Web Page Recommendation Model for Web Personalization

Abdul Manan Ahmad[1] and Mohd. Hanafi Ahmad Hijazi[2]

[1] Department of Software Engineering, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
manan@fsksm.utm.my
[2] School of Engineering and Information Technology, Universiti Malaysia Sabah, Locked Bag No. 2073, 88999 Kota Kinabalu, Sabah, Malaysia
hanafi@ums.edu.my

Abstract. Web usage mining has gained more popularity among researchers in discovering the users browsing behavior mining the web server log that records all the users' transactions activities. In this paper, we developed a usage model for predictions based on association rule. Similarity between items contained in the active user profile will be calculated upon the matched rules and finally the top-*N* most similar items are then recommended to the user. We used the time spent on each page for weighting the pages instead of binary. Two evaluation metrics were applied to evaluate the accuracy of the recommendations, namely precision and coverage.

1 Introduction

Web personalization can be described as any action that can customized the content or structure of a Web site to the user's taste or preferences has widely been utilized by e-commerce organizations to better serve their customers. The actions can be made by highlighting the hyperlinks, inserting new hyperlinks that seem to be of interest for the current user dynamically, and the creation of new index pages. The common personalization systems for the Web can be categorized into three groups, which is manual decision rule system, content-based filtering agents and collaborative filtering systems [4]. Of all, the collaborative filtering system has become the predominant approach in furnishing the e-commerce system with an intelligence to capture user profiles and recommending relevant pages to the users. However, it suffers some limitations, which are mostly related to the scalability and efficiency [7]. Item-based similarity [8] and dimension reduction was proposed by some researchers to overcome the drawback.

Web usage mining [1], the most recent method in personalizing Web page has gained more popularity among researchers in discovering the users browsing behavior. The common data mining technique used is varied from clustering, association rule mining [3] and sequential pattern. Association rule mining has been successfully applied in the pages recommendation system [3, 4, 9].

In this paper, we developed a usage model for predictions based on association rule. The methodology is like this: first, the association rules of each URL will be

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 587–593, 2004.

extracted from the Web log data and similarity between items will be calculated upon the rules instead of user sessions. Secondly, the recommendation engine will then search the *top-N* most similar items to the active user click-stream before generating recommendation for the user. Furthermore, we used the time spent on each page for weighting the pages instead of binary. The rational is time spent on pages is a good implicit interest indicators of a user on that pages [10].

2 Web Usage Mining for Web Personalization

Web personalization constructed from several steps includes the collection of Web data, the modeling and categorization of these data, the analysis of the collected data and the determination of the actions that should be performed. The most popular method in analyzing the collected data is collaborative filtering (CF) [4]. CF uses rating to predict which information is of interest to the users. The goal is to find the most similar groups of users for the active user and recommending pages that have been rated by users in the group and not yet viewed by the active user, which is called user-based systems. However, user-based personalization suffers limitation in term of scalability and flexibility. This is because the calculation of similarities between active user and the usage profile has to be done online, which the time required for searching may become prohibitive [9]. Item-based systems have been proposed [8] to overcome the scalability problem as it performs the calculation of item similarities in an offline basis.

Web usage mining has been defined as the process of applying data mining techniques to the discovery of usage patterns from web usage data [1], is the latest technologies emerge for personalization. Users left behind his footprints on browsing the web and this information collected in server access logs. There will be very large volumes of data that form an important and priceless knowledge buried inside. Lots of researches have contributed in evaluating the effectiveness of Web usage mining for Web personalization [3, 4].

3 Association Rule Based Page Recommendation Model

In this paper, the system recommends relevant pages to users based on the probability of the pages to be clicked by user. The recommendation will be depending absolutely on the web server logs. But first, we have to clean the log in order to allow data mining perform on the validated access log. Then, we have to identify the user transactions. Finally, the association rule mining will complete the pattern extraction task.

The access log recorded every request made by user. Various kind of files available, which is some of them are not appropriate for the purpose of mining navigations patterns. The objectives of this phase were to identify only the HTML documents. Therefore, we have deleted unnecessary files such as graphic and sound files. Beside that, users might sometimes request the page that does not exist. This results in error entries being recorded. Since we are requiring existing web URLs only, the error entries are therefore deleted.

The goal of transaction identification is to create meaningful clusters or references of each user [6]. It show pages browsed by a single user in a sequential manner, which is important in generating the user browsing profiles. We define a transaction as a set of web pages that was requested by a user in a single session. Two steps involve in identifying the transactions. First, we identify the user sessions, which are greatly complicated by the existence of local caches, firewalls and proxy server. The goal of user identification is to ease the obligation of generating user profiles during the mining process. Several heuristics are commonly used to help identify unique users such as using the access log in conjunction with the referrer log and site topology to construct browsing path for each user. Accessed pages which are not linked together will be assumed by the heuristics that there exists another user with the same IP. However, since we are only depending on the access log alone and the FSKSM's site is mostly accessed by students in the computer laboratories without passing any proxy servers, we simply identified the users by referring to the IP address.

Once user has been identified, the next task is to identify the user sessions. The purpose is to extract single user transactions. The user session shows the pages accessed by a user sequentially. Each transaction commonly distinguished using the 30 minutes time out. Users in a single IP address that required more than 30 minutes before accessing other pages will be assumed to be different user using the same IP address.

After the pre-processing, a set of m page, $P = \{p_1, p_2, .., p_m\}$ and n user transactions, $T = \{t_1, t_2, ..t_n\}$ generated. Each $t_i \in T$ is a subset of P. We define each transaction as $t_i = \langle (p_1, w_1), (p_2, w_2), .., (p_3, w_3) \rangle$ where w_i represent the weight associated with the page p_i in transaction t_i. In this paper, we are using time spent weights on each page instead of binary which is typically used in other researches [3, 4, 9]. The time spent by a user for viewing a page is a very important piece of information in measuring the user's interest on the page [10]. Here, the rating of a page in a transaction is presented by the duration of a user viewing that page. However, the size of a file and the network traffic may affect the actual viewing time. Therefore, we take the time spent for each byte of the file by dividing the time spent to the size of the file as a weight and we assume that the network traffic is same for all transactions.

3.1 Association Rule Mining of Web Usage Log

Association rule mining [5] is commonly utilized in finding relationship between stored data. Association rule is an expression A => B, where A and B are itemsets. The objective of this rule is to calculate the probability of a transaction contain item B, given that it contains A, which is known as the *confidence* of the association rule. *Confidence* is given as *support(A∪B)/support(A)*, whereas *support* is the sum of stated item in the database.

Some researches have considered the use of association rule mining in recommendation systems [7, 9] and Web usage mining [3]. The combination of similarity measurements and association rule has been discovered in [9]. Instead of matching the active user with the rules, [9] measures the similarity of active user with the rules to produce recommendations. However, none of this research took the time spent on

each page as an implicit rating for that page. But, they did use the time spent for pre-processing phase.

In this paper, besides the *confidence* and *support*, we also calculate the sum of time spent for each page during rules generation. The result, R of association rule mining conceptually described as $r_j = \langle (p_1, p_2, .. p_n), (w_1, w_2, .. w_n), \sigma, \alpha \rangle \in R$ where σ represent the support and α represent the confidence of the rule. Any pages in a transaction that has a *support* < minimum support will be deducted. In this paper, we are utilizing the most well known Apriori algorithm [5] to extract rules from the access log.

3.2 Similarity Measurement

The rules produced after the mining process is representing the behavior of user's navigation on the Web site. The recommendation engine then ranked the top-*N* most similar items for each item in the log from the rules. The formulation described as:

$$rank(i, j) = similarity(i, j) \times confidence(i, j) \tag{1}$$

We calculate the probability of both items *i* and *j* occurred together instead of the confidence of the rules as we are interested in focusing on the similarity between items. The rational behind this formulation is that some users may accidentally rated item *i* and *j* high because of external disturbance such as a phone call, making a coffee and the network latency. Recommendation engine might make a wrong assumption and thus confidence is appropriate to neutralize the error. The similarity computation of item *i* and *j* is given by:

$$similarity(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \cdot \vec{j}}{\|\vec{i}\|_2 * \|\vec{j}\|_2} \tag{2}$$

In order to compute the top-*N* recommended items for active user, *U* we first identify the set of candidate recommended items, *C* by taking the union of *m* most similar items for each item $c \in C$ which is not yet available in *C*. Then, we compute the similarity of item $c \in C$ to the set of *U* as the sum of similarities given by the user to all the items $i \in U$ multiply by the confidence of *i* existed in rule *u*. The formulation as been proposed by [8] described as follows:

$$recommend(u, i) = \frac{\sum_{all\ similar\ items}^{N} (s_{i,N} * R_{u,N})}{\sum_{all\ similar\ items}^{N} |s_{i,N}|} \times confidence(u, i) \tag{3}$$

4 Experiment and Evaluation

We are using the access logs of the Faculty of Computer Science and Information System, Universiti Teknologi Malaysia dated July, 1 2003 until December, 7 2003. After removing all the unnecessary files, the data was then pruned by eliminating both

files that appear less in 0.5% or greater than 80% of transactions, to remove noise and avoid the large items from dominating the rules construction. The pruning left 4, 630 transactions and 140 unique URLs for the mining process with data sparseness of 0.9522. Randomly, 70% of the data selected for training while another 30% for testing.

Each transaction, t in testing data set is divided into two parts. The first n pages in t are taken as an active user session, ua_t while the next t-n pages will be used to evaluate the recommendations, denoted as E_t. Set of candidate recommendation that match the user active session and satisfying the recommendation threshold, τ will be generated by the recommendation engine. This candidate recommendation set is denoted as $cand(ua_t, \tau)$. However, only the top-N rank items will be recommended to the user, which is here denoted as $R(cand(ua_t, \tau), N)$. Our method of evaluation is almost similar as [4], with additional of selecting the top-N items based on different criteria as been described above.

Two different standard measures applied on each method for evaluation, namely *precision* and *coverage*. *Precision* measures the degree of the recommendation engine produce accurate recommendations. Formally, *precision* can be defined as:

$$precision(R(cand(ua_t, \tau), N)) = \frac{\left|R(cand(ua_t, \tau), N) \cap E_t\right|}{\left|R(cand(ua_t, \tau), N)\right|} \tag{4}$$

Coverage measures the ability of the recommendation engine to produce all the pages that might be visited by the user. Thus, *coverage* is defined as:

$$coverage(R(cand(ua_t, \tau), N)) = \frac{\left|R(cand(ua_t, \tau), N) \cap E_t\right|}{\left|E_t\right|} \tag{5}$$

Finally, the mean over all transactions in the evaluation set for a given recommendation threshold was computed as an overall score for each *precision* and *coverage*.

4.2 Results

In all experiments, we used recommendation thresholds varying from 0.1 to 1.0 to measure the precision and coverage. The window size is set to 2 as the mean transaction length of the data is 3. As been expected, the high minimum support produces fewer recommendations candidate, which affects the coverage while lower minimum support cause lots of irrelevant recommendations generated so as with the recommendation threshold. Thus, the right selection of both values is critical in providing the recommendation engine with appropriate numbers of rules for better recommendation.

Fig. 1 show the superior of using time spent for viewing 1 byte of file size compared to full file size as weighting value. Further observation on this matter revealed that the bigger the size of the file requested, the longer it takes to download that file. Unfortunately, as server can only recorded the time once a page requested, the time spent for each page will be incorrectly gathered. Therefore, by calculating the time spent for each byte of the file size, we can get the better precise time spent for each page.

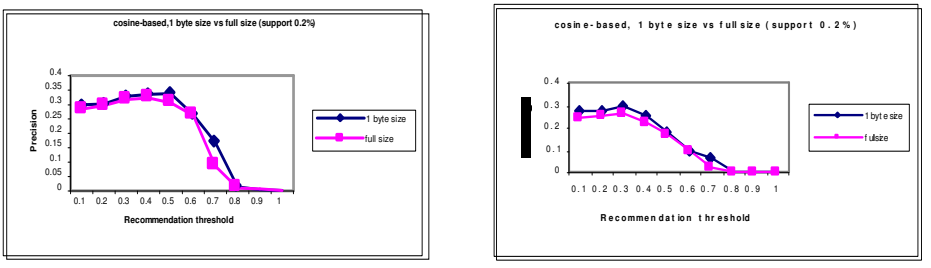

Fig. 1. Impact of using time spent for viewing 1 byte of file size and full file size as weighting value

Fig. 2 shows the impact of using similarity together with confidence of the rules (ARsim) in generating recommendations. Even though traditional associations rule have defeat ARsim in precision, but ARsim is found to be more superior in coverage. Overall, we found that ARsim is better by calculating the effectiveness of each technique using F1 metrics (not shown here). By combining similarity between items and confidence of the rules, the recommendation engine has selected only the most relevant items by taking into account the relationship of an item towards the active user click-stream and also the similarity between items in term of its usage and concepts. Therefore, it increases the effectiveness of the recommendation engine.

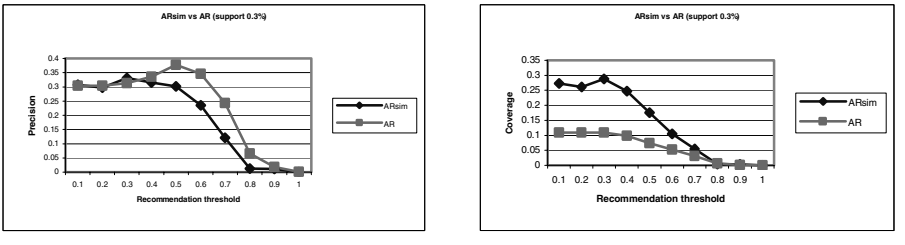

Fig. 2. ARsim versus traditional association rule

5 Conclusion

Web usage mining has sparked the new method for recommendation system. It has the advantages to discover usage patterns from the web server logs, to develop the user navigations model and finally applied the model for recommendation. In this paper, we concentrated on exploiting the advantages buried in the web server log as we have observed that there are still certain interesting parts that gain less popularity but considerably have the potential to improve the recommendation quality. We used the time spent on each page to determine the importance of the pages. Some alteration on the recorded time spent has been made and we have successfully showed the effectiveness of our method in improving the accuracy of the recommendation. We even compared

the traditional association rule recommender system with our model and found that the latter is better and more effective for Web personalization.

References

1. Srivastava, J., Cooley, R., Deshpande, M., and Tan, P. N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations (2000) 1(2):12-23.
2. Cooley, R., Mobasher, B. and Srivastava, J.: Web Mining: Information and Pattern Discovery on the World Wide Web. In Proceedings of the International Conference on Tools with Artificial Intelligence, Newport Beach (1997) 558-567.
3. Mobasher, B. Dai, H., Luo, T., and Nakagawa, M.: Effective Personalization Based on Association Rule Discovery from Web Usage Data. In Proceedings of the 3rd ACM Workshop on Web Information and Data Management, Atlanta, Georgia (2001) 9-15.
4. Mobasher, B., Cooley, R., and Srivastava, J.: Automatic Personalization based on Web Usage Mining. Communications of the ACM (2000) 43(8):142-151.
5. Agrawal, R. and Srikant, R.: Fast Algorithm for Mining Association Rules. In Proceedings of 20th International Conference on Very Large Data Bases (1994) 487-499.
6. Cooley, R., Mobasher, B. and Srivastava, J.: Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems (1999) 1:1-27.
7. Sarwar, B., Karypis, G., Konstan, J., and Riedl, J.: Analysis of Recommender Algorithms for E-Commerce. In Proceedings of the 2nd ACM E-Commerce Conference, Minneapolis (2000).
8. Sarwar, B., Karypis, G., Konstan, J. and Riedl, J.: Item-Based Collaborative Filtering Recommendation Algorithms. In Proceedings of the 10th International World Wide Web Conference. Hong Kong (2001).
9. Demiriz, A. 2002. Enhancing Product Recommender Systems on Sparse Binary Data. Accepted to be published in the Journal of Data Mining and Knowledge Discovery (2003).
10. Shahabi, C., Zarkesh, A. M., Adibi, J., and Shah, V.: Knowledge Discovery from Users Web Page Navigation. In Proceedings of 7th International Conference on Research Issues in Data Engineering (1997) 20-29.