ISSN:0975-9646

K Sudheer Reddy et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012,3664-3667

An Effective Methodology for Pattern Discovery in Web Usage Mining

K Sudheer Reddy[#1], Dr. G. Partha Saradhi Varma[*2], M. Kantha Reddy[#3]
[#] *Dept. of CSE, Acharya Nagarjuna University, GUNTUR, AP, INDIA*
[*] *Dept. of IT, SRKR Engineering College, Bhimavaram, WG Dist, AP, INDIA*
[#]*Director Operations, Indo US Collaboration for Engineering Education – Hyderabad, AP, INDIA,*

***Abstract*— One of the important steps in the data mining is Sequential pattern mining and it comes after the pre-processing phase of Web Usage Mining. Pattern discovery deals with the sorted set of data items presented as part of a sequence. Using this sequential pattern mining, users can easily recognize the web paths that users commonly follow on a web site. The aim of this research work discovers the patterns which are most relevant and interesting by using a Web usage mining process. The web log files serves as an input to this process. Our target is to discover user's behaviours, who have visited the web sites for a lesser number of times. We have employed a method for clustering, which is based on pattern summaries. We have conducted vivid experiments and the results are shown in this paper.**

***Keywords*— Web usage mining, pre-processing, usage patterns, pattern discovery, sequential patterns, clustering, patterns summary.**

I. INTRODUCTION

Analyzing the web user's behavior is also known as Web Usage Mining (WUM). WUM is an active research area which entails in adapting the mining methods to the records of web access log files. These web log files collect numerous types of data include, host IP address, the URL requested, the date and the other information about the user navigation of the web. The techniques of Web Usage Mining provide interesting knowledge about the web user's behavior in order to excerpt relationships in the recorded data. Amongst the techniques available, the sequential patterns are predominantly well adapted to the web log study. Sequential patterns extraction on a web log file, is hypothetical to provide the thoughtful relationship:

On SRKREC Web Site, 23% of users visited the homepage consecutively, the available resources page, the RSC offers, the RSC missions and finally the past RSC competitive selection". Exhibiting this type of behaviour is an assumption, because sequential patterns extraction on a web log file also infers managing several problems, as listed below:

- The number of records in the web access log file, are lowered due to the user's computer cache and the proxies.
- On this site, the diversity is more.
- The web log file entries can be reduced and also reduce the user navigations is possible with the aid of research engines. As a result the user can directly access a definite portion of the web site.
- The number of portions visited on the site compared to the entire site.
- The representativeness of web users who navigate the web through that part, compared to the whole site users.

If the problems of web caching can be solved [5], the representativeness requires a strong study. To exemplify our goal, let's consider sequential patterns we are supposed to get. Due to the minor size of the "job offer" part of the web site, users requesting a web page on that part represent only 0.3% of users on the whole site. In the similar way, users navigating on the "research" part of the research assignment represent only 0.003% of all the users. So, the study of Web Usage Mining on this type of Web site has to manage this specific representativeness in order to provide sufficient results. Our objective is to show that a traditional sequential pattern mining is unable to provide web users behaviors with such a weak support.

Web Usage Mining (WUM) entails in extracting most interesting information from files which catalogue Web log files. Most of the research methods in this domain take into an account the entire period during which usage traces were recorded, the results found certainly being those which conquer over the total period. Therefore, certain types of user behaviors, which take place through short sub periods are not detected and accordingly remain unknown by out dated methods. It is important to study such behaviours and consequently carry out an analysis related to time sub-periods. It will then be possible to study the temporal evolution of users' profiles by providing descriptions that can integrate the temporal aspect. Furthermore, as the volume of mined data is great, it is important to define summaries to represent user profiles.

Further, we present a method for discovering behaviour of all web users of a Web site. We label our test and experiments and then conclude the paper. The following diagram shows our principle of pattern discovery which has the log division functionality.

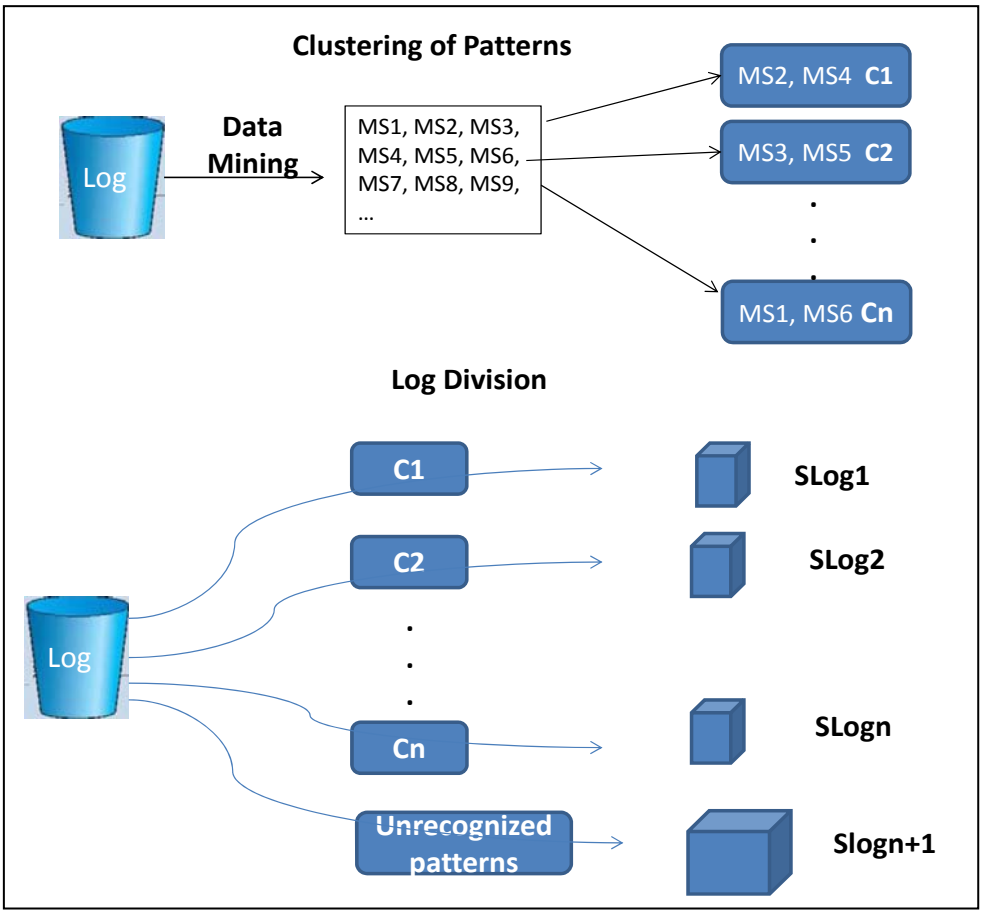

Fig. 1 The Principle of Discovery

II. PRINCIPLE

We propose a methodology and described the outline as under: discovering the clusters of web users (users are typically grouped by a behavior) and then analyzing the user navigations by means of the sequential pattern mining process. Our methodology therefore relies on two steps. The first step targets at dividing the web log into sub-logs, theoretical to represent separated actions. The second step targets at analyzing the user behavior recorded in each sub-log.

The key principle involved in our method is described as given below:

1. Extracting the sequential patterns on the original log.
2. These sequential patterns can be clustered.
3. Dividing the web log according to the clusters obtained. Each sub-log encompasses user sessions from the original log, approving at least one of the behavior of the cluster which permitted to create this sub-log. A distinct sub-log is created then to collect the user sessions from the original sub-log which doesn't correspond to a cluster from the earlier step.
4. Apply the whole process recursively, for each sub-log.

The above Figure 1 graphically illustrates the proposed method. Initially, sequential patterns are to be obtained and the same to be clustered, in the figure it is shown from C1 to Cn. Then, the web log is split into various sub-logs, from SLog1 to SLogn upon these clusters. Finally a sub-log SLogn+1 is created for the user sessions which cannot be harmonized with a behavior from the original web log. The quality of the produced results by our methodology will depend on this sub-log file. In detail, the first sub-logs comprise the most represented categories of the web users. Hence, they are interesting, but the most interesting patterns discovery will derive from the research study of the uncluster sessions of the sub-log SLogn+1. Seeing this sub-log as a new original web log, and recursively repeating the process will allow users to discover behavior with the minimal representativeness. To acquire reliable results, our method suitable on a "the quality of the split proposed for a log". This split relies on the clustering done on the discovered patterns in the original log. In the next section, we described briefly about the method that we used to cluster patterns.

III. CLUSTERING BASED ON PATTERN GENERALISATION

We have deliberate several methods of clustering for sequential patterns discovery. We propose and describe here the utmost efficient method for sequential pattern clustering that we used. The clustering method used in this research study is grounded on a method developed by [8] in 2000 for indexing the web sequences in the perspective of Web-based recommender systems. The efficiency of is based on the neural approach for such method and its effectiveness relies on usage of summarized descriptions for sequential patterns:

these descriptions are based on generalization of Web access sequences.

A. Neural Method

The proposed neural clustering method based on [8] a framework for supporting the reuse of past experiences using the integrated object oriented structure. This methodology was successfully applied on browsing behaviors of thematic repertory, for large organizations. This method is based on a hybrid model and composed from the connexionist part [5] and pure flat memory compound of patterns' groups.

A threshold TSi is related to each prototype, which will be altered during the learning step. For such a threshold governs an influence region in an input space. If a pattern introduced in the network drops in the influence region of a prototype, then this prototype will be activated. Such region is determined by set of input vectors adequate a distance measure lower than the threshold. In case, if there is an inactivated prototype, a new prototype is created.

Hence, the structure of a prototype-based network is an evolutionary in the sagacity that the numerous prototypes at the hidden level is not the priori fixed and might be enlarged during learning step. A prototype is characterized by using its reference vector, a set of representing patterns and an influence region.

IV. Experiments

The methods of extraction are written in an object oriented programming language, C++ on a Intel Pentium (3.2 Ghz) PC running a latest version of Linux system, with 2 GB Random Access Memeory. The algorithm we employed is the PSP algorithm [12] for extracting sequential patterns. The neural method and GUI are comprehended in Java. For the SRKREC's site, the data was composed over a period of 45 days, while for the other intranet sites, over a period of 70 days. The narrative of the characteristics (refer Table I) is: the number of lines in the web log is indicated by N, number of user sessions is S, the number of filtered URLs is U, the average session length is donated as L, the average number of session URLs is SU. Through our experiments, we could be able to bring into respite frequent behaviors, with a comparative representativeness getting feebler and weaker, depending on the sub-log's depth.

TABLE I
Log file characteristics

	www.srkrec.ac.in	**www.srkrec.ac.in/intranet**
N	12 57 24	17 167 81
S	287 493	437 648
U	46 218	61 398
L	3.5	2.9
SU	4.6	3.2

C1: The user behavior given here is purely related to the higher education prospects offered by the SRKREC. The users visit and read the higher education page, and then the web page describing the competitive selection and lastly the web pages describing the education opportunities.

<(trv/higheredu/educon.html)
(trv/higheredu/educon/oppot.html)
(highedu/inplo/index.html)
(highedu/inplo/listings/index.html) >

(support: 0.28%).

Fig. 2 User behaviors for C1

C2: This user behavior is a search for a security fleabag in the system. Generally, these web attacks are programmed once and further shared and used by different bodies.

<(lscripts/root.exe) (c/winnt/system32/cmd.exe)
(..%255c../..%255c../winnt/system32/cmd.exe)
(..%255c../..%255c/..%c1%1c../..%c1%1c../..%c1%1c../winnt/system32/cmd.exe) (winnt/system32/cmd.exe)
(winnt/system32/cmd.exe) (winnt/system32/cmd.exe)>
(support: 0.04%)

Fig. 3 User behaviors for C2

The discovered user behaviors by employing our method cover more than 75 surfing goals on SRKREC main web site and more 130 goals on the intranet site of SRKREC. We stated the three goals here, from job opportunities requests to activities of hacking. Thus, these discovered behaviors demonstrate the success of our methodology in discovering the behaviors.

V. Conclusions

In this research paper, we offered a sophisticated method for extracting of the all web user's behavior of a Web site. Our methodology has the distinguishing feature to divide the log file recursively in order to discover the behaviors and to characterize them as clusters (similar behaviors are grouped into a cluster). For this perseverance, we had to offer a detailed clustering method, which is devoted to sequential patterns. The key advantage of our proposed method is to study the Web Usage Mining with minimal support as a composite problem that can be solved by succeeding divisions. The problem therefore, shifts from one single open problem to n number of problems we can solve and the one problem that has to be recursively divided. By furthering in this approach, we could establish that the boundary between the data quantity and the quality of results can sometimes be pushed vertebral by extracting behaviors with a minimal representativeness.

Though there are several powerful knowledge discovery methods or techniques have been proposed for Web Usage Mining, but, very little work has been devoted to handling problems related to data that can evolve over a period of time.

The future research work could involve applying various clustering based methods and implementing various clustering techniques that enable the automatic pattern discovery of the number of clusters as well as identifying fusions and splits over a time period.

References

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2] J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.
[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
[8] FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[12] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.