International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 4, April - 2013

Efficient Web Page Mining for Dynamic Web Site

A.Thomas Andrews[1], P. Michel Dominique[1], J. Balaji[1]

[1]Department of Information technology,Pondicherry Engineering College,Pondicherry.

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

I.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

G. Bhanu et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, 1951-1953 1951

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.

II.Existing System:.

Typically, discovering the Web usage patterns, such as profiles or prediction models, consists of three steps: preprocessing the raw usage data, discovering patterns from the pre-processed data, and analyzing these discovered patterns. There are two primary tasks in preprocessing: data cleaning, and transaction identication also known as sessionization. Data cleaning eliminates irrelevant items such as image requests and search engine requests from the server log. The transaction identification process groups the sequences of page requests into

logical units, each of which is called a session which is the set of pages that are visited by a single user within a predefined period of time.After pre-processing, the Web sessions are used as an input to patten discovery methods
that are typically rooted in areas such as data mining, artificial intelligence, or statistics.
These discovery methods may include: Statistical Analysis, Sequential Pattern mining ,

Path Analysis, Association Rule Mining, Classification , and Clustering .After discovery, the usage patterns are analyzed to better understand and interpret them, using a variety of analysis tools from the ideas of statistics, graphics, visualization, or database querying. Examples of analysis tools can be found in Technical applications, such as robotic applications or human-computer interfaces require a fast solution to generate accurate estimations of image motion.
Fast approaches often fail to disambiguate motion and accurate solutions often don't run in real time.Many approaches fail to generate fast and accurate results.The major bottleneck in the development of a reliable biologically inspired technical system with real-time motion analysis capabilities based on this neural model is needed huge amount of memory.
The accuracy of extracting image motion to combine observations from different image locations to overcome ambiguities which inherently cannot be solved by purely local information. Global integration is critical since it makes it impossible to distinguish between different moving objects.

IV.Proposed System:

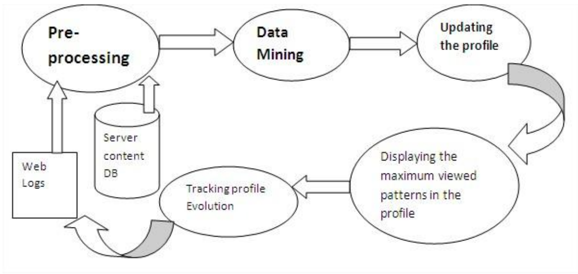
Pre-processing
Data Mining
Updating the profile
Server content DB
Web Logs
Displaying the maximum viewed patterns in the profile
Tracking profile Evolution

Fig 1.System Architecture

The web usage mining process is traditionally done in several steps with only a few variations. It starts with preprocessing the log files, discovering the usage patterns using a web usage mining algorithm, and then interpreting the discovered patterns. These steps have been used to discover usage patterns, predominately, within one specific period of time, but they can arguably be reapplied periodically, over several periods, to capture the changes in navigation patterns. However, there are some concerns using this approach, as explained below.
A) Reapplying the steps periodically can either be performed on the entire historic data, or on the new log files only. The former approach reduces the probability of discovering new trends because of their small weight compared to older trends, while the latter approach completely forgets all previous patterns which may not be reasonable or e-client, since some of these patterns might still be significant in the new period, and would have to be rediscovered again.
B) Trying to always discover the new behaviors from all the accumulated log files up to
the current period will require significant computational resources, which defines the
scalability requirement.
C)All the above approaches do not capture the changes in the usage behaviors in detail,
i.e. we do not know which URLs have changed or have become more interesting from one period to another.

The proposed framework, depicted in Figure 1, overcomes the above issues and can be
summarized as follows, assuming that we start with a set of initial (previous or seed) profiles
mined from an initial period:
1. Preprocess the new web log data to extract the current user sessions,
2. Update the previous profiles using the extracted user sessions,
3. Re-apply clustering to the distinct user sessions only (i.e. the ones not used in step 2
to update the previous profiles),
4. Post-process the distinct (new) profiles mined in step 3,
5. Combine the updated profiles with the distinct profiles to create the new seed profiles
for future periods,
6. Interpret and evaluate the discovered profiles

7. Go to step 1 as data from a new period becomes ready

HIERARCHICAL UNSUPERVISED NICHE CLUSTERING AND ITS AP-PLICA TION TO WEB USAGE MINING

We retain the principal structure of UNC (Nasraoui and Krishnapuram, 2000),except for a few differences that result from the distinct nature of the session data:

The solution space for possible session prototypes consists of binary chromosome

strings which are defined to be the binary session attribute vectors and the new Web session dissimilarity measure is used instead of the Euclidean distance to take the Web site topology in account.UNC's computational time can be significantly reduced if we perform cluster-ing in a hierarchical mode. In other words, we could cluster smaller subsets of the data using a smaller population size at multiple levels, instead of clustering the entire data set on a single level which would necessitate a larger population size.

V. Conclusion:

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.

VI.Reference:

[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.