Techniques for Understanding User Usage Behavior on the Internet

Aparna Ranade
Department of Computer
DJSCOE
Mumbai, India

Abhijit R. Joshi, Ph. D
Head of the Department, Information Technology
DJSCOE
Mumbai, India

ABSTRACT

Web usage mining (WUM) is one of the type of data mining method which is used for analysing web usage patterns with the help of users' session and behavior. It is the technique to classify the web pages and internet users by taking into consideration the contents of the page and behavior of internet user in the past. Web mining extracts data accumulated in server access logs, referrer logs, agent logs, client-side cookies, user profile and Meta data. Usually the WUM techniques study the visitors' browsing behavior to obtain interesting knowledge. There are existing different techniques for web usage mining. Those existing techniques have their own advantages and disadvantages. Here we present a survey on some of the existing web usage mining techniques. First technique that we discuss is based on Sequential Pattern Algorithm. Sequential mining is the process of applying data mining techniques to a sequential database for the purposes of discovering the correlation relationships that exist among anordered list of events. Sequential mining techniques is web usage mining technique, where the sequences of web page accesses made bydifferent web users over a period of time, through a server, are recorded. Ford Lumban Gaol has proposed a web log sequential pattern mining using Apriori-all algorithm. Second technique that we discuss in this paper is based on distance between 2 sequences using no-Euclidean distance formula.Peiqian Liu and Wei Lihave proposed an improved Ward's methodfor web user clustering. They have given a formula, to calculate distance between elements which is a no-Euclidean distance measure. Since it is not a Euclidean distance the output preserves the ordering of events. In the last section of this document we have proposed a new web mining algorithm for analysing usage pattern of users. Proposed algorithm is based on WebLog Sequential Pattern Algorithm[Ford Lumban Gaol] and DBS[Peiqian Liu and Wei Li 2011].)

Index Terms

Web Usage Mining, Web Log sequential, no Euclidean distance formula

1.INTRODUCTION

The World Wide Web consists of billions of web pages.A huge amount of information is available within these web pages. Now a days World Wide Web is the most popular mean of transferring the information.The Web has opened a new wayof doing businesses; amazon.com is one of the examples.The web is huge, diverse and active. The growth of the web has resulted in a huge amount of information. The several kinds of data have to be handled and organized in a manner over a web so that it can be accessed by many users effectively and efficiently. One important issue here is how to deal with an overwhelming amount of information.The search engines like google, yahoo, altavista etc. use algorithms based on keywords. When a user clicks a web link, he needs to check if this information is relevant for him or not. Improving the web site usability, structure and content to keep the visitors interested on it is a challenging task. Many techniques like Web Text Mining (WTM), Web Structure Mining (WSM), Web Usage Mining (WUM), WebPersonalization (WP), etc. are used to help managers and web masters to improvea web site or automatically giving an on-line recommendation directly to the visitors.

So, the usage of datamining methods and knowledge discovery on the web is now on the spotlight of a boosting number of researchers. Web usage mining is a kind of data mining method that can be useful inrecommending the web usage patterns with the help of users' session and behaviour. [Sisodia, D.S.; Verma, S, 2012] Web usagemining includes three process, namely, pre-processing, pattern discovery and pattern analysis.

2.OUTLINE OF THE DOCUMENT

This paper is organized as follows:

Section III reviews two existing algorithms on web mining. In Section IV, we describe proposed algorithm on web mining. A new algorithm discussed in section IV has overcome all drawbacks of both algorithms discussed in section III. Finally, we end this report with conclusion in section V and references in section VI.

3. RELATED WORK

Sequential pattern is a sequence of itemsets that sequentially occurred in a specific order, all items in the same itemsets are supposed to have the same transaction-time value or within a time gap. [Kirti S. Patil, Sandip S. Patil, 2013]The information obtained from sequential pattern mining can be used in marketing, medical records, sales analysis, and so on. In Sequential Pattern Mining every single page access of a website can be recorded automatically in the web logs by the web server. Usually all the transactions of a customer are together viewed as a sequence, usually called customer-sequence, where each transaction is represented as an itemsets in that sequence, all the communications are list in a certain order with regard to the transaction-time. The basic idea of sequential pattern mining is given a set of sequences, called data-sequences, as the input data. Each data-sequence is a list of transactions, where each transaction contains a set of literals, called items. Given a user-specified minimum support threshold, sequential pattern mining finds all of the sequential subsequence's in the sequence database, i.e. the subsequence's whose ratios of appearance exceed the minimum support threshold.

3.1 Weblog Sequential Pattern Algorithm

1) *It identifies patterns in sequences from web log data in a specific period.*

2) *P is a set of literals, called web pages or clicks*

3) *U is a set of users.*

4) *A log is a set of triples (u1,p1,t1),...,(un,pn,tn)} where ui*

∈U, pi ∈P, &ti is time.

Following are the steps for finding sequential usage patterns based on time.

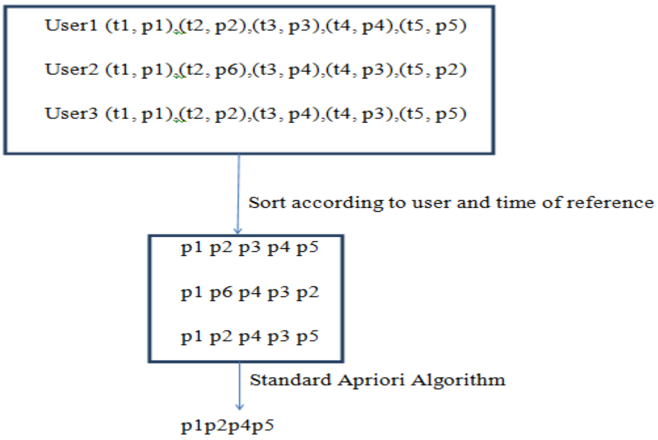
Step 1: *D = Sorting on User ID and time of reference on the first page in each session.*

Step 2: *Find all sequences (s1, s2 ,..., sk) from D such that each one has a minimum Support(s).*

Step 3: *Find L1 in D*

Step 4: *L = AprioriAll (D, s, L1)*

Step 5: *Find a maximal reference sequence of the L, using standard Apriori algorithm.*


User1 (t1, p1),(t2, p2),(t3, p3),(t4, p4),(t5, p5)
User2 (t1, p1),(t2, p6),(t3, p4),(t4, p3),(t5, p2)
User3 (t1, p1),(t2, p2),(t3, p4),(t4, p3),(t5, p5)
Sort according to user and time of reference
p1 p2 p3 p4 p5
p1 p6 p4 p3 p2
p1 p2 p4 p3 p5
Standard Apriori Algorithm
p1p2p4p5

Example of WebLog Sequential Pattern algorithm

The advantage of WebLog Sequential Algorithm is it preserves the time of access of a web page and identifies patterns in sequences from web log in a specific period. However this algorithm does not consider the order in which the web pages are accessed. The order in which web pages are accessed is also an important criterion for checking the usage pattern of a user. In the next section we talk about second web minig algorithm DBS algorithm, which is a modified version of existing Ward method.

3.2 DISTANCE BETWEEN SEQUENCEALGORITHM (DBS)

DBS isproposed by Peiqian Liu and Wei Li [2011]. It is a Web user clustering method. It is a modified existing ward method – cluster analysis method.DBS is non-Euclidean distance reflecting the order of elements. It takes care of duplicate sequences and subsequences in the database. [Peiqian Liu, Wei Li, 2011] has proposed following DBS algorithm.

Euclidean distance between sequence s1 and s2

$$d(s_1, s_2) = \sum_{i=1}^{n} f_i \qquad f_i = \begin{cases} 1 \\ 0 \end{cases}$$

n is the maximum length of s1 and s2.

DBS distance between two sequences S1 and S2 is calculated using

dDBS (S1,S2)= wdD + wiI + ♌R

Where D is the number of deletion operations.

I is the number of insertion operations.

R is the number of reordering operations.

wd is the weight value of deletion operations.

wi is the weight value of insertion operations.

♌ is the weight value of reordering operations.

This algorithm works as follows.

Firstly, theraw Web logs collected on Web server are cleaned to obtain aaccess sequences database. Secondly, the distance betweenevery two sequences is calculated. Thirdly, the DBS method is applied to get clusters.

3.2.1 Example of DBS [Peiqian Liu and Wei Li , 2011]

Suppose: *wd* = *wi* = 1 and *♌*= *wd* + *wi*

S1: {1,4,7,8} , S2: {1,2,3,4,8,7}

To equalize S1 with S2, requires **2 insertion operation**

S1: {1,2,3,4,7,8} S2: {1,2,3,4,8,7}

The equalization process continues with **one reordering**

common element 7 (or 8) in S1 or S2:

S1: {1,2,3,4,8,7} S2: {1,2,3,4,8,7}

Equalizing S1 with S2 took 2 insertion and 1reordering

operation, which gives us:

dDBS (S1,S2)=4

The advantage of this algorithm is, it preserves the order in which web pages are accessed. However this algorithm does not consider time reference of the web pages. Time reference is the important criteria in the analysis of usage pattern of users. Time reference exactly tells us the usage pattern at different times.

A closure look at both algorithms, WebLog Sequential Pattern algorithm and DBS algorithm discussed in section 2.3 and section 2.4 respectively shows that none of the algorithm gives the correct output for ordered sequence of web pages for a given time reference.[Ford Lumban Gaol, 2010]

4. PROPOSED METHODOLOGY

Many interesting patterns are available in the web log data. But it is very complicated to extract the interesting patterns without pre-processing phase. Pre-processing phase helps to clean the records and discover the interesting user patterns which are done using techniques suggested by [P.Nithya and Dr.P.Sumathi 2012]. Figure1 shows basic block diagram of proposed methodology. The processed log data further goes for pattern discovery. Various usage patterns are discovered using Web Log sequential and DBS algorithm. These patterns are then analysed and a real time module is recommended.

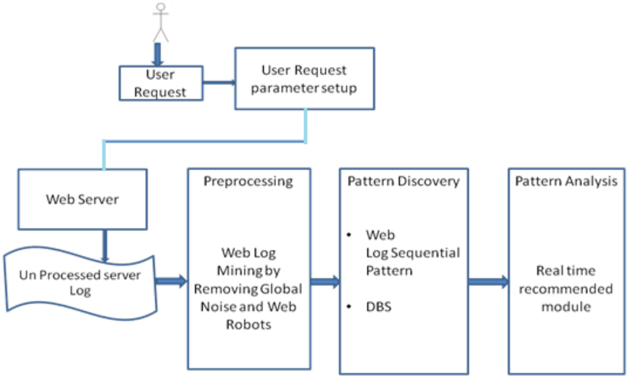

User Request
User Request parameter setup
Web Server
Un Processed server Log
Preprocessing
Web Log Mining by Removing Global Noise and Web Robots
Pattern Discovery
• Web Log Sequential Pattern
• DBS
Pattern Analysis
Real time recommended module

Figure 1: Basic block Diagram of proposed System

4.1 Proposed Algorithm

1. *Let D be the database of all transactions sorted on user ID and time of reference on first page in each session.*

2. *Order the pages as per the time of reference for each user.*

3. *Find DBS between all users in D – ordered in step 2 using formula.*

    dDBS (S1,S2)= wdD + wiI + ♌R

    Where D is the number of deletion operations.

    I is the number of insertion operations.

    R is the number of reordering operations.

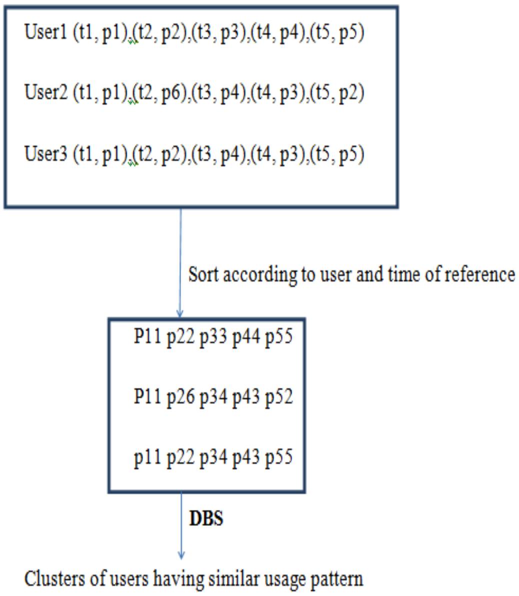
    wd is weight value of deletion operation

    wi is the weight value of insertion operations.

    ♌ is the weight value of reordering operations

4. *Generate distance matrix for DBS distance calculated in step 5. the user visiting sequences is then clustered into several groups.*

Example of proposed algorithm is discussed below.


User1 (t1, p1),(t2, p2),(t3, p3),(t4, p4),(t5, p5)
User2 (t1, p1),(t2, p6),(t3, p4),(t4, p3),(t5, p2)
User3 (t1, p1),(t2, p2),(t3, p4),(t4, p3),(t5, p5)
Sort according to user and time of reference
P11 p22 p33 p44 p55
P11 p26 p34 p43 p52
p11 p22 p34 p43 p55
DBS
Clusters of users having similar usage pattern

Example of Proposed Algorithm

5. CONCLUSION

The increasing popularity of the Web has greatly attracted the Web mining technology. A vital research area in Web mining is Web usage mining which mainly focuses on the discovery of patterns in the browsing and navigation data of Web users. WUM has been a potential technology for understanding behaviour of the user on the Web. There are several techniques proposed by different researchers for the web usage mining. This paper discussed about various techniques available for web usage mining.In this work, we propose a new web miningalgorithm for finding usage pattern of users on the internet. Proposed algorithm not only takes care of ordering of events but also the time period in which the sequence is present. Companies have to implement Web mining systems to understand their customers' profiles, and to identify their own strength and weakness of their E-marketing efforts on the web through continuous improvements [Shailey Minocha, Nicola Millard, Lisa Dawson 2003]. Internet is a gold mine, but only for those companies who realize the importance of Web mining and adopt a Web mining strategy now.

The proposed algorithm gives the clusters of users having similar browsing pattern at some time of the day.The information obtained by the proposed algorithm will be used to design websites according to user habit patterns at particular time of the day that will be able to further increase the number of user visits to the website.

6. ACKNOWLEDGMENTS

We would like to thank Dr. Hari Vasudevan, Principal and Professor and Head of the Dept of Production Engineering, DJSCOE for his support.

7. REFERENCES

[1] Adel T. Rahmani and B. Hoda Helmi, “EIN-WUM an AIS-based Algorithm for Web Usage Mining”, Proceedings of GECCO’08, Atlanta, Georgia, USA, ACM978-1-60558-130-9/08/07, Pp. 291-292, 2008.

[2] Anuradha Yadav, Satbir Jain “Analyses of Web Usage Mining Techniques To Enhance the Capabilities of E-Learning Environment”, 2011 IEEE

[3] Ford Lumban Gaol, “Exploring the Pattern of Habits of Users Using Web Log Squential Pattern” 2010 Second International Conference on Advances in Computing, Control, and Telecommunication Technologies

[4] Jinguang Liu & Roopa Datla, “Web Usage Mining -- Pattern Discovery and its applications”**,** 2013

[5] Kirti S. Patil, Sandip S. Patil “Sequential Pattern Mining Using Apriori Algorithm & Frequent Pattern Tree Algorithm”, IOSR Journal of Engineering (IOSRJEN), Vol. 3, Issue 1 (Jan. 2013)

[6] P.Nithya, Dr.P.Sumathi “Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots” 2012 National Conference on Computing and Communication Systems (NCCCS)

[7] P.Nithya, Dr.P.Sumathi “A Survey on Web Usage Mining: Theory and Applications” IJCTA | July-August 2012

[8] Peiqian Liu, Wei Li “Navigation Pattern Discovery on Web Site Based on the Distance Between Sequences”, 2011 IEEE

[9] Sisodia, D.S. ; Verma, S “Webusage pattern analysis through web logs: A review”, Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference

[10] Varnagar, C.R. ; Madhak, N.N. ; Kodinariya, T.M. ; Rathod, J.N. ‘ “Webusagemining: A review on process, methods and techniques” Information Communication and Embedded Systems (ICICES), 2013 International Conference

[11] Shailey Minocha, Nicola Millard, Lisa Dawson, “Integrating Customer Relationship Management Strategies in (B2C) E-Commerce Environments”, IFIP Conference on Human-Computer Interaction-INTERACT, 2003.