Study And Implementation Of LCS Algorithm For Web Mining

Vrishali P. Sonavane

Department Of Computer Engineering, SSBT's COET, North Maharashtra University
Jalgaon, Maharashtra, INDIA

Abstract

The Internet is the roads and the highways in the information World, the content providers are the road workers, and the visitors are the drivers. As in the real world, there can be traffic jams, wrong signs, blind alleys, and so on. The content providers, as the road workers, need information about their users to make possible Web site adjustments. Web logs store every motion on the provider's Web site. So the providers need only a tool to analyze these logs. This tool is called Web Usage Mining.[8] Web Usage Mining is a part of Web Mining. It is the foundation for a Web site analysis. It employs various knowledge discovery methods to gain Web usage patterns.

Keywords: *Web mining, LCS, Sessionization, Clustering, WUM.*

1. Introduction

In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task[1].
Web is the single largest data source in the world. Due to heterogeneity and lack of structure of web data, mining is a challenging task[3]. And here comes the idea of Web mining. Web mining is a technique to automatically discover and extract information from Web documents/services[2].Web mining is defined as the "discovery and analysis of useful information from the World Wide Web"[6].

Basically Web mining is divided into 3 parts:

1. Web Usage Mining
2. Web Content Mining
3. Web Structure Mining

In this project I have basically focussed on the web usage mining. Discovering hidden information from Web log data is called Web usage mining[6]. The aim of discovering frequent patterns in Web log data is to obtain information about the navigational behavior of the users.

Web content mining describes the discovery of useful information from the web contents. Web structure mining tries to discover the model underlying the link structures of the web. This model is based on the topology of the hyperlinks with or without the description of the links [7]. Web usage mining tries to make sense of the data generated by the web surfer's session or behaviors. While the web content and structure mining utilize the real or primary data on the web, web usage mining mines the secondary data derived from the interactions of the users while interacting with the web. The goal of WUM is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. The Web Usage mining includes the data from the web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls, and any other data as the results of interactions.

2. Soueces Of Data For Web Usage Mining

Basically there are three sources of data for web usage mining as follows[1]

- Web servers
- Proxy servers
- Web clients

3. User Behavior Prediction Process and Its Implementation

User Behavior Prediction Process consist of following phases [7]

3.1 Data Pretreatment
3.2 Navigation Pattern Mining
3.3 Navigation Pattern Modeling
3.4 Clustering
3.5 Prediction Engine

3.1 Data Pretreatment

In Data Preparation phase the web log data must be cleaned, filtered and reformatted to identify all web access session as shown in Fig. 1 and Fig. 2.

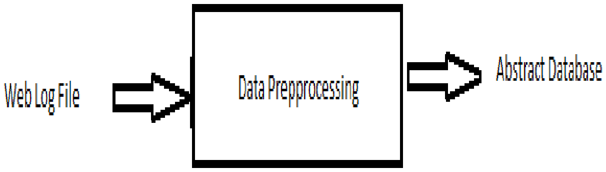
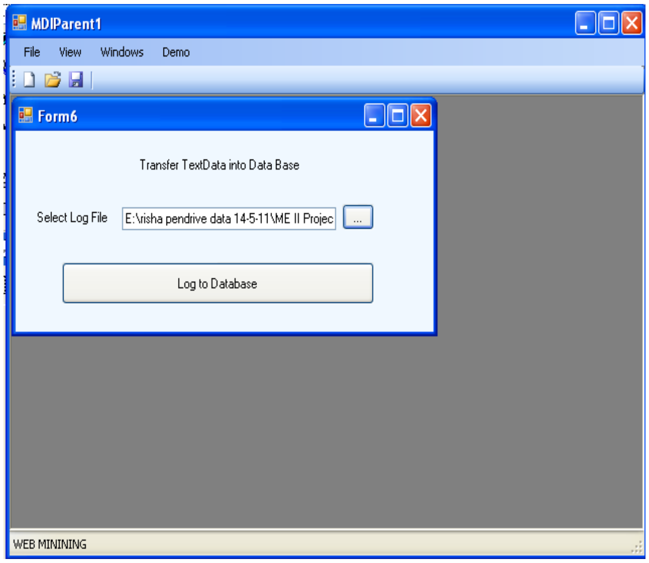


Fig. 1 Model of Data Preprocessing



Transfer TextData into Data Base

Select Log File E:\visha pendrive data 14-5-11\ME II Projec ...

Log to Database

WEB MININING

Fig. 2 Data Cleaning

After data cleaning session identification is done. Here we used threshold of 30 sec foe each session. The session identification is done as follows Fig. 3.

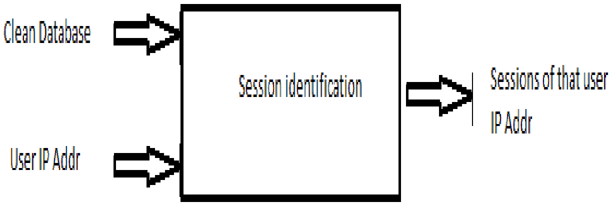
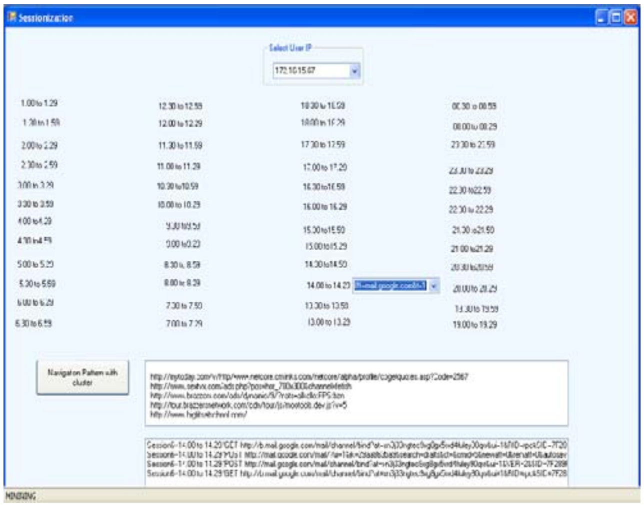


Fig. 3 Model of Session Identification

In the pre-processing stage, the clickstream data is cleaned and partitioned into a set of user transactions representing the activities of each user during different visits to the site. The raw data is converted into the data



172.10.15.67

1.00 to 1.29
1.30 to 1.59
2.00 to 2.29
2.30 to 2.59
3.00 to 3.29
3.30 to 3.59
4.00 to 4.29
4.30 to 4.59
5.00 to 5.29
5.30 to 5.59
6.00 to 6.29
6.30 to 6.59

12.30 to 12.59
12.00 to 12.29
11.30 to 11.59
11.00 to 11.29
10.30 to 10.59
10.00 to 10.29
9.30 to 9.59
9.00 to 9.29
8.30 to 8.59
8.00 to 8.29
7.30 to 7.59
7.00 to 7.29

18.30 to 18.59
18.00 to 18.29
17.30 to 17.59
17.00 to 17.29
16.30 to 16.59
16.00 to 16.29
15.30 to 15.59
15.00 to 15.29
14.30 to 14.59
14.00 to 14.29 [illegible]mail.google.com[illegible]
13.30 to 13.59
13.00 to 13.29

00.30 to 00.59
00.00 to 00.29
23.30 to 23.59
23.00 to 23.29
22.30 to 22.59
22.00 to 22.29
21.30 to 21.59
21.00 to 21.29
20.30 to 20.59
20.00 to 20.29
19.30 to 19.59
19.00 to 19.29

http://mytoday.com/v/http/www.netcore.cmlinks.com/netcore/alpha/profile/cogenquotes.asp?Code=2567
http://www.sextvx.com/ads.php?poswhat_700x300&channelfetish
http://www.brazzers.com/ads/dynamic/9/?rots=allicdlo:FPS.htm
http://tour.brazzersnetwork.com/cdn/tour/js/mootools.dev.js?v=5
http://www.highwebschool.com/

Session6-14.00 to 14.29 GET http://b.mail.google.com/mail/channel/bind?at=[illegible]
Session6-14.00 to 14.29 POST http://mail.google.com/mail/?ui=1&ik=[illegible]
Session6-14.00 to 14.29 POST http://mail.google.com/mail/channel/bind?at=[illegible]
Session6-14.00 to 14.29 GET http://b.mail.google.com/mail/channel/bind?at=[illegible]

MINIING

Fig. 4 Result of Session Identification

abstraction necessary for the further applying the data mining algorithm.

3.2 Navigation Pattern Mining

After the data pretreatment step, we perform navigation pattern mining on the derived user access sessions

3.3 Navigation Pattern Modeling

To model navigational patterns an algorithm is used for modeling the pages accesses information as an undirected graph G= (V, E). The set V of vertices contains the identifiers of the different pages hosted on the Web server.

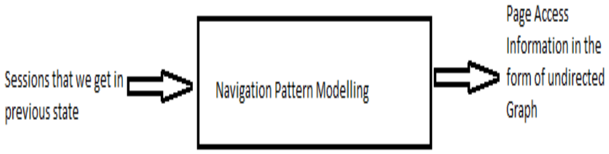


Fig. 5 Navigation Pattern Modeling

3.4 Clustering

In this phase we try to find out groups of strongly correlated pages using clustering.

3.5 Prediction Engine

The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users' future requests.

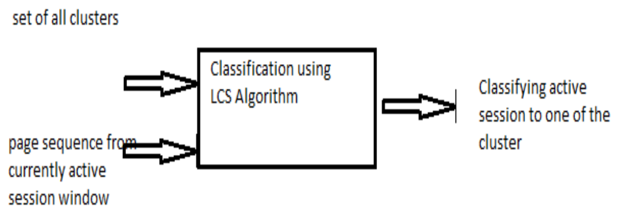
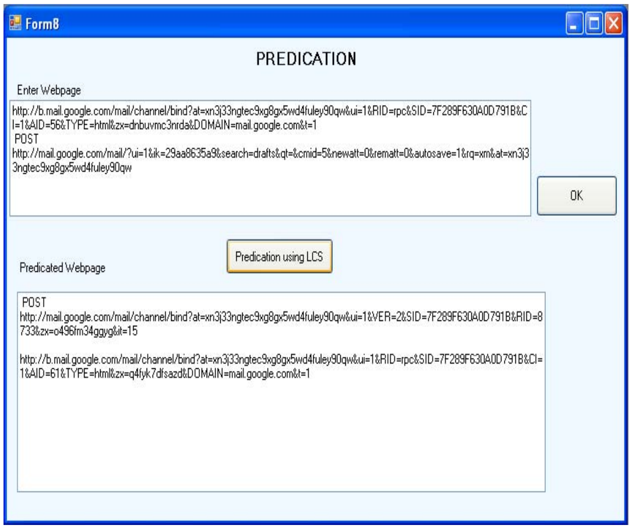

set of all clusters
Classification using LCS Algorithm
Classifying active session to one of the cluster
page sequence from currently active session window

Fig. 6 Model of classification using LCS algorithm

Results of prediction are as shown in Fig. 7.



```
http://b.mail.google.com/mail/channel/bind?at=xn3j33ngtec9xg8gx5wd4fuley90qw&ui=1&RID=rpc&SID=7F289F630A0D791B&C
I=1&AID=56&TYPE=html&zx=dnbuvmc3nrda&DOMAIN=mail.google.com&t=1
 POST
http://mail.google.com/mail/?ui=1&ik=29aa8635a9&search=drafts&qt=&cmid=5&newatt=0&rematt=0&autosave=1&rq=xm&at=xn3j3
3ngtec9xg8gx5wd4fuley90qw
```

```
 POST
http://mail.google.com/mail/channel/bind?at=xn3j33ngtec9xg8gx5wd4fuley90qw&ui=1&VER=2&SID=7F289F630A0D791B&RID=8
733&zx=o496fm34ggyg&it=15

http://b.mail.google.com/mail/channel/bind?at=xn3j33ngtec9xg8gx5wd4fuley90qw&ui=1&RID=rpc&SID=7F289F630A0D791B&CI=
1&AID=61&TYPE=html&zx=q4fyk7dfsazd&DOMAIN=mail.google.com&t=1
```

Fig. 7 Resultes of Prediction.

4. Longest Common Subsequence Algorithm

The problem of comparing two sequences $\vec{\alpha}$ and $\vec{\beta}$ to determine their similarity is one of the fundamental problems in pattern matching. One of the basic forms of the problem is to determine the longest common subsequence (LCS) of $\vec{\alpha}$ and $\vec{\beta}$. The LCS string comparison metric measures the subsequence of maximal length common to both sequences [10].

Formally, given a sequence $\vec{\alpha} = \langle \alpha_1 , \alpha_2 , \ldots, \alpha_n \rangle$ another sequence $\vec{\gamma} = \langle \Upsilon_1 , \Upsilon_{2,\ldots,} \Upsilon_n \rangle$ is a subsequence of $\vec{\alpha}$ if there exists a strictly increasing sequence <j1, j2,…,jn> of indices of $\vec{\alpha}$ such that for all i=1,2,..,l, we have aji $= \Upsilon_i$. Given two sequence $\vec{\alpha}$ and $\vec{\beta}$ we say that $\vec{\gamma}$ is common subsequence of $\vec{\alpha}$ and $\vec{\beta}$ if $\vec{\gamma}$ is a subsequence of both $\vec{\alpha}$ and $\vec{\beta}$. We are interested in finding the maximum-length or longest common subsequence (LCS) given two paths or sequence of page-visits $\vec{\alpha} = \langle \alpha_1 , \alpha_2 , \ldots, \alpha_n \rangle$, $\vec{\beta} = \langle \beta_1, \beta_2, \ldots, \beta_m \rangle$.

The LCS has a well-studied optimal sub-structure property as given by the following:

Theorem

Let $\vec{\alpha} = \langle \alpha_1 , \alpha_2 , \ldots, \alpha_n \rangle$ and $\vec{\beta} = \langle \beta_1, \beta_2, \ldots, \beta_m \rangle$ be sequences, and let $\vec{\gamma} = \langle \Upsilon_1 , \Upsilon_{2,\ldots,} \Upsilon_n \rangle$ be any LCS of $\vec{\alpha}$ and $\vec{\beta}$.

1. If $\alpha_n = \beta_m$, then $\Upsilon_1 = \alpha_n = \beta_m$ and $\vec{\gamma}_{l-1}$ is a LCS of $\vec{\alpha}_{n-1}$ and $\vec{\beta}_{m-1}$.
2. If $\alpha_n \neq \beta_m$, then $\Upsilon_1 \neq \alpha_n$ implies $\vec{\gamma}$ is a LCS of $\vec{\alpha}_{n-1}$ and $\vec{\beta}$.
3. If $\alpha_n \neq \beta_m$, then $\Upsilon_1 \neq \beta_m$ implies $\vec{\gamma}$ is a LCS of $\vec{\alpha}$ and $\vec{\beta}_{m-1}$.

Where $\vec{\alpha}_{n-1} = \langle \alpha_1, \alpha_2, \ldots, \alpha_{n-1} \rangle$, $\vec{\beta}_{m-1} = \langle \beta_1, \beta_2, \ldots, \beta_{m-1} \rangle$ and $\vec{\gamma}_{l-1} = \langle \Upsilon_1 , \Upsilon_{2,\ldots,} \Upsilon_{n-1} \rangle$.

5. Conclusions

In this paper we used LCS algorithm for improving accuracy of recommendation. The Excremental results show that the approach can improve accuracy of classification in the architecture. Using LCS algorithm we can predict user's future request more accurately.

Acknowledgments

It gives me immense pleasure to acknowledge my deep sense of gratitude to my respected guide Mr. K.P.Adhiya, Associate Professor and Head of Computer Engineering Department for his valuable guidance, complete technical assistance, critical interruptions and offering many innovative ideas and helpful suggestions. I am thankful to all teaching and Non-teaching staff of the Computer Engineering Department of SSBTs College Of Engineering & Technology Bambhori, Jalgaon, for their timely help, co-operation and support. I would like to thank North Maharashtra University for their educational and financial support.

References

[1] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed Computer Science & Engineering Shahjalal University of Science & Technology Sylhet, Bangladesh "Pattern Discovery of Web Usage Mining",International Conference on Computer Technology and Development, 2009 .

[2] Li Pingxiang Chen Jiangping and Bian Fuling, "A DEVELOPED ALGORITHM OF APRIORI BASED ON ASSOCIATION ANALYSIS", 2008.

[3] Renáta Iváncsy, and Sándor Juhász, "Analysis of Web User Identification Methods"World Academy of Science, Engineering and Technology 34, 2007.

[4] Renáta Iváncsy, István Vajk Department of Automation and Applied Informatics, and HAS-BUTE Control Research Group Budapest University of Technology and Economics Goldmann Gy. tér 3, H-1111 Budapest, Hungary, "Frequent Pattern Mining in Web Log Data", Acta Polytechnica Hungarica Vol. 3, No. 1, 2006

[5] Paolo Giudici(Università di Pavia)and Claudia Tarantola(Università di Pavia) "Web Mining pattern discovery", Data Mining Laboratory, Department of Economics and Quantitative Methods, University of Pavia, Italy September 12, 2003.

[6] Mathias G´ ery, Hatem Haddad Information Technology Department VTT Technical Research Centre of Finland, Espoo, Finland, "Evaluation of Web Usage Mining Approaches for User's Next Request Prediction", New Orleans, Louisiana, USA, *WIDM'03*, November 7–8, 2003.

[7] Bomshad Mobasher, Namit Jain, Eui-Hong (Sam) Han, Jaideep Srivastava, "Web Mining: Pattern Discovery from World Wide Web Transactions", Technical Report 96- 050, University of Minnesota, Sep, 1996.

Vrishali P. Sonavane received a Batchler of Engineering degree from the North Maharashtra University in 2005. She is currently a student of Master Of Engineering degree course in the Department of Computer Engineering at the SSBT's COET, Jalgaon, North Maharashtra University, Jalgaon.