An Approach for Restructuring of Web Pages

[*]Chennupati.R.Prasanna, [*]M.Venkata Kishore , [*]P. Srinivasa Rao, [*]L.Mohana sandeep, [*] Dr.D.Rajya Lakshmi
*DEPARTMENT OF INFORMATION TECHNOLOGY, GITAM UNIVERSITY,ANDHRA PRADESH, INDIA

SUMMARY
Today the internet has become one of the biggest information provider where web sites are playing a wide role, web mining has gained more important .Web mining can be classified into three main areas: web usage mining, Web content Mining and Web Structure mining .Web usage mining is a kind of web mining, which exploits data mining techniques to discover valuable information from navigation behavior of World Wide Web users. There are generally three tasks in Web Usage Mining: Preprocessing, Pattern analysis and Knowledge discovery .Preprocessing cleans log file of server by removing log entries such as error or failure and repeated request for the same URL from the same host etc. The main task of Pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. The information collected from the log file can help to discover the knowledge. This knowledge collected can be used to take decision on various factors like Class1, Class 2, users and Eminent, Average and Delicate web pages based on hit counts of the web page in the website. The topology of the website is reconstructed based on hit counts which provide quick response to the web users. This paper addresses challenges in three phases of Web Usage mining along with Web Structure Mining.

Key words:
Web mining, web site, hit count, log file, HTTP, URL, topology.

1. INTRODUCTION

Web mining technology provides techniques to extract knowledge from web data. Researchers on web mining have already distinguished three main areas, namely web content mining, web usage mining and web structure mining[1].Web structure mining deals with the discovery of structures from the web topology. Recent publication by Miller and Remington [2] pointed out that the structure of linked pages has a decisive impact on the usability. Previous studies including Shneiderman[3], and Larson and Czerwinski [4] also provided suggestions on how to create the best structure .Larson and Czerwinski [4] found that users took significantly longer time to find items in a structure with depth than breadth. Web usage mining (WUM)[5] is a new research area which can be defined as a process of applying data mining techniques to discover interesting patterns from web usage data. Web usage mining provides information for better understanding of server needs and web domain design requirements of web-based applications.Web usage data contains information about the identity or origin of web users with their browsing behaviors in a web domain.Web pre-fetching[6,7],link prediction[8,9,10] site reorganization [11,12]and web personalization[13,14,15,16] are common applications of WUM .We will concentrate on Web Usage Mining and Web Structure Mining in the following The web site is a collection of web pages .A web page is a page with HTML (Hyper Text Markup Language) tags. The web pages in the given web site can be arranged in different fashions i.e. either breadth-wise, depth –wise or combination of both which tells about structure or topology of the website .Web site will be kept under observation for some period of time. Every time when web user requests for a particular transaction, web server will record requested transaction entry in its log files. Log file located in web server includes access log, referrer log and agent log. Access log which is also named as CLF (common Log format) has entry format as follows. HostID rfc931 authuser[date] "method URL protocol" status bytes.

After every fort night log file is accessed and various statistics can be collected such as how many users have visited web pages, time spent on each page, number of bytes downloaded etc

Web usage mining extracts user's navigation patterns by applying data mining techniques to server logs, together with employing some topology of the web site, Web structure. Web Usage Mining deals with three main steps: Preprocessing, Knowledge discovery and pattern analysis

The remaining of the paper is organized as follows:

Section 2 contain preprocessing; section 3 contain pattern analysis section 4 contain knowledge discovery section 5 contains web structuring mining .section 6 describes the conclusion

2. PREPROCESSING

Real-world data tend to be dirty, incomplete and inconsistent. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, since quality decisions must be based on quality data .Data used in preprocessing cover server log files, web page[17] content web site structure and hit counts of pages in the website. Fig 1 shows restructuring of web pages.

Data cleaning removes entries unhelpful to data analyzing and mining based on various data cleaning techniques. It has to remove log entries that have status code as "failure" or "error". Secondly some automatic search engines generate some access records, those have to be identified and removed from the log file .Some of other common indicators such as (a) the repeated request for the same URL from the same host;(b) a time interval between requests too short to apprehend the contents of a page; The dynamic behavior can be used to construct more complex navigation behaviors in a single session. These four basic behaviors constructing complex navigations are given below:

1. A web user can start session with anyone of the possible entry pages of a web site. This behavior includes new page which is

Manuscript received February 5, 2010
Manuscript revised February 20, 2010

not requested by any other previous page accessed from the same domain in near-time.
2. A web user can select the next page having a link from the most recently accessed page.
3. A web user can press the back button one more time and thus selects as the next page a page having a link from any one of the previously browsed pages.
4. A web user can terminate the session.

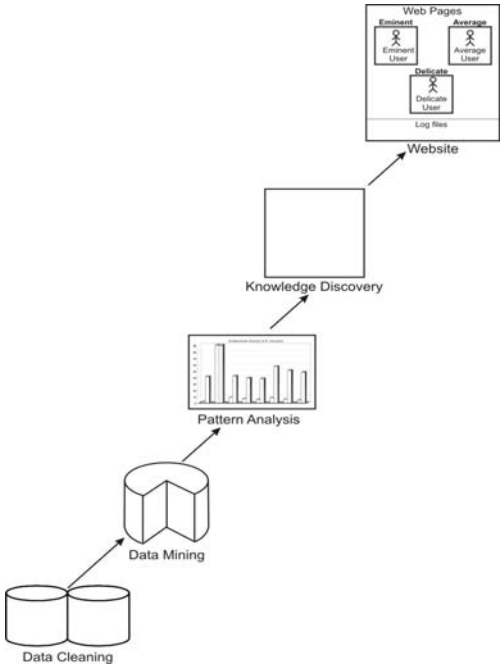

Fig.1 Restructuring of Web pages.

3. PATTERN ANALYSIS

Pattern analysis is to filter uninteresting information and to visualize and interpret the interesting pattern to users. Visualization assist an analyst to better apprehend navigation patterns and to predicate trends of data knowledge about content and structure also contribute to filtering un-useful knowledge. Many web tools provide some objective criteria supporting and confidence .such criteria are helpful to manually filter some believed unimportant knowledge.
webViz tool has done some pioneering work in visualizing of access patterns. It display access pattern of user as directed graph, with nodes representing page of the access pattern and links representing the hyperlinks between pages

Web pages in the website can be classified into
Eminent- the web pages with highest hit counts
Average- the web pages with average hit counts
Delicate – the web pages with least hit count
Web users who visit the web sites can be classified as
Class1 user (Eminent), class2 user (Average), class3 user (Delicate)

4. KNOWLEDGE DISCOVERY

The information collected from the web site can help in discovering the knowledge. This knowledge obtained can be used to take decision on various factors like

1. The web pages with highest hit counts will be the popular pages.
2. what is possible the navigation patterns of users.
3. The time spent on each page which tells about importance of the web page.
4. If time spent on particular web page is negligible it indicates that the web page does not contain important information.
5. The web pages for which no user's request is there indicates that page must be modified.
6. If log file entry says repeatedly for particular web page "redirect", it should be notified to web site designer/owner/maintainer.

5. WEB STRUCTURING MINING

After the pattern analysis is done on web pages, the important decision can be done regarding structure of the website .The eminent web page will be moved very near to the home page, at next level average web pages will be moved and soon. The pages with more hit count can be given the preference to be brought closer to the home page provided web site owner/designer agrees. The heap tree can be generated based on the hit counts available in the log file during particular session. The heap tree generated will help us make decision about the topology of web site during next interval so that the web pages which are more popular can be brought very near to the home web page .with this restructuring, the web users can gain quick access to the web pages along with best utilization of bandwidth and server memory space since every HTTP request will be entered into log file of server

6. CONCLUSIONS

We described importance of web usage mining and its relationship with web structure mining. All the three phase of Web Usage Mining provide good log file which is free from inconsistent, un-useful data .it helps in filtering unwanted access web pages/patterns. The web structure mining plays an important role with web users, reducing lot of HTTP
Transactions between users and servers thus saving memory space of server, better utilization of bandwidth along with server processor time

References

[1] B.Liu Web Data Mining: Exploring Hyperlinks, contents, and Usage Data .Springer, 2006
[2] Miller, C.S. and Remington, R.W.,"Modeling Information Navigation: Implications for Information Architecture", Human –Computer Interaction, Vol.19, No.3, 2004
[3] Shneiderman, B.,"Designing the User Interface", Strategies for Effective Human-Computer Interaction 3rd ed reading MA: Addison-Wesley, 1998

[4] Larson,K.,&Czerwinski,M., "Web page design:Implications of memory ,structure and scent for information retrieval",CHI'98:Human Factors in Computing Systems ,New York:ACM press 1998,pp.25-32
[5] J.Srivastava, R.Cooley, M.Deshpande, and P.N .Tan. Web Usage mining: Discovery and applications of usage patterns from web data .SIGKDD
[6] P.P.J.E.Pitkow.Mining longest repeating subsequences to predict World Wide Web surfing .In USENIX, 1999
[7] S.E.Schecter, M.Krishnan, and M.D.Smith .Using path profiles to predict http requests, Computer Networks, 30(1-7):457-467, 1998
[8] S.G "und" uz and M.T. "Ozsu".A web page prediction model based on click-stream tree representation of user behavior. In KDD, pages535-540, 2003
[9] E.Frias-Martinez and V.Karamcheti .A customizable behavior model for temporal prediction of web user sequences. In WEBKDD, pages 66-85, 2002.Explorations, 1(2):12-23, 2000
[10] Y.M.A.Nanopoulos, D.Katsaros.Effective prediction of web-user accesses: A data mining approach. In WEBKDD, 2001
[11] M.Spiliopoulou.Web usage mining for web site evaluation .Commun.ACM,43(8):127-134,2000
[12] R.Srikant and Y.Yang .Mining web logs to improve website organization. In WWW, pages 430-437,2001
[13] B.Mobasher ,R.Cooley ,and J.Srivastava.Automatic personalization based on web usage mining.Commun.ACM,43(8):142-151,2000
[14] B.Mobasher, H.Dai ,T.Luo and M.Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. Data Min.Knowl.Discov. ,6(1):61-82, 2002
[15] O.Nasraoui and R.Krishnapuram .An evolutionary approach to mining robust multi –resolution web profiles and context sensitive url associations. International Journal of Computational Intelligence and Applications, 2(3):339-348, 2002
[16] D.Pierrakos, G.Paliouras, C.Papatheodorou, and C.D.Spyropoulos. Web usage mining as a tool for personalization :A survey .User Model.User Adapt.Interact., 13(4):311-372,2003
[17] Jiawei Han and Micheline kamber. Data mining concepts and technique, second edition,pp.50-51

CHENNUPATI R PRASANNA pursing M.Tech in information technology from GITAM University.

M.VENKATA KISHORE pursing M.Tech in information technology from GITAM University.

P.SRINIVAS RAO pursing M.Tech in information technology from GITAM University.

L.MOHANA SANDEEP pursing M.Tech in information technology from GITAM University.

Dr.D. RAJYA LAKSHMI. M.Tech Ph.D, HOD Dept of information technology, GITAM University, over 17 years of teaching experience, published 12 papers in various national and international conference and journals. She has been awarded the doctorate degree from JNT University Hyderabad, India