Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 4, No. 2, May 2013), Pages: 15-24
www.jacr.iausari.ac.ir

Mining users` navigation patterns for building web pages recommendation system

Farzad Soleymani Sabzchi[1]✉, Shahram Jamali[2], Maryam Jafari[3]

(1) Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran
(2) Computer Engineering Department, University of MohaghehArdabili, Ardabil, Iran
(3) Department of Computer Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran

f_soleymani63@yahoo.com; jamali@iust.ac.ir; m123_jafari@yahoo.com

Received: 2013/03/19; Accepted: 2013/05/02

Abstract

Due to the quick growth of the World Wide Web, retrieval of useful information from the Internet for a particular web user or a group of users becomes very difficult. Recommendation systems using web usage mining help providing an adaptive web environment for the web users. This paper presents a novel approach for page recommendation using fuzzy association rule mining algorithm. This method extracts previous users` access patterns and then employs them to recommend appropriate web pages for the active user. An illustrative example explains this method in details.

Keywords: *Web recommender, Web usage mining, Fuzzy association rule mining, Pattern discovery*

1. Introduction

Over the recent decades, fast growth of World Wide Web has led to huge amount of available information that is not simply accessible for the users. Recommender Systems (RS) help us to cope with confusion among the huge volume of information, and recommend pages that may fit our interests [1,2]. Web recommender systems are one of the applications obtained from Web Usage Mining (WUM) technique. WUM can elicit the web user's navigational behavior from secondary data, such as web server access logs, proxy server logs, browser logs, user profiles, user sessions, and user queries.Then it models the behavior as navigation patterns for designing the intelligent web sites [3,4,5]. Some examples of these web sites are recommending books, CDs and other products at Amazon.com [6], recommending movies by MovieLens [7], recommending books at LIBRA [8], and recommending electronic television program guides [9]. Several WUM techniques have been used to effectively develop recommendation systems. There have been efforts to use association rules [10], sequential patterns [11], and Markov models [12] in recommender systems.

Two effective factors in recommend pages to a web user are previous visited pages of the active user and the history of browsing the web pages of previous users. Hence, the main purpose of this paper is to recommend the appropriate pages based on discovered patterns from previous users` sessions (off-line) and active user`s navigational behavior (on-line). Patterns of previous users are extracted using Fuzzy Association Rule Mining

(FARM) technique in off-line part. Afterward, on-line part of proposed recommendersystem gets the active user's request, searches the antecedent of the rules for a match. In the case of finding a perfect match, recommends the pages in the consequence.

The rest of this paper is as follows. In the next section motivation of the paper is presented. In section 3 related works on web recommendation systems are reviewed. The proposed approach in this paper is described in section 4. An example to illustrate the proposed approach is given in section 5. Finally a conclusion of this work is presented in section 6.

2. Motivation of the research

Most of the web users complain about finding useful information on web sites. Web recommender systems predict the information needs of users and provide them with recommendations to facilitate their navigation.Recommender systems have been extensively explored in web mining. However, the quality of recommendations and the user satisfaction with such systems are still not optimal.All recommender systems based on web usage mining techniques have strengths and weaknesses. One of these weaknesses is lack of consideration of an appropriate measure to calculate the users` interest degree of pages. Applying fuzzy association rule mining to a recommender system leads to calculate users` interest of pages more accurately. Consequently generated recommendations will be more desirable.

3. Related Works

In this section some widely used techniques for implementation of recommender systems are described. The first technique clusters pages found from web server log files. A simple k-means algorithm clusters user sessions based on vector distances. This model is based on clusters of user sessions and uses the visiting time of pages and does not consider the visiting order of the pages [13]. In the click-stream tree model, the recommendations are generated using the recommender technique proposed in [14]. This technique uses the sequence of pages visited in a user session. In this method pairwise similarities between user sessions are calculated by a new similarity measure. Afterward the user sessions are clustered using a graph-based clustering algorithm. Nakagawa and Mobasher [15], proposed a hybrid recommender system that can intelligently switch among various recommendation models. Switching criterion is based on the degree of hyperlink connectivity and the neighborhood of a user’s current location within the site to select the best recommendation model. This system uses three recommendation models are based on Association Rules, Sequential Patterns and Contiguous Sequential Pattern to generate navigation patterns of Web users.

In another work, the recommender system proposed by Mobasher et al., [16] is based on association rule mining method for web personalization. This approach present a data structure for storing the discovered frequent item sets from clickstream data. Presented recommendation algorithm applies this data structure to generate recommendations in real-time, without the need to extract all association rules from frequent item sets. This system produces recommendations based on matching the current user navigational behavior against patterns discovered through association rule mining.

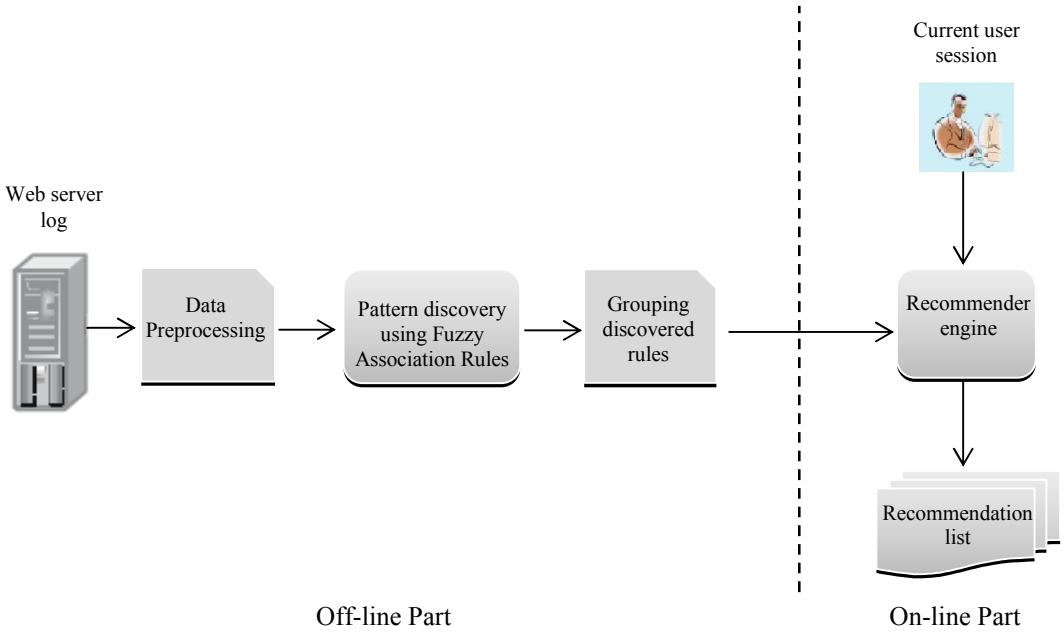

Figure 1. Recommender system architecture

4. Proposed Approach: Web Recommender System Based on FARM

The framework of usage based recommendation system shown in Fig. 1 which includes off-line part and on-line part. This system includes different phases such as preprocessing, pattern discovery, off-line part and on-line part, which are described in following subsections.

4.1 Preprocessing

Data preprocessing is the first phase of this proposed system. The results of data preprocessing directly impact the results of next steps. This phase consists of data cleaning, user and session identification and generating the database of log file. Usually, preprocessing is done by considering the visit time of a page.

A weight measure presented in [17] is used to approximate the interest degree of a web page for a user. Here, two concepts related to this measure are introduced: "*Frequency*" and "*Duration*". Frequency is the number of times that a page is accessed by different users. A parameter that must be considered to calculate the frequency of a page is the In-degree of that page (e.g. the number of incoming links to the page). The formula of "*Frequency*" is given in (1) which is normalized by the total number of visits of web pages in the session:

$$Frequency(P) = \frac{Numberofvisit(P)}{\sum_{Q \in T} Numberofvisit(Q)} \times \frac{1}{Indegree(P)} \quad (1)$$

"*Duration*" is defined as the time the user spent on a page. It is guessed that the longer time a user spends on a page means that the user is more interested in that page. "*Duration*" of a web page is shown in (2).

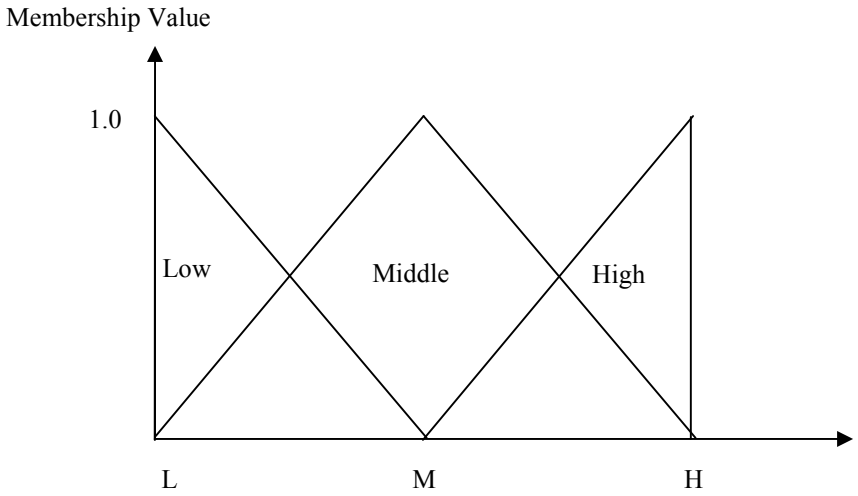

Figure 2. The triangle membership function used for partitioning interest degree of the page

$$Duration(P) = \frac{\frac{Total\ duration(P)}{Size(P)}}{max_{\ Q \in T} \frac{Total\ Duration(P)}{Size(P)}} \quad (2)$$

"*Frequency*" and "*Duration*" are considered two strong indicators of users' interest. The harmonic mean of these two measures is used to represent the interest degree of a web page to a user in the session, shown as below:

$$Importance(P) = \frac{2 \times Frequency(P) \times Duration(P)}{Frequency(P) + Duration(P)} \quad (3)$$

Equation (3) guarantees that "*Interest*" of a page is high only when "*Frequency*" and "*Duration*" are both high.

4.2 Pattern Discovery

Pattern discovery is a phase which extracts the user behavioral patterns from the formatted data. In this phase, several data mining techniques are applied to obtain hidden patterns reflecting the typical behavior of users. Some important techniques for this phase are: path analysis, standard statistical analysis, clustering algorithms, association rules, classification algorithms, and sequential patterns. In this paper fuzzy partition method and fuzzy association rule mining technique are used to extract interested patterns.

4.2.1 Fuzzy partition method

The concepts of fuzzy sets and linguistic variables were proposed by Zadeh [18,19] and it is reasonable that different linguistic values can indicate diversity of degree of the attribute in many applications [20]. In the simple fuzzy partition method, *K* several linguistic values are defined in each quantitative attribute. *K* is also predefined before performing the method. Triangular and trapezoid membership functions are usually used for the linguistic values.

In this paper, the quantitative attribute "*Interest degree*" is partitioned by simple fuzzy partition method with triangle membership function. This attribute is divided into 3 linguistic values (*Low, Middle, High*) and its membership function is predefined in

Fig. 2. These membership functions can be represented by triangle fuzzy numbers. Limits of each linguistic variable can be displayed as below: $A_{Low}^{Interest}$ = (*l, m, m*) and $A_{Middle}^{Interest}$ = (*l, m, h*) and $A_{High}^{Interest}$ = (*m, m, h*).

Using formulas (4), (5) and (6), the attribute is partitioned into suitable linguistic value and obtained its membership in accordance with membership function, respectively.

$$\mu_{\text{Low}}^{\text{Interest}}(x) = \begin{cases} 1, & x < l \\ \frac{m-x}{m-l}, & l \le x \le m \\ 0, & x > m \end{cases} \quad (4)$$

$$\mu_{\text{Middle}}^{\text{Interest}}(x) = \begin{cases} \frac{x-l}{m-l}, & l < x \le m \\ \frac{h-x}{h-m}, & m < x < h \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$$\mu_{\text{Low}}^{\text{Interest}}(x) = \begin{cases} 0, & x < m \\ \frac{x-m}{h-m}, & m \le x < h \\ 1, & x \ge h \end{cases} \quad (6)$$

4.2.2 Fuzzy association rules

In general, the interesting or preference of each user is difficult to describe clearly. In order to generate recommendations, fuzzy association rules are used to provide behavioral patterns. Association rules indicate how much the degree of user interest in a page will effect on the next visits of pages in the web site. A rule consists of an antecedent and a consequent. The rule expresses that when the antecedent happens the consequent will also happen. The importance of each rule is measured using its support and confidence. The support of a rule is the number of instances that contain both the itemset of its antecedent and the itemset of its consequence. The conditional probability of the occurrence of the consequent given the antecedent is referred to as the confidence of the rule.

4.3 Off-line part of the Recommender System

The off-line part, models usage patterns from the web server access log data and builds a predictive model based on the extracted usage patterns. After generating frequent patterns by FP-Growth algorithm [21], the procedure of fuzzy association rules mining is executed. Steps that are performed in off-line part are as follows:

Off-line part algorithm:

Step 1. Preparation and preprocessing of web log data.
Step 2. Attributes will be partitioned into appropriate linguistic value using fuzzy partition method.
Step 3. Frequent patterns are extracted by using fuzzy FP-Growth algorithm
Step 4. All the fuzzy association rules are generated from available frequent patterns.
Step 5. The general fuzzy association rule is indicated by (7). The Fuzzy Support (FS) for all obtained fuzzy association rules are calculated using (8) as follows:

$$R: A_{s_1}^{x_1}, A_{s_2}^{x_2}, \ldots, A_{s_k}^{x_k} \rightarrow A_{s_\alpha}^{x_\alpha} \quad (7)$$

$$FS(A_{s_1}^{x_1}, A_{s_2}^{x_2}, \ldots, A_{s_k}^{x_k}) = \frac{1}{n}\sum_{j=1}^{n} [\mu_{s_1}^{x_1}(d_j^{x_1}) \times \mu_{s_2}^{x_2}(d_j^{x_2}) \times \ldots \times \mu_{s_k}^{x_k}(d_j^{x_k})] \quad (8)$$

Where, x_i is the ith attribute, s_i is the term of degree in x_i, $A_{si}^{x_i}$ is the linguistic value of s_i in the ith attribute, $\mu_{s_i}^{x_i}$ is the membership function of s_i in the ith attribute, $d_j^{x_i}$ is the original value of the ith attribute in the jth tuple.

Step 6. If the FS value of fuzzy association rule is larger than or equal to the user-specified min FS, this rule will be accepted.

Step 7. Ineffective fuzzy association rules will be filtered out with fuzzy confidence. The Fuzzy Confidence (FC) of R is defined as follows:

$$FC(R) = FS(A_{s_1}^{x_1}, A_{s_2}^{x_2}, \ldots, A_{s_k}^{x_k}, A_{s_\alpha}^{x_\alpha}) / FS(A_{s_1}^{x_1}, A_{s_2}^{x_2}, \ldots, A_{s_k}^{x_k}) \quad (9)$$

When the FC(R) of rule R is less than the user-specified min FC, the rule R will be regarded as ineffective fuzzy association rule and filtered out.

Step 8. The useless fuzzy association rules are Pruned. If there exist two rules, denoted by R and S, possessing same consequence, and the antecedence of R is subset of S, then R is useless and can be removed.

Step 9. The interested discovered association rules are then grouped based on the number of items in their antecedent. The result is a group of rules with one item in the antecedent, a group with two items and so on.

4.4 On-line part of the Recommender System

The recommendation system takes the user's request (the active user session) and searches the items on the antecedent of each rule for finding matches. This process starts from the group of rules with one-item in the antecedent and then the next group with two items in the antecedent and so on. If matches are found, these rules are sorted according to their confidence values in descending order. Then the top *N* items of consequents from this list are recommended to the user.

When the user accesses two web pages, system searches the one-item group for the second web page and next, searches the two-items group for these two pages. When no match is found the system removes the user's browsing history and again the process starts from the beginning when a match is found. In the next section, an example to illustrate the method proposed in this paper is presented.

5. Illustrative Example

In this section, the function of the proposed recommendation system is illustrated with an example. This system uses a database of users` sessions. This database shown in Table 1 that consists of ten users` sessions and six web pages denoted P_1, P_2, P_3, P_4, P_5 and P_6.Each item is represented by a tuple (page name: visit time). Features of these pages are shown in Table 2 that includes In-degree and the size of each page. The membership function used in this example is drawn in Fig. 2. In this membership function, $l = 0.15$, $m = 0.45$ and $h = 0.75$.

Table 1. The ten sessions used in the example

Session ID	Page Access Sequence
1	(P_1: 83), (P_3: 355), (P_4: 191), (P_6: 378)
2	(P_2: 207), (P_3: 267), (P_4: 455), (P_5: 407), (P_6: 411)
3	(P_1: 59), (P_2: 181), (P_5: 233), (P_6: 256)
4	(P_1: 67), (P_2: 391), (P_3: 271), (P_4: 241), (P_5: 291), (P_6: 501)
5	(P_2: 173), (P_5: 302), (P_6: 598)
6	(P_1: 120), (P_3: 107), (P_4: 219), (P_5: 391)
7	(P_2: 132), (P_3: 62), (P_4: 141), (P_5: 358), (P_6: 311)
8	(P_1: 95), (P_2: 283), (P_3: 211), (P_6: 279)
9	(P_1: 87), (P_3: 93), (P_4: 257), (P_6: 351)
10	(P_1: 139), (P_2: 98), (P_3: 148), (P_5: 183), (P_6: 393)

Table 2. Features of pages

Page	In-degree	Size
P_1	1	293 KB
P_2	2	382 KB
P_3	2	170 KB
P_4	1	131 KB
P_5	3	111 KB
P_6	3	209 KB

Table 3. The amount of interest to page in user`s session

Session ID	Page Access Sequence
1	(P_1: 0.225), (P_3: 0.571), (P_4: 0.646), (P_6: 0.445)
2	(P_2: 0.207), (P_3: 0.414), (P_4: 0.734), (P_5: 0.734), (P_6: 0.385)
3	(P_1: 0.167), (P_2: 0.273), (P_5: 0.374), (P_6: 0.397)
4	(P_1: 0.156), (P_2: 0.368), (P_3: 0.482), (P_4: 0.646), (P_5: 0.52), (P_6: 0.452)
5	(P_2: 0.218), (P_5: 0.37), (P_6: 0.462)
6	(P_1: 0.199), (P_3: 0.247), (P_4: 0.53), (P_5: 0.374)
7	(P_2: 0.164), (P_3: 0.176), (P_4: 0.429), (P_5: 0.374), (P_6: 0.363)
8	(P_1: 0.361), (P_2: 0.429), (P_3: 0.559), (P_6: 0.462)
9	(P_1: 0.248), (P_3: 0.325), (P_4: 0.75), (P_6: 0.444)
10	(P_1: 0.371), (P_2: 0.186), (P_3: 0.429), (P_5: 0.364), (P_6: 0.462)

In this example, values are represented by three fuzzy regions: *Low, Middle* and *High*. Thus, three fuzzy membership values are produced for each page in a session according to the predefined membership functions (4), (5) and (6).

Here, assume that the predefined min FS and min FC are 0.20 and 0.60 respectively. The off-line part method can be described as follows:

Step 1,2: The results of these steps are presented in Table 3.

Step 3: using fuzzy FP-Growth algorithm, frequent patterns were extracted in Table 4.

Step 4,5: Fuzzy association rules from the results of step3 are discovered .

Step 6,7: Comparing FS and FC of the rules with predefined min FS and min FC, interested rules obtained and are shown in Table 5.

Step 8,9: By applying the results of Step 7, the pruned rules are grouped in Table 6.

Table 4. Conditional pattern base

Frequent patterns	Fuzzy values	Frequent Patterns	Fuzzy values
P_4.High	3.52	P_2.Low, P_3.Middle	2.98
P_2.Low	4.34	P_2.Low, P_5.Middle	4.27
P_5.Middle	4.51	P_5.Middle, P_6.Middle	3.76
P_1.Low	4.74	P_3.Middle, P_5.Middle	3.03
P_3.Middle	4.94	P_1.Low, P_5.Middle	2.98
P_6.Middle	8.15	P_1.Low, P_6.Middle	3.9
P_4.High, P_6.Middle	3.25	P_1.Low, P_3.Middle	3.8
P_3.Middle, P_4.High	3.52	P_3.Middle, P_6.Middle	4.62
P_1.Low, P_4.High	2.57	P_6.Middle, P_3.Middle, P_1.Low	2.96
P_2.Low, P_6.Middle	4.34	P_6.Middle, P_3.Middle, P_4.High	3.25
P_6.Middle,P_1.Low, P_4.High	2.3	P_3.Middle,P_5.Middle,P_2.Low	2.91
P_3.Middle, P_1.Low, P_4.High	2.57	P_6.Middle,P_3.Middle, P_5.Middle	2.28
P_6.Middle,P_3.Middle,P_2.Low	2.98	P_3.Middle,P_1.Low, P_5.Middle	2.23
P_6.Middle,P_5.Middle, P_2.Low	4.27	P_6.Middle, P_3.Middle,P_1.Low, P_4.High	2.3
		P_6.Middle, P_3.Middle,P_5.Middle, P_2.Low	2.91

Table 5. Fuzzy association rules

Association rules	FS	FC
P_4.High → P_6.Middle	0.3	0.86
P_4.High → P_3.Middle	0.25	0.7
P_3.Middle → P_6.Middle	0.43	0.87
P_1.Low → P_6.Middle	0.37	0.78
P_5.Middle → P_6.Middle	0.33	0.74
P_2.Low → P_6.Middle	0.37	0.85
P_2.Low → P_5.Middle	0.26	0.6
P_2.Low, P_5.Middle → P_6.Middle	0.22	0.87
P_1.Low, P_3.Middle → P_6.Middle	0.21	0.87
P_3.Middle, P_4.High → P_6.Middle	0.22	0.88

Table 6. Fuzzy association rules

Association rules	FS	FC	Group
P_4.High → P_3.Middle	0.25	0.7	group1
P_2.Low → P_5.Middle	0.26	0.6	group1
P_2.Low, P_5.Middle → P_6.Middle	0.22	0.87	group2
P_1.Low, P_3.Middle → P_6.Middle	0.21	0.87	group2
P_3.Middle, P_4.High → P_6.Middle	0.22	0.88	group2

The on-line part of the recommender system works as follow:

Assume three users visited these pages in this manner: user1 {P_2, P_5}, user2 {P_1}, user3 {P_2, P_3}. For user1, recommender system searches P_2 in the antecedents of rules in group1. Next it searches P_2, P_5 in the antecedents of rules in group2. Sorted rules according to FC are recommended to this user. These recommendations are P_6.Middle and P_5.Middle. For user2 who visited P_1, recommender system could find no appropriate page and waits until this user visit the second page without removing the user`s browsing history. Recommender system could not find any matches in group2 for user3 with two visited pages $P_2$ and $P_3$. In this situation, the recommender system will drop the user`s browsing history.

6. Conclusion

In this paper, a framework was proposed to generate recommendations for web users. For this purpose it employed web usage data. Our recommender system integrates the results from two distinct parts, off-line and on-line parts. It generates a single recommendation list for current user. Off-line part of the system discovers frequent patterns by identifying popular pages in user`s navigation path based on fuzzy association rule mining. Next, on-line part uses these navigation patterns and current user's request to generate recommendations for the same current user. An illustrative example showed how the proposed method recommends pages based on the system history.

7. References

[1] Campos L, Fernandez-Luna J, Huete J. A collaborative recommender system based on probabilistic inference from fuzzy observations. Fuzzy Sets and Systems 2008;159:1554–1576.
[2] Goksedef M, Oguducu S. A Consensus Recommender forWeb Users. Springer-Verlag Berlin Heidelberg 2007;287–299.
[3] Buchner A, Mulvenna M. Discovering Internet marketing intelligence through online analytical web usage mining. ACM SIGMOD Record 1998;27:54–61.
[4] Lin S, Chen M, Ho J, Huang Y. ACIRD: Intelligent Internet document organization and retrieval. IEEE Transactions on Knowledge and Data Engineering 2002;14:599–614.
[5] Srivastava J, Cooley R, Deshpande, Tan P. Web usage mining: discovery and applications of usage patterns from web data. ACM SIGKDD Explorations Newsletter 2000;1:12–23.
[6] Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing 2003;7:76–80.
[7] Miller B, Albert I, Lam S, Konstan J, Rield J. Movielens unplugged: experiences with an occasionally connected recommender systems. Intelligent User Interfaces 2002;3:263–266.
[8] Mooney R, Roy L. Content-based book recommending using learning for text categorization. Digital Libraries 2000;5:195–204.
[9] O'Sullivan D, Smyth B, Wilson D, McDonald K, Smeaton A. Improving the quality of personalized electronic program guide, User modeling User-Adapted Interact 2004;14:5–35.
[10] Nanopoulos A, Katsaros D, Manolopoulos Y. Effective prediction of web-user accesses: a data mining approach. Conference on Mining Log Data Across All Customer Touchpoints (WebKDD'01), San Francisco 2001.
[11] Agrawal R, Srikant R. Mining sequential patterns. Eleventh International Conference on Data Engineering 1995;95:3-14.
[12] Deshpande M, Karypis G. Selective Markov models for predicting web-page accesses. ACM Transactions on Internet Technology (TOIT) 2004;4:163-184.
[13] Mobasher B, Dai H, Luo T, Nakagawa M. Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery 2002;6:61–82.
[14] Demir G, UyarA, Gunduz S. Multiobjective evolutionary clustering of Web user sessions: a case study in Web page recommendation. Soft Computing - A Fusion of Foundations, Methodologies and Applications 2010;14:579-597.
[15] Nakagawa M, Mobasher B. A hybrid web personalization model based on site connectivity. WEBKDD Workshop (Washington, DC) 2003;5:59-70.
[16] Mobasher B, Dai H, Luo T, Nakagawa M. Effective personalization based on association rule discovery from web usage data. Web information and data management 2001;3: 9–15.
[17] Liu H, Keselj V. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. Data & Knowledge Engineering 2007;61:304–330.
[18] Zadeh L. The concept of a linguistic variable and its application to approximate reasoning. Information Science 1975;8:199-249.
[19] Zadeh L. The concept of a linguistic variable and its application to approximate reasoning. Information Science 1975;8:301-357.

[20] Hu Y, Hu J, Chen R, Tzeng G. Assessing weights of product attributes from fuzzy knowledge in a dynamic environment. European Journal of Operational Research 2004;154:125-143.

[21] HAN J, PEI J, YIN Y, MAO R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 2004;8:53–87.