Volume 3 No. 1, January 2013 ISSN 2223-4985

International Journal of Information and Communication Technology Research

http://www.esjournals.org

A New Model for Recommender Systems based on Data Sources Integration

Ali Hashemi, Mohammad Nadimi, Mohammad Naderi
Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University

ABSTRACT

The goal of Web recommender system is the process of selecting web pages shown to user based on his navigation patterns and interests. In this paper, a new model for recommender system is proposed to increase the accuracy of recommendations. In this model, some effective data sources are integrated to know the user interestingness. The sources used the proposed model are user spent times on pages, the count of each page views per session, user's location and data referred extracted from search engines. This data sources, combined through proposed model and then clustering operation is performed on it and recommendations are presented to the user through classification operation. In this paper some algorithms are proposed to extract user's interest from each of data sources. The approach is implemented as an experimental system, and its accuracy is evaluated based on F1 criterion.

Keywords: *Recommender system, Web personalization, Web usage mining, Data sources integration, Users' location, Search engines.*

1. INTRODUCTION

Nowadays, interactional websites having voluminous data and also large number of viewers are found enormously all over the webs. Through some approaches, these websites are trying to meeting users satisfaction and demands, getting much more income and saving their costs. In order to meeting user's satisfaction, they attempt to discover users attributes somehow, helping them while finding their interested data [1]. To discover these attributes faces to some challenges. One of these challenges is increasing daily trend of sites data enters volume and contents which entails to getting more accurately users attributes. User attributes discovery is done through some data provided by whom directly and indirectly to site. The data acquired from users here is called data sources. In present paper a model has been developed by which several present and user available data sources can be integrated and through them recommender system accuracy might be accrued. The sources which we are going to integrate them in ingoing paper are included user spent times on pages, the count of each page views per session, user's location which aiding us to get some users attributes and characteristics and some data might be obtained from referrers. Firstly in section 2 we explain motivation and the subject issue with some examples, in section 3, the offered recommender system is illustrated separately in terms of performance. An experiment to compare and assess accuracy of classic and modern systems outputs is presented in section 4 and eventually conclusion remarks are expressed.

2. MOTIVATION

Web usage mining refers to the automatic discovery and analysis of patterns in clickstreams, user transactions and other associated data collected or generated as a result of user interactions with Web resources on one or more Web sites [2]. in other word, web usage mining is a branch of web mining that mainly focusing on secondary data obtained from users interaction to web pages instead of that on web main data(pages structures and contents) [3], so that it applies data mining technique on such data to discover interesting usage patterns. Namely it serves as user data mining process, user searches and access patterns to one or more websites. These access patterns usually are saved within web server log. Three general phases in web usage mining application are include Data preparation and then its conversion into a satisfactory format, pattern discovery from previous phase data and offering a recommend to user [4]. Transaction registration time, IP and user browser agent are the main data stored in web log by which users session could be recognized [5]. Each session is consisted of several transactions; each of them represents a user access to a set of Webpages. Through IP address, one can find user's location and having finding location, user attributes can be predicted. One of the most vital data stored in web log is thought of to be referrer field. Each transaction referrer field represents a URL, by which user has been entered to our webpage. It might be previous user accession to our websites pages, searching engines or others sites. In respect to referrer field, one can find that what subject user follows and why user has been entered to our site. Hence, based on its data, firstly, we can offer more accurate recommends to user and secondly we can have recommends for a new user who did not browse our web site. We need to get more accurate user attributes recognition to give much more precise recommends. Amount of time spent on each page by user for each session might be varying. Also, it might be some pages reviewed up to one time, presumably both factors can be more effective and promising to get through user interest level to each page [6]. These sources might have better outcomes for various websites and functions. Our hypothesis for purposed system is that integrating these data sources can strengthen recommending accuracy. The crucial point on achieve this goal, is ways and manner of integrating these data so that it will extract and emerge user real characteristics.

There have been conducted outstanding researches on this context. Some just considered click stream regardless to the spent time on pages as a criterion [7-9], others just considered time regardless to pages repetition in a session [10]. In some cases, the time spent on pages and page repetition has integrated and led to increased recommends accuracy [11].

Among offered recommenders system, there have been little attention on user's location as a parameters to get through

user's attributes. Interestingly considering this parameter can prevent giving incorrect recommend to a user. Assume that, for instance, we have a bilingual website in which each language involves some pages. The first language pages, account for more number of sites pages having prefix F and the second language pages have prefix E. the main page and more contributions of site data have been managed and constructed in respect to the first languages contents. Now a user located on a location related to the second language turn to site and firstly reviews pages F1 and F2 respectively to get information about it and be able to find pages on his language, then looks up page E1. Meanwhile our site recommending system has user browser field E1, F1 and F2, while just page E1 must be considered for recommending user interested pages related to the user's language.

Considering that where newcomer user is from and what for he or she has viewed our site, also can be very effective on offering recommends related to the user's interest. One of the main sites referrer is that through search engines. Search engines act based on keywords, it means that first user enters some keywords as input to search engine and search engine site in turn recommends some sites pages to user based on its index pages. By referring user to each site, related site can find user searched keywords through referrer filed. Now assume that site user in previous case, searches keywords K6 related to page F6. K6 Keyword also exists in page F1 but page F1 has not specific relation to keyword K6 conceptually and just this keyword has been existed in its content. Since most of recent search engines have not some tools to acquire concepts from sites, just their content is considered and page F1 for keyword K6 is purposed to user. Now user sees page F1 and our site recommending system having page F1, offers recommends regardless to keyword K6, while page F6 has been the correct recommend for user.

3. PURPOSED RECOMMENDER SYSTEM

Proposed recommender system is designed based on partitioning methods exploits algorithm k-mean to cluster user session. Data sources to which we are going to use in present paper are included user spent times on pages, the count of each page views per session, user's location which aiding us to get some users attributes and characteristics and some data might be obtained from referrers. Ways of integrating these sources and their application outstands and increases recommender system accuracy. This system is consisted of two main components of offline and online.

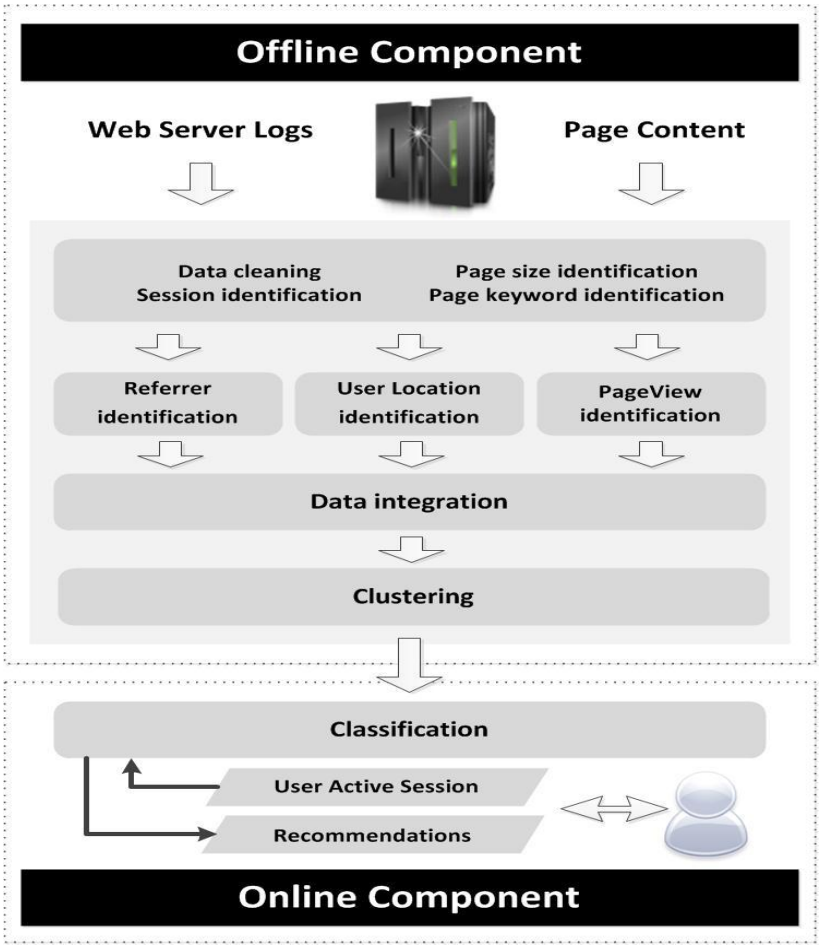

Figure 1 :the schematic architecture of purposed system

Figure 1 shows this system architecture. Out of three general phases of web usage mining functions previously discussed in section 2, the first two phases included Data preparation and then its conversion into a satisfactory format and useful pattern discovery are exist in offline component and the final phase which is giving recommend to user are conducted in online component. In order to obtain and discover patterns, recommender systems require train set by which, former user's patterns can be extracted and new user's demands might be predicted. Train set in this system involves a series of web server log raw set.

3.1. Data cleaning, Users identification and Session identification

Web server log raw set are entered into this part as an input and some trivial data which confuses us in getting users attributes are eliminated. Some of such data include images, scripts and requests created by web robots [12]. In present paper user IP address, browser type and users operation systems are used to users identifications. A session is defined as a group of a user activities entered to site in a given time interval until it get out of there. Therefore, session identification is discovery process and user accesses classification for each independent session [5]. There have developed two different approaches to user session identification [13]: session duration based and page stay time based. Here we use the page stay time based method. In this method if interval between two accesses of a user is less than a threshold, the both accesses will fall into a single session. Interval threshold of 10 minutes has been reported in different literatures [13, 14].

3.2. Identification of each page keywords and size

Page user accessibility sets are considered to discover user's attributes and interests. Such pages in turn involve some keywords illustrate page key content and concepts. Let now in order to get a user's attribute, we consider user interested keywords instead of a user accessibility set. Result is that an overlap occurs among some page common keywords, so user intends for browsing and reviewing our pages will be elucidated. To access page keywords is possible through some manners which some of them are automated like N-grams [15] and others are manually through knowledge domain and classifications present in site. The time interval spent by user on each page serves as one of the data sources considered to identify user's attributes as a parameter. Abovementioned interval relays on directly on each page contents. So we need to know each page size that it is carried out in page size identification section.

3.3. User's interest identification based on page view

Pattern discovery phase entails data with appropriate format. For algorithm k-mean, such data usually come into matrix form [11]. A matrix has been used in [16] where its rows and columns represent sessions and sites pages respectively. These matrix elements might be in two forms. They may take values 0 or 1(i.e. whether user has saw interested page or not) or it take values range between 0 and 1(i.e. user interest rate to that page). A weighting criterion has been applied in [11] estimating user attribute in respect to two parameters of pages view count and time spent on the pages. Indeed, the main purpose of doing so is to get a solution to integrate data source together [6, 17]. In order to calculate index of pages repetition in user session, following equation is used:

$$Frequency(Page) = \frac{NOV(Page)}{\sum_{Page \in VisitedPages}(NOV(Page))}, \quad (1)$$

Where, NOV function is count of visits of a specific page, indeed, for individual pages, per page repetition is divided by sum of all pages repetition. To calculate factor of user spent time on each page, following formula has been used:

(2)

Where TD function is Total Duration and L function is Length of content, indeed, total spent time on each page is calculated in respect to that pages content, then it is divided by maximum value in all pages. Using these two equation user interest level can be calculated from equation (3):

1 (3)

By having user interest to each page, session-pages matrix might be filled in. on the other hand there are matrix pages-keywords previously obtained in section 3.2. By multiplying these two matrixes, the session-keywords matrix is obtained which can be entered to data source integrating phase as an input.

3.4. User's interest identification based on location

User's location may be effective to user attributes identification and subsequently obviates probable user's mistakes. One of the most popular cases of using location occurs within multilingual websites. The websites that extract viewer user IP address as online acquire their user's language in terms of IP address range. As a result, they show users some part of sites in related to their range predominant language. Here, we are going to create a session-keywords matrix indicating each user's interest to one or more keyword based on user's location. Aforementioned matrix development takes places in three phases:

The first phase: firstly all users' session IP address is elicited from data in part of user session identification. These IP addresses are classified in some classes based on their first two octets, presumably, users IP scope list in train set is obtained. By inferring to these data the session-IP scope matrix is created. Given that session user location in IP scope, associated element is valued by 1. In resultant matrix, each row involves one and only one element is observed as 1 and others take zero.

The second phases: a matrix called IP scope-keywords is created in this phase in which the rows are the same IP address scope list in previous phase and its columns represent keywords came from pages keywords identification part. It explicitly indicates that to which keywords the users in each scope are the highest interest. There are different ways to complete elements of this matrix. It might be fill in respect to domain knowledge. It means that, the managers can associate each keyword to one or more IP scope in respect to their attitude into their sites keywords and content. These tools also can be used to advertisement targeting. The other way is to fill matrix automatically such that the calculated value is incorporated in element in which rows and columns represent session and keywords respectively in respect to each keywords repetition on viewed pages per session. Then the matrix is normalized and considering a given threshold, the numbers less and more than threshold are considered 0 and 1 respectively.

The third phase: in this phase, the obtained matrixes in two previous phase are multiplied together to obtain session-keywords matrix.

3.5. Identification of user's interest based on referrer field

The page address through which user has been entered to our site is thought of another parameter by which once can find user attribute and demands. The search engines are of the main tools guide users forwards to sites. As user follows search engines for a given purpose, the search engine give a series page to user by entering one or more keywords. For instance, as keyword Isfahan, is entered, the searching engine address will be as following:

https://www.google.com/search?q=isfahan&ie=utf-8&oe=utf-8&aq=t&rls=org.mozilla:en-US:official&client=firefox-a

And display some pages contains keyword Isfahan. By a click on each of these links, search engine address is sent to destination site as a referrer field. Through processing this field, the destination sites extracts interested keywords and recognize users' demands. In this section, we are going to develop session-keywords matrix representing each session requires keyword extracted through referrer field. A comparison is done between session's referrer field and all keywords extracted from pages keywords identification part. In case, referrer field include each kind of these keywords, the element related to such referrer field session and present keyword will amounted to 1, otherwise it considered as zero. The phases for this matrix development are shown in Algorithm 1.

Algorithm 1 the manner of matrix [session-keywords] creation from referrer field data

Input: Sessions List, Keywords List
Output: SessionsKeywords Matrix

Fill SessionsKeywords rows with Sessions
Fill SessionsKeywords columns with Keywords
For all ($S_i \epsilon$ Sessions) **do**
  If (S_i have referrer field) **then**
    For all (keyword $K_j \epsilon$ Keywords) **do**
      If (S_i [referrer] include K_i) **then**
        SessionsKeywords [$S_{i,}$ K_j] = 1
      Else
        SessionsKeywords [$S_{i,}$ K_j] = 0
      End if
    End for
  End if
End for

3.6. Data Sources Integration

Data related to three sections (identification based on page view, identification based on user physical location and identification based on referrer field) are entered into part of data integration as an input. The inputs and outputs for three sections are session-keywords matrix with similar data structure. The output matrix from section of identification based on page view is called Pageview matrix. The elements of the matrix are filled with values between 0 and 1. The output matrix from referrer identification and user location identification are respectively called location and referrer matrix that their elements filled by zero or 1.

Every user who view our site surly has an IP address but the user may have referrer field or not. For this reason we carry out classification and clustering operations separately on two matrixes, namely one for user who have not referrer field, hence data integration must be landed on both pageview and

location matrixes, another is for users who came into our sites through search engines, and in this case data integration must be conducted on all of three matrixes. In order to integrate data for both pageview and location matrixes which are fuzzy and binary respectively, we apply Element-wise scalar multiplication operation (.*). These operations reduce noises and user's mistakes present in pageview matrix. The result is called user without referrer matrix.

(4)

Where, SK_{UWOR} is session-keywords matrix for user without referrer field, SK_p is pageview matrix and SK_l is location matrix. To integrate data for all of three matrixes, firstly user without referrer matrix is multiplied by Element-wise scalar multiplication matrix. Then the final matrix is pluses into referrer matrix (equation (5)). The main reason for this is the great deal importance of user referrer for his attribute identification. Also applying keyword is associated to referrer for users who have not seen the pages related to referrer keywords. For supplement this explanation, assume that a user has entered keyword K1 in searching engine and guided into a page in our site irrelevant to this keyword. In case we just use element-wise scalar multiplication, the element related to columns keyword K1 for that user session in matrix is considered as 0, whereas user interest and attribute have been completely related to keyword K1.

(5)

Where, SK_{UWR} is session-keywords matrix for user with referrer field, SK_p is pageview matrix, SK_l is location matrix and SK_r is referrer matrix.

3.7. Pattern discovery

According to figure 1, session-keywords matrixes which entered from previous phase as an input to present phase, involve elements ranges 0 and 1, implying users interest rate to each keyword. These matrixes are clustered using algorithm k-mean so that user with similar interest can fall into same groups. To promote clustering quality, session vectors of these matrixes are normalized individually. Normalization operations are conducted in following manner: each element per row is divided by maximum value element per row. It is obvious that for each row, at least one element is created with value 1 and there will not any element much more than this. Partitioning algorithm like k-mean requires a series of initial data as an input to conduct clustering operations. The number of clusters is served as one of these inputs. This number usually is selected manually or given to domain knowledge. This trend has been automate for system purposed in [11], since it represents the number of site accessibility patterns, illuminated that how many aggregate usage profiles must be developed. The optimized number means that what the best session clustering number is given to their distribution [18].

Following doing clustering process by algorithm k-mean, output must be ended up as the aggregate usage profiles. Aggregate usage profiles is operations of users patterns discovery based on clusters, allowing recommenders operations to be much more efficient [8]. In [16], Mobasher explains this approach:

l (6)

Where pr_c is usage profile for cluster C. $weight(k, pr_c)$ is weighted mean value of keyword k in cluster C that calculated through following equation:

(7)

In above equation, t represents one of the session transactions existed in cluster C and $w(k,t)$ is weighted mean value in transaction t.

3.8. Prediction System

The input of this section is two clustered matrixes that are from pattern discovery section and also current user session vector. The output also is list of recommends must be supplied to user. Usage profile matrix is constructed for individually clustered matrixes.

That whether user has entered through search engines to our site, ways of finding recommends will be different. Following phases show the manner of this trend. In case, user has forwarded to our site via search engine, the fourth phase, otherwise the third one is conducted regardless the fourth phase.

The first phase: first, the current session-IP scope matrix is constructed. This matrix just has one row and its column is the same of session-IP scope matrix columns in previous section 3.4. It determines that which IP scope user belong to.

The second phase: the resultant matrix is multiplied by IP scope-keywords matrix from section 3.4, give current session-keywords matrix. Since this matrix has one raw it is called location vector.

The third phase: current session and location vectors multiplied by together in scalar element manner to form final current session. This process is the same as that was conducted in offline phase (equation 4).

The fourth phase: if user has been entered our sites through search engine, this phase would occurred, otherwise would not. The current session-keywords matrix as referrer one is constructed in this phase given to approach that stated in section 3.5. Since it has just one raw that it may be called referrer vector. Finally using equation 5, final current session is developed.

The fifth phase: considering that if user has entered to our site via search engine or not, the usage profile matrix is selected proportional to it. Classification algorithm KNN [16] is used to finding the closest usage profile to current session and giving recommends to user. This algorithm calculates the same score for each elite usage profile keywords:

(8)

Where S is user's current session and C is usage profile. S_k represent user interest rate to keyword k and w_k^C is weighted value in usage profile C for keyword k. this equation is calculated for all usage profiles and the more much value is obtained, the more similarity of usage profile to user current session. Now, similarity index for each keyword to current user's keywords is calculated by selecting similar usage profile:

(9)

Recommending weight for keyword k to current session S is obtained in light of the above equation. The more this weighted value, the more prioritization of that keyword to offer user will be. The set of UREC(S) is considered to obtain final keywords.

(10)

Where, a minimum weight is incorporated for elite keywords (ρ) enhancing recommender system accuracy and precision.

4. EVALUATION

In order to evaluate recommender system, the accesses set of Isfahan virtual tourism website [19] was considered for five months. It includes 500 sessions whose 80% records for train set and rest is considered in test set. As it was previously stated in section 3, for train set, clustering operations are conducted on two matrixes.

Test set is applied to evaluate prediction system accuracy [10]. For each session of test sets, first two transactions are entered to prediction system from T transaction as an input. This set is called W. the output for prediction system is called R set. Using it, recommender system efficiency is calculated for each method. Individually methods efficiency is measured using three criteria precision, coverage and F1. Benchmark for R precision in transaction T is as following:

(11)

Coverage R in transaction T is as follow:

(12)

Where T-W is set of page views which recommender system must provide. R ∩ (t-w) represents set of purposed correct recommends. As for recommender system, precision criteria the degree to which the recommendation engine produces accurate recommendations (i.e., the proportion of relevant recommendations to the total number of recommendations). On the other hand, coverage criteria the ability of the recommendation engine to produce all of the pageviews that are likely to be visited by the user (proportion of relevant recommendations to all pageviews that should be recommended). None of these criteria cannot represent accuracy of recommender system function which most often affect together. Idealistically, we are going to increase both coverage and precision. To do so, a united criterion known as F1 is used [20]:

(13)

According to Formula (13), F1 will be maximized when the benchmark measure precision and coverage are the maximum, too. Given to cases mentioned on two precision and coverage interactions, F1 presumably may be appropriate criterion to calculation of recommender system accuracy and efficiency inter alia. F1 is measured on elite dataset based on two different approaches, one from previous, and another from proposed recommending system. The system we are going to compare serves as one of the most outstanding operations on recommenders system filed, integrating user time spent on pages and the count of user pageview in a session to find out attributes and characteristics [11].

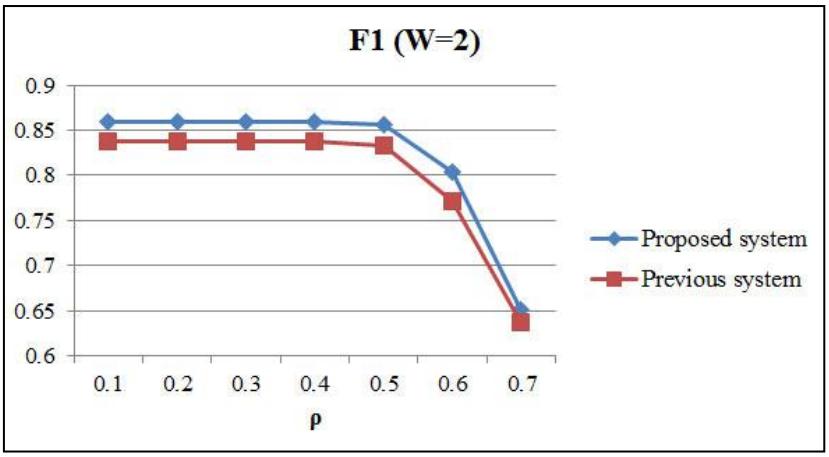

Figure 2 Result for experiment of purposed system efficiency using Criterion F1

Figure 2 shows results of abovementioned experiment. This experiment was landed using a hardware having processor intel-i5-3.30GHz, RAM about 4 Gig and software has run by programming language C#. Different values are considered as the minimum recommender score and F1 is calculated for previous and new purposed system. As it is seen on figure, the new recommender system indicates higher efficiency and accuracy rate.

5. DISCUSSION

In this paper, a new model for recommender system is proposed to increase the accuracy of recommendations by integrating some effective data sources. With respect to the increasing content of websites and the intense need to rediscover the real interests of the users, the lack of accuracy of recommends has become a serious challenge to recommendation systems.

Previously, the data sources of page view count and the time spent on pages have been integrated. Here, two data sources of the user's location and the data extracted from referrer filed have been considered. Useful information can be extracted from user's location. One of them is user's dominant language. In proposed model, we have been trying to eliminate user's mistakes and wrong behaviors in some pages that users visits. For example, it allowing much more accurate clustering operations and subsequently, due to more exact user's interests recognition. On the other hands, when a user looks up our site from the other sites, inadvertently has some potential data helping us to recognize its real interests. In proposed model, the keywords used by users in search engines to lead to our site's pages, were extracted from their referrer field and taken as criterion for their interest recognition. To evaluate efficiency of new model, an experiment was designed and compared with one on the previous classic systems. The new recommender system based on new model indicates higher efficiency and accuracy rate.

REFERENCES

[1] Vishal, S., B. Rajesh, and V. Bhupendra, *A Framework for Improving Target Marketing Using Collaborative Data Mining Approach.* International Journal of Information and Communication Technology Research, 2011. **1**(2).

[2] Srivastava, T., P. Desikan, and V. Kumar, *Web mining–concepts, applications and research directions.* Foundations and Advances in Data Mining, 2005: p. 275-307.

[3] Hashemi, A. and M. Nadimi, *Recommender Systems Based on Web Usage Mining: A Survey.* International Review on Computers and Software (IRECOS), 2012. **7**(6).

[4] Mobasher, B., *Data mining for web personalization.* The Adaptive Web, 2007: p. 90-135.

[5] Bamshad, M., R. Cooley, and J. Srivastava, *Data preparation for mining world wide web browsing patterns.* Journal of knowledge and Information Systems, 1999. **1**(1): p. 5-32.

[6] Dumais, S., et al. *SIGIR 2003 workshop report: implicit measures of user interests and preferences*. 2003.

[7] Yan, T.W., et al., *From user access patterns to dynamic hypertext linking.* Computer Networks and ISDN Systems, 1996. **28**(7-11): p. 1007-1014.

[8] Mobasher, B., R. Cooley, and J. Srivastava, *Automatic personalization based on Web usage mining.* Communications of the ACM, 2000. **43**(8): p. 142-151.

[9] Baraglia, R. and F. Silvestri. *An online recommender system for large web sites*. 2004. IEEE.

[10] Moh, T.S. and N.S. Saxena, *Personalizing Web Recommendations Using Web Usage Mining and Web Semantics with Time Attribute.* Information Systems, Technology and Management, 2010: p. 244-254.

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

[12] Tan, P.N. and V. Kumar. *Modeling of web robot navigational patterns*. 2000.

[13] Spiliopoulou, M., et al., *A framework for the evaluation of session reconstruction heuristics in web-usage analysis.* INFORMS Journal on Computing, 2003. **15**(2): p. 171-190.

[14] Berendt, B., et al. *Measuring the accuracy of sessionizers for web usage analysis*. 2001.

[15] Cavnar, W., *Using an n-gram-based document representation with a vector processing retrieval model.* NIST SPECIAL PUBLICATION SP, 1995: p. 269-269.

[16] Mobasher, B., et al., *Discovery and evaluation of aggregate usage profiles for web personalization.* Data Mining and Knowledge Discovery, 2002. **6**(1): p. 61-82.

[17] Chan, P.K. *A non-invasive learning approach to building web user profiles*. 1999.

[18] He, J., et al., *On quantitative evaluation of clustering systems.* Clustering and Information Retrieval, 2003. **11**: p. 105-133.

[19] *Isfahan Virtual Tourism*. Available from: http://tourism.isfahancht.ir

[20] Lewis, D.D. and W.A. Gale. *A sequential algorithm for training text classifiers*. 1994. Springer-Verlag New York, Inc.