IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4753

Recommendation System Using Fuzzy C- Means Clustering

[1]Kwoting Fang, [2]Chung-Yuan Liu
[1,2]Department of Information Management, National Yunlin University Science of Technology
[1]fangkt@mis.yuntech.edu.tw; [2]jonen@ms19.hinet.net

ABSTRACT

The call for customer with personalized web pages has become loud. To achieve personalized web pages, this study proposes a recommendation system with two approaches on user access behavior by fuzzy clustering method. The aim of this study is to show similar preference to the given users and recommends what they have liked. An empirical case from National Museum of History is illustrated. The recall value was used to evaluate recommended accuracy with two algorithms (item-to-item and user-to-user). It is hopefully that our proposed method may provide valuable information for administrator to manage the web site more efficiency and for users to assess the interested pages more quickly.

1. INTRODUCTION

Personalized recommendation system is an important issue to provide one-to-one guidance to the users (Resnick et al., 1994) and it's become more and more important for information provider or electronic commerce. Until now, 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 (Lang, 1995); in collaborative technique approach, it defines other users that have showed similar preference to the given users and recommends what they have liked (Shardanand and Maes, 1995). Fu et al., 2000; Nasraoui et al., 1999; Resnick et al., 1994; Shardanand and Maes, 1994 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 (Chen et al., 1996; Cooley et al., 1997; Mannila and Toivenen, 1996; Yan et al., 1996; Zaiane et al., 1998), there are few researches for applying on actual web page and product recommendations. Accordingly, this study proposed a recommendation system with two different approach (item-based and user-based) based on users access behavior by fuzzy clustering.

2. RESEARCH METHOD AND ARCHITECTURE

This study proposed a web page recommendation system, which constructed in two modules (offline and online). The online module included three steps (data preparing, preprocessing and usage mining stage); the other is practice online recommendation and evaluates the recommended accuracy by recall value. The mainly two modules are described as follows.

2.1 Offline module

There are three stages in offline module, the first one is prepare and process data, secondly is to identify a user, the third is to mine user usage. 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. The method of identification used in this paper is to set the 30 min timeout. We proposed that an IP represents a single user to determine the order in which page are browsed. Since the pages have names of different lengths, each page must be assigned an identifying label such as 1, 2,…, n. Then, there are two algorithms in usage mining stage: item-based and user-based approach to implement. In addition, to obtain user access behavior, WebTrend was used (Software, 1995).

2.1.1 Item-based approach

The aim is to extract the high association browsed pages that user have not clicked previous. This study uses directed graph (known as digraph) to represent the relationship between pages. The form is G(V,E), V(vertex) = <url_A, url_B,…, url_N>, where E(edge) = <$link_{AB}$, $link_{BC}$,…,$link_{MN}$>, V is page identification and N is the number of all the pages that can be browsed.

Then, based on digraph, similar/dissimilar matrix are construed as follows: (1) extracting user access pattern (user session) from web log, such as <A, B, C, B, D>; (2) constructing N x N similar/dissimilar matrix for further research on fuzzy C-means, where N is as same in digraph; (3) determining the elements in similar matrix; for one user, if A_{ij} in matrix A equals 1, where i and j represent the vertex V and that means page i and page j are clicked together by user, otherwise equals 0; for all users, A_{ij} is accumulated by individual element value and it may be greater than 1, the value is larger, the concurrence frequency between page i and page j is larger too.

Unfortunately, there is a defect for above representation. Example of user session <A, B, C, B, D>, we can explain that user from A forward to B and C, but may be user want to hyperlink to D. However, limited to the web site structure, user could not do this directly, just backward to B then forward to vertex D only. Although it seemed reasonable, it's hard to represent the actual needs for user himself. Until now, maximal forward reference (MFR; Chen et al., 1996) proposed the common method to conquer this question. Therefore, dividing user session into number of transactions is the first thing to do. The detail algorithm for MFR would refer to Chen et al. (1996).

In this study, fuzzy C-means is used to cluster page items, which have similar concurrence clicked frequency thru similar/dissimilar matrix. The clustering result would show the membership degree of each page item to represent the relative weight (url, C_i), where C_i is i^{th} cluster.

2.1.2 User-based approach

The aim is to group users who have similar access behavior and then share common characteristics for other users. The key steps are as follows:

(1) A user session s is expressed as a bit vector.

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

Since only the click numbers are considered, the importance of the page cannot be determined. Therefore, the method is refined:

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

(2) To construct similar/dissimilar matrix, cosine of the angle between two users is used to measure. Each element represents the similarity between users. A larger value implies with greater similar assess behavior between users.

$$S_{xy} = \frac{\sum_{i=1}^{Nu} s_i^{(x)} s_i^{(y)}}{\sqrt{\sum_{i=1}^{Nu} (s_i^{(x)})^2} \sqrt{\sum_{i=1}^{Nu} (s_i^{(y)})^2}}$$, where N_u is the number of unique URL; $s_i^{(x)}$ is the frequency that user x access and $s_i^{(y)}$ is the frequency that user y access.

As same as item-based approach, fuzzy C-means method is also used. In user-based approach, the number of clicks is used to evaluate the relative weighting of the pages,

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

The membership degree is used to represent the similarity between i^{th} user session (s_i) and the cluster c_i, in which $c_i = \{< s_1, \deg ree >, < s_2, \deg ree >, ..., < s_n, \deg ree >\}$. This result directly relates user behavior. To determine a user should belong to which cluster, we compared with the membership degree of user session in each cluster and then select the max one.

2.2 Online recommendation

Sliding window (Mobasher et al., 2000) is used to display the most recent web page items in which the active user is most interested. For example, when the number of sliding window equals 3, the active session is <A, B, C>, and once the user requests page D, the active session becomes<B, C, D>.

No absolute method governs the decision of the appropriate number of page items. In this study, the ten is given.

To recommend the top-N page items for other users, item-based and user-based approaches are described separately as follows:

2.2.1 Item-based recommendation procedure

A set of similar page items have cluster thru fuzzy C-means. Therefore if we can show out the active user session, the nearest cluster would be determined.

(1) Active session A is expressed as a bit vector: $\overrightarrow{A} = (url_1^A, url_2^A, ..., url_n^A)$, where

$$url_i^A = \begin{cases} 1 & ,if \quad url_i \in A \\ 0 & ,otherwise \end{cases}$$

(2) The similar score between active session and each cluster was computed and then selected the maximal score to be the nearest cluster, e.g.,

$$\max(match(C_i) = \sum_{j=1}^{n} (url_j^{\deg ree} \times url_j^A)$$

2.2.2 User-based recommendation procedure

Three steps are shown as follows:

(1) This step is as same as step 1 in Section 2.2.1.

(2) The similarity between the active user and a group of users is computed by $SC_j^s = \sum_{i=1}^{n} (URL_i^j \times URL_i^A)$, where SC_j^s is the matched score between each user session and active user; SC_j^s is multiplied by the membership degree in each j^{th} cluster, e.g., $match(C_i) = \sum_{j=1}^{n} (S_j^{\deg ree} \times SC_j^s)$.

(3) $\max\{match(C_1), match(C_2), ..., match(C_i)\}$, 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.

2.3 Evaluating the Result

Finally, Recall and precision (Kowalski, 1997) are used to evaluate the accuracy of the recommendation system. These indices are generally used in information retrieval. Since to record the top N recommended links in each page lacks efficiency, only the recall value is considered for the evaluation, as shown in below:

$$\text{Re}call = \frac{\|actually\ browse \cap recommended\ \ page\ items\|}{\|actually\ browse\|}$$

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

For example, recommendation system recommended three page items: T1, T2 and T3. The user actually browsed only T1 and T3 and thus recall value 0.5 was given.

When the recall value equals 1, means user is interested in all of the recommended page items; on the contrary, 0 represents the recommendation system does not work for users.

3. EXPERIMENTAL RESULTS

3.1 Data

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.

However, to identify a user is difficult. In this paper, 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 page items.

3.2 Item-based approach

46 x 46 similar matrix was constructed. In this paper, we adopted compactness and Separation Validity Function (S) to determine the appropriate cluster number (Xie and Beni, 1991) and four was given. According to fuzzy C-means algorithms, 46 page items are assigned to each cluster by the objective function

$$J(U, v) = \sum_{i=1}^{c} \sum_{k=1}^{n} (\mu_{ik})^m d^2 (X_k, V_i).$$

The result was shown in table 1.

3.2 User-based approach

The cosine angle is used to compute the similarity to construct a 143*143 matrix. In this matrix, all the elements on the diagonal are 1.

We also used Fuzzy C-means is used to cluster similar users and Compactness and Separation Validity Functional (Xie and Beni, 1991) are used to measure the clusters' validity, then 14 groups of users were obtained (shown in Table 2).

Table 1: Fuzzy C-means Clustering for item-based approach

Cluster no.	Similar page items	Items no.
1	2,3,4,5,6,7,9,10,12,13,14,15,16,17	14
2	1,23,25,26,28,29,30,31,32,36,39,41	12
3	8,11,18,19,20,21,22,24,27,33,34,35,37,38	14
4	40,42,43,44,45,46	6

3.4 Demonstration site

Either item-based or user-based, at the online recommendation stage, we used the Cookie Technique to record the active session and then to determine the top-10 recommended pages for users who are interested in a web site, as shown in Fig. 1. There are two frames in each window. The top frame contains the actual page contents that the user clicks on a link node in either frame. The bottom frame contains the top-10 page links, those were determined dynamically by our recommendation system.

3.3 Evaluation results

In evaluation stage, this study adopted volunteer users who have network and interested in this experiment for convenient sampling. 108 users were invited to visit and 1398 records were recorded. However, 25 users whose browsed pages number is less than 3, deleted later; 77 web logs may have been created by reloading events, were removed also.

Finally, 83 users and 131 are obtained for analysis, in which, 39 users clicked 404 pages and 324 browsed from recommended page items based on item-based approach; Recall value equals 324/404=80.2 percent. In addition, 44 users with a clickstream of 917, in which 865 were the recommended page items, the recall value was 865/917=94.32%.

4. CONCLUSION

To achieve personalized web pages, this study proposed a recommendation system with two different approach based on user behavior oriented by using the web log files from National Museum of History. Finally, to evaluate and compared with these two algorithms, the index of recall value was used. It is hopefully that the recommendation algorithms presented in this paper may provide valuable information for administrator to manage the web site more efficiency and for users to assess the interested pages more quickly.

The results showed that user-based recommendation algorithm recall value is good at 94.32 percent, which is better than item-based approach (80.2 percent). Despite of this, our recommendation system with different approaches is useful. Finally, some issue and questions remain outstanding: first, our experiment was just on a dataset; therefore, the suitability of the recommendation system for other web sites cannot be evaluated and lack reliability. Next, our recommendation system is based on user usage behavior and content mining will be included to improve the accuracy of recommendation and fit the users' potential preference.

Table 2: Fuzzy C-means clustering of users

Cluster no.	User	Item no.
1	3,4,13,19,21,52,56,58,59,67,75,83,84,102,112,114,123,138	18
2	2,16,36,45,54,77,105,107,111,127,133,134	12
3	9,32,49,50,51,70,98,131	8
4	29,132	2
5	63	1
6	1,5,6,7,8,10,11,15,17,24,25,28,31,33,35,38,39,40,41,42,43,44,46,47,48,55,61,62,64,66,69,72,78,79,80,81,86,88,91,92,93,94,97,100,103,106,108,109,110,113,115,116,117,118,119,120,121,125,128,135	62
7	12,90,124,130	4
8	60,71	2
9	76,137	2
10	27,34,101	3
11	14,23,26,53,57,74,82,87,99,122,126,136	12
12	18,20,30,37,68,73,85,129	8
13	65,89,95,96,104	5
14	22	1

Figure 1: Example of the demonstrate sites for user-based

REFERENCE

1. Chen M.S., Park J.S. and Yu P.S. (1996), Data Mining for Path Traveral Patterns in a Web Environment, In Proceedings of the 16th International Conference on Distributed Computing Systems, pp. 385-392.

2. Cooley R., Mobasher B. and Srivastava J. (1997), 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.

3. Fu X., Budzik J., Kristian J.H. (2000), Mining Navigation History for Recommendation, ACM, pp. 106-112.

4. Kowalski, G. (1997), Information Retrieval Systems — Theory and Implementation, Kluwer Academic Publishers.

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

6. Mannila H. and Toivonen H. (1996), 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.

7. Mobasher B., Cooley R. and Srivastava J. (2000), Automatic Personalization Based on Web Usage Mining, Communications of the ACM, 43 (8), pp. 142-151.

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

9. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Reidl, J. (1994), Grouplens: an open architecture for collaborative filtering of netnews. Proceedings of the ACM Conference onComputer Supported Cooperative Work.

10. Shardanand, U. and Maes, P. (1995), Social Information Filtering: Algorithms for Automating 'Word of Mouth', In Processings of CHI95 (Denver CO), ACM Press, pp. 210-217.

11. Software Inc. (1995), Webtrends, http://www.webtrend.com.

12. Xie X.L. and Beni G. (1991), A Validity Measure for Fuzzy clustering, IEEE Transacions on Pattern Analysis and Machine Intelligence, 13 (8), pp. 841-847.

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

14. Zaiane O.R., Xin M. and Han J. (1998), 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.