ISSN:0975-9646



# Evolving User Profiles in Dynamic Web Sites

G BHANU*, A .V. RAMANA#

**Department of computer science. GMRIT, Rajam, (A.P.), India*
*# Department of Information Technology (IT). GMRIT, Rajam, (A.P.), India*

***Abstract* – Most of the companies have the web sites for their business. Most of the customers of the organization register their details as user profiles. These user profiles have the personal details and their interesting habits of the customer. When the customer visits our web sites the log file is created in the server. By associating the user profiles and web log file we can find out the frequently visited customers. From the frequently visited customer, we can find out when they are visited by clustering the user profiles with web log files. In our work we explain how to understand "who" the users were, "what" they looked at, and "how their interests changed with time, "when" they visit all of which are important questions in Customer Relationship Management (CRM). In our study we present clustering the user profiles. We also describe how the discovered user profiles can be enriched with explicit information.**

**Key Terms- Web usage mining, Web logs, User profiles, Click streams**

## 1 INTRODUCTION

Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. The World Wide Web is an invaluable tool for researchers, information engineers, health care companies and practitioners for retrieving knowledge. However, the extraction of information from web resources is a difficult task due to their unstructured definition, their untrusted sources and their dynamically changing nature. Web mining technologies are the right solutions for knowledge discovery on the Web. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

Web content mining is a process of extracting useful information from the web content. Google or Yahoo search that we do, and the resultant links listing page we get is an example of content mining. The search is done by search engine which includes a spider. The search can be for text or image or multimedia.

Web structure mining is done at the hyper link level. A relevant example can be Google's Page rank. HITS and Page rank are applied web structure mining uses. Web structure mining, is a tool used to identify the relationship between Web pages linked by information or direct link connection.

Web usage mining process involves the log time of pages. The world's largest portal like yahoo, msn etc., needs a lot of insights from the behaviour of their users' web visits. Without this usage reports**,** it will be difficult to structure their monetization efforts.

Customer Relationship Management (CRM) can use data from outside an organization to allow an understanding of its customers on an individual basis or on a group basis such as by forming customer profiles. An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective CRM. Mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files.

Web usage mining has several applications in e-business, including personalization, traffic analysis, and targeted advertising. The development of graphical analysis tools such as Webviz popularized Web usage mining of Web transactions. The main areas of research in this domain are Web log data preprocessing and identification of useful patterns from this preprocessed data using mining techniques. Most data used for mining is collected from Web servers, clients, proxy servers, or server databases, all of which generate noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. Some of the research is available for data preprocessing into subtasks and note that the final outcome of preprocessing should be data that allows identification of a particular user's browsing pattern in the form of page views, sessions, and clickstreams. Clickstreams are of particular interest because they allow reconstruction of user navigational patterns. Some of the research provide Web logs for usage mining and suggests novel ideas for Web log indexing. Such preprocessed data enables various mining techniques.

## 2 DATA SOURCES OF WEB MINING

When a user agent (Internet Explorer, Mozilla, Netscape, etc.) hit an URL in a web server's domain, the information related to that operation is recorded in that web server's access log file. An access log file contains its information in Common Log file Format (CLF). In CLF, each client request for any URL corresponds to a record in access log file. Each CLF record is a tuple containing seven attributes that are given below:

- Client machine's IP address
- Access date and time
- Request method (GET or POST),
- URL of the page accessed
- Transfer protocol (HTTP 1.0, HTTP 1.1,)
- Success of return code
- Number of bytes transmitted

User session reconstruction, IP address, request time, and URL are the only information needed from the user web access log in order to obtain users' navigation paths.

Important data for effective CRM and E-business listed below.

1. Server data. Customers will leave their respective log data on Web servers when visiting these sites. These log data are usually stored in server in the form of document files, generally including server logs, error logs, and cookies logs and so on.

2. Query data. Query data is a typical kind of data produced on e-business Web servers. For example, customers stored on line perhaps search for some

Products and some advertisement information, and this query information is just related to the server log through cookies or register information.

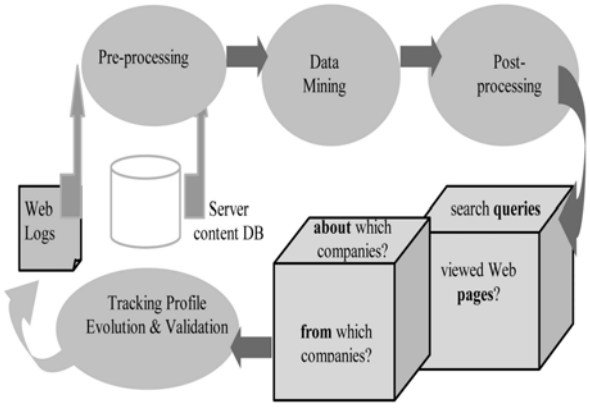
3. On-line market data. The major part of the data is about e-business websites, purchases of customers, merchandises and so on, which is stored in traditional relational databases.

4. Web pages. Web pages include HTML or XML pages, which comprise texts, pictures, audio, and video and so on.

5. Hyperlinks between Web pages. It is an important resource, which indicates the relation of hyperlinks between pages.

6. Customer registration information**.** It is the information that customers have to input via a Web page and submit to the server. It is usually about the demographic characteristics of users. In Web mining, customer registration information should be integrated with visiting logs to improve the accuracy of data mining and produce more knowledge about customers.

## 3. WEB USAGE MINING PROCESS DIAGRAM




## 4. IMPROVE CUSTOMER RELATIONSHIP MANAGEMENT USING WWW

The core of CRM, on one hand, is to discover potential markets and customers by collecting effective data. Collecting customers and their activities; and on the other hand, is to meet customers' needs and to realize customers' lifetime value by improving the customer service and a deep analysis of customers. CRM provides traditional enterprises with management systems and technical artifices for their survival in the network economy era. Product-oriented model to customers oriented.

### 4.1. APPLICATION OF WEB MINING TO CRM

Web mining can help enterprises identify customers' features, which enables enterprises to provide targeted services for customers. Web mining used in CRM of ebusiness has several aspects, such as the acquisition and maintenance of customers, identification of the value of customers, analysis of customers' satisfaction, and improvement of site structure and so on.

With Web mining, we can understand the dynamic behavior of visitors and optimize the operation mode of e-business websites. We can put the large number of customers acquired into different categories and provide personalized services for customers from different categories to improve the satisfaction of customers and to maintain old customers consequently. We can determine which category a new visitor belongs to and whether he or she is potentially profitable by analyzing the records of the pages that the visitor browses, so that we can deal with different customers. Customers with similar browsing behaviors are grouped together and their common features are extracted, so that customers can be clustered, which can help e-business enterprises to better understand customers' interests, consuming habits and trends, predict customers' needs, recommend specific commodities to them accordingly and realize crossselling . The trading volume and the rate of successful trades will be increased and the efficiency of distribution will be improved.

The structure and content of the site is the key to customers' interest. With the discovery of association rules, we can rearrange them dynamically for different customers, and put together the commodities with some degree of support and trust to promote sales; By the means of path analysis we can identify the paths along which a category of customers visit the site frequently. These paths reflect the sequence and habits of such customers visiting pages of the site. We can hyperlink the related documents customers have visited in order that they can access their favored easily. Such a site will leave a good impression on customers.

By Web mining, we can acquire reliable market feedback to evaluate the rate of return on advertisement investment and decide whether the online marketing mode is successful or not; According to the browsing mode of visitors interested in a certain product, we can determine the location of the advertisement to increase the pertinence and the rate of return on advertisement investment and reduce companies' operating costs.

## 5. ALGORITHM

Algorithm: Clustering Algorithm for webusuage mining.

Input: List Of Urls related to different categories of web
Output: Clustered urls according to their usage

1. Take the initial cluster value is 1 for each cluster (ex: cluster1_count=0)
2. Track the URL visited by the user by using predefined list of urls.
3. Increment the count of a cluster when the URL is matched with a particular cluster.

```
If (visit_url->cluster1)
 Cluster1_count++;
Else
Check for next_cluster;
```

4. Repeat this process until all URLS are clustered.

Finally each cluster will have the similar kind of urls. From this we can easily identify the usuage of a webpage or a category.

## 6. ENRICHING USER PROFILES WITH SEARCH QUERY TERMS

In addition to the relevant URLs that are extracted from the sessions assigned to each profile, we can extract information about the explicit information need of the users in each profile from the queries that they could have typed prior to visiting the Web site when this information is available from the readily available REFERRER field in the Web log files. Hence, for each profile, we accumulate all the search phrases extracted from the REFERRER fields of the assigned user sessions. This allows us to describe each profile in terms of either a set of significant URLs or a set of explicit search query phrases and terms.

### 6.1 ENRICHING USER PROFILES WITH INQUIRING COMPANY INFORMATION

In addition to the relevant URLs that are extracted from the sessions assigned to each profile, we can extract information about which companies or organizations tend to visit the Web site and fall in this profile. We extract this information from two complementary sources: 1) by getting the
Company information that corresponds to an ID in the server content Database, where the ID is extracted from the Web log file in case the visitors register and sign in through the registration page, or 2) if the visitors did not sign in through the registration page, then an attempt is made to obtain the company affiliation from some of specialized Web services. This can be queried with an IP address via an API to determine not only what information was found relevant on the Web site but also to whom it was relevant to help support further personalization efforts.

### 6.2 ENRICHING USER PROFILES WITH QUERIED COMPANY INFORMATION


Extracted information about which companies have been inquired about by visitors in this profile in case a user searches and clicks on one of the listed companies contact information on the Web site. We parse the identity of the company from the Web log file and map it to a specific company via the server content database.

## 7 CONCLUSIONS

We presented a framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, including evolving user profiles and access patterns, dynamic Web pages, and external data describing an ontology of the Web content. A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries and inquiring and inquired companies. The choice of the period length for analysis depends on the application or can be set, depending on the cross-period validation results. Even though we did not focus on scalability, the latter can be addressed by following an approach similar to ,where Web click streams are considered as an evolving data stream, or by piping some new sessions to persistent profiles and updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.

## REFERENCES

[1] O. Nasraoui, R. Krishnapuram, H. Frigui, and A. Joshi, "Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering,"
[2] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introuduction to data mining".
[3] O. Zaiane, M. Xin, and J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs," Proc. Advances in Digital Libraries
[4] Bing Liu,"Web Data Minig-Exploring Hyperlinks, Contents and Usage Data"
[5] Arun.k.Pujari "Data Mining Techniques"
[6] Jiawei Han and Micheline Kamber "Data Mining -Concepts and Techniques"
[7] M.A. Maloof and R.S. Michalski, "Selecting Examples for Partial Memory Learning," Machine Learning, vol. 41, no. 11, pp. 27-52, 2000.
[8] T. Mitchell, R. Caruana, D. Freitag, J. McDermott, and D. Zabowski, "Experience with a Learning Personal Assistant," Comm. ACM, vol. 37, no. 7, pp. 80-91, 1994.
[9] D. Billsus and M.J. Pazzani, "A Hybrid User Model for News Classification," Proc. Seventh Int'l Conf. User Modeling (UM '99), J. Kay, ed., pp. 99-108, 1999.
[10] J. Schlimmer and R. Granger, "Incremental Learning from Noisy Data," Machine Learning, vol. 1, no. 3, pp. 317-357, 1986.
[11] G. Widmer and M. Kubat, "Learning in the Presence of Concept Drift and Hidden Contexts," Machine Learning, vol. 23, pp. 69-101, 1996.