PERSONALIZATION WITH WEB MINING

[1] Ms. Madhuri Anwat A., [2] Prof. Umesh Pawar.

[1]PG Student, [2] Assistant Professor Department of computer Engineering
S.N.D.COE and RC, Yeola

Abstract: In today's competitive environment learners from any area are growing and they are thirsty for getting knowledge in all looks and corners. Day by day their self learning habits are increasing and they are becoming explorers of the knowledge. The subjective initiative of learners is strengthening and they want a dynamic mechanism to fulfill their needs as quick as possible. Most of them uses internet as a primary source of information but this source is very big and vast, merely it provides more accurate results as per learners interest and expectations. Personalized learning system is developed in this dissertation to provide the learning services to the learners to fulfill their needs, interests and habits through the use of web services and web mining technology. The developed system provides personalized learning environment and it is able to find out the individual differences, individual characteristics, and habits to provide results according to users interests and habits. Web mining technology is used for implementation of system aim to identify the relevant results. Basically, personalized learning system consists of five major steps to construct a personalized learning environment are: Data collection, Data preprocessing, Data analysis, Result determination, Personalized interface. All these modules use web mining techniques to achieve system goals. Finally, a typical system is constructed using .net framework to get satisfactory results.

***Index Terms*— Learning System, Personalization, Recommendations, Dynamic Interest Links, and Clustering, etc.**

1. INTRODUCTION

In today's competitive environment user's subjective initiative, learning eagerness and thirst of knowledge is increased. Their subjective Initiative is strengthening. Personalized Learning System is a true attempt to simplify these efforts.

1.1 Personalization

Personalization means persons would get the things or results according to their interests and expectations without giving much more input. Personalization systems are a subclass of information filtering system that seek to predict the 'ratings' or 'preferences' that a user would give to an items, they had not yet considered, using a model built from the characteristics of an item (content-based approaches or collaborative filtering approaches). Personalization systems analyzes the individual characteristics and habits without expecting much more input from user and constructs an automated responses to fulfill individual needs. Personalization systems have become very common in recent years. These systems are more flexible, reliable, and dynamic to provide personalized results[11].

1.2 Existing Systems

Modern learning technologies have become very popular because of increased competitions and thirst of knowledge. The important characteristic of modern learning system is the subjective initiative of learners is strengthening. They become self learners and explorers of the knowledge. Hence, to provide flexible learning environment various computing systems are developed and these systems are built over web technologies. Online learning systems provide various services beyond schools and colleges at any time. At present the online learning systems are categorized into three classes[5]: Customized systems, Usage based systems, and Personalized systems. Here we need to understand the basic difference between above classes of learning systems. (a) Customized systems are based on layout customization. In layout customization, according to users interest and preferences web pages are adjusted manually or semi-automatically. (b) Usage based systems relies on the application of statistical and data mining methods to the web log data resulting in a set of useful patterns that indicate user's navigational behavior[14]. (c) Personalized systems combine the techniques of usage mining, content mining, and structure mining to provide more personalized results. Personalized learning systems may have characteristics such as, (i) It is able to find users interests, preferences, choices, and learning orientations from vast and huge amount of data, (ii) On the basis of log records these systems can dynamically adjust the user interfaces, (iii) On the basis of analysis of user interests and preferences it can provides dynamic recommendations and links, (iv) It can provide study materials and course wares according to results, (v) It helps to all power users such as administrators to adjust teaching plans, policies, rights, and teaching aids and methods specifically on the basis of user's interest[7].

1.3 Web Mining Taxonomy for Personalization

Web mining technology is emerging field of data mining for WWW based information and resources. The basic focus of web mining is to use data mining techniques and algorithms to extract useful and hidden patterns from unstructured and huge web data or resources. Web mining taxonomy is divided into three categories according to sources of web data[10]. These categories are Web content mining, Web usage mining, and Web structure mining shown in Figure 1.1. As you can see from this Figure;

1. Web content mining means the extraction of useful information and web knowledge from web sources or web contents such as text, Image, audio, video, and structured records[1].
2. Web usage mining is the application of data mining techniques to find out interesting patterns from web usage data. It mainly tries to extract useful and interesting patterns from usage data such as server logs, client browser logs, proxy server logs, cookies, user sessions, registration data, mouse clicks, user queries, bookmarks etc. and any other data as the results of user interactions[14].
3. Web structure mining tries to identify the structure of hyperlink in html documents and deduce knowledge [8].

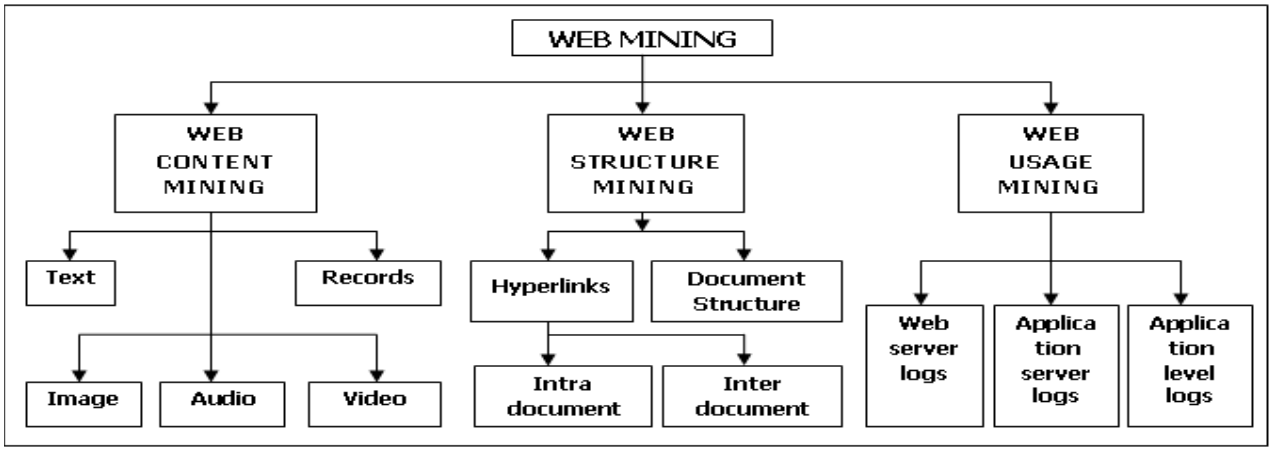

Figure 1.1: Web Mining Taxonomy

Paper is organized as follows. Section 2 describes about the related work done earlier for the system to be developed. Section 3 presents methodology and algorithms used for system development. Section 4 presents experimental results for customized system and personalized system. Finally, Section 5 presents conclusion.

2. RELATED WORK

Now a day, Web personalization systems are becoming very popular because of its ability to deliver contents according to user's preferences and needs. Hence, Number of researchers has done much on personalization. This chapter covers some interesting research efforts of various researchers related to personalization.

2.1 Web Personalization Approaches

The requirement for predicting user needs in order to improve the usability and user retention of a web site can be addressed by personalizing it. Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users navigational behavior and individual interests, in combination with the content and the structure of the web site. The objective of a web personalization system is to provide users with the information they want or need, without expecting from them to ask for it explicitly [7].

2.1.1 Dissimilarity Matrix for user session files

Nasraoui and Krishnapuram et al. [2000] discovered the user session files and formulated groups on the basis of similar characteristics using fuzzy algorithms[12]. According to their research a user or a page can have more than one cluster. In their proposed approach, after preprocessing of usage data dissimilarity matrix of preprocessed data is created. This is used by fuzzy algorithms in order to cluster typical user session.

2.1.2. WebPersonalizer: a Framework for Mining Web Logs

Mobasher et. al. [2000] proposed most advanced system, "WebPersonalizer"[3]. It is a powerful framework for mining web log files to extract the useful information for the purpose of recommendations based on the browsing similarities of current user to previous user. After collecting and cleaning of usage data (creating various abstractions of collected data), data mining techniques such as association rule mining, sequential pattern discovery, clustering, and classification are applied in order to discover interesting usage patterns.

2.1.3 Interval based coarsening for usage mining

The most important contribution of Berendt [2001] in the area of web usage mining is STRATDYN (Strategic and Dynamic)[2] add-on module. It determines the differences between navigational patterns of user and then it exploits the site semantics in the visualization of the results. In this approach, web pages are grouped together on the basis of concept hierarchies. He focused on "interval based coarsening" technique for usage data at different levels of abstraction.

2.2 Personalization using Data Mining Methods

Data mining is an extension of traditional data analysis and statistical approaches in that it incorporates analytical techniques drawn from a range of disciplines including, numerical analysis, pattern matching and areas of artificial intelligence such as machine learning, neural networks and genetic algorithms [13]. It provides various methods such as clustering, classification, association rule mining, sequence discovery, and pattern matching for extracting useful hidden knowledge from vast and big database. These algorithms can be used for web site personalization.

2.2.1 Data mining methods for finding navigational behavior

Magdalini Eirinaki et. al. [2003] focused on web usage mining. This process relies on the application of statistical and data mining methods to the web log data, resulting in a set of useful patterns that indicate user's navigational behavior [7].

2.3 Personalization of Education Systems

The users from academic institutes are growing day by day. They are becoming a self learners and explorers of the knowledge. Subjective awareness and initiatives of these user's are increased [5]. So, to fulfill the knowledge acquiring eagerness of these user's some researchers have focused on learning systems personalization using web mining strategies.

2.3.1 Existing Web Mining Tools for Learning Personalization

Yuewu Dong et. al.[2010], presented a simplified architecture of distance education system based on web usage mining and content mining to realize personalization[5]. They have constructed a basic user interface based on database of BUPT-SK and external data mining tools to realize personalized environment in distance education system.

3. METHODLOGY & ALGORITHMS

To design a general interactive online framework to find individual preferences, habits, behavior, and access patterns. Generate and provide personalized outcomes using web mining techniques to meet individual needs without expecting much more input from user.

3.1 SYSTEM ARCHITECTURE

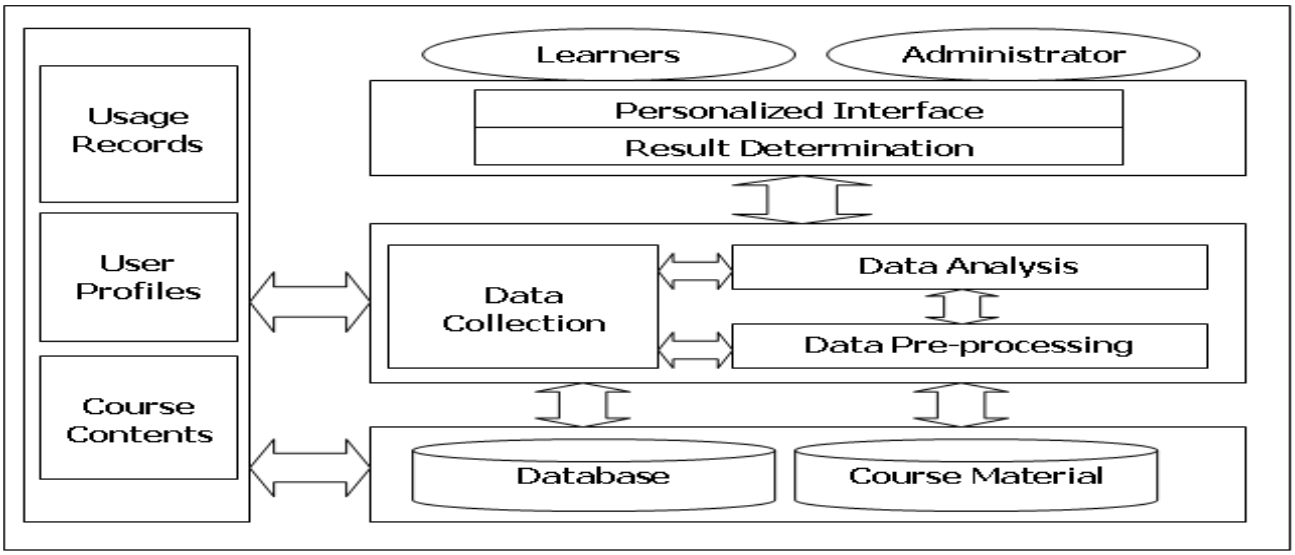

Figure 3.1: Block diagram of Personalized Learning System

3.2 Mathematical Formulae

The different mathematical formulae used in different algorithms are shown here:

1. The Cluster mean of $k_i = t_{i1}, t_{i2}, t_{i3}, \ldots\ldots, t_{im}$ is defined as:

$$m_i = \frac{1}{m}\sum_j^m t_{ij} \quad [1]$$

2. Cosine similarity:

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}} \quad [2]$$

3. Accuracy:

$$AC = \frac{a + d}{a + b + c + d} \quad [3]$$

4. Precision:

$$P = \frac{d}{b + d} \quad [4]$$

5. Recall:

$$R = \frac{d}{c + d} \quad [5]$$

6. Fall-Out:

$$FO = \frac{a}{a + b} \quad [6]$$

7. F-measure:

$$F = \frac{2 \cdot (P \cdot R)}{(P + R)} \quad [7]$$

8. Overall System Performance:

$$point_i = \frac{|relevantItems_i|}{|evaluationItems_i|} \times \frac{|visitedBefore_i|}{|domains_i|} \quad [8]$$

3.3 Mathematical Modeling

A mathematical model is a description of a system using mathematical concepts and language. The process of developing a mathematical model is termed mathematical modeling. A mathematical model here is used to explain system, to study the different effects of different components, and to make the prediction about behavior.

Set Theory: Let S be a set of input parameters to mine so that it will generate set of personalized results. Such that;

S = {I, F, O} where;

I represent the set of inputs;

I = {I1, I2, I3}

I1: Set of usage records.

I2: Set of user profile information.

I3: Set of course materials.

And F is a set of functions;

F = {F1, F2, F3, F4, F5, F6, F7, F8}

F1: Login to the system

F2: Collection of data from web sources.
F3: Data preprocessing and session identification.
F4: Apply clustering algorithms to form initial clusters.
F5: Apply clustering algorithms and usage mining to form more relevant clusters.
F6: Discover and rank pages relevant for a particular topic.
F7: Determine useful sequences for output generation.
F8: Generate and provide personalized output.

Finally, O is a set of outputs;

O = {O1, O2, O3, O4, O5}
O1: Personalized Learning GUI.
O2: Clusters of Relevant Documents.
O3: Personalized Recommendation.
O4: Dynamic Interest Links.
O5: Optimized Course Structure.

Functions

F1: Login to the system.
If x = login.
F(x)= login successful.
If U[A-Z a-z] and P [A-Z a-z 0-9] and Length (P)
Login successful
Else
Login fails
Where, P: Password and U: Username

F2: Data collection.
X: collection of information.
F(X) = usage records, profile information, or contents.

F3: Data preprocessing and session identification.
X: collected information such as usage records or contents.
F(X) = provides meaningful data for particular user session.

F4: Apply clustering algorithms to form initial clusters.
X: User profile information and contents.
F(X) = provides clusters of relevant documents.

F5: Apply clustering algorithms and usage mining to form more relevant clusters.
X: set of initial clusters and usage records.
F(X) = provides more appropriate and relevant clusters.

F6: Apply HITS to discover and rank pages.
X: set of different relevant clusters.
F(X) = provides ranked pages relevant to a particular topic.

F7: Determine useful sequences.
X: set of relevant documents with page ranking.
F(X) = determines best possible sequences.

F8: Generate and provide personalized results.
X: best possible sequences.
F(X) = generate personalized GUI, interest links, recommendations.

3.4 Algorithms

K-means Algorithm: K-means is an iterative clustering algorithm[6] in which items are moved among sets of clusters until the desired set is reached. The cluster mean of $k_i = t_{i1}, t_{i2}, t_{i3}, \ldots\ldots, t_{im}$ is defined as:

$$m_i = \frac{1}{m}\sum_j^m t_{ij}$$

Input: D = $t_1, t_2, t_3, \ldots\ldots, t_n$ /* Set of elements */
k /* Number of desired clusters */
Output: K // Set of clusters
algoKMeans:

1. Assign initial values for means $m_1, m_2, m_3, \ldots\ldots, m_k$
2. Repeat
   - Assign each item t_i to the cluster which has closest mean
   - Calculate new mean for each cluster
   - Go to step 2, until convergence criteria is met
3. Stop

LINGO Algorithm: The general idea behind LINGO is to first find meaningful descriptions of clusters, and then, based on the descriptions, determine their content. To assign documents to the already labeled groups LINGO could use the Latent Semantic Indexing in the setting for which it was originally designed for given a query, retrieves the best matching documents.

Input: User search query (q), term (t), word (w)
k, teaming words St_{sw}, set of stopwords

$S_{sw} = \{'i','a','about','an','and','as',\ldots..\}$
Output: K // Set of clusters
algoLingo:

1. Preprocessing
   a. Remove stop words
   if (q contains S_{sw}) then remove words
   b. if (q contains St_{sw}) then rewrite word by removing steaming
   c. if (q contains phrase) then extract that phrase
2. Find clusters and label using vector space model along with the Latent Semantic indexing (LSI) technique.
   a. Covert input query to tf - idf
   $tf_{t,d}$ = number of occurrence of term in document d
   $idf_t = \log_{10} \frac{N}{dft}$
   where dft is document d that contain a term t
   $tf - idf_{t,d} = tf_{t,d} \times idf_t$
   b. Create data structure indexed by document (d)
   c. For each t
   Update entry in score
   $score[d] = score[d] + tf = idf_{t,d} \times tf - idf_{t,d}$
   Normalize score
   $magnitude[d] = magnitude[d] + tf - idf_{t,d}^2$
   End for
   d. For each d
   Calculate cosine similarity
   $\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\cdot\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$
   Where, q_i is the tf - idf weight of term i in the query.
   d_i is the tf - idf weight of term i in the d.
   End for
3. Stop

HITS Algorithm: Hyperlink Induced Topic Search (HITS) algorithm[6,8] made to use of the link structure of the web in order to discover and rank pages relevant for a particular topic. HITS is applied on a sub graph after a search is done on the complete graph.

1. Sampling step:
   $B(p) = \sum_{i=0}^{n} R(q)_i$
   Where, B(p):set of relevant pages & R(q):result of given query
2. Iterative step: For each result of query;
   $H_p = \sum_{q \in I(p)} A_q$
   $A_p = \sum_{q \in B(p)} H_q$
   Where, H_q:hub score of page & A_q:authority score of page,
   I(p):set of reference pages & B(p):set of referrer pages.
   Input: W //www viewed as a directed graph
   q //query
   s //support
   Output: A // Set of authority pages
   H //Set of hub pages
   algoHITS:
   1. $R = SE(W,q)$
   2. $B = R \cup$ {pages link to from R} $\cup$ {pages that link to pages in R}
   3. $G(B,L)$ = sub graph of W induced by B
   4. $G(B,L^1)$ = delete links in G within same site
   5. $X_p = \sum_{q\ where\ (q,p)\in L^1} Y_q$ //authority weights
   6. $Y_p = \sum_{q\ where\ (p,q)\in L^1} X_q$ //hub weights
   7. $A = \{p|p$ has one of the highest $X_p\}$
   8. $H = \{p|p$ has one of the highest $Y_p\}$
   9. Stop

4. RESULTS AND DISCUSSIONS

The Figure 4.1 take a closer look at the overall system performance applying the previously determined parameter configuration. A typical measure used so far is an additional curve, which incorporates the results of earlier experiments

regarding the system evaluation. The single points of this curve are calculated by multiplying the percentage of relevant documents and the percentage of domains viewed before. A more formal description of this calculation is given in the next equation.

4.1 Overall System Performance:

$$point_i = \frac{|relevantItems_i|}{|evaluationItems_i|} \times \frac{|visitedBefore_i|}{|domains_i|}$$

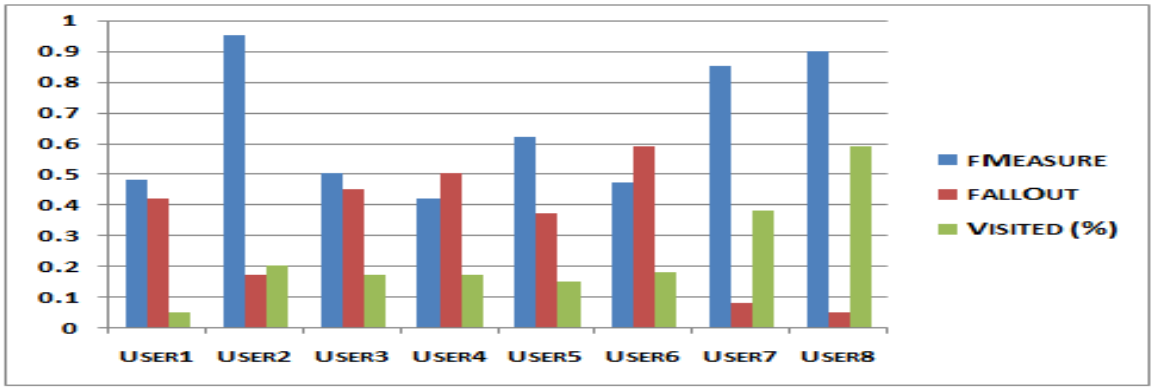

Figure 4.1: System performance for eight users

4.2 Personalized Learning System vs. Customized System

The comparative study between personalized learning system and customized learning system is shown here. Table 4.1 and Figure 4.2 describe the performance analysis of personalized system versus customized system.

Sr. No.	Attribute	Performance Measure	Personalized Learning	Customized Learning
			Points out of 10	
1	Reaction	Satisfaction of user requirements	7.5	5
2	Learning	Improvement in user knowledge	6.8	4.3
3	Behavior	How System Responds to user	8	5
4	Result	Benefits to improve users performance	8	5.3
5	Ratings	Preference of user to system	9	6.2

Table 4.1: Personalized Learning System vs. Customized System

Figure 4.2: Personalized Learning System vs. Customized System

5 CONCLUSION

In this paper, a simplified architecture of personalized learning system using web mining is implemented to provide simple, flexible, and modularized learning environment to users of learning system. For conceptual clarity this report focuses on process of personalization using mining algorithms. With the help of SRS, design and modeling concepts personalized learning system is implemented using .net technology. Based on the results of data collection, pretreatment, and analysis phases, a personalized learning environment is constructed to realize personalization. This environment has characteristics such as Personalized User Interface, Dynamic Interest Links, Personalized Recommendations, and Optimized Course Structure etc.

1. REFERENCES

1. Michael Azmy. Web content mining research: A survey. ACMSIGMOD Explorations, 01(01):203{212, November 2005.
2. Berendt B. Understanding web usage at different levels of abstraction: Coarsening and visualizing sequences. ACM SIGKDD Knowledge discovery & Data mining, 04(07):104{108, August 2001.
3. Mobasher B., Cooley R., and Srivastava J. Automatic personalization based on web usage mining. ACM Communication, 43(08):142{151, August 2000.
4. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 01(01):84{89, February 1999.
5. Yuewu Dong and Jiangtao Li. Personalized distance education system based on web mining. IEEE Education and Information Technology, 02(05):187{191, August 2010.
6. Margaret H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson, 01 edition, April 2006.
7. Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. ACM Transactions on Internet Technology, 03(01):1{27, February 2003.
8. Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 03 edition, May 2007.
9. Borges J. and Levene M. Data mining of user navigation patterns. Springer-Verlag, 1836(08):92{111, April 1999.
10. Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. SIGKDD Explorations, pages 95{104, July 2000.
11. Maurice D. Mulvenna, Sarabjot S Anand, and Alex G. Buchner. Personalization on the net using web mining. ACM Communication, 43(08):122{128, August 2000.
12. Nasraoui O., Frigui H., Krishnapuram R., and Joshi A. Extracting web user profiles using relational competitive fuzzy clustering. IJAI Knowledge Discovery, 09(04):8{14, April 2000.
13. Zaiane O. R., Xin M., and Han J. Discovering web access patterns and trends by applying olap and data mining technology on web logs. Advances in Digital Libraries (ADL), 02, April 1998.
14. Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pangning Tan. Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorations, 01(03):187{192, January 2000