IJETCR Open Access

ISSN: 2348 - 2117

International Journal of Engineering Technology and Computer Research (IJETCR)
Available Online at www.ijetcr.org
Volume 3; Issue 2; Page No. 164-168

Extracting Weighted Frequent Patterns from Web Log Data

Priyanka Baraskar[1], Supriya Chavan[2], Dipshree Dhage[3], Samruddhi Giri[4], Jayshree Jha[5]

[1] Student, Pursuing B.E., Information Technology, Atharva College of Engineering, Mumbai, India

piya.baraskar@gmail.com

[2] Student, Pursuing B.E., Information Technology, Atharva College of Engineering, Mumbai, India

supriyachavan321@gmail.com

[3] Student, Pursuing B.E., Information Technology, Atharva College of Engineering, Mumbai, India

dipshree1004@gmail.com

[4] Student, Pursuing B.E., Information Technology, Atharva College of Engineering, Mumbai, India

09samruddhigiri@gmail.com

[5] Professor, Information Technology, Atharva College of Engineering, Mumbai, India

09samruddhigiri@gmail.com

ABSTRACT

World Wide Web (WWW) today is growing into infinity and it has massive wealth of information. Mining the web is very essential in order to retrieve the necessary information for any user. Sequential Pattern Mining using weights involves applying data mining methods to large web log data to extract most weighted frequent patterns. In this paper, Sequential Pattern Mining is performed using weighted graph algorithm. This paper investigates algorithm, on the basis of various other algorithms which are designed to increase efficiency of mining. In this project we perform usage analysis which includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through Website. The Weighted graph algorithm is used to find the frequent browsing patterns of the user. The experimental results show that the performance of the weighted graph algorithm is relatively better than GSP in spite of having complex calculations involving weights.

Key Words: Generalized Sequential Pattern, Graph Based Web, Sequential Pattern Mining, Weighted Graph Web Usage Mining

INTRODUCTION

Sequential Pattern Mining finds interesting sequential patterns among the large database. It finds out frequent sub sequences as patterns from a sequence database [1]. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining sequential patterns from their database. Web mining approaches such as web content mining, web structure mining and web usage mining are available which help to extract and produce useful knowledge from the web.

Web Content Mining includes applying data mining techniques to different contents namely unstructured document (text) or semi-structured document (html) or structured document to retrieve the needed and semantic information from the web is called as Web Content Mining (WCM). Web Content Mining focuses on the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders.

Web structure mining is based on the link analysis and topology of the web graph. Web graph is the collection of web pages where each page is represented as a node in the graph and the nodes are interconnected via edges. Here edges represent the hyperlinks that are the interconnection between the web pages. Web graph is the edge weighted, directed graph [2].

Web usage mining consists of three steps: Preprocessing, Pattern discovery, and Pattern analysis. Web usage data can be collected from three sources: Server level, Client level and Proxy level. Web usage mining uses SLF (Server Log File) that is a server level data source. Depending on

Corresponding author: Priyanka Baraskar

the configuration of the web server, the SLF comes in various formats. Basically, it has two formats: a common log format and an extended common log format. The common log format includes the following fields. Remote host field, Date/Time field and HTTP request field. A referrer is the URL of a previous item which led to the web request. For each HTTP request, one record, containing request details, is appended to the end of the SLF [4].

1. PROPOSED SYSTEM:

With the growth and rapid multiplication of web-based systems, the volumes of Web Usage Data collected by web servers have reached huge proportions. Analyzing such data can help website owners to optimize the functionality, content, and structure of their websites. Web usage mining is a method to analyze Web usage data. Given a sequence database where each sequence is a list of transactions ordered by transaction action time and each transaction consists of a set of items. Web usage mining applies data mining methods on Web usage data to discover web usage patterns. Graph mining is a kind of data mining methods.

The graph mining method that is applied on Web usage data is called Graph Based Web Usage Mining (Gweb usage mining).Graph based web usage mining that is applied on a weighted graph is called Weighted Graph Web Usage Mining (WGweb usage mining)[3]. WGweb usage mining perceives the website structure as a vertex weighted graph where each vertex represents a web page, each edge represents the link between web pages, and the vertex's weight represents the numerical value assigned to the web page. The vertex's weight is used to distinguish between vertices. This study proposes a WGweb usage mining method that covers all Web usage data sources and takes page browsing time into account. In addition, this paper proposes a users' browsing behavior analysis approach which is based on applying web usage mining techniques. The proposed users' browsing behavior analysis approach is beneficial for the area of website design improvement.

The click stream simply does not reflect the user's path if only the temporal order is considered. We have developed the WGweb usage mining method, which takes such parallel browsing behavior into account and use it to develop a graph structure approach encompassing all paths a user could have taken. A graph structure contains all paths a user might have taken and uses it to reconstruct session data. It finds all sequential patterns with a user specified minimum support, where the support is the number of data sequences that contain the pattern.

We propose the robust concept of mining weighted approximate sequential patterns. Based on the framework of weight based sequential pattern mining, an approximate factor is defined to relax the requirement for exact equality between weighted supports of sequential patterns and a minimum threshold. After then, we address the concept of mining weighted approximate sequential frequent patterns to find important sequential patterns. We analyze the characteristics of weighted approximate sequential patterns and run extensive performance tests [5].

2. ARCHITECTURE OF THE PROPOSED WGWEB USAGE MINING METHOD:

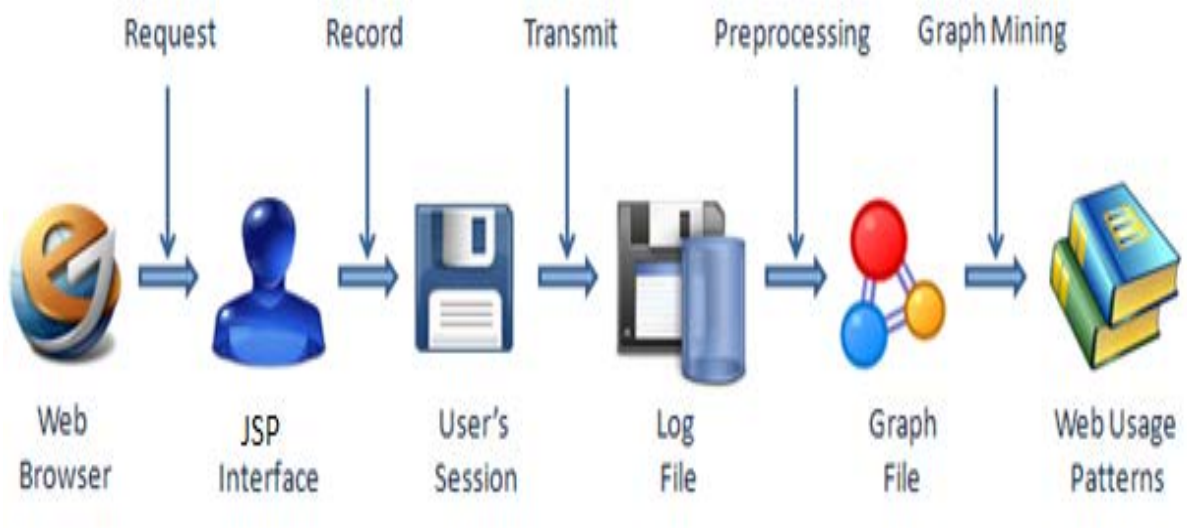

Figure 1: Architecture of proposed WGweb

The above figure presents architecture of the proposed WGweb usage mining method. A JSP interface is placed between the browser and web server to monitor and record the user's web browsing behaviors; it includes requests that are sent to the server and activities that are done on the client side such as switching between tabs or windows.

3. Phases of the Proposed Method:

The proposed WGweb usage mining method consists of three phases:

- Data Collection
- Data Pre-processing
- Pattern Discovery

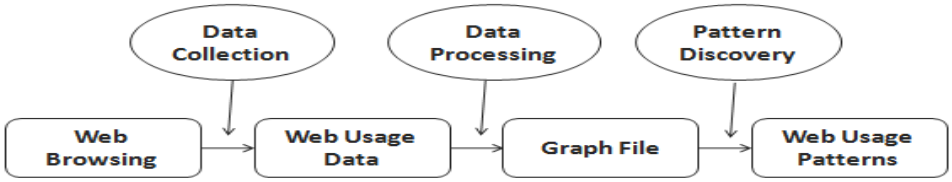

Figure 2: The phases of the proposed WGweb usage mining method.

Data Collection

This phase is related to the monitoring and recording of the user's web browsing behaviors. To do this, we designed a JSP interface which helps us to obtain both client and server side data. It controls five events:

(1) On Session Start
(2) On Session End
(3) On Page Request
(4) On Page Load
(5) On Page Focus

'On Session Start', 'On Session End' and 'On Page Request' are server side events while 'On Page Load' and 'On Page Focus' are client side events.

Data Pre-processing

This phase is related to the converting of the CLF data to the graph structure. In this phase, web usage data is converted to the graph structure in a way which can be used for the graph mining method. In this phase, a base graph is constructed, the users' session is separated, the users' traversal path is discovered a traversal database is constructed and assigned to a corresponding vertex in the base graph.

Pattern Discovery

This phase is related to applying the graph mining method to discover web usage patterns. The Seong and Hyu graph mining method in (Seong and Gyu, 2009) is applied to discover weighted frequent patterns.

4. EXPERIMENTAL DATASET:

Visit Date	IP	Source	Destination	Location	Browser	Version	Code Name	Platform	User
Mon Mar 23 11:34:23 EDT 2015	127.0.0.1	index	index	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; MALNJS; rv:11.0) like Gecko	Mozilla	Win32	User1
Mon Mar 23 11:34:32 EDT 2015	127.0.0.1	index	link	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; MALNJS; rv:11.0) like Gecko	Mozilla	Win32	User3
Mon Mar 23 11:34:33 EDT 2015	192.168.2.3	link	client	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; MALNJS; rv:11.0) like Gecko	Mozilla	Win32	User1
Mon Mar 23 11:34:36 EDT 2015	127.0.0.1	client	service	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; MALNJS; rv:11.0) like Gecko	Mozilla	Win32	User2
Mon Mar 23 11:34:40 EDT 2015	192.168.2.2	service	client	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729; MALNJS; rv:11.0) like Gecko	Mozilla	Win32	User1
Mon Mar 23 11:35:34 EDT 2015	127.0.0.1	index	index	en-US	Netscape	5.0 (Windows NT 6.3; WOW64; Trident/7.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR 3.0.30729;	Mozilla	Win32	user2

Figure 3: Experimental dataset

The above dataset has the attributes such as visit date, IP address, source destination, location, browser, version, code name, platform and user. Data set was initially in MS-Excel format which was later converted into SQL format.

5. ALGORITHM:

Sequence Weighted Graph Algorithm

Step 1: Set the minimum threshold limit.
Step 2: Count the number of visits from source to destination.
Step 3: if the count of number of visits is more than minimum support
Insert the node from source to destination into table or array.
Step 4: Sort the values into descending order.
Step 5: Generate the vertex weighted graph
Assign weight to each vertex.
Step 6: Generate traversal weighted graph
if(i=j)
{
}
else
{traversal weight = vertex weight(source) + vertex weight(destination).
Step 7: Scount.
Count from of each node.
Step 8: Find different users from the log.
Step 9: d=No. of different users.
Sbound = [(vertex weight * (d))]/(vertex weight+ maximum weight).
Wbound = [(vertex weight) +(last 3 maximum weight)].
Step 10: if((Scount[i] >= Sbound[i])
{ weighted graph[i] = 'T'} // to check feasibility
else
{ not feasible }
Step 11: if feasible
Sbound = (vertex weightS * d)/((vertex weight) + maximum weight)
Wbound = (vertex weights + vertex weightD) + (last 3 maximum weights)
Step 12: Sort scount in descending order.

6. EXPERIMENTAL RESULTS:

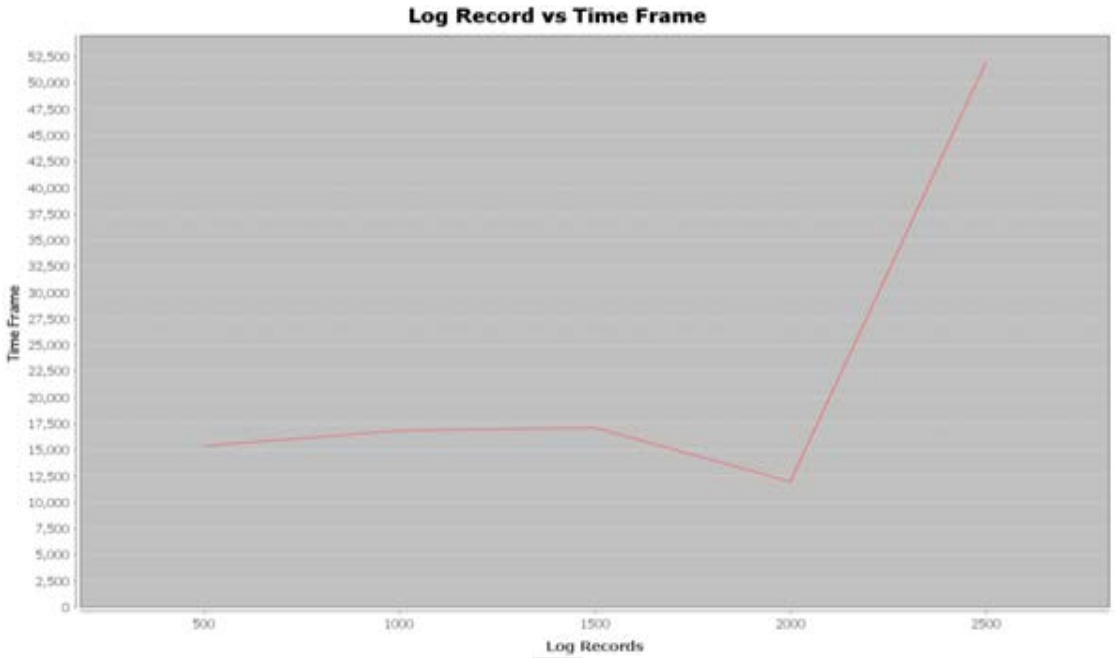

Figure 4: GSP Graph

The above graph is generated on the basis of timeframe and number of log records in the database. The above graph displays the running time required by the GSP algorithm to extract the frequent patterns for specific the log records.

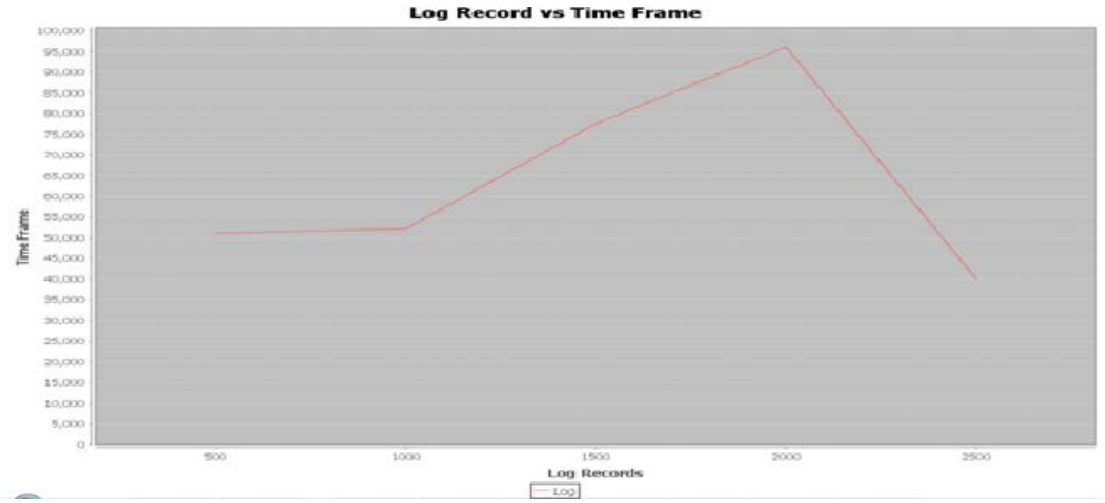

Figure 5: SWG Graph

The above graph is generated on the basis of timeframe and number of log records in the database. The above graph displays the running time required by the sequence weighted graph algorithm to extract the frequent patterns for specific the log records.

7. CONCLUSION:

This paper deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of the project is to introduce the process of extracting frequent sequences from web log data using Weighted Graph, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior.

8. REFERENCES:

1. Chetna Chand, Amit Thakkar, Amit Ganatra "Sequential Pattern Mining: Survey and Current Research Challenges" IJSCE, ISSN: 2231-2307, Volume-2, Issue-1, March 2012
2. V. Uma, M. Kalaivany, G. Aghila "Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm", IJARCCE, Vol. 3, Issue 11, November 2014
3. Seong Dae Lee and Hyu Chan Park "Mining Weighted Frequent Patterns from Path Traversals on Weighted Graph", IJCSNS, VOL.7 No.4, April 2007.
4. Mehdi Heydari, Raed Alsaqour, Khairil Imran,Kamelia Vaziry "A Weighted Graph Web Usage Mining Method to Evaluate Usage of Websites"-ISSN 1991-8178
5. Unil Yun, Keun Ho Ryu"Approximate weighted frequent pattern mining with/without noisy environments", Knowledge-Based Systems archive Volume 24 Issue 1, February, 2011