International Journal of Computer and Communication Technology

Volume 8 | Issue 1 Article 1

January 2017

An Approach for Customer Behavior Analysis Using Web Mining

Preeti Sharma
NIT Raipur, Raipur, Chattishgarh, India, preeti.nitu@gmail.com

SANJAY KUMAR
Dept. of Electrical Engineering NIT Raipur, India, skumar.it@nitrr.ac.in

An Approach for Customer Behavior Analysis Using Web Mining

Preeti Sharma & Sanjay Kumar

NIT Raipur, Raipur, Chattishgarh, India
E-mail: preeti.nitu@gmail.com & Skumar.IT@nitrr.ac.in

Abstract - Customer satisfaction is the key secret of success for all industries regardless of whether it is web enabled or not. This paper focuses the role of web mining in achieving a viable edge in business. Web mining is becoming the tool for success for those who adopt electronic means of operation for conducting their business. Web mining is the application of data mining techniques to discover patterns from the Web through content mining, structure mining, and usage mining. Web mining can contribute to a large extent in gaining a competitive advantage in business. Business goals should be well understood.

Keywords - *Web Minning, Log data, Data Mining, Sesssionzation, Frequent path.*

I. INTRODUCTION

Customer relationship is one of the major applications of Web mining. A website should be designed to entice the customers. Web Mining analyses visitor's behavior and makes predictions on their future interaction. This can be exploited to improve website performance and to recommend products or links based on user's behavior. Visitors entering the site exhibits different behavior. They might just surf through or the process might end up in a purchase. For understanding customer behavior and thus improve the performance of your web site, certain standards should be used like perform mining on web log data

A. WEB MINING:

The information space known as Web is a collection of resources (Web resources) residing on the Internet, that can be accessed using HTTP and protocols that derive from it. A resource "can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., "today's weather report for Los Angeles"), as well as a collection of other resources. Not all resources are network "retrievable"; e.g., human beings, corporations, and bound books in a library can also be considered resources".

The most important concept regarding the Web is of course the resource that a server makes available to clients spread everywhere on the Internet, without any resource, the whole system won't have any sense. When a resource is accessed by a client at a specific time and space, we talk of resource manifestation [5]. The general definition for client is "the role adopted by an application when it is retrieving and/or rendering resources or resource manifestations", whereas the specific one for the Web defines the Web client as an application "capable of accessing Web resources by issuing requests and render responses containing Web resource manifestations"[8] On the other hand, the server is "the role adopted by an application when it is supplying resources or resource manifestations" to the requesting client.

Web mining involves a wide range of applications that aims at discovering and extracting hidden information [4] in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files for example for predictive Web caching. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage Mining [2]. While web structure and content mining utilize primary data on the web, web usage mining works on the secondary data such as web server access logs, proxy server logs, referrer logs, browser logs, error logs, user profiles, registration data, user sessions or transactions, cookies, user queries, and bookmark data. Through analyzing these log files [2] and documents we can access to interesting usage patterns and information. The Various Business Areas Where Web Mining has helped in Improving the Business Decision Making

B. WEB SITE SERVICE QUALITY IMPROVEMENT

The World Wide Web is one of the most used interfaces to access remote data and commercial, noncommercial services and the number of actors involved in these transactions is growing very quickly[4]. Everyone using the Web Experiences knows that how the connection to a popular website may be very slow during rush hours and it is well known that web users tend to leave a site if the wait time for a page to be served exceeds a given value[7]. Therefore, performance and service quality attributes have gained enormous relevance in service design and deployment. This has led to the development of web benchmarking tools that are largely available in the market. One of the most common criticism to this approach is that synthetic workload produced by web stressing tools is far from realistic. Moreover, websites need to be analyzed for discovering commercial rules and user profiles and models must be extracted from log files and monitored data[7].

C. *WEB DATA*

Web data are those that can be collected and used in the context of Web personalization. These data are classified in four categories according to [SC+00]

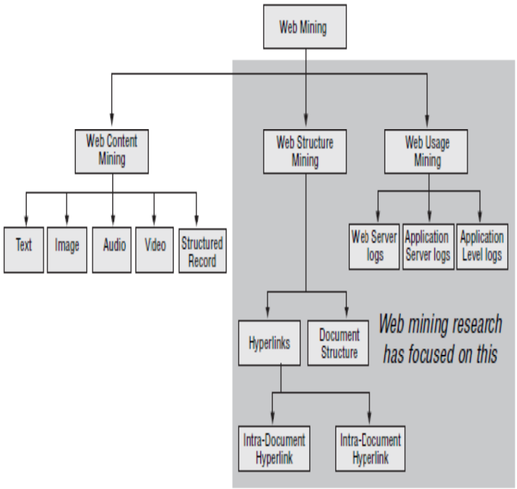


Fig. 1: Web Mining Taxonomy

- Content data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.
- Structure data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.
- Usage data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.
- User profile data provide information about the users of a Web site. A user profile contains demographic information for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

usage mining. Log files [20] are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing[18]. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered.

II DISCOVERY OF USAGE PATTERNS FROM HTTP SERVER LOGS

Before going on, I would like to introduce another term, episode it is "a subset of related user clicks that occur within a user session". It is the related word that has to catch the attention of the Web usage miner. In order to discover usage patterns from the available data described above, it is necessary to perform three steps

A. *Pre-processing*

This phase is probably the most complex and ungrateful step of the overall process. Its main task is to "clean" the raw web log files and insert the processed data into a relational database or a data warehouse, in order to make it appropriate to apply the data mining techniques in the second phase of the process.

B. *Pattern discovery*

The goal of this stage is to find hidden relationships in the data. Typically the first technique applied to the data is statistical analysis. With this technique, the type of information extracted is

- Most frequently requested pages;
- Average access time;
- Most common error coded, etc.

Although this kind of information can be valuable for the systems administrator, in terms of a business perspective it has limited interest. Other methods like clustering, Hidden Markov Models or Bayesian Belief Networks are usually applied to make classification and discover dependency between data[19].

C. *Pattern analysis*

The various patterns can be obtained after patterns discovery process. All of these patterns can not be interesting, pattern analyzer find the interestingness among patterns and only choose some of the interesting patterns, and then rest of the patterns can be ignored.

These phases define the Web mining process and can also be used for the general case, not only the Web server one[16]. The first phase is the most important of all, because of the complex nature of the Web architecture, and therefore it is the most difficult one. Raw data coming from the Web server is unfortunately incomplete, and only a few fields are available for discovering patterns (IP address, time, user agent); besides, it is almost impossible to identify a user (unless an authentication check has been issued) and more often a single client (when this is behind a proxy). Thus this preparatory step is essential, and its aim is to build an as complete and robust as possible data file (server session file), by gathering information from the different available sources shown in the previous section. And this task is anything but easy[10].

The pattern discovery phase, consists of different techniques derived from various fields such as statistics, machine learning, data mining, pattern recognition, etc. applied to the Web domain and to the available data.

D. *Statistical analysis*

This kind of analysis is performed by many tools, available also for free, and its aim is to give a description of the traffic on a Web site, like most visited pages, average daily hits, etc.;

III. ASSOCIATION RULES

The main idea is to consider every URL requested by a user in a visit as basket data (item) and to discover relationships with a minimum support level between them; this is the case discussed in this paper from next section; Association rule Mining is as the task of finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories[11]. This field of data mining was originally developed to perform Market Basket Analysis, where the idea was to find the items that were bought most frequently together. In the web mining context the idea is to find the related pages that are accessed in the click stream, a give it a certain measure of probability, for example

- "If a person visits the CNN Web site, there is 60% chance the person will visit the ABC News Web site in the same month".
- Several algorithms exist to perform association rule mining apriori-like algorithms, Eclat, Frequent-Pattern tree algorithms, and many others.

A. *Sequential patterns*

The attempt of this technique is to discover time ordered sequences of URLs followed by past users, in order to predict future ones (this is much used for Web advertisement purposes). Sequence mining is the task of finding temporal patterns over a database of sequences, in this case a data base of click streams. Sequence mining is considered to be an extension of association mining that only finds no temporal patterns. This technique can have a very important role in knowledge discovery in web log data, due to the (temporally) ordered nature of click-streams. The type of patterns that results form the application of this technique, can have an example like this

If user visits page X, and then page Y, it will visit page Z with c% of chance".

The algorithms for sequence mining inherited much from the association mining algorithms, and many of them are extensions of the firsts, where the main difference is that in sequence mining inter-sequence patterns are searched, where in the association mining the patterns searched are intra-sequence patterns.

B. *Clustering:*

Meaningful clusters of URLs can be created by discovering similar characteristics between them according to users behaviours. In the next sections, I consider only the case of the discovery of association rules from an HTTP server data.

IV. THE A-PRIORI ALGORITHM:

Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems.

Find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. Generate association rules from those large item sets with the constraints of minimal confidence [12].

Suppose one of the large item sets is Lk = {I1,I2,...,Ik}; association rules with this item sets are generated in the following way the first rule is {I1,I2,...,Ik − 1} = > {Ik}. By checking the confidence this rule can be determined as interesting or not. Then, other rules are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. This process iterates until the antecedent becomes empty.

Since the second sub problem is quite straight forward, most of the research focuses on the first sub problem. The Apriori algorithm finds the frequent sets L in Database D.

Find frequent set Lk − 1.

Join Step.

Ck is generated by joining Lk − 1with itself

Prune Step.

Any (k − 1) -itemset that is not frequent cannot be a subset of a frequent k -itemset, hence should be removed. Where (Ck Candidate itemset of size k),(Lk frequent itemset of size k)

A. Apriori Pseudo code

1) Apriori

```
$L_1$←large 1-itemsets that appear in more
than ε transactions }
    $k \leftarrow 2$
    while $L_{k-1} \neq$
    $C_k \leftarrow \text{Generate}(L_{k-1})$
  for transactions $t \in T$
    $C_t \leftarrow \text{Subset}(C_k, t)$
    for candidates $c \in C_t$
    $\text{count}[c] \leftarrow \text{count}[c] + 1$
    $L_k \leftarrow \{c \in C_k | \text{count}[c] \geq \varepsilon\}$
    $k = k + 1$

    $\text{return} \bigcup_k L_k$
```

V. EXPERIMENTAL RESULTS

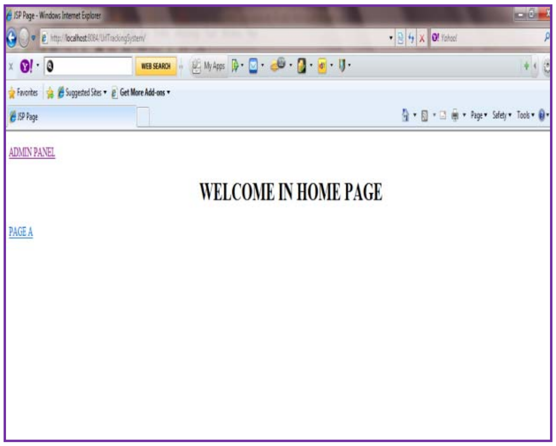
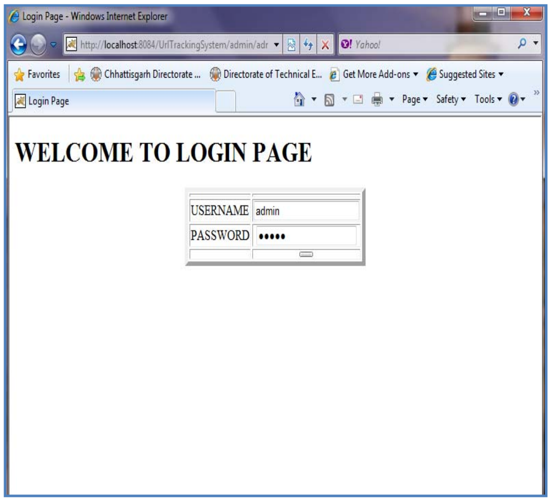


Fig. 1 : Home Page


Login Page - Windows Internet Explorer
http://localhost:8084/UrlTrackingSystem/admin/adr
Yahoo!
Favorites
Chhattisgarh Directorate ...
Directorate of Technical E...
Get More Add-ons
Suggested Sites
Login Page
Page
Safety
Tools
WELCOME TO LOGIN PAGE

USERNAME admin

PASSWORD •••••

Fig. 2: Login page

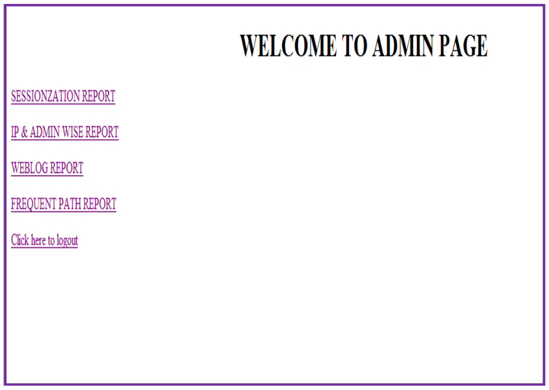
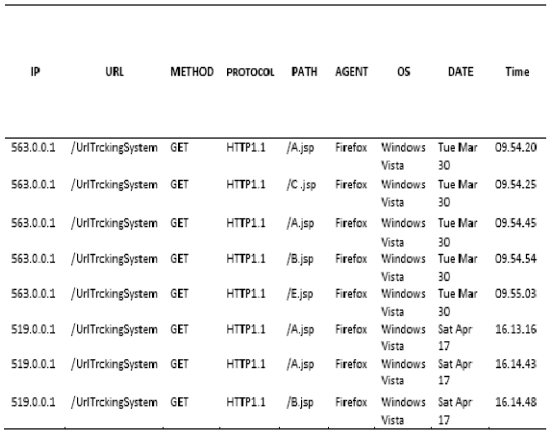


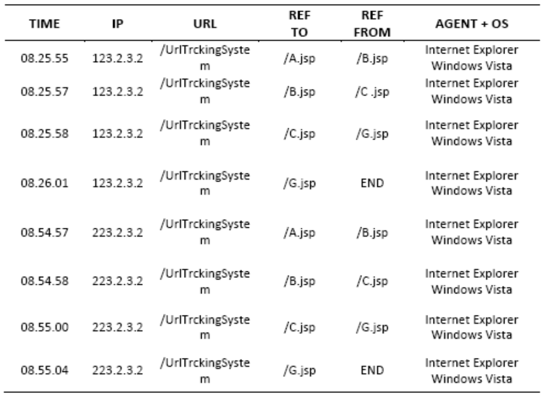
Fig. 3 : Administrative Page

Table 1 : Web Log Report



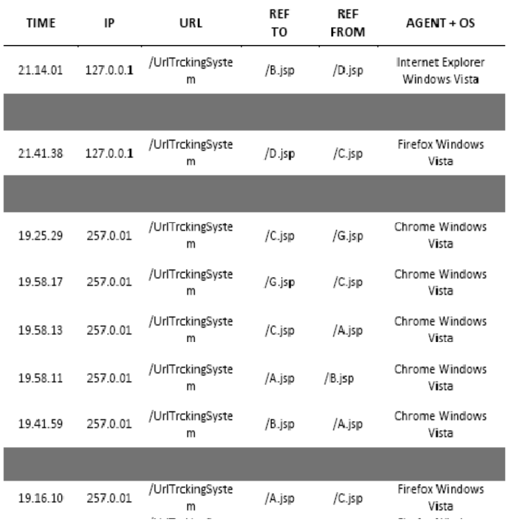
IP	URL	METHOD	PROTOCOL	PATH	AGENT	OS	DATE	Time
563.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/A.jsp	Firefox	Windows Vista	Tue Mar 30	09.54.20
563.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/C .jsp	Firefox	Windows Vista	Tue Mar 30	09.54.25
563.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/A.jsp	Firefox	Windows Vista	Tue Mar 30	09.54.45
563.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/B.jsp	Firefox	Windows Vista	Tue Mar 30	09.54.54
563.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/E.jsp	Firefox	Windows Vista	Tue Mar 30	09.55.03
519.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/A.jsp	Firefox	Windows Vista	Sat Apr 17	16.13.16
519.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/A.jsp	Firefox	Windows Vista	Sat Apr 17	16.14.43
519.0.0.1	/UrlTrckingSystem	GET	HTTP1.1	/B.jsp	Firefox	Windows Vista	Sat Apr 17	16.14.48

Table 2: Sessionzation Report



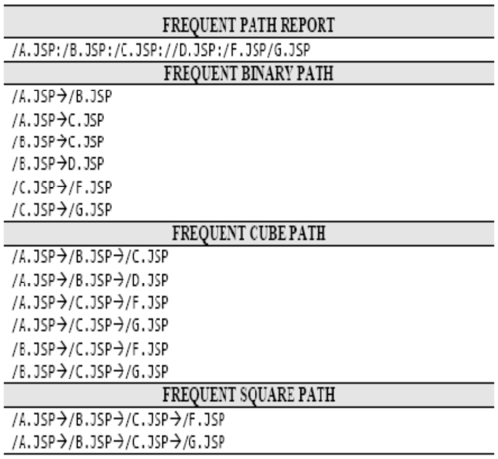
TIME	IP	URL	REF TO	REF FROM	AGENT + OS
08.25.55	123.2.3.2	/UrlTrckingSystem	/A.jsp	/B.jsp	Internet Explorer Windows Vista
08.25.57	123.2.3.2	/UrlTrckingSystem	/B.jsp	/C .jsp	Internet Explorer Windows Vista
08.25.58	123.2.3.2	/UrlTrckingSystem	/C.jsp	/G.jsp	Internet Explorer Windows Vista
08.26.01	123.2.3.2	/UrlTrckingSystem	/G.jsp	END	Internet Explorer Windows Vista
08.54.57	223.2.3.2	/UrlTrckingSystem	/A.jsp	/B.jsp	Internet Explorer Windows Vista
08.54.58	223.2.3.2	/UrlTrckingSystem	/B.jsp	/C.jsp	Internet Explorer Windows Vista
08.55.00	223.2.3.2	/UrlTrckingSystem	/C.jsp	/G.jsp	Internet Explorer Windows Vista
08.55.04	223.2.3.2	/UrlTrckingSystem	/G.jsp	END	Internet Explorer Windows Vista

Table 3: IP+ AGENT Wise Report



TIME	IP	URL	REF TO	REF FROM	AGENT + OS
21.14.01	127.0.0.1	/UrlTrckingSystem	/B.jsp	/D.jsp	Internet Explorer Windows Vista
21.41.38	127.0.0.1	/UrlTrckingSystem	/D.jsp	/C.jsp	Firefox Windows Vista
19.25.29	257.0.01	/UrlTrckingSystem	/C.jsp	/G.jsp	Chrome Windows Vista
19.58.17	257.0.01	/UrlTrckingSystem	/G.jsp	/C.jsp	Chrome Windows Vista
19.58.13	257.0.01	/UrlTrckingSystem	/C.jsp	/A.jsp	Chrome Windows Vista
19.58.11	257.0.01	/UrlTrckingSystem	/A.jsp	/B.jsp	Chrome Windows Vista
19.41.59	257.0.01	/UrlTrckingSystem	/B.jsp	/A.jsp	Chrome Windows Vista
19.16.10	257.0.01	/UrlTrckingSystem	/A.jsp	/C.jsp	Firefox Windows Vista

Table 4: Frequent Path Report



FREQUENT PATH REPORT
/A.JSP:/B.JSP:/C.JSP://D.JSP:/F.JSP/G.JSP
FREQUENT BINARY PATH
/A.JSP→/B.JSP
/A.JSP→C.JSP
/B.JSP→C.JSP
/B.JSP→D.JSP
/C.JSP→/F.JSP
/C.JSP→/G.JSP
FREQUENT CUBE PATH
/A.JSP→/B.JSP→/C.JSP
/A.JSP→/B.JSP→/D.JSP
/A.JSP→/C.JSP→/F.JSP
/A.JSP→/C.JSP→/G.JSP
/B.JSP→/C.JSP→/F.JSP
/B.JSP→/C.JSP→/G.JSP
FREQUENT SQUARE PATH
/A.JSP→/B.JSP→/C.JSP→/F.JSP
/A.JSP→/B.JSP→/C.JSP→/G.JSP

VI. CONCLUSION

This research work is introducing all the data pre-processing activities, so that data can be prepared for applying the algorithm. For the discovery of most frequent associated pages, Association Mining Rule (Apriori algo.) is used, so that most frequent navigation pages can be retrieved. Pattern analyzer can use these patterns for performing some important applications like system Improvement by page caching, site modification, page personalization, , website restructuring etc. the brief description about these applications are as follows

A. *System Improvement*

Performance and other service quality attributes axe crucial to user satisfaction from services such as databases, networks etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to under- standing Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission, load balancing, or data distribution. Security is an acutely growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins etc.

B. *Site Modification*

The attractiveness of a Web site, in terms of both content and structure, is crucial to many applications, e.g. a product catalog for e-commerce. Web usage mining provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign decisions. While the results of any of the projects could lead to re-designing the structure and content of a site, the adaptive Web site project focuses on automatically changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages is used to determine which pages should be directly linked.

C. Business Intelligence

Information on how customers are using a Web site is critical information for marketers of e-tailing businesses. Many web miners have presented a knowledge discovery process in order to discover marketing intelligence from Web data. They define a Web log data hypercube that will consolidate Web usage data along with marketing data for e-commerce applications. They identified four distinct steps in customer relationship life cycle that can be supported by their knowledge discovery techniques, customer attraction, customer retention, cross sales and customer departure. There are several commercial products, such as SurfAid [11], Accrue [1], Net- Genesis [7], Aria [3], Hitlist [5], and WebTrends [13] that provide Web traffic analysis mainly for the purpose of gathering business intelligence. Accrue, NetGenesis, and Aria axe designed to analyze e-commerce events such as products bought and advertisement click-through rates in addition to straight forward usage statistics. Accrue provides a path analysis visualization tool and IBM's SurfAid provides OLAP through a data cube and clustering of users in addition to page view statistics.

REFERENCES

[1] Kim,Wooju. Song,Yong U. Hong ,June S. "Web enabled expert systems using hyperlink-based inference". Expert Systems with Applications. 2004. pp1-13.

[2] Michele Facca, Federico. Luca Lanzi, Pier. "Mining interesting knowledge from web logs a survey". Data & Knowledge Engineering. Accepted for publication. 2004.

[3] Hsu, Jeffrey. "Data mining trends and developments The Key Data Mining Technologies and Applications for the 21st Century". Proc. of ISECON 2002.

[4] Chakrabarti, Soumen. "Mining the web discovering knowledge form hypertext data". San Francisco, CA. Morgan Kaufmann Publishers An imprint of Elsevier Science 2003. pp 1-13.

[5] Arotaritei, Dragos. Mitra, Sushmita. "Web mining a survey in the fuzzy framework". Fuzzy Sets and Systems vol. 148, 2004. pp 5–19.

[6] Larsen, Jan. Lars Hansen, Kai. Szymkowiak Have, Anna. Christiansen,Torben. Kolenda, Thomas. "Webmining learning from the World Wide Web". Computational Statistics & Data Analysis. 38. 2002. pp 517–532.

[7] Eirinaki, Magdalini. Vazirgiannis, Michalis. "Web Mining for Web[19]. Gatetrade.net. Information on gatetrade.net and some of their solutions Marketplace Personalization". ACM Transactions on Internet Technology vol. 3, no.1, 2003. pp 1–27.

[8] DeYoung, Colin G. Spence, Ian. “Profiling information technology users en route to dynamic personalization". Computers in Human Behavior.. Vol. 20. 2004. pp 55–65.

[9] Rakesh Agrawal, Sakti Ghosh, Tomasz Imielinski, Bala Iyer, and Arun Swami,An Interval Classi_er for Database Mining Applications", VLDB-92, Vancouver, British Columbia, 1992, pp 560-573.

[10] Dina Bitton, Bridging the Gap Between Database Theory and Practice", Cadre Technologies, Menlo Park, 1992.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Wadsworth, Belmont, 1984.

[12] B. Falkenhainer and R. Michalski, Integrat-ing Quantitative and Qualitative Discovery The ABACUS System", Machine Learning, 1(4) 36701.

[13] M. Kokar,Discovering Functional Formulas through Changing Representation Base", Proceedings of the Fifth National Conference on Artificial Intelligence, 1986, 455.

[14] P. Langley, H. Simon, G. Bradshaw, and J. Zytkow, Scientific Discovery Computational Explorations of the Creative Process, The MIT Press, Cam- bridge, Mass., 1987.

[15] Heikki Mannila and Kari-Jouku Raiha, Dependency Inference", VLDB-87, Brighton, England, 1987.

[16] J. Ross Quinlan, \Induction of Decision Trees", Machine Learning, 1, 1986.

[17] G. Piatetsky-Shapiro, Discovery, Analysis, and Presentation of Strong Rules, 229-248.

[18] G. Piatetsky-Shapiro (Editor), Knowledge Discovery in Databases, AAAI-MIT Press, 1991.

[19] L.G. Valiant,A Theory of Learnable", CACM, 27,1134.

[20] L.G. Valiant, Learning Disjunctions and Conjunctions", IJCAI-85, Los Angeles, 1985, 560-565.

❑❑❑