ISSN:0975-9646
J.Isabella et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (6) , 2011, 2725-2727

Improving User Experience by Mining Usage Patterns

[1]J.Isabella
Research Scholar, Sathyabama University,
Chennai, INDIA

[2]R.M.Suresh
R.M.D.Engineering College,
Chennai, INDIA

***Abstract*— With rapid development of the Internet, many websites lack interactivity for both naïve and advanced users alike. With dynamic websites becoming a reality, creating websites according to the customers requirement improves the user experience. User experience typically concentrates on the user in human computer interaction (HCI) which not only involves the content but also on the context and quality of the web site. Content, hypertext / linking structure and user profile play a crucial role to improve the quality of experience for a website. The challenges faced by website administrators managing a large number of hits are the difficulty in obtaining feedbacks from the users on their experiences with the website. In this paper we propose to extract pages requested by each user along with session id of an user from web logs, perform usage mining and identify the optimal structuring of the website.**

Keywords: Usage mining, Web log, User experience, Human computer interface, Association rule discovery.

I. INTRODUCTION

The large volume of Internet usage and web browsing in recent years has resulted in generation of large amount of log files by web servers that is potentially valuable for understanding the behavior of visitors. This knowledge can be applied in various ways, such as enhancing the. effectiveness of websites through user personalization, improving user experience or developing directed web marketing campaigns[1].

The ease and speed with which business transactions can be carried out over the Web have been a key driving force in the rapid growth of electronic commerce. Business-to-business e-commerce is the focus of much attention today, mainly due to its huge volume. While there are certainly gains to be made in this arena, most of it is the implementation of much more efficient supply management, payments, etc. On the other hand, e-commerce activity that involves the end user is undergoing a significant revolution. The ability to track users' browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before. It is now possible for a vendor to personalize his product message for individual customers at a massive scale, a phenomenon that is being referred to as mass customization. Though the scenario outlined here is from e-commerce, the type of personalization described is applicable to any Web browsing activity. Web personalization can be described, as any action that makes the Web experience of a user personalized to the user's taste. The experience can be something as casual as browsing the Web or as significant as trading stocks, purchasing a camera or using an online software service. The actions can range from simply making the presentation more pleasing to an individual to anticipating the needs of the user and providing the right information as well as performing a set of routine book-keeping functions automatically[2].

User experience more concentrates on the user in human computer interaction (HCI). Early writings on usability already expressed that the primary of usability is the person's experience at the moment experienced. Disorientation, or the tendency to lose one's sense of location in a Web site, can cause users to become frustrated, lose interest, and experience a measurable decline in efficiency. Nevertheless, user experience in the sense of a positive HCI would, thus, focus on how to create outstanding quality experiences rather than merely preventing usability problems[3]. The concepts of user experience and quality of experience were originally promoted by human-computer interaction researchers to emphasize concern with the *outcomes* of people's experience with — or through — technology. QoE should concern user performance based on actual usage. The main point so far has been that objective measures of QoE can and should be collected from user tests and that these measures enable us to extend beyond user perception to user experience. However, understanding user opinion remains important, and a combination of objective and subjective variables should better reflect the complexity of QoE [4].

II. PREVIOUS WORK

Dimitrijević et al., proposed a system for the discovery of association rules in web log usage data as an object orient application. Pruning was used to eliminate directly linked pages out of the rule set. In the proposed system the lift outperformed confidence after the minimum confidence threshold was taken into account [5].

Zaki et al.[6] evaluate sampling for association rule mining. It was observed that for a given itemset, sample sizes as required by chernoff bounds is independent of the size of the data for analysis. It is shown that for number of tuples that are less than 400000, reasonable accuracy is achieved through chernoff bounds based sample size selection. The sampling technique reduces the size of data needed for computation.

Cadez et al. propose a new methodology for visualizing navigation in web sites. In the proposed method the users are partitioned into clusters with each cluster containing similar navigation patterns. Clusters are created using first order Markov models with the Expectation Maximization algorithm[7]. The advantage of the proposed method is the linear scaling for both users and clusters.

III. METHODOLOGY

In this paper it is proposed to investigate an association rule based approach to usage-based web personalization to enhance user experience using data mining techniques extensively. Association rule mining over the *basket data model* was introduced by Agrawal *et al* [8]. It allows designers to infer useful information on web click patterns, browsing criterion for any large website. The web log essentially consists of a large number of individual records called *transactions* and each transaction is a list of pages visited in that particular transaction. Consider for example, the log of all the transactions that take place in a webserver. The goal of association rule mining is to discover rules of the type, "whenever a transaction includes a particular set *W* of links, it is likely to contain a specific link I ∉ W". In case of a website, such rules can be used to arrange the links on the website to increase the quality of experience. Informally, the input to association rule mining consists of the collection of transactions and two parameters $\theta \leq 1$, the required *support* and $\gamma \leq 1$, the desired *confidence*. It consists of two steps, namely *frequent itemset mining* in which itemsets with frequency of at least θ are identified, and *association rule mining* in which the association rules of the type $W \Rightarrow I$, with I ∉ W, are identified. The itemset *W U I* should have a support of at least θ , and of all the transactions containing *W*, the fraction of the transactions that contain *I* should be at least γ . Agrawal and Srikant [9] present the *Apriori* algorithm for frequent itemset mining and *FastGenRules* heuristic to generate the association rules. Importance of usage mining has been investigated in [10,11,12]. Various proposals have been made on the importance of user experience and usage mining[13,14,15].

The popular apriori algorithm can be represented as follows :

Let C represent the candidate itemset

Let L represent the frequent itemset

Pass 1

1. Generate the candidate itemsets in C_1
2. Save the frequent itemsets in L_1

Pass *k*

1. Generate the candidate itemsets in C_k from the frequent itemsets in L_{k-1}
   a. Join L_{k-1} *p* with L_{k-1}q, as follows: **insert into** C_k **select** *p*.item$_1$, *p*.item$_2$, . . . , *p*.item$_{k-1}$, *q*.item$_{k-1}$ **from** L_{k-1} *p*, L_{k-1}q **where** *p*.item$_1$ = *q*.item$_1$, . . . *p*.item$_{k-2}$ = *q*.item$_{k-2}$, *p*.item$_{k-1}$ < *q*.item$_{k-1}$
   b. Generate all (*k*-1)-subsets from the candidate itemsets in C_k
   c. Prune all candidate itemsets from C_k where some (*k*-1)-subset of the candidate itemset is not in the frequent itemset L_{k-1}
2. Scan the transaction database to determine the support for each candidate itemset in C_k
3. Save the frequent itemsets in L_k

To implement the proposed technique DePaul CTI Web server dataset was used. The original data contains random sample of users visiting this site for a 2 week period during April of 2002. The original data contained a total of 20950 sessions from 5446 users. The filtered data files were produced by filtering low support page views, and eliminating sessions of size 1. A subset of this data set containing 1119 sessions with 15 page links. Each page view is assigned a numerical value as shown partially in table I.

Table I. Pages assigned unique numbers

0	/admissions/
1	/admissions/career.asp
2	/admissions/checklist.asp
3	/admissions/costs.asp
4	/admissions/default.asp
5	/admissions/general.asp
6	/admissions/helloworld/arabic.asp

Each session is represented in row format with each attribute representing a column with value "t" that the page has been viewed in the particular session or "f" otherwise. The data representation format is shown in table II.

Table II : Pages visited in a session

/admission/	*/admissions/ career.asp*	*/admissions/ checklist.asp*	-
t	*F*	*t*	-
t	*T*	*t*	-
t	*F*	*t*	-
t	*F*	*t*	-
t	*T*	*t*	-

IV. EXPERIMENTAL RESULTS

Preparation of data : The CTI web server dataset file contains the full (unfiltered) preprocessed sessionized data. Each session in begins with a line of the form: SESSION #n (USER_ID = k) where n in the session number, and k is the \ user id. Within a given session, each line corresponds to one \ pageview access. Each line in a session is a tab delimited sequence of 3 fields: time stamp, pageview accessed, and the referrer. The time stamp represents the number of seconds relative to January 1, 2002. In order to prepare the web log data for the mining process, the web log file needed to be cleared of irrelevant requests, each relevant request needed to be assigned to a visit session, and the resulting file had to be transformed to a format that could be fed into the mining algorithm. Since a pageview / admissions has been extensively used by most of the users, this work considered all the sessions containing this pageview and its subsequent links.

For the association rule discovery the support was measured at 0.75. Over 493 rules was discovered. Some of the results discovered are tabulated below.

/admin/general.asp /admin/international.asp /admin/statuscheck.asp ==> /admin/default.asp confidence:(0.98)

/admin/general.asp /admin/mailrequest.asp /admin/statuscheck.asp ==> /admin/default.asp confidence:(0.98)

/admin/general.asp /admin/i20visa.asp /admin/statuscheck.asp ==> /admin/default.asp confidence:(0.98)

/admin/general.asp /admin/mailrequest.asp ==> /admin/international.asp confidence:(0.94)

/admin/i20visa.asp /admin/statuscheck.asp ==> /admin/mailrequest.asp confidence:(0.94)

/admin/default.asp /admin/mailrequest.asp ==> /admin/international.asp confidence:(0.94)

/admin/mailrequest.asp ==> /admin/costs.asp /admin/default.asp /admin/i20visa.asp confidence:(0.81)

/admin/default.asp ==> /admin/general.asp /admin/i20visa.asp /admin/statuscheck.asp confidence:(0.81)

/admin/default.asp ==> /admin/general.asp /admin/international.asp /admin/mailrequest.asp confidence:(0.81)

Figure 1 Shows the distribution of rules with respect to the confidence level. From the graph it can be seen that strong associations are formed between the various links. However it has to be noted that most well designed websites automatically follow certain well known design principles where in certain rules are established in placing the links and has a high probability of the web user to follow.

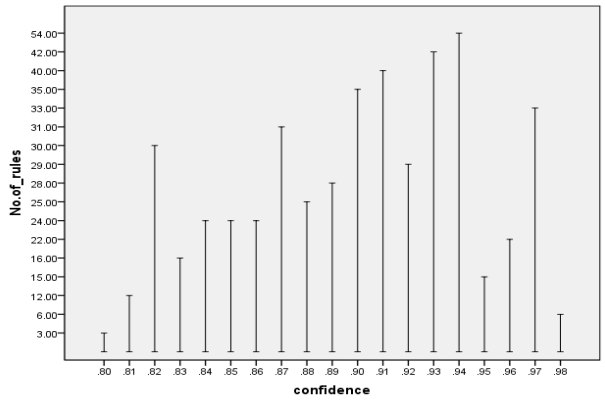

Figure 1. Plot of rule discovered versus the confidence level.

Focusing on the bottom 10% of the confidence level the rules that were discovered are

/admin/international.asp=f 1035 ==> /admin/general.asp=f /admin/i20visa.asp=f /admin/statuscheck.asp=f 843 conf:(0.81)

/admin/i20visa.asp=f 1050 ==> /admin/costs.asp=f /admin/general.asp=f 855 conf:(0.81)

/admin/i20visa.asp=f 1050 ==> /admin/general.asp=f /admin/international.asp=f /admin/mailrequest.asp=f 854 conf:(0.81)

/admin/international.asp=f 1035 ==> /admin/costs.asp=f /admin/statuscheck.asp=f 841 conf:(0.81)

/admin/mailrequest.asp=f 1035 ==> /admin/costs.asp=f /admin/default.asp=f /admin/i20visa.asp=f 839 conf:(0.81)

/admin/default.asp=f 1061 ==> /admin/general.asp=f /admin/i20visa.asp=f /admin/statuscheck.asp=f 860 conf:(0.81)

/admin/default.asp=f 1061 ==> /admin/general.asp=f /admin/international.asp=f /admin/mailrequest.asp=f 859 conf:(0.81)

/admin/i20visa.asp=f 1050 ==> /admin/costs.asp=f /admin/statuscheck.asp=f 848 conf:(0.81)

/admin/default.asp=f 1061 ==> /admin/costs.asp=f /admin/i20visa.asp=f /admin/international.asp=f 856 conf:(0.81)

The associations above can be used to provide insight to the web administrator which can lead to improvement in the user experience of the web user.

V. CONCLUSION

In this paper a web log was preprocessed and useful rules were mined using association rule discovery, which ultimately can be used to improve the quality of browsing and hence the user experience. In this work it is assumed that rules discovered with confidence level of 0.82 and above would already been implemented in the website and concentrate on rules discovered below the cut off level. Further work needs to be done to find useful rules discovered from huge amount of discovered rules. The proposed work can merge with clustering based techniques already proposed in literature.

REFERENCES

[1] Maja Dimitrijević, Zita Bošnjak, "Discovering Interesting Association Rules in the Web Log Usage Data", Interdisciplinary Journal of Information, Knowledge, and Management . Volume 5, 2010 pg 191-207

[2] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Automatic Personalization Based on Web Usage Mining" . Dept. of Computer Science, DePaul University. 1999. maya.cs.depaul.edu/mobasher/papers/MCS00.pdf

[3] Paurobally, S. Turner, P.J. Jennings, N.R. Automating negotiation for m-services.Systems. IEEE TransactionsMan and Cybernetics, Part A: Systems and Humans .2003.Volume : 33 , Issue:6, On page(s): 709.

[4] Jie Yang,Shiwu Zhang,Jiming Liu. Characterizing Web Usage Regularities with Information Foraging Agents. IEEE Transactions on Knowledge and Data Engineering , 2004.566-584

[5] Maja Dimitrijević, Zita Bošnjak. Web Usage Association Rule Mining System. Interdisciplinary Journal of Information, Knowledge, and Management Volume 6, 2011. pg 137-150]

[6] M. Zaki, S. Parthasarathy, W. Li, and M. Ogihara. Evaluation of sampling for data mining of association rules. In RIDE, 1997.

[7] Igor Cadez, David Heckerman, Christopher Meek, Padhraic Smyth, Steven White. Visualization of Navigation Patterns on a Web Site Using ModelBased Clustering.Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining

[8] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*.

[9] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487-499, 1994.pg.2000

[10] Esin Saka,Olfa Nasraoui,Richard Germain,Antonio Badia,Maha Soliman. A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites.IEEE Transactions on Knowledge and Data Engineering.2008.202-215

[11] Dimitrios Pierrakos,Georgios Paliouras.Personalizing Web Directories with the Aid of Web Usage Data.IEEE Transactions on Knowledge and Data Engineering. 2010 pp. 1331-1344.

[12] Shin-Yi Wu,Yen-Liang Chen.Mining Nonambiguous Temporal Patterns for Interval-Based Events.IEEE Transactions on Knowledge and Data Engineering. 2007. pp. 742-758.

[13] Yang, C. C.; Ng, T. D.Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.2011. Issue:99 On page(s): 1

[14] Lancieri, L.; Durand, N. Internet user behavior: compared study of the access traces and application to the discovery of communities. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans. 2006, Volume 36,Issue:1 On page(s): 208

[15] Awad, M.A. Khan, L.R. Web Navigation Prediction Using Multiple Evidence Combination and Domain Knowledge.IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans.2007. Volume:37 ,Issue:6, On page(s): 1054.