User-Based Web Recommendation System: A Case study of the National Museum of History

Kwoting Fang, Jonen Liu
Department of Information Management
National Yunlin University Science of Technology
Yunlin, Taiwan
jonen@ms19.hinet.net

Abstract

With the explosion and the rapidly growing market of the Internet, it is imperative that managers re-think to using technology, especially internet, to deliver services faster, cheaper, and with better quality than their competitors do. The web site provides a communication way that reveals real-time assess data and fruitful information of customers. Therefore, the call for customer with personalized web pages has become loud. To achieve personalized web pages, this study proposes recommendation algorithm of user behavior oriented by using the web log files from National Museum of History.
Keywords: Web usage mining, recommendation system, Fuzzy C-means.

1. Introduction

The largest web sites are offering an increasing amount of information. However, the World Wide Web is a Computer-Mediated Environment [9] in which uses the hyperlinks to connect related content. Since the hyperlink is a non-linear structure [5][16], most of users waste much time in retrieving information that they do not want and thus result in reducing the number of times web sites are revisited.

Therefore, a web site manager not only need to offer complete information for visitors, users' access behavior also need to be analyzed for increasing competitive advantages. Accordingly, personalized recommendation system is an important issue to provide one-to-one guidance to the users [17] and it's become more and more important for information provider or electronic commerce.

There are two major approaches to provide personalized recommendation: content based approach and collaborative technique approach. In former, it recommends items that are similar to what the user has liked in the past [11]; in collaborative technique approach, it defines other users that have showed similar preference to the given users and recommends what they have liked [19]. [6][15][19] have proposed the earliest and the most successful collaborative recommendation technologies.

Although most of researchers have addressed the development of recommendation systems to share information among interested parties[3][4][11][21][22], there are few researches for applying on actual web page and product recommendations. Accordingly, this study proposed a user-based recommendation system.

2. Related Work

Data mining is the process of extracting unknown but useful information from a very large database in order to enable a manager to make a decision [2]. An increasing number of enterprises are trying to understand customer behavior and provide a quality service to increase their competitive advantages using data mining technology. Accompany with the growth of the Internet has resulted in the development of web mining. Web mining extracts the information from an unlabeled and semi-structured dataset. A web site manger can use mining to elucidate and analyze user behavior to set a marketing strategy.

Until now, [4] addresses separated web mining applications into web content mining and web usage mining. In former is based on content, hyperlinks, structure, keywords, and so on, a search engine is an example; the other is based on a web log, user preference, register, clickstream, and so on, in which users access pattern are recorded.

Data that support to web mining are usually stored in a web log which file exists on web server that interacts with the user and the web server [18][15][4]. However, when they were be used, preprocessed at first is necessary. [14] divided the preprocessed procedure into five steps: fire step is data cleaning, followed by how to define user session, web page, completely access pattern and storing.

There are two major approaches to provide personalized recommendation: content based approach and collaborative technique approach. In former, it recommends items that are similar to what the user has liked in the past [11].

Content-based method using the user access pattern or previously purchased products to compute the relevance of other page items or products. Most researchers in this area were use distance to represent the relevance of similarity among page items. The NewsWeeder system is an example [11]; Collaborative technique approach is that products or page items are ranked according to a user's preference then similarities among users are determined. When a customer enters a store, the system refers to the user's preference to find similar users and recommends products or page items in which the new user may have an interest, but which he or she has never bought, GroupLens is an example [8]. These methods depend on users' purchase experience, a user must input preferences among items and specify related items.

Furthermore, [15][18][21] also considered the clustering of users with similar access patterns, using clustering to predict another user's behavior. [15] developed robust fuzzy clustering, applied it to web mining, who used the leader algorithm proposed by [7] to cluster the similar user sessions and thus identify users with similar interests or browsing paths. [21] applied clustering approach to user access patterns, and used a method of web page recommendation developed by WebWatcher [1] and Litizia [12].

3. Research Method and Architecture

The study proposes a web page recommendation system. This system with two phases -- offline module and online module. The offline module prepares and preprocesses data and mines usage; the online module is involved with online recommendation and evaluates the accuracy of recommendation.

3.1 Preparing and Preprocessing Data

The original web log used in the experimented was generated from httpd server. The preprocessing includes three parts – cleaning data, identifying user sessions and determining usage behavior. In order to clear the web log, .gif, .jpeg and .cgi et al., but not .html are removed.

3.2 Identifying a user

Identifying a user directly from a web log is difficult. The method of identification used in this paper is to set the 30 min timeout and suppose that an IP represents a single user, to determine the order in which page are browsed.

3.3 User Usage Mining

This paper considers web usage information obtained using the tool WebTrend, developed by Software Inc. Then, Fuzzy C-means clustering is to group users with similar access patterns. Accordingly similarities among users can be computed and related page items recommended to another user. The method of user-based recommendation is collaborative clustering users with similar access behavior. Common characteristics are extracted for sharing. The key steps are as follows:

First, a set of n unique URL is assumed as a vector:

$$\vec{D} = \{url_1, url_2, ..., url_n\} \quad (1)$$

Next, a user session S is expressed as a bit vector.

$$\vec{s} = (url_1^s, url_2^s, ..., url_n^s), \; url_i^s = \begin{cases} 1 & ,if \quad url_i \in s \\ 0 & ,otherwise \end{cases} \quad (2)$$

Only the click numbers are considered, so the importance of the page cannot be determined. The method is refined using the view number.

$$Url_i^s = \begin{cases} \sum count, & |Url_i \in s, count++, initial \;\; count = 0 \\ 0 & ,otherwise \end{cases} \quad (3)$$

Then, Fuzzy C-means clustering is used to determine the common characteristics of user behavior on which to base a recommendation. Before the fuzzy clustering is applied, the cosine of the angle between the two users is used to measure the similarity matrix. The element represents the similarity between users. A larger value implies greater similarity.

$$S_{xy} = \frac{\sum_{i=1}^{Nu} s_i^{(x)} s_i^{(y)}}{\sqrt{\sum_{i=1}^{Nu} (s_i^{(x)})^2} \sqrt{\sum_{i=1}^{Nu} (s_i^{(y)})^2}} \quad (4)$$

N_u, the number of unique URL; $s_i^{(x)}$, the frequency that user x access; $s_i^{(y)}$, the frequency that user y access.

The number of clicks is used to evaluate the relative weighting of the pages,

$$Weight(Url_i, c) = \sum Url_i^s, \; |s \in c \quad (5)$$

This result directly relates user behavior. The cluster to which each user belongs must be clearly determined. In this paper, the membership degrees of all clusters are compared and each user assigned to the highest.

Not all the work is included in a single phase, to ensure that the online recommendation is efficient. The work is separated into two phases – online and offline. The offline work is processed once a week. When the active user proposes a request, the online work is to trace who was browsed and give the user some feedback in which he may be interested.

3.4 Recommendation Phase

Sliding windows [14] are used to display the most recent web page items in which the active user is very interested. For example, when the number of sliding windows equals 3, the active session is <A, B, C>, and once the user requests page D, the active session becomes<B, C, C>.

No absolute method governs the decision of the appropriate number of page items. In this study, the default value is set to ten. Recommending the related page items for different users. There are three key steps:
First, the bit vector is S_A used to represent the active user's trace.

Next, the similarity between the active user and a group of users is computed.

$$SC_j^s = \sum_{i=1}^{n} (URL_i^j \times URL_i^A) \quad (6)$$

SC_j^s is the score for each session that are compared with active session S_A;

SC_j^s is multiplied by the membership degree in each jth cluster. The formula is following:

$$match(C_i) = \sum_{j=1}^{n} (S_j^{degree} \times SC_j^s) \quad (7)$$

Finally, $\max\{match(C_1), match(C_2), ..., match(C_i)\}$, for all $C_i \in C$,. The result is the cluster has the highest similarity to the active user. From the C_i, the top ten web page items are recommended according to the relative weights.

3.5 Evaluating the Result

Recall and precision [10] are used to evaluate the accuracy of the recommendation system. These indices are generally used in information retrieval. Recording the top ten recommended pages links for each page is difficult. Therefore, only the recall value is considered in the evaluation according to the following formula.

$$\text{Re}\,call = \frac{\left\| actually\ \ browse \cap recommended\ \ page\ \ items \right\|}{\left\| actually\ \ browse \right\|} \quad (8)$$

$\| \|$: number of web pages items

4. An experiment and Result

One million web log records were collected from the National Museum History from 1999/7 to 2001/11. A mirror based on the FreeBSD 4.5 platform was used to avoid influencing the original web site. Graphical records and those of clicks of about ten web page in 30 seconds were removed.

Identifying a user is difficult. The timeout was set to 30 minutes. The percentage of users who clicked less than ten pages was nearly 96%. Only the 143 user sessions of browsing of over 20 web pages were extracted.

The web site has a total of 46 web pages, six pages were clicked more than remnant (1327).

Before similar user characteristics can be clustered, the similarity between users must be computed. First, a bit vector is used to identify a user who clicks the page. Secondly, the cosine angle is used to compute the similarity to construct a 143*143 matrix, which is clustering base. In this matrix, all the elements on the diagonal are 1.

We use Fuzzy C-means is used to cluster similar users and Compactness and Separation Validity Functional [20] are used to measure the clusters' validity.

Figure 1 presents an example of the recommendation result. Frame is used to show the recommended pages dynamically.

52 users were invited to click the web pages and 929 records were recorded. Eight users' clickstreams were under three and 12 web logs may have been created by reloading events, and were removed. Finally, 44 users will with a clickstream of 917, in which 865 were the recommendation page items, the recall value was 865/917=94.32%.

5. Conclusion and recommendation

To achieve personalized web pages, this study proposed recommendation algorithm of user behavior oriented by using the web log files from National Museum of History and to evaluate the algorithm according to the recall value.

Some issue and questions remain outstanding: first, obtaining web log information is difficult. Therefore, the suitability of the recommendation system for other web sites cannot be evaluated. Next, the recommendation system involves mining user behavior. In the future, content mining will be included to improve the accuracy of recommendation and fit the users' interests. Then, Scanning the databases many times increases the time taken, extends the computation period, and increases the number of users. Finally, making the algorithms efficient and determining the optimal number of clusters remains important outstanding issues.

Figure 1: The recommendation result

Reference

[1] Armstrong R., Freitag T., Joachims and Michell, T. (1995), WebWatcher: A Learning Appretice for The World Wie Web, AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, March 1995.

[2] Cabena Peter, Hadjinian, Pablo, Stadler, Rolf, Verhees, Jaap and Zanasi, Alessandro, Discovering Data Mining from Concept to Implementation, New Jersey, Prentice Hall PTR, 1998.

[3] Chen M.S., Park J.S. and Yu P.S., Data mining for path traveral patterns in a Web environment, In Proceedings of the 16th International Conference on Distributed Computing Systems, pp. 385-392, 1996b.

[4] Cooley R., Mobasher B. and Srivastava J., Web mining: Information and Pattern Discovery on the World Wide Web, Proceedings of the 9the International Conference on Tools with Artificial Intelligence (ICTAI'97), pp. 558-567, 1997.

[5] Delisle, N. and Schwartz, M., Contexts: A Partitioning Concept for Hypertext, ACM Transactions on Office Information Systems, pp. 168--186, April 1987.

[6] Fu X., Budzik J., Kristian J.H., Mining Navigation History for Recommendation, ACM, pp. 106-112, 2000.

[7] Hartigan J., Clustering Algorithms, John Wiley., 1975.

[8] Herlocker J., Konstan J., Borchers A. and Riedl J., An Algorithmic Framework for Performing Collaborative filtering, To appear in Proceedings of the 1999 Conference on Research and Development in Information Retrieval, August, 1999.

[9] Hoffman D.L. and Novak T.P., Marketing in Hypermedia Computer-Mediated Environment: Conceptual Foundations, Journal of Marketing, pp. 50-68, July 1996.

[10] Kowalski, G., Information Retrieval Systems -- Theory and Implementation, Kluwer Academic Publishers, 1997.

[11] Lang K. Newsweeder: Learning to Filter Netnews, In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, Calif, 1995.

[12] Lieberman H.L. (1995), An Agent that Assists Web Browsing, In International Joint Conference on Artificial Intelligence, Montreal, August.

[13] Mannila H. and Toivonen H., Discovering generalized episodes using minimal occurrences, In Proc. Of the Second Int'l Conference on Knowledge Discovery and Data Mining, pp. 146-151, Portland, Oregon, 1996.

[14] Mobasher B., Cooley R. and Srivastava J., Automatic Personalization Based on Web Usage Mining, Communications of the ACM, Vol. 43, No. 8, August 2000,pp. 142-151.

[15] Nasraoui O., Krishnapuram R. and Joshi A., Mining Web Access Logs Using a Fuzzy Relational Clustering Algorithm based on a Robust Estimator, Proceedings of WWW8, August, 1999.

[16] Nielsen J., Hypertext and Hypermedia, USA Academia Press, Inc, 1990.

[17] Resnick, P., Neophytos, I., Miteth, S., Bergstrom, P. and Riedl, J., GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of CSCW94, Conference on Computer Supported Cooperative Work (Chapel Hill NC). Addison-Wesley, pp. 175-186, 1994.

[18] Shahabi C., Zarkesh A.M., Abidi J. and Shah V., Knowledge Discovery from User's Web-page Navigation, Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE), pp. 20-29, 1997.

[19] Shardanand, U. and Maes, P., Social Information Filtering: Algorithms for Automating 'Word of Mouth', In Processings of CHI95 (Denver CO), ACM Press, pp. 210-217, 1994.

[20] Xie X.L. and Beni G., A Validity Measure for Fuzzy clustering, IEEE Transacions on Pattern Analysis and Machine Intelligence, Vol. 13, No. 8, pp. 841-847, August 1991.

[21] Yan T., Jacobsen M., Garcia-Molina H. and Dayal U., From user access patterns to dynamic hypertext linking, In Fifth International World Wide Web Conference, Paris, France, 1996.

[22] Zaiane O.R., Xin M. and Han J., Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs, in Proc. Advances in Digital Libraries ADL'98, pp. 19-29. 1998.

[23] Software Inc. Webtrends. http://www.webtrend.com, 1995.