GLOBAL JOURNAL OF ENGINEERING, DESIGN & TECHNOLOGY
(Published By: Global Institute for Research & Education)
www.gifre.org

A COGNITIVE APPROACH IN PATTERN ANALYSIS TOOLS AND TECHNIQUES USING WEB USAGE MINING

M.Gnanavel[1]& Dr.E.R.Naganathan[2]
[1.] Research Scholar, SCSVMV University, Kanchipuram,Tamil Nadu,India.
[2.] Professor & Head, Department of CSE, Hindustan University, Chennai, Tamil Nadu, India

Abstract

Web mining is an enormous quantity of information available on the web and increasingly vital role that the web plays in today's social society. An enormous amount of knowledge in respect of pattern analysis of web mining shall be provided in this paper. Web mining primarily focus with pattern discovery and analysis of usage patterns in order to provide the needs of web based applications. Web mining contemplates on the techniques that could forecast the navigational pattern of the user and sequential pattern of the user while the user interacts with the World Wide Web.

This paper is concerned with analysis of patterns tools and techniques for web mining. The pattern analysis tools distributed with different independent activities Knowledge Query Mechanism (KQM), OLAP, Visualization techniques and Intelligent Agents. Once access patterns have been identified, discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns. The WEBMINER system proposes an SQL like query mechanism for querying the discovered knowledge in the form of association rules and sequential patterns.

Keywords: Web mining, Pattern discovery and analysis, Knowledge Query Mechanism (KQM), OLAP and Visualization techniques

I .Introduction

The beginning of the World-Wide Web (WWW) has plagued the typical home computer user with an enormous flood of information. To be able to survive with the plenty of available information, users of the WWW need to rely on intelligent tools that assist them in finding, sorting, and filtering the available information. The data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of Web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular in text documents that are published on the web. Depending on the nature of the data, one can distinguish three main areas of research within the Web mining constituency;

1. **Web Content Mining:** application of data mining techniques to unstructured or semi-structured data, usually HTML-documents.
2. **Web Structure Mining:** use of the hyperlink structure of the Web as an (additional) information source.
3. **Web Usage Mining:** analysis of user interactions with a Web server (e.g., click-stream analysis) i.e. collecting data from web log records.

II. Research Implication of Web Usage Mining

Web usage mining is the type of data mining techniques that involves the automatic discovery of user access patterns from one or more web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. For any organizations often generate and collect large volumes of data in their day to day activity or daily operations. Most of this information is usually generated automatically by web servers and collected in server access logs. Other sources of user information include referrer logs, which contain information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts. Analyzing such data can help these organizations to determine the life time value of customers, marketing strategies across products, and effectiveness of advancement campaign among other things. Analysis of server access logs and user registration data can also provide valuable information on how to better structure a Web site in order to create a more effective presence for the organization. In organizations using intranet technologies, such analysis can shed light on more effective management of workgroup communication and organizational infrastructure. Finally, for organizations that sell advertising on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users.

Most of the existing Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools, for example, it is possible to determine the number of accesses to the server and the individual files within the organization's Web space, the times or time intervals of visits, and domain names and the URLs of users of the Web server. However, in general, these tools are designed to deal handle low to moderate traffic servers, and furthermore, they usually provide little or no analysis of data relationships among the accessed files and directories within the Web space.

III. Web Usage Mining Architecture

While extracting simple information from web logs is easy, mining complex structural information is very challenging. Data cleaning and preparation constitute a very significant effort before mining can even be applied. The relevant data challenges include: elimination of irrelevant information such as image files and CGI scripts, user identification, user

session formation, and incorporating temporal windows in the user modeling. After all this pre-processing, one is ready to mine the resulting database. We have developed a general architecture for Web usage mining. The WebMiner is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main stages. The first stage includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes pre-processing, transaction identification, and data integration components. The second stage includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web mining process is depicted in Figure1.

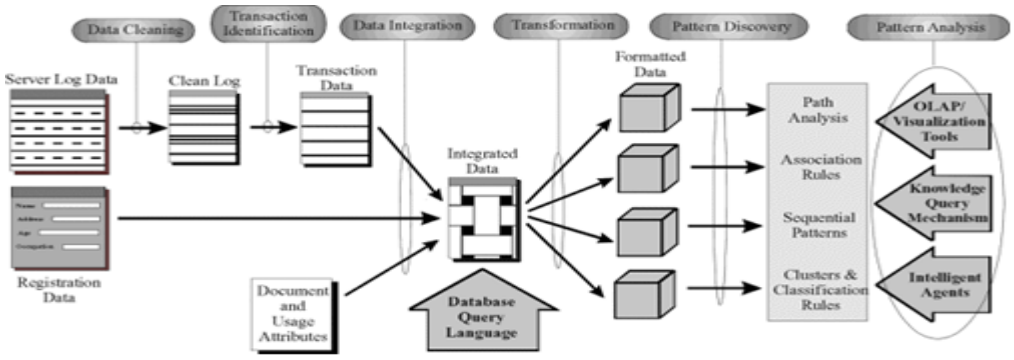

Figure1: Web Usage Mining Architecture

Data cleaning is the first step performed in the Web usage mining process. Any of the cleaning techniques can be used to pre-process a given Web server log. Currently, the WebMiner system uses the simplistic method of checking filename suffixes. Some low-level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. Behind the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The clean server log can be thought of in two ways; either as a single transaction of many page references, or a set of many transactions each consisting of a single page reference. The goal of transaction identification is to create meaningful clusters of references for each user. Therefore, the task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. This process can be extended into multiple steps of merge or divide in order to create transactions appropriate for a given data mining task. A transaction identification module can be defined as either a merge or a divide module. Both types of modules take a transaction list and possibly some parameters as input, and output a transaction list that has been operated on by the function in the module in the same format as the input. The requirement that the input and output transaction format match allows any number of modules to be combined in any order, as the data analyst sees fit.

Access log data may not be the only source of data for the Web mining process. User registration data, for example, is playing an increasingly important role, particularly as more security and privacy conscious client-side applications restrict server access to a variety of information, such as the client user IDs. The data collected through user registration must then be integrated with the access log data. There are also known or discovered attributes of references pages that could be integrated into a higher level database schema. Such attributes could include page types, classification, usage frequency, page meta information, and link structures. While WEBMINER currently does not incorporate user registration data, various data integration issues are being explored in the context of Web usage mining. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data-mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns.

Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints. The emerging data mining tools and systems lead naturally to the demand for a powerful data mining query language, on top of which many interactive and flexible graphical user interfaces can be developed. Such a query mechanism can provide user control over the data mining process and allow the user to extract only relevant and useful rules. In WebMiner, a simple query mechanism has been implemented by adding some primitives to an SQL-like language. This allows the user to provide guidance to the mining engine by specifying the patterns of interest.

IV. Pattern Analysis Tools and Techniques

Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns. Examples of such tools include the WebViz system for visualizing path traversal patterns. Others have proposed using OLAP techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from server access logs. The WebMiner system proposes an SQL like query mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns). Pattern Analysis is a final stage of the whole Web usage mining. The main motto of this process is to remove irrelevant patterns rules and to extract the interesting patterns or rules from the output of the pattern discovery process. These queries require analysis of the structure of hyperlinks as well as the contents of the page which is done with the help of some analysis tools and methodologies. (Figure 2)

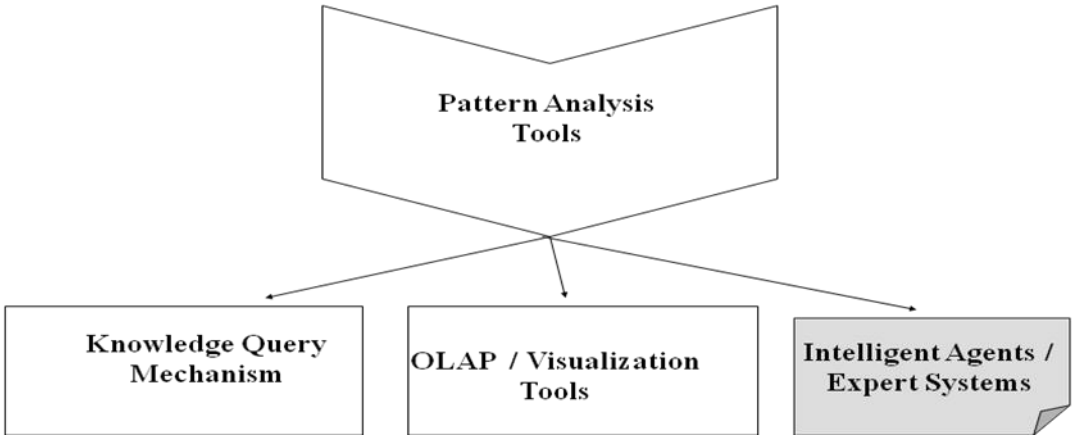

Figure 2: Pattern analysis tools classification methods:

There are three important techniques used in the pattern analysis; i) Knowledge Query Mechanism ii) OLAP / Visualization Tools and iii) Intelligent Agents / Expert Systems.

i) **Knowledge Querying Mechanisms**: One of the reasons attributed to the great success of relational database technology has been the existence of a high-level, declarative, query language, which allows an application to express what conditions must be satisfied by the data it needs, rather than having to specify how to get the required data. Given the large number of patterns that may be mine d, there appears to be a definite need for a mechanism to specify the focus of the analysis. First, constraints may be placed on the database to restrict the portion of the database from which to mine for. Second, querying may be performed on the knowledge that has been extracted by the mining process, in which case a language for querying knowledge rather than data is needed. SQL-like Knowledge Query Mechanism,SELECT association-rules (A*B*C*) FROM "rules.out" WHERE time >= 970101 AND domain = "edu" AND support >= 0.01 AND confidence >.90.

ii) **OLAP / Visualization Tools**; The term On-Line Analytic Processing – (OLAP) refers to technology that allows users of multidimensional databases to generate on-line descriptive or comparative summaries ("views") of data and other analytic queries. On-line Analytical Processing (OLAP) is emerging as a very powerful paradigm for strategic analysis of databases in business settings. Some of the key characteristics of strategic analysis include: very large data volume, explicit support for the temporal dimension, support for various kinds of information aggregation and long-range analysis in which overall trends are more important than details of individual data items.

Visualization has been used very successfully in helping people understand various types of phenomena, both real and abstract. Hence it is a natural choice for understanding the behavior of web users. According Growth the visualization is simply the graphical presentation of data. Software used in Web Usage Mining: The Web Miner is an introduces a general architecture for Web usage mining, automatically discovering association rules and sequential patterns from server access logs and proposes an SQL-like query mechanism for querying the discovered knowledge in the form of association rules and sequential patterns. A framework for Web mining, the applications of data mining and knowledge discovery techniques, association rules and sequential patterns, to Web data:

Pattern Analysis Tools	Features
Webviz	Analyze the patterns and provides them in the form of graphical patterns
Naviz	Visualization tools that combines 2D graph of visitor access and graphing of repeated pages.
Webminer	Mined the useful pattern and provides the user specific information

V. Implementation of Association rules using Apriori Algorithm

Association rules using Apriori algorithm: Association rule discovery is to find the relationships between the different items in a data base. It is normally express in the form X [??] Y, where X and Y are sets of attributes of the dataset which implies that transactions that contain X also contain Y.
In Association rules using Apriori algorithm in the context 40% of clients who accessed the Web page with URL /company/products/product1.aspx, also accessed /company/products/product2.aspx.

Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).A large computer laptop show room tracks sales data by Stock-keeping Unit (SKU) for each item, and thus is able to know what items are typically purchased together. Apriori is a moderately efficient way to build a list of frequent purchased item pairs from this data. Let the database of transactions consist of the sets {1,2,3,4}, {1,2}, {2,3,4}, {2,3}, {1,2,4}, {3,4}, and {2,4}. Each number corresponds to a product such as "Sony" or "Dell". The first step of Apriori is to count up the frequencies, called the supports, of each member item separately: This table explains the working of apriori algorithm.(Table 1).

Item	Support
1	3
2	6
3	4
4	5

Table 1: Sales data items by SKU

We can define a minimum support level to qualify as "frequent," which depends on the context. For this case, let min support = 3. Therefore, all are frequent. The next step is to generate a list of all 2-pairs of the frequent items. Had any of the above items not been frequent, they wouldn't have been included as a possible member of possible 2-item pairs. In this way, Apriori prunes the tree of all possible sets. In next step we again select only these items (now 2-pairs are items) which are frequent: (Table 2)

Item	Support
{1,2}	3
{2,3}	3
{2,4}	4
{3,4}	3

Table 2.Pairs of frequent data items generated.

To generate a list of all 3-triples of the frequent items (by connecting frequent pair with frequent single item). In the example, there are no frequent 3-triples. Most common 3-triples are {1,2,4} and {2,3,4}, but their support is equal to 2 which is smaller than our min support.As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length $k-1$. Then it prunes the candidates which have an infrequent sub pattern. (Figure. 3)

Algorithm Apriori(*T*)

$C_1 \leftarrow$ init-pass(T);
$F_1 \leftarrow \{f \mid f \in C_1, f.count/n \geq minsup\}$; // n: no. of transactions in T
for ($k = 2$; $F_{k-1} \neq \emptyset$; k++) do
  $C_k \leftarrow$ candidate-gen(F_{k-1});
  for each transaction $t \in T$ do
    for each candidate $c \in C_k$ do
      if c is contained in t then
        c.count++; end
  end
  $F_k \leftarrow \{c \in C_k \mid c.count/n \geq minsup\}$ end
return $F \leftarrow \bigcup_k F_k$;

Figure 3: Apriori Algorithm

iii) Intelligent Agents: An intelligent agent (IA's or i-agents) are a computer programs that assist the user with their tasks and its capable of flexible autonomous action in some environment. i- agents may be on the Internet or they can be used any types of architectures. The idea of intelligent software agents captures the popular imagination. Tell the agent what you want done, set it free, and wait for it to return results sounds too good to be true. In the context of this research, however, the tasks that we are primarily concerned with include reading, filtering and sorting and maintaining information.

VI. Current Issues on Pattern Analysis Tools

There are some current issues on pattern analysis in the research i) an identification of exact user not possible and Exact sequence of pages referenced by a user not possible due to caching. ii) Session not well defined and Security, privacy, and legal .iii) New type of knowledge. iv) Improved mining algorithms v) Incremental web mining.

VII. Conclusion and Future Work

When the web mining and its usage continues to cultivate, so too cultivate the opportunity to analyze web data and extract all manner of useful knowledge from it. The past ten years have seen the emergence of web mining as a rapidly growing area, due to the efforts of the research people as well as various organizations that are practicing it. In this paper we analyzed the research area of Web mining, focusing on the pattern analysis tools and techniques of Web mining. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research. We have tried to give a obvious understanding of the pattern analysis tools and techniques. Web usage patterns and web mining can be foundation for great deal of future research.

References

[1] K. J. Cios, W. Pedrycz, R. W. Swiniarski, and L. A. Kurgan, "Data mining: A knowledge discovery approach," Springer, New York, 2007.
[2] J. T. Tou and R. C. Gonzalez, "Pattern recognition principles," Addison-Wesley, London, 1974.
[3] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern classification," Wiley, 2001.
[4] V. S. Tseng and S. C. Yang, "Mining multi-level association rules from gene expression profiles and gene ontology," in Proceedings IEEE Workshop Life Science Data Mining (held with IEEE ICDM), UK, November 2004.
[5] R. J. Kuo, S. Y. Lin, and C. W. Shih, "Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan," Expert Systems with Applications, Vol. 33, pp. 794-808, 2007.
[6] M. L. Antonie, O. R. Zaiane, and A. Coman, "Application of data mining techniques for medical image classification," in Proceedings Second International Workshop on Multimedia Data Mining, pp. 94-101, 2001.
[7] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, and R. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, Vol. 32, No. 2, pp. 71-83, 2004.
[8]Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
[9]Dr.A.Venumahav,Web Usage Mining has pattern discovery,IJER Volume 4,Issue 11,November 2013.
[10].Kamika Chaudhary,Santhoshkumar Gupta,Web Usage Mining Tools and Techniques-A Survey,IJSER Volume 4,Issue 6,June 2013.