

📝 Data clear out (cleaning)
⌨️ IP address identification
🔍 Session of usage identification
🌟 Web link visited identification
📊 Data integration
🔍 Data makeover
🔍 Data diminution
🔍 Usage mining

---

III. FIGURES AND TABLES

![Diagram of Web usage mining Preprocessing Framework](image-url)
3.2 Research work representation

User identification is identifying each user accessing Web site
User IP+Browser (User ID, IP, OS or cookie) => Identify the users
Session identification is to find each user’s access pattern and frequency path.

3.3 Sequential steps which followed for the research work

- Data clear out
- IP address identification
- Session of usage identification
- Web link visited identification
- Data integration
- Data makeover
- Data diminution
- Usage mining
IV. CONCLUSION

There are several data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in data. Data integration merges data from multiple sources into a logical data store such as a data warehouse. Data diminution can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering. Usage of data mining method and data extraction on the web is now highlighting of the boosting number of researchers. web usage mining is a type of data mining technique that can be useful in suggesting the web usage type with the identification of user session, this research also help in identification of the end user behavior with the help of the identification of the user (IP address + web link navigation + session usage) and combine them together for identification user behavior more accurately for preprocessing data. This research mainly concentrates on providing another approach for preprocessing the data more accurately. Experimental results suggest the significance of the proposed approach.

REFERENCES