WEB PERSONALIZATION USING CLUSTERING OF WEB USAGE DATA

Bhawesh Kumar Thakur [1], Syed Qamar Abbas[2] and Mohd. Rizwan Beg [3]

[1]Department of Computer Science & Engineering, B.I.E.T., U.P.T.U, Lucknow, India
[2]A.I.M.T.,U.P.T.U Lucknow, India
[3]Department of Computer Science & Engineering, Integral University, Lucknow, India

ABSTRACT

The exponential growth in the number and the complexity of information resources and services on the Web has made log data an indispensable resource to characterize the users for Web-based environment. It creates information of related web data in the form of hierarchy structure through approximation. This hierarchy structure can be used as the input for a variety of data mining tasks such as clustering, association rule mining, sequence mining etc.

In this paper, we present an approach for personalizing web user environment dynamically when he interacting with web by clustering of web usage data using concept hierarchy. The system is inferred from the web server's access logs by means of data and web usage mining techniques to extract the information about users. The extracted knowledge is used for the purpose of offering a personalized view of the services to users.

KEYWORDS

Web Mining, Web Usage Data, Clustering, Personalization.

1. INTRODUCTION

Data mining gives capabilities to analyzing large amount of data and extract information that are useful for statistical and business purposes as well. Now a day's our capability of generating and collecting data have been increasing exponentially. Contributing factors include, the computerization of many business, scientific, and government transactions, and advances in data collection, wide spread use of bar codes, satellite remote sensing data etc. Besides this, heavy access of the World Wide Web as a knowledge repository provides a large amount of data and information. This extra ordinary growth in collected data has generated an immediate requirement for new automated platform that can intelligently help us in converting the huge collection of data into valuable information and knowledge.

Data mining aims at discovering valuable information that is hidden in conventional databases. Data mining applied to the web has the potential to be quite beneficial. The emerging field of web mining aims at finding and extracting relevant information that is hidden in Web-related data. Etzioni (1996) is believed to be the inventor of 'Web Mining' and he described it as: "Web Mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services."

Web mining is the mining of data related to the World Wide Web. It is categorized into three active research areas according to what part of web data is mined, of which Usage mining, also known as web-log mining, which studies user access information from logged server data in order to extract interesting usage patterns [1].

Web mining is commonly divided into the following three sub-areas:

1. Web Content Mining: It consists of a number of techniques for searching the Web for documents whose content meets a certain criterion. It is an Application of data mining techniques to unstructured or semi structured text.
2. Web Structure Mining: It tries to extract information from the hyperlink structure between different pages. It uses the hyperlink structure of the web as an (additional) information sources.
3. Web Usage Mining: Web Usage Mining is defined as "the application of data mining techniques to discover usage patterns from web data". It is an analysis of user interaction with a web server.

The commonly used Web mining technologies are user access pattern analysis, Web document clustering, classification and information filtering. There are some deficiencies such that most of Web search systems are just in a simple search model, focused on the retrieval efficiency and ignored the retrieval accuracy due to inherent subjectivity, inaccuracy and uncertainty of user queries. The soft decision-making is not available. The relevance of the documents is only a part of attributes and has not very clear boundaries.

Clustering is the process of grouping the data into classes or clusters so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Dissimilarities are assessed based on the attribute values describing the objects. Clustering has its roots in many areas including data mining, statistics, biology and machine learning. The entity in data mining could have large number of attributes. Clustering of said high dimensional data creates major difficulties, more than in supervised learning. In decision trees irrelevant attributes will not be selected for decision nodes. In clustering high dimensionality faces a different problem. First, the irrelevant attributes dilutes any indication on clustering nature. This could also possible with low dimensional space, the probability of occurrence and number of irrelevant attributes increases with dimension. The second problem is the dimensionality curse that is a lack of separation of data in a high dimensional space.

Web mining is used to automatically discover and extract information from Web-based data sources such as documents, log, services, and user profiles.[8] A usage-based Web personalization system utilizes Web data in order to modify a Web site.[9]

According to Personalization Consortium the main reason of using information technology to provide personalization are [10]:

- By anticipating needs to cater the customer requirements in a better manner;
- Make the interaction efficient and satisfying for both parties;
- Build a relationship with customer that encourages them to return for consequent visits.

Personalization is a process of collecting and storing information about website users, analyzing the information relatd to them, and, on the basis of analysis, giving the proper information to each visitor at the right time.[11]

2. PREPROCESSING OF WEB DATA

In the WWW (World Wide Web) environment, a large amount of traffic can be generated between clients and web server. The traffic from a client to a server is a URL. The traffic from server to client is a named HTML file that will be interpreted and displayed on the client screen.
The web usage log probably is not in a format that is usable by mining applications. As with the data may need to be reformatted and cleaned. Steps that are part of the preprocessing phase include cleansing, user identification, session identification, path completion, and formatting.

- Let P be a set of literals, called pages or clicks, and U be a set of users. A Log is a set of triplets $\{(u_1,p_1,t_1)\ldots(u_n,p_n,t_n)\}$ where $u_i \in U$, $p_i \in P$, and t_i is a time stamp.

Standard log data consist of the following:-Source site, destination site, and time stamp. A common technique for a server site to divide the log records into sessions.

A session is a set of page references from one source site during one logical period. Historically, a user logging into a computer, performing work, and then logging off the login would identify a session and log off represent the logical start and end of the session.

- Data Cleaning

To clean a server log to remove irrelevant items is very important for Web log analysis. The discovered information are only useful if the data represented in the server log gives a correct scenario of the user accesses to the Web site. The HTTP protocol maintains an individual connection for each file requested from the Web server.

- User Identification

Next, unique users must be identified. This is a very complex task because of the existence of caches, firewalls, and proxy servers. The Web Usage Mining methods that rely on user co-operation or by the automatic identification of web user are the ways to deal with this problem.

- Session Identification

For logs that cover long duration of time, it is most probable that users will visit the Web site multiple times. The objectives of session identification are to break the web page visits of each user into individual sessions. The simplest method for this is through a timeout, where if the time between page requests crossed a certain threshold, it is supposed that the user is starting a new session. Many commercial products use 15-30 minutes as a default timeout.

- Path Completion

Another problem in reliably identifying unique user sessions is determining if there are vital accesses that are not recorded in the web access log. This problem is referred to as path completion. Methods similar to those used for user identification can be used for path completion.

- Formatting

Once the appropriate pre-processing steps have been applied to the server log, a final preparation module can be used to properly format the sessions or transactions for the type of data mining to be accomplished.

Let $P = (p_1,p_2\ldots..p_u)$ is the sequence of Web pages accessed from a certain IP between t_1 and t_u. Then , a user session $v(t_1,t_f),t_f$ t_u, is defined as: $v(t_1,t_f)=(p_1,p_2,\ldots.p_f)$: $(\Delta t=t_j-t_i-1\ \delta,\ 1<j$ f) $\wedge$ (f = $u \vee t_f+1-t_f>\delta$), where δ is a predefined time threshold [3].

There are many problems associated with the pre-processing activities, and most of these problems centre around the correct identification of the actual user. User identification is complicated by the use of proxy servers, client side caching, and corporate firewalls. Cookies can be used to assist in identifying a single user regardless of machine used to access the web.

3. CLUSTERING HIGH DIMENSIONAL WEB USAGE DATA

Hundreds of properties are there in data mining objects. A terrific complexity is faced in high dimensional spaces clustering which is much more than predictive learning. For node splitting, however in decision trees, unnecessary properties would not be selected and those unnecessary properties would not show any impact on Naïve Bayes. A twofold problem is introduced through high dimensionality in clustering. Firstly, if there is any desire on clustering tendency, unnecessary properties occurrence removes it.under whatever definition of similarity,.However, if

there are no amount of clusters present then it is a useless process to conduct a search for clusters. The low dimensional data also follows the same process, Unnecessary properties always rises in dimension with the propability of its number and occurrence. Secondly, dimensionality curse that is a loose way of speaking about a lack of data separation in high dimensional space.

Scientifically, nearest neighbour query becomes unbalanced: the distance to the bulk of points is almost similar to the distance of nearest neighbor. For the dimensions higher than fifteen, this impact starts showing its brutal effect. Therefore, related to the theory of distance, creation of clusters is becoming unsure in such conditions.

- Dimensionality Reduction

Most of the clustering algorithms based on spatial datasets indices which helps in rapid search of the nearest neighbours. As a result, indices provides fine proxies related to dimensionality curse performance impact. Indices whose dimensions ranges under sixteen work successfully, these indices are used in clustering algorithms. For a dimension $d > 20$ their performance degrades to the level of sequential search (though newer indices achieve significantly higher limits). Hence, possibly it states that more than sixteen properties data is high dimensional.

Web clustering totally depends on the contents of the site resulting two hundred to one thousand properties (pages/contents) for reserved Web sites. Dimensions in genomic and biology figures without any problem goes beyond two thousand to five thousand properties. Finally, thousands of properties are dealt in data mining and information recovery. Two approaches are used to handle high dimensionality:

(1) Attributes transformations and
(2) Domain decomposition.

4. CLUSTERING OF WEB USAGE DATA USING CONCEPT HIERARCHY

Clustering of data in a large dimension space is of great interest in many data mining applications. In this, I present a method for personalizing web user environment using clustering of web usage data in a high dimensional space based. In this approach, the relationship present in the web usage data are mapped into a fuzzy proximity relation of user transactions. For this, we cluster the user transactions using the concept hierarchy of web usage data and similarity relation of user transactions.

• A user transaction is a sequence of URLS. Let the users' transaction be:

$\tau = \{T_1, T_2, T_3 \ldots T_m\}$

• Let U be the set of Unique URLS appearing in the log data for all transactions

$U = \{URL_1, URL_2, URL_3 \ldots, URL_n\}$

Here, each $T_i \in \tau$ is a non-empty subset of U, which is a set of unique URLs appearing in the pre-processed log.

4.1. Concept hierarchy of web usage data

Concept hierarchy plays an important role in knowledge discovery process as it specifies background and domain knowledge and helps in better mining of data. In proposed approach, clusters of user transaction are formed from pre-processed log of web usage data, which can be put in a hierarchy.

• A concept hierarchy is a poset (partially ordered set) $(H, <)$ where H is a finite set of concept and $<$ is a partial order on H. [4]

4.2. Clustering user transactions

Clustering of user transactions tend to establish user transactions into groups having similar characteristics that exhibit similar patterns. In this method, transactions related to user are mapped into a multi-dimensional space as vectors of URL. Normally clustering technique partition this space into groups of URL that are nearer to each other based on a distance measure. With reference to web based transactions, each cluster represents a set of transactions that are similar based on co-occurrence patterns of URL references [5]. Such knowledge is especially useful in order to perform market segmentation in E-Commerce application or provide personalized web content to the users.

• A pair (A, B) is a (formal) concept of (D, T, R) if and only if
$A \subseteq D$, $B \subseteq T$, $A' = B \wedge B' = A$
Where,
$A' = \{t \in T \mid (d,t) \in R \text{ for all } d \in D\}$
$B' = \{d \in D \mid (d,t) \in R \text{ for all } t \in T\}$[6]

• A proximity relation is a mapping $S: D \times D \rightarrow [0, 1]$ such that for $x, y, z \in D$, the following rules hold:
$S(x, x) = 1$ (reflexivity)
$S(x, y) = S(y, x)$ (symmetry)
A fuzzy relation that is reflexive and symmetric is called fuzzy-similarity relation [2].

• α- Similarity:-

If S is a proximity relation on D, then given an $\alpha \in [0,1]$, two elements $x, z \in D$ are α -similar (denoted by $x\ S_\alpha\ z$) if and only if $S(x, y) >= \alpha$, definition (α Similarity) are said to α-proximate (denoted by $x\ S_\alpha + z$) if and only if either $x\ S_\alpha\ z$ or there exist a sequence $y1, y2, y3, \ldots yn \in D$ such that

$x\ S_\alpha\ y_1,\ y_1\ S_\alpha\ y_2\ ,\ y_2\ S_\alpha\ y_3 \ldots\ldots\ldots,\ y_n\ S_\alpha\ z$

• Indiscernibility:-

If $S: D \times D \rightarrow [0, 1]$ is a proximity relation, then S_α^+ is an equivalence relation.

From proximity relation concept, S_α^+ fulfill the property of reflexivity and symmetry. Based on α Similarity condition, the clusters (transactions) will merges, and it appears as it is due to transitivity.

Hence, $S_\alpha +$ is an equivalence relation.

Let the children of a transaction T be denoted as child (T). The similarity of two transactions T_i and T_j is defined as: - [4]

$$Sim(T_i, T_j) = \left| Child(T_i) \cap Child(T_j) \right| \Big/ \left| Child(T_i) \cup Child(T_j) \right|$$

Let $\tau = \{T_1, T_2, T_3 \ldots T_m\}$ be the user transactions and $U = \{URL_1, URL_2, URL_3 \ldots URL_n\}$ be the set of unique URLs in all the transactions here $T_i \subseteq U$ for I =1 to m. For all $T_i, T_j \in \tau$, I define a fuzzy relation R on τ as

$$R(T_i, T_j) = Sim(T_i, T_j)$$

From the concept of Indiscernibility, it can be seen that the similarity of two transactions Ti and Tj is a number between 0 and 1.When the similarity of two transactions Ti and Tj is 0, then Ti and Tj are completely dissimilar. On the other hand, if the similarity is 1, then the two transactions are completely similar.

The measure of similarity gives information about the users' browsing pattern. Actually, the browsing of web by any two users may not be exactly similar but may have common interesting sites. Based on the said definition, a similarity matrix can be calculated and used as the base for

clustering. In the context of clustering user transactions, the huge numbers of dimensions are the URLs present in the access logs.

We consider the fuzzy relation to decide the amount of similarity between the user transactions. For all $T_i, T_j \in \tau$, I define a fuzzy relation R on τ as:-

$R(T_i, T_j) = Sim(T_i, T_j)$

Relation R formed by the similarity of user transaction $\tau = \{T_1, T_2, T_3 \ldots T_m\}$ is a fuzzy proximity relation and is given by R:

Where $\mu_{ij} = Sim(T_i, T_j)$

Now, for any given value $\alpha >= 0$, $R+_\alpha$ is an equivalence relation. The corresponding partition represents transaction clusters. These partitions represent the transaction Clusters. The user specified α value is an important factor based on which the performance of cluster depends. A domain expert can choose the value of $\alpha \in [0,1]$ based on her/his experience. Algorithm -1 represents the overall approach used for clustering

Table 1. Algorithm-1 for cluster generation based on concept hierarchy using fuzzy similarity

INPUT:

1. $\tau = \{T_1, T_2, T_3 \ldots T_m\}$ // Set of transactions for unique user
2. $U = \{URL_1, URL_2, URL_3 \ldots URL_n\}$ // set of unique URLs appearing in the preprocessed log.
3. Value of α (by domain expert)

OUT PUT:

1. Number of clusters
2. Clusters (Set of Transactions)

Step-1: Find the children (all URLs) in a single transaction T as Child (T)

Step-2: Find the similarity of two transactions. The similarity of two transactions T_i and T_j is defined as: -

$Sim(T_i,T_j) = |Child(T_i) \cap Child(T_j)| / |Child(T_i) \cup Child(T_j)|$

Step-3: Compute similarity matrix for all T_i and T_j, where each value in matrix $\mu_{ij} = Sim(T_i, T_j)$. Proximity relation is defined as

$Sim(T_i,T_j) \in [0,1]$

$Sim(T_i,T_i) = 1$ (Reflexivity)

$(T_i,T_j) = Sim(T_j,T_i)$ (Symmetry)

Value of similarity measure always lies between 0 and 1 **(fuzzy similarity)**

Step-4 : Compute α- Similarity.

For a given value (provided by domain expert) ,here $\alpha \in [0,1]$, two elements T_i, T_z are α -similar if $Sim(T_i, T_j) >= \alpha$, definition (α Similarity) are said to α-proximate if there exist a sequence $y_1, y_2, y_3, \ldots y_n$ such that $Ti\ Sim_\alpha\ y_1, y_1\ Sim_\alpha\ y_2, y_2\ Sim_\alpha\ y_3 \ldots\ldots, y_n\ Sim_\alpha\ T_z$. (Transitivity)

Step-6: Proximity relation between T_i and T_j fullfil the following properties

1. Reflexivity
2. Symmetry
3. Transitivity

Hence , $Sim(T_i, T_j)$ proves the **Equivalence Relation**. This relation generates the equivalence classes. For given value $\alpha >= 0$, the corresponding partition (equivalence class) represent transaction clusters

4.3. Experimental Results & Analysis for Clustering of High Dimensional Data

We assessed the effectiveness of presented new clustering approach by experimenting with real data sets. The data sets came from MCU Web-site's server log file. We show similarity matrix related to one data set and their clusters and partition coefficients in table-I and table-II. The log data from accesses to the server during a period of 30 days is used for the purpose of generation of user sessions profiles.

In this model clustering quality is purely dependent on user specific value $\propto \in [0,1]$. Cluster members vary accordingly variation in the value of $\propto \in [0,1]$. Validity of cluster goodness produced by this model is highly subjective matter depends upon domain expert.Algorithm-2 gives an approach to check cluster validity.

We evaluate cluster goodness on different value of $\propto$ on different data sets (six data sets) using Partition coefficient, which is simple method for evaluating cluster goodness and applicable on high dimensional data (Web Data). The objective is to seek clustering schemes where most of the vectors of the dataset exhibit high degree of similarity in one cluster. We note, here, that a fuzzy clustering is defined by a matrix $U = [u_{ij}]$ where u_{ij} denotes the degree of similarity of the vector x_i in the j cluster.

Partition Coefficient (Bezdek proposed in Bezdeck et al.(1984) the partition coefficient, which is defined as[7]

The PC index values range in [1/nc, 1], where nc is the number of clusters. The closer to unity the index the "crisper" the clustering is. In case that all membership values to a fuzzy partition are equal, that is, $u_{ij}=1/nc$, the PC obtains its lower value. Thus, the closer the value of PC is to 1/nc, the fuzzier the clustering is. Furthermore, a value close to 1/nc indicates that there is no clustering tendency in the considered dataset.

Table 2. Cluster validity Algorithm

INPUT:
Nc – Number of clusters
N-Total number of transactions in all clusters
OUTPUT:
PC-Partition coefficient index value
ALGORITHM:
For i =1 to N do
For j = 1 to nc
$U = \mu^2_{ij}$
μ_{ij} denotes the degree of similarity of the vector x_i in the j^{th} cluster
end for j
end for i
PC = U/N

4.4. Snapshot of Web Log data of MCU Server:

2013-12-30 02:47:36
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie) cs(Referer) sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken

2013-12-30 02:47:36 W3SVC8521 NS1 70.84.233.186 GET /robots.txt - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 404 0 2 1814 153 46

2013-12-30 02:47:36 W3SVC8521 NS1 70.84.233.186 GET /forum/Default.asp - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 200 0 0 14115 225 421

2013-12-30 03:01:44 W3SVC8521 NS1 70.84.233.186 GET /Study_Institue_Scheme_2005.htm - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 404 0 2 1814 249 46

2013-12-30 03:40:04 W3SVC8521 NS1 70.84.233.186 GET /forum/www.satpuralawcollege.org - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 404 0 2 1814 250 93

2013-12-30 11:43:44 W3SVC8521 NS1 70.84.233.186 GET /robots.txt - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 404 0 2 1814 153 62

2013-12-30 11:43:44 W3SVC8521 NS1 70.84.233.186 GET /downloads.htm - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 200 0 0 8004 232 140

2013-12-30 13:27:27 W3SVC8521 NS1 70.84.233.186 GET /robots.txt - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 404 0 2 1814 153 62

2013-12-30 13:27:27 W3SVC8521 NS1 70.84.233.186 GET /index.htm - 80 - 65.54.188.68 HTTP/1.0 msnbot/1.0+(+http://search.msn.com/msnbot.htm) - - 200 0 0 16766 219 187

After processing (as per algorithm-1) we obtained following set of transactions (period of 15 minutes) and set of unique URLs for input to algorithm-2.

Table 3. User Transactions Detail.

User1	**Transaction**	**URLs**
65.54.188.68	T1	/robots.txt /forum/Default.asp /Study_Institue_Scheme_2013.htm
,,	T2	/forum/www.satpuralawcollege.org
,,	T3	/robots.txt /downloads.htm
,,	T4	/robots.txt /index.htm
,,	T5	/ResearchProjectPhoto/a2.html /Media_courses_syllabus_Download_Page.htm

Table 4. Unique URLs Detail.

URLs	**URL Name**
URL1	/robots.txt
URL2	/forum/Default.asp
URL3	/Study_Institue_Scheme_2005.htm
URL4	/forum/www.satpuralawcollege.org
URL5	/downloads.htm
URL6	/index.htm

URL7	/ResearchProjectPhoto/a2.html
URL8	/Media_courses_syllabus_Download_Page.htm

Computation of similarity matrix based upon above 5 transactions and 8 unique URLs

Table 4. Similarity Matrix obtained from Test data set-I (text) file, IP Address-65.54.188.68, Date –30-12-13

	T1	**T2**	**T3**	**T4**	**T5**
T1	1	0	0.25	0.25	0
T2	0	1	0	0	0
T3	0.25	0	1	0.33	0
T4	0.25	0	0.33	1	0
T5	0	0	0	0	1

Table 5. Test Data Set-I for IP Address 65.54.188.68.

IP Address-65.54.188.68, Date –30-12-13		
Total Transactions-5, Number of Unique URLS-8		
$\propto$ >=	**Number of clusters**	**Partition Coefficient (PC)**
0.5	5	1.0
0.6	5	1.0
0.7	5	1.0
0.9	5	1.0

5. WEB PERSONALIZATION ARCHITECTURE

The overall architecture of the web personalization model is depicted in figure below. There are three main components which are emphasized in this figure i.e. Client, Recommendation Engine and the Server. The client only request for particular URL and in response server provides the personalized environment i.e. the set of recommended URLs. When a user starts its active session then the list of URLs gets stored at server's log file. At the server side, data pre processing is done using data cleaning, session identification and transaction identification. Structured data output from the processing of server content data and usage data in a form of transaction file is used to obtain the transaction cluster using fuzzy logic based similarity relation. This transaction cluster is then fed to the recommendation engine component. In this component the recommendation set

is generated using its logic. The result of this component provides the set of recommendations that will be used to personalize the user interface to the server.

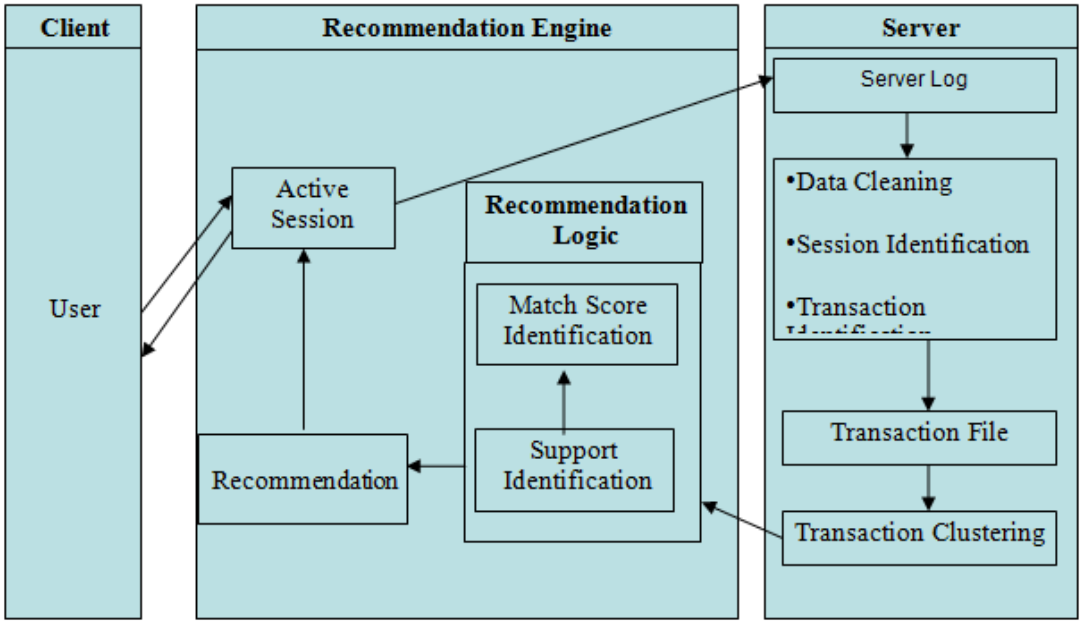

Figure 1. Architecture for Web Personalization

6. RECOMMENDATION ALGORITHM

For our approach, we use transaction clusters which are obtained from a fuzzy logic based method. Transaction clustering will result in a set $C = \{c_1, c_2, \ldots, c_k\}$ of clusters, where each C_i is a subset of T, i.e., a set of user transactions. Each cluster represents a group of users with "common" access patterns. Generally transaction clusters are not an effective way of capturing a combined view of common user profiles. Each transaction cluster may contain thousands of transactions involving hundreds of URL. We present a method to personalize the user session by providing a set of URL links dynamically from the highest match score clusters down to lowest match score cluster.

Let $A = \{T_1, T_2, T_3, \ldots\ldots Tm\}$ be the set of user transactions and
$U = \{URL_1, URL_2, URL_3, \ldots URL_n\}$ be the set of unique URLs in all the transactions here $T_i \subseteq U$ for I =1 to m.For all $T_i, T_j \in A$.
Let $C_1, C_2, \ldots, C_m$ be the transaction clusters obtained from user transaction A with the help of fuzzy logic based similarity relation.

STEP-1

For the given input i.e user transactions, accessed URLs and clusters, calculate value of URL_i in Cluster C_k. The value of URL_i in Cluster C_k is defined as.
Definition 1: The proportion of number of occurrence of URL_i across transactions in the cluster C_k to the total number of transactions in the cluster is known as the Value of URL_i in cluster C_k.[12]

$$Val(URL_i, C_k) = \frac{\{T | URL_i \in Child(T), T \in C_k\}}{}$$

$$\sum T \quad \{Child(T) | \ T \in C_k$$

Where Child(T) is the set of children of transaction T.
The measure of value of an URL in a cluster gives the information on the weight of that URL in that cluster
Let a new transaction is started ,say T_{new}. Assign T_{new} itself a cluster,that is, $C_{new} = \{T_{new}\}$
The potential usefulness of C_k with respect to C_{new} is a factor defining its interestingness can be estimated by a utility function such as support.

STEP-2

Definition 2: The support of new a cluster C_{new} to cluster C_i is defined as

$$Support(C_{new}, C_i) = \frac{\sum URL(Val(URL,C_{new}) - Val(URL, C_i)}{\{ Child(T) \mid T \in C_i \cup C_{new} \}}$$

Where Child(T) is the set of children of transaction T. $Val(URL,C_{new})$ is the value of URL in the cluster C_{new} and $val(URL,C_i)$ is the value of URL in the cluster C_i.
Once a new user starts a session, at every step, the objective is to match, the half-way user session with the suitable clusters and provide correct recommendations to the user.

STEP-3

Definition 3:The match score between C_i and C_{new} is defined as
$Match(C_{new},C_i) = 1- Support(C_{new},C_i)$
The preference set of URLs can be made from the highest match score clusters down to lowest match score cluster. We can also enforce a threshold ϕ on the matching score to reduce the dimensional space.

6.1. Experimental Result (User IP Address (65.54.188.68))

We used the access logs from the Web site of the MCRPV (http://www.mcu.ac.in) for our experiment. The log data from accesses to the server during a period of 30 days was used for the purpose of generation of user session profiles. The site includes a number of pages related to university profile and activities held in university campus. After pre processing we prepared several data sets for experiment purpose.

1. Log Data File: This is pre processed file taken from MCU web log file for the IP address 65.54.188.68.One transaction contains URL accessed by user within 15 minutes.

Table 6. Transactions with their URLs

Transactions	**URLs**
T_1	Url_1
T_1	Url_2
T_1	Url_3
T_2	Url_4
T_3	Url_1
T_3	Url_5
T_4	Url_1
T_4	Url_6
T_5	Url_7
T_5	Url_8

2. Input Clusters:-These clusters are taken as input using clustering approach having fuzzy logic based proximity relation.

Table 7.Clusters with their transaction

Clusters	Transactions
C_1	T_1
C_2	T_2
C_3	T_3
C_4	T_4
C_5	T_5

3. New Transaction:- From the web log file I extract a transaction which to be assumed a new transaction for which we have to provide personalized environment. Initially transaction contain following unique URLs.
$\{Url_1, Url_3\}$

4. Unique URLs:- These are the unique URLs accessed by user within different transactions. Each URL i.e. URL_1, URL_2, URL_3 etc. represent a URL entry of web log file given in table 4.

5. Values of URLs in Clusters:-
Value of URL_i in cluster C_i is calculated as follows

Table 8. Value of URL_i in Cluster C_i

URLs	$Cluster_1$	$Cluster_2$	$Cluster_3$	$Cluster_4$	$Cluster_5$	$Cluster_{NEW}$
Url_1	0.333333	0.000000	0.500000	0.500000	0.000000	0.500000
Url_2	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000
Url_3	0.333333	0.000000	0.000000	0.000000	0.000000	0.500000
Url_4	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
Url_5	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000
Url_6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Url_7	0.000000	0.000000	0.000000	0.500000	0.500000	0.000000
Url_8	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000

URLs	$Cluster_1$	$Cluster_2$	$Cluster_3$	$Cluster_4$	$Cluster_5$	$Cluster_{NEW}$
Url_1	0.333333	0.000000	0.500000	0.500000	0.000000	0.500000
Url_2	0.333333	0.000000	0.000000	0.000000	0.000000	0.000000
Url_3	0.333333	0.000000	0.000000	0.000000	0.000000	0.500000
Url_4	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000
Url_5	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000
Url_6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Url_7	0.000000	0.000000	0.000000	0.500000	0.500000	0.000000
Url_8	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000

6. Support of new cluster to existing Clusters:-

The support calculation is as follows:.

Table 9. Clusters with their Support

Clusters	Support
Cluster 1	0.222222
Cluster 2	0.666667
Cluster 3	0.333333
Cluster 4	0.333333
Cluster 5	0.500000

7. Match score to $Cluster_{NEW}$ to $Cluster_i$:-

The Match score calculation is as follows:

Table 10. Clusters with Match Score

Clusters	Match Score
Cluster 1	0.777778
Cluster 2	0.333333
Cluster 3	0.666667

Cluster 4	0.666667
Cluster 5	0.500000

Here, highest Match Score corresponds to $Cluster_1$. Hence, user 65.54.188.68 will be provided Cluster 1 as preference set. i.e. URLs of transactions belongs to cluster1 will be suggested as preference links for providing the personalized environment.

6.2 Evaluation of Recommendation Effectiveness

In order to evaluate the recommendation effectiveness for given model, We measured the performance using three different standard measures, namely, precision, coverage, and the F1 measure.[13] These measures are adaptations of the standard measures, *precision* and *recall*, often used in information retrieval.

Table 11. Measures values for Data Sets

Data Set	Precision	Coverage	F1 Measure
Data Set 1 IP Address: 65.54.188.68	0.33	1.0	0.4962
Data Set 2 IP Address: 59.94.43.47	0.5714	1.0	0.7261146
Data Set 3 IP Address:70.84.233.186	0.4705	1.0	0.6394558
Data Set 4 IP Address:61.0.56.2	0.5625	0.9	0.690411
Data Set 5 IP Address:67.180.89.140	0.63636	1.0	0.773

In the context of personalization, precision is used to identify the degree to which accurate recommendations can be produced by the recommendation engine, while coverage identifies the ability of the recommendation engine to produce all of the URLs that are likely to be visited by the user. A low precision in this context will likely result in irritated visitors who are not interested in the recommendation list, while low coverage will show the inability of the site to generate relevant recommendations at the moment the user interact with the site. Personalization systems are often evaluated based on two statistical measures, namely precision and coverage. In our case of personalization approach, the proposed model is capable of producing recommendation sets with high coverage and high F1 measure with moderate precision.

7. CONCLUSIONS

In this paper, we proposed a method to construct clusters based on the concept hierarchy and α-similarity. The new model to cluster the web user transactions based on the concept hierarchy of web usage data and fuzzy proximity relations of user transactions was tested for its effectiveness. The fuzzy concept ensemble to identify amount of similarities between user transaction clusters. The user specified α value is an important factor based on which the performance of cluster depends. A domain expert can choose the value of $\alpha \in [0,1]$ based on her/his experience. That is, the approach makes use of domain knowledge for forming clusters and it is very helpful especially for e-commerce application that uses these user transaction clusters to make personalized environment for the Web users.

Employing Web usage mining techniques for personalization is directly related to the discovery of effective profiles that can successfully capture relevant user navigational patterns. In this paper, we also proposed an approach for automatic discovery of user interest domain. A web site should permit users to make their choices in natural way. This is possible by giving broad coverage and understandable choices, so that the user can formulate an easy choice that best suited to their needs. The proposed model is capable of producing recommendation sets with high coverage and high F1 measure with moderate precision.

REFERENCES

[1] Lin HuaXu, Hong Liu. Web User Clustering Analysis based on KMeans Algorithm, 2010 International Conference on Information, Networking and Automation (ICINA).

[2] Abdolreza Mirzaei and Mohammad Rahmati A Novel Hierarchical-Clustering-Combination Scheme Based on Fuzzy-Similarity Relations, IEEE Transactions On Fuzzy Systems, VOL. 18, NO. 1, FEBRUARY 2010.

[3] Dimitrios Pierrakos and Georgios Paliouras, Personalizing Web Directories with the Aid of Web Usage Data, IEEE Transactions On Knowledge And Data Engineering, Vol. 22, NO. 9, SEPTEMBER 2010.

[4] Quan, Thanh Tho, Hui, Siu Cheug and Cao, Tru Hoang. A Fuzzy FCA-based Approach to Conceptual Clustering for Automatic Generation of Concept Hierarchy on Uncertainty Data, V. Sn´a¡sel, R. B¡elohl´avek (Eds.): CLA 2004, pp. 1–12, ISBN 80-248-0597-9.V¡SB – Technical University of Ostrava, Dept. of Computer Science, 2004.

[5] B.Mobasher, R.Colley and J.Srivastava. Automatic personalization based on web usage mining, Commun. ACM 43,8 (2000) 142-151.

[6] Yun Zhang, Boqin Feng, Yewei Xue, A New Search Results Clustering Algorithm based on Formal Concept Analysis, Fifth IEEE International Conference on Fuzzy Systems and Knowledge Discovery, 2008

[7] Halkidi , Maria ,Batistakis, yannis ,et.al., On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17:2/3, 107– 145, 2001 [2] Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.

[8] Dragos Arotariteia, Sushmita Mitra. “Web mining: a survey in the fuzzy framework” ELSEVIER, Fuzzy Sets and Systems 148 (2004)

[9] Magdalini Eirinaki And Michalis Vazirgiannis .”Web Mining for Web Personalization” ACM Transactions on Internet Technology, Vol. 3, No. 1, February 2003

[10] The Personalization Consortium http://www.personalization.org/personalization.html

[11] Web site personalization http://www 128.ibm.com/developerworks/websphere/library /techarticles/hipods/personalize.html

[12] B.Mobasher, R.Colley and J.Srivastava. “Automatic personalization based on web usage mining” Commun. ACM 43,8 (2000)

[13] Bamshad Mobasher, Honghua Dai, Tao Luo and Miki Nakagawa “Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization” School of Computer Science, Telecommunications, and Information Systems DePaual University, Chicago, Illinois, USA. http://ai.stanford.edu/users/ronnyk/WEBKDD-DMKD/Mobasher.pd

Authors

Bhawesh Kuamr Thakur received the M.Tech degree in CSE from the M.C. University, Bhopal, India. He is currently working toward the PhD degree in computer science at the Department Of Computer Science & Engineering, Integral University, Lucknow, India. He is working as a faculty member in the Bansal Institute of Engineering & Technology, Lucknow, an UPTU, Lucknow affiliated Engineering College. His research interests lie in the areas of Web mining and Web personalization.

Prof. Dr. Syed Qamar Abbas completed his Master of Science (MS) from BITS Pilani. His PhD was on computer-oriented study on Queueing models. He has more than 20 years of teaching and research experience in the field of Computer Science and Information Technology. Currently, he is Director of Ambalika Institute of Management and Technology, Lucknow

Prof. Dr. M. Rizwan Beg is M.Tech & Ph.D in Computer Sc. & Engg. Presently he is working as Controller of Examination in Integral University Luck now, Uttar Pradesh, India. He is having more than 16 years of experience which includes around 14 years of teaching experience. His area of expertise is Software Engg., Requirement Engineering, Software Quality, and Software Project Management. He has published more than 40 Research papers in International Journals & Conferences. Presently 8 research scholars are pursuing their Ph.D in his supervision.