IMPROVED WEB PAGE RECOMMENDER SYSTEM BASED ON WEB USAGE MINING

Yahya AlMurtadha, Md. Nasir Sulaiman, Norwati Mustapha and Nur Izura Udzir

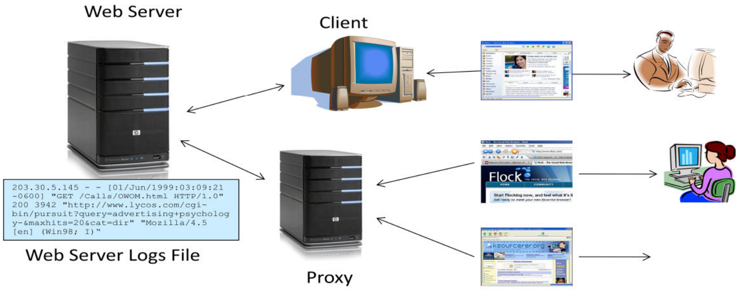
Universiti Putra Malaysia, Malaysia, y.murtadha@gmail.com, {nasir,norwati,izura}@fsktm.upm.edu.my

ABSTRACT. Web now becomes the backbone of the information. Today the major concerns are not the availability of information but rather obtaining the right information. Mining the web aims at discovering the hidden and useful knowledge from web hyperlinks, contents or usage logs. This paper focuses on improving the prediction of the next visited web pages and recommends them to the current anonymous user by assigning him to the best navigation profiles obtained by previous navigations of similar interested users. To represent the anonymous user's navigation history, we used a window sliding method with size *n* over his current navigation session. Using CTI dataset the experimental results show higher prediction accuracy for the next visited pages for anonymous users compared to previous recommendation system.

Keywords: web mining, web page recommender

INTRODUCTION

Web mining aims to discover useful information or knowledge from Web hyperlinks, page contents, and usage logs (Liu, 2007). Yet an important problem is how to mine complex data formats including Image, Multimedia, and Web data (Yang & Wu, 2006). Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three main types: Web structure mining, Web content mining and Web usage mining (Liu, 2007). Web structure mining discovers knowledge from hyperlinks, which represent the structure of the Web. Web content mining extracts useful information/knowledge from Web page contents. Web usage mining (WUM) mines user access patterns from usage logs, which record clicks made by every user. Fig.1 shows the web usage mining recording process of the users' browsing activities either from direct client-server browsing or proxy-server browsing in terms of IP address, date, method, file required, protocol, browser types,… et which stored at the web server logs files. The output of the WUM is some patterns that may be the input to the Recommendation systems Engine which is one of the application areas of the Web usage gives the ability to predict the next visited page for a given user.



```
203.30.5.145 - - [01/Jun/1999:03:09:21
-0600] "GET /Calls/OWOM.html HTTP/1.0"
200 3942 "http://www.lycos.com/cgi-
bin/pursuit?query=advertising+psycholog
y-&maxhits=20&cat=dir" "Mozilla/4.5
[en] (Win98; I)"
```

Figure 1. Web Usage Mining Recording Process of the users' browsing Activities

Related works

Recently, many researches tried to improve the prediction accuracy of the recommendation systems. (Mobasher, Cooley, & Srivastava, 2000; Nakagawa & Mobasher, 2003) , presented WebPersonalizer a system which provides dynamic recommendations, as a list of hypertext links, to users. (Mobasher, Dai, Luo, & Nakagawa, 2002) presented and experimentally evaluate two techniques, based on clustering of user transactions and clustering of pageviews, in order to discover overlapping aggregate profiles that can be effectively used by recommender systems for real-time Web personalization. (Zhou, Hui, & Chang, 2004) proposed an intelligent web recommender system known as SWARS (Sequential Web Access based Recommender System) that uses sequential access pattern mining. (Liu & Kešelj, 2007) proposed a novel approach to classifying user navigation patterns and predicting users' future requests based on the combined mining of Web server logs and the contents of the retrieved web pages. (Baraglia & Silvestri, 2004) (Baraglia & Silvestri, 2007) proposed a WUM system called SUGGEST, that provide useful information to make easier the web user navigation and to optimize the web server performance. (Cornelis, Lu, Guo, & Zhang, 2007) developed a conceptual framework for recommending one-and-only items. It uses fuzzy logic, which allows to reflect the graded/uncertain information in the domain, and to extend the CF paradigm, overcoming limitations of existing techniques. A possible application in the context of trade exhibition recommendation for e-government is discussed to illustrate the proposed conceptual framework. (Jalali, Mustapha, Sulaiman, & Mamat, 2010) Proposed WebPUM, a recommender system based on Common Sequences algorithm (LCS). (AlMurtadha, Sulaiman, mustapha, & Udzir, 2010) proposed a method for Learning and mining the web navigation profiles to provide an appropriate model to recommend to the anonymous user. (Castellano, Fanelli, & Torsello, 2011) presented NEWER as a usage-based Web recommendation system that exploits the potential of Computational Intelligence techniques to dynamically suggest interesting pages to users according to their preferences. (Almurtadha, Sulaiman, Mustapha, & Udzir, 2011) Presented iPACT an improved recommendation system using Profile Aggregation based on Clustering of Transactions

The Methodology

As shown in Fig.2, the proposed architecture consists of two main components, namely the offline and online. In the offline component tow three important processes are taken. First, preprocess the web server logs by allying data cleaning techniques and then partition the web navigations into sessions determined by the period of browsing. Second, partition the filtered sessionized page views into clusters of use's navigation patterns with similar pageviews browsing activities using K-mean algorithm. Finally, generate web navigation profiles based on the preformed clusters. The online component does the matching of the new anonymous user request (current active session) to the profile shares common interests to the user. The details will be discussed in the following sections.

This study used the clusters produced by the clustering step (previous step) to build the usage or the navigation profile with one profile for each cluster. The navigation profile contains only those web pages that passed certain confidence support and weights values. The confidence support determines the frequency occurrence on those pages in the cluster. These profiles don't consider specific users since this study don't take the users history in account during obtaining the profiles. To summarize we construct a navigation profile as a set of pageview-weight pairs:

$$profile = \{ p, weight(p) \quad | p \in P, weight(p) \geq min\_weight \}.$$

where $P =\{p1, p2, . . . , pn\}$, a set of n pageviews appearing in the transaction file with each pageview uniquely represented by its associated URL and the weight(p) is the (mean) value of the attribute's weights in the cluster. Fig.3 shows a navigation profile Database snapshot of two profiles obtained for two clusters 1 and 2 where each profile contains related page views. For example, profile 1 represents the activity of a user interested in the courses and the programs offered while profile 2 represents the activity of a user interested in the pageviews related to the admission and advising.

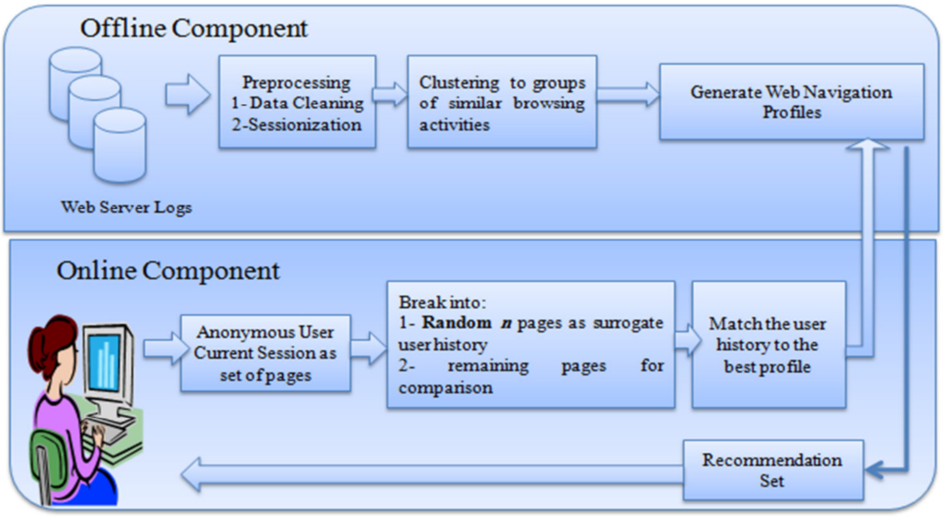

Offline Component
Web Server Logs
Preprocessing
1- Data Cleaning
2-Sessionization
Clustering to groups of similar browsing activities
Generate Web Navigation Profiles
Online Component
Anonymous User Current Session as set of pages
Break into:
1- Random *n* pages as surrogate user history
2- remaining pages for comparison
Match the user history to the best profile
Recommendation Set

Figure 2. the Proposed Architecture Overview

When the user navigates the internet, the web server will start to keep his logs on a file. This file can be accessed to extract the current active navigation web pages called the Active Session. Using this active session, the online component is responsible for assigning this user activities to the best navigation profile where by a recommendation list is to be created and attached to the user navigation list. A statistical analysis for the matching purpose was used in this study. Since the active session and the choose profile can be represented as vectors; the cosine coefficient commonly used in information retrieval was used to do the matching purpose (Mobasher, et al., 2002). A recommendation score is computed for those items not already visited by the user in the active session in order to recommend them based on their scores.

$$\text{RecScore} = \sqrt{[\text{PageWeight} * \text{ProfileMatch}]} \qquad (1)$$

Two factors are used in determining this recommendation score: the overall similarity of the active session to the profile as a whole, and the average weight of each item in the profile computed during obtaining the profiles (offline component).

Results and Discussion

Experimental Setup

Based on the proposed Architecture, a recommendation system is developed using Microsoft VC++ connected to Microsoft Access database through an Open Database Connection (ODBC). We used CTI dataset which contains 13745 sessions with 683 pageviews for the experiments with 70% for training and 30% for testing. We used the precision, coverage and F1 In order to evaluate the recommendation effectiveness. Assume that we have active current session *A* taken from the evaluation set and we have *R* as a recommendation set using the prediction engine over the navigation profiles. *W* represents the items that already visited by the user in *A*. The precision is defined as:

$$\text{Precision}(R, A) = \frac{|R \cap (A - w)|}{|R|}$$

the coverage is defined as $\text{coverage}(R, A) = \dfrac{|R \cap (A - w)|}{(A - w)}$

Finally, the harmonic mean of both is defined as $F1(R,A) = \frac{2 \times \text{Precision}(R,A) \times \text{coverage}(R,A)}{\text{Precision}(R,A) + \text{coverage}(R,A)}$

Evaluation of the recommendation Accuracy

Since the current user is anonymous to the recommender system with no previous navigation history, hence a sliding window technique over the current user session was used to represent the user history. To do so, the user current session is broken into two parts; the first part with size *n* pages is used as the surrogate user history which is matched against the web navigation profiles then produces a recommendation list from the selected profile. The remaining pages form the second part which is used for the comparison purpose to evaluate the recommendation accuracy. we used a window size equal to 2 to represent the surrogate user history and the rest pages are used for the evaluation purpose. Figures 3,4 and 5 relate the recommendation effectiveness for our system compared to the findings of previous methods namely, PACT and Hypergraph (Mobasher, et al., 2002). With a recommendation score's threshold varies from 0.1 to 1.0, the F1 measurement as a performance evaluation shows that our system performs better and achieves higher prediction accuracy. This improvement is due to the mining processes applied to the extracted navigation profiles by the offline component followed by a better classification of the current user to the best web navigation profiles.

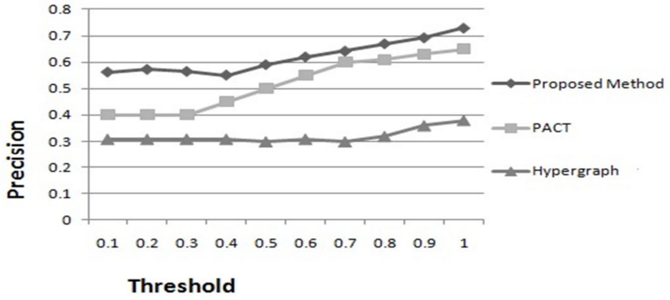


Figure 3. Precision Accuracy with Window Size=2

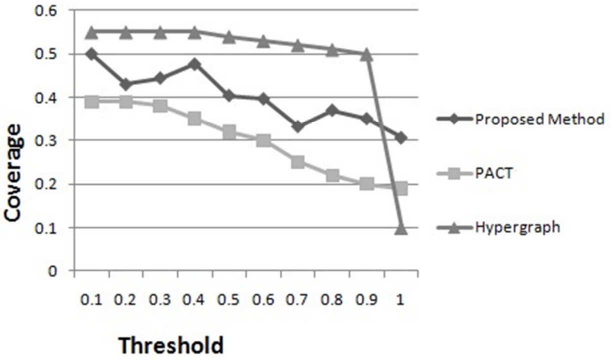


Figure 4. Coverage with Window Size=2

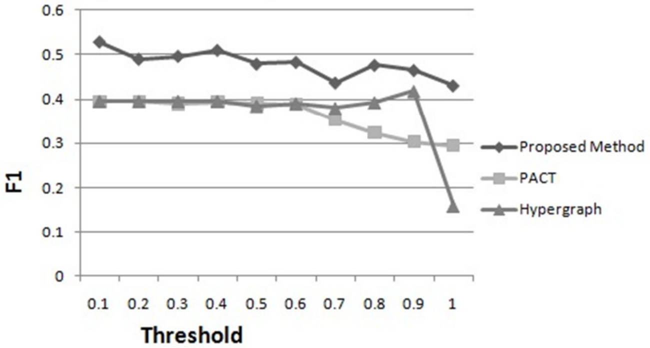


Figure 5. F1 Recommendation Measurement with Window Size=2

Conclusion and Future Works

The ability of predicting the next visited pages and recommending it to the user is highly recommended specially in e-commerce applications. This study proposed a recommender system to predict the anonymous user next navigation by assigning the user to the best web navigation profiles for similar interested users. The results showed higher prediction accuracy. Rebuilding the profiles is a time consuming process. Adaptive profiles are one of the future interests.

References

Almurtadha, Y., Sulaiman, M. N. B., Mustapha, N., & Udzir, N. I. (2011). IPACT: Improved Web Page Recommendation System Using Profile Aggregation Based On Clustering of Transactions. *American Journal of Applied Sciences, 8*(3), 277-283.

AlMurtadha, Y. M., Sulaiman, M. N. B., mustapha, N., & Udzir, N. I. (2010). Mining web navigation profiles for recommendation system. *Information technology Journal, 9*, 790-796.

Baraglia, R., & Silvestri, F. (2004). *An online recommender system for large Web sites*. Paper presented at the Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence

Baraglia, R., & Silvestri, F. (2007). Dynamic personalization of web sites without user intervention. *Communications of the ACM, 50*(2), 67.

Castellano, G., Fanelli, A. M., & Torsello, M. A. (2011). NEWER: A system for NEuro-fuzzy WEb Recommendation. *Applied Soft Computing, 11*(1), 793-806.

Cornelis, C., Lu, J., Guo, X., & Zhang, G. (2007). One-and-only item recommendation with fuzzy logic techniques. *Information Sciences, 177*(22), 4906-4921.

Jalali, M., Mustapha, N., Sulaiman, M. N., & Mamat, A. (2010). WebPUM: A Web-based recommendation system to predict user future movements. *Expert Systems with Applications, 37*(9), 6201-6212.

Liu, B. (Ed.). (2007). Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data Springer

Liu, H., & Kešelj, V. (2007). Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data & Knowledge Engineering, 61*(2), 304-330.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM, 43*(8), 142–151.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2002). Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. *Data Mining and Knowledge Discovery, 6*(1), 61 - 82.

Nakagawa, M., & Mobasher, B. (2003). *A hybrid web personalization model based on site connectivity*. Paper presented at the Proceedings of the 5th International WebKDD.

Yang, Q., & Wu, X. (2006). 10 Challenging Problems In Data Mining Research. *International Journal of Information Technology & Decision Making, 5*(4), 597–604.

Zhou, B., Hui, S. C., & Chang, K. (2004). *An intelligent recommender system using sequential web access patterns.* Paper presented at the 2004 IEEE Conference on Cybernetics and Intelligent Systems.