“Survey of Various Techniques for WebUsage Minning”

Navjot Kaur[1], Dr. Himanshu Aggarwal[2]
Department of Computer Engineering, University
College of Engineering,Punjabi University Patiala(India)
[1]navjot_anttal@yahoo.co.in, [2]himansu.pup@gmail.com

Abstract: *Web usage mining is the area of data mining which deals with the discovery and analysis of usage patterns from Web data, specifically web logs, in order to improve web based applications. Web usage mining consists of three phases, pre-processing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use these information for the specific needs. This paper presents several data preparation techniques in order to identify unique users and user sessions. It also provides a detailed information of various Web usage mining areas to work in. The survey of the existing work is also provided. Finally, a brief overview of the applications of Webusage mining (Letizia, WebSIFT , Adaptive Websites).*

Keywords: *World Wide Web, data mining, web usage mining, information retrieval, information extraction.*

I. Introduction

Web mining [1] is the application of data mining techniques to discover patterns from the Web. The World Wide Web is an immense source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world. Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web.

According to analysis targets, web mining can be divided into three different types[2], which are Web content mining, Web structure mining and Web usage mining. The World Wide Web is an immense source of data that can come either from the Web Content, represented by the billions of pages publicly available or from the Web usage represented by the log information daily collected by all the servers around the world.

Web Usage Mining is often regarded as a part of the Business Intelligence in an organization rather than the technical aspect. It is used for deciding business strategies through the efficient use of Web Applications. It is also crucial for the Customer Relationship Management (CRM) as it can ensure customer satisfaction as far as the interaction between the customer and the organization is concerned.

The major problem with Web Mining in general and Web Usage Mining in particular is the nature of the data they deal with. With the upsurge of Internet in this millennium, the Web Data has become huge in nature and a lot of transactions and usages are taking place by the seconds. Apart from the volume of the data, the data is not completely structured. It is in a semi-structured format so that it needs a lot of preprocessing and parsing before the actual extraction of the required information.

Web Usage mining is the application of data mining techniques to large Web data repositories in order to produce results that can be used in the designing tasks of any website. Some of the data mining algorithms that are commonly used in Web Usage Mining are association rule generation, sequential pattern generation, and clustering. Association Rule mining techniques discover unordered correlations between items found in a database of transactions. In the context of Web Usage Mining a transaction is a group of Web page accesses, with an item being a single page access. A Web usage mining system performs five major tasks: Data gathering, Data preparation, Navigation pattern discovery, Pattern analysis and visualization, and Pattern applications.

II. Data Sources

In Web Usage Mining, data can be collected in server logs (Access Log, Referrer Log, Agent Log), proxy logs, Web clients or obtained from an organization's database. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation.

There are many kinds of data that can be used in Web Mining.

Content: The visible data in the Web pages or the information which was meant to be imparted to the users. A major part of it includes text and graphics (images). Typical applications are Content-based categorization and content based ranking of Web pages.

Structure: Data which describes the organization(or Structure) of the website. Source data mainly consist of the structural information present in web pages(e.g., links to other pages); It is divided into two types. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kind of inter-page structure information are the hyper-links used for site navigation. Applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure and reverse engineering of Web site models.

Usage: Data that describes the usage patterns of Web pages. Data is extracted from server log files Usage pattern, such as name and IP addresses of the remote host, page references, and the date and time of accesses and various other information depending on the log format(Common Log Format, Extended Log Format, LogML). Applications are those based on user modelling techniques, such as Web personalization, adaptive Web sites and user modelling.

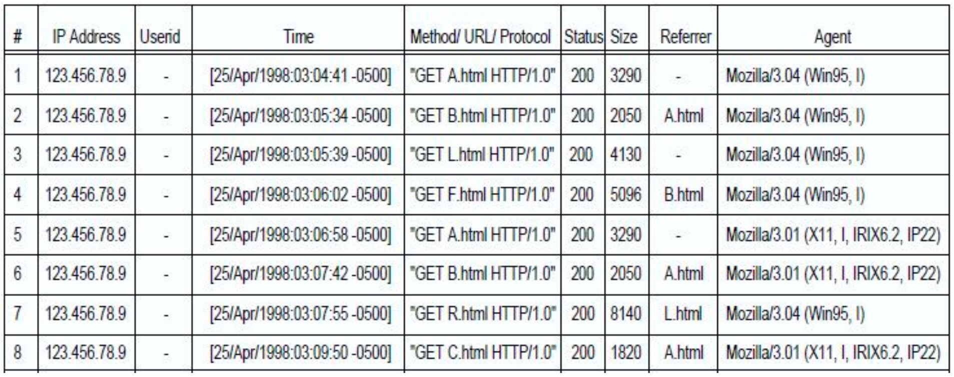
Web Usage Mining applications are based on data collected from following main sources:

A. Web Server Logs

These are logs which maintain a history of page requests. The W3C maintains a standard format for web server log files, but other proprietary formats exist. More recent entries are typically appended to the end of the file. As shown in Fig1 information about the request[3], include client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are typically added. These data can be combined into a single file, or separated into distinct logs, such as an access log, error log, or referrer log. However, server logs typically do not collect user-specific information. These files are usually not accessible to general Internet users, only to the webmaster or other administrative person. When exploiting log information from Web servers, the major issue is the identification of users_sessions, i.e., how to group all the users_ page requests so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use cookies to track down the sequence of users_ page requests (see [4] for an overview of cookie standards). If cookies are not available, various heuristics [5] can be employed to reliably identify users_ sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the back button is not tracked at the server level [6]. Apart from Web logs, users_ behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of users_ sessions is still an issue, but the use of packet sniffers provides some advantages [7]. In fact: (i) data are collected in real time; (ii) information coming from different Web servers can be easily merged together into a unique log; (iii) the use of special buttons (e.g., the stop button) can be detected so to collect information usually unavailable in log files. Notwithstanding the many advantages, packet sniffers are rarely used in practice. Packet sniffers raise scalability issues on Web servers with high traffic [7].

B. Proxy Server Logs

A Web proxy is a caching mechanism which lies between client browsers and Web servers side. Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. This may serve as a data source to discover the usage pattern of a group of anonymous users, sharing a common proxy server. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers. Even in this case, session reconstruction is difficult and not all users_ navigation paths can be identified. However, when there is no other caching between the proxy server and the clients, the identification of users_ sessions is easier.



#	IP Address	Userid	Time	Method/ URL/ Protocol	Status	Size	Referrer	Agent
1	123.456.78.9	-	[25/Apr/1998:03:04:41 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.04 (Win95, I)
2	123.456.78.9	-	[25/Apr/1998:03:05:34 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.04 (Win95, I)
3	123.456.78.9	-	[25/Apr/1998:03:05:39 -0500]	"GET L.html HTTP/1.0"	200	4130	-	Mozilla/3.04 (Win95, I)
4	123.456.78.9	-	[25/Apr/1998:03:06:02 -0500]	"GET F.html HTTP/1.0"	200	5096	B.html	Mozilla/3.04 (Win95, I)
5	123.456.78.9	-	[25/Apr/1998:03:06:58 -0500]	"GET A.html HTTP/1.0"	200	3290	-	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
6	123.456.78.9	-	[25/Apr/1998:03:07:42 -0500]	"GET B.html HTTP/1.0"	200	2050	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)
7	123.456.78.9	-	[25/Apr/1998:03:07:55 -0500]	"GET R.html HTTP/1.0"	200	8140	L.html	Mozilla/3.04 (Win95, I)
8	123.456.78.9	-	[25/Apr/1998:03:09:50 -0500]	"GET C.html HTTP/1.0"	200	1820	A.html	Mozilla/3.01 (X11, I, IRIX6.2, IP22)

Fig 1, Sample Web Server Log.

C. *Client side*

Usage data can be tracked also on the client side by using Javascript, Java applets, or even modified browsers. These techniques avoid the problems of users_ sessions identification and the problems caused by caching (like the use of the back button). In addition, they provide detailed information about actual user behaviors [6]. However, these approaches rely heavily on the users cooperation and raise many issues concerning the privacy laws, which are quite strict. Client-side collection scores over server-side collection because it reduces both the bot and session identification problems.

III. The Webminner System

The WEBMINER system [8] divides the Web Usage Mining process into three main parts, as shown in Figs1. Input data consists of the three server logs - access, referrer, and agent, the HTML files that make up the site, and any optional data such as registration data or remote agent logs. The first part of Web Usage Mining, called preprocessing, includes the domain dependent tasks of data cleaning, user identification, session identification, and path completion. Data preprocessing has a fundamental role in Web Usage Mining application. The preprocessing of Web logs is usually complex and time demanding .It comprises of several steps: data collection, data pretreatment, knowledge discovery and pattern analysis.

Data collection is the first step of web usage mining, the data authenticity and integrality will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data. At present, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

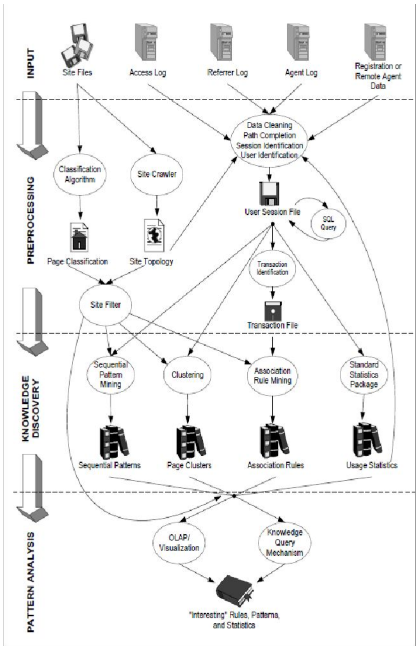


Fig. 2, Architecture of Web inner

A. Data pre-processing

Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion.

1) Data Cleaning: The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed:

-The records of graphics, videos and the format information The records have filename suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record;

-The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information.

2) *User and Session Identification:* The task of user and session identification is find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is

proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

- The different IP addresses distinguish different users;

- If the IP addresses are same, the different browsers and operation systems indicate different users;

- If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;

- The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

3) Path completion: Another critical step in data pre-processing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data pre-processing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to the mined.

B. Knowledge Discovery

Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have Association Rules, Clustering, Classification, Sequential Patterns , Dependency Modeling ,the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present**.**

C. Pattern analysis

Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

IV. WEB Usage Mining Areas

As shown in Figures 3, usage patterns extracted from Web data have been applied to a wide range of applications. The Letizia, WebSIFT and Adaptive Website project is discussed in last section.

A. Personalization

Personalization the Web experience for a user is the holy grail of many Web-based applications, e.g. individualized marketing for e-commerce. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage behavior is very attractive to many applications, e.g. cross-sales and up-sales in e-commerce. Web usage mining is an excellent approach for achieving this goal, as illustrated in Existing recommendation systems, do not currently use data mining for recommendations, though there have been some recent proposals [9].

TheWebWatcher, SiteHelper, Letizia, and clustering work by Mobasher et. al. [10] and Yan et. al. have all concentrated on providing Web Site personalization based on usage information. Web server logs were used by Yan et. al. to discover clusters of users having similar access patterns. The system proposed in consists of an offline module that will perform cluster analysis and an online module which is responsible for dynamic link generation of Web pages. Every site user will be assigned to a single cluster based on their current traversal pattern. The links that are presented to a given user are dynamically selected based on what pages other users assigned to the same cluster have visited.

The SiteHelper project learns a users preferences by looking at the page accesses for each user. A list of keywords from pages that a user has spent a significant amount of time viewing is compiled and presented to the user. Based on feedback about the keyword list, recommendations for other pages within the site are made. WebWatcher follows" a user as he or she browses the Web and identifies links that are potentially interesting to the user. The WebWatcher starts with a short description of a users interest. Each page request is routed through the WebWatcher proxy server in order to easily track the user session across multiple Web sites and mark any interesting links.

WebWatcher learns based on the particular user's browsing plus the browsing of other users with similar interests. Letizia is a client side agent that searches the Web for pages similar to ones that the user has already viewed or bookmarked. The page recommendations in [10] are based on clusters of pages found from the server log for a site. The system recommends pages from clusters that most closely match the current session. Pages that have not been viewed and are not directly linked from the current page are recommended to the user. [11] attempts to cluster user sessions using a fuzzy clustering algorithm. [11] allows a page or user to be assigned to more than one cluster.

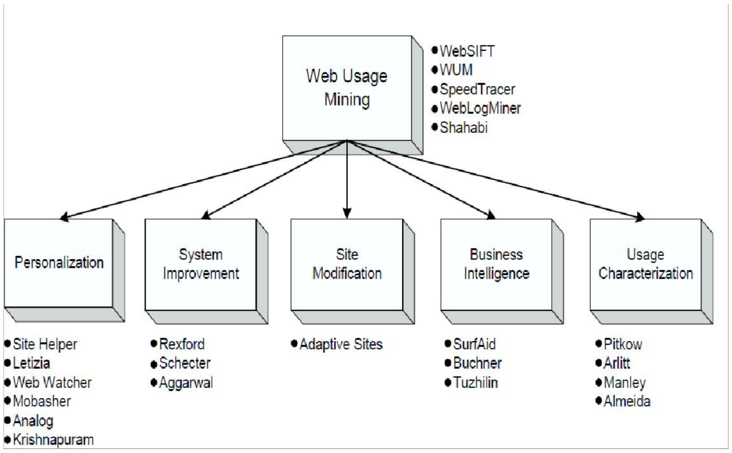


Fig. 3, Major Application Areas for Web Usage Mining

B. System Improvement

Performance and other service quality attributes are 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 understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission [14], 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 [16]. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc.

Almeida et al. [12] propose models for predicting the locality, both temporal as well as spatial, amongst Web pages requested from a particular user or a group of users accessing from the same proxy server. The locality measure can then be used for deciding pre-fetching and caching strategies for the proxy server. The increasing use of dynamic content has reduced the benefits of caching at both the client and server level. Schechter et. al. [15] have developed algorithms for creating path profiles from data contained in server logs. These profiles are then used to pre-generate dynamic HTML pages based on the current user profile in order to reduce latency due to page generation.

C. 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 (SCML algorithm) [22] 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.

D. Business Intelligence

Information on how customers are using a Web site is critical information for marketers of e-tailing businesses. Buchner et al [18] 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, Accrue, Net-Genesis, Aria, Hitlist, and WebTrends that provide Web traffic analysis mainly for the purpose of gathering business intelligence. Accrue, NetGenesis, and Aria are 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. Padmanabhan et. al. use Web server logs to generate beliefs about the access patterns of Web pages at a given Web site. Algorithms for finding interesting rules based on the unexpectedness of the rule were also developed.

E. Usage Characterization

While most projects that work on characterizing the usage, content, and structure of the Web don't necessarily consider themselves to be engaged in data mining, there is a large amount of overlap between Web characterization research and Web Usage mining.

Catledge et al. [19] discuss the results of a study conducted at the Georgia Institute of Technology, in which the Web browser Xmosaic was modified to log client side activity. The results collected provide detailed information about the user's interaction with the browser interface as well as the navigational strategy used to browse a particular site. The project also provides detailed statistics about occurrence of the various client side events such as the clicking the back/forward buttons, saving a _le, adding to bookmarks etc.

Arlitt et. al. [17] discuss various performance metrics for Web servers along with details about the relationship between each of these metrics for different workloads. Manley [21] develops a technique for generating a custom made benchmark for a given site based on its current workload. This benchmark, which he calls a self-configuring benchmark, can be used to perform scalability and load balancing studies on a Web server. Chi et. al. [20] describe a system called WEEV (Web Ecology and Evolution Visualization) which is a visualization tool to study the evolving relationship of web usage, content and site topology with respect to time.

V. Techniques

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of Web usage data. Three main paradigms are: association rules, sequential patterns, and clustering (see [29] for a detailed description of these techniques)

A. Statistical

Statistical techniques are the most common method to extract knowledge about visitors to a Web site. By analysing the session file, one can perform different kinds of descriptive statistical analyses (frequency,

Authors are responsible for any plagiarism issues.

mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Many Web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This report may include limited low-level error analysis such as detecting unauthorized entry points or finding the most common invalid URI. Despite lacking in the depth of its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions.

B. Association Rules

Association rule generation can be used to relate pages that are most often referenced together in a single server session. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. Aside from being applicable for business and marketing applications, the presence or absence of such rules can help Web designers to restructure their Web site. The association rules may also serve as a heuristic for prefetching documents in order to reduce user-perceived latency when loading a page from a remote site.

Apriori algorithm[23] is one of the prevalent techniques used to find association rules (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994). Apriori operates in two phases.

-In the first phase, all item-sets with minimum support (frequent item-sets) are generated.

-The second phase of the algorithm generates rules from the set of all frequent item-sets.

The Apriori heuristic achieves good performance . However, in situations with a large number of frequent patterns, long patterns, or quite low minimum support thresholds, an Apriori-like algorithm may suffer from the two nontrivial costs.

In order to overcome the drawback inherited in Apriori, J.Han develop an efficient FP-treebased mining method, FP-growth,which contains two phases, where the first phase constructs an FPtree, and the second phase recursively projects the FPtree and outputs all frequent patterns.
AprioriTID [25] is an extension of the basic Apriori approach. Instead of relying on the raw database AprioriTID internally represents each transaction by the current candidates it contains. With AprioriHybrid both approaches are combined, c.f. [25]. To some extent also SETM [26] is an Apriori(TId)-like algorithm which is intended to be implemented directly in SQL. DIC is a further variation of the Apriori Algorithm[23]. The Partition-Algorithm[27] is an Apriori-like algorithm .In [28] the algorithm Eclat is introduced.

C. Clustering

Clustering[30] is a technique to group together a set of items having similar characteristics. In the Web Usage domain, there are two kinds of interesting clusters to be discovered : usage clusters and page clusters. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in E-commerce applications or provide personalized Web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for Internet search engines and Web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs.

Clustering algorithms may be classified as : Exclusive Clustering, Overlapping Clustering, Hierarchical Clustering, Probabilistic Clustering.

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure 4, where the separation of points is achieved by a straight line on a bi-dimensional plane. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted.

Finally, the last kind of clustering use a completely probabilistic approach.

Commonly used clustering algorithms are: K-means, Fuzzy C-means, Hierarchical clustering, Mixture of Gaussians

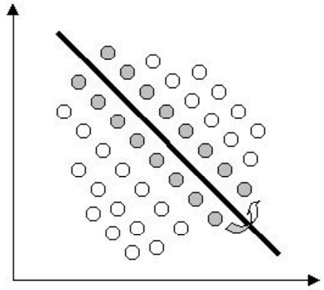


Fig. 4, Exclusive Clustering

Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an *exclusive clustering* algorithm, Fuzzy C-means is an *overlapping clustering* algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a *probabilistic clustering* algorithm. We will discuss about each clustering method in the following paragraphs.

Clustering algorithms can be applied in many fields, for instance:

-*Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;

-*Biology*: classification of plants and animals given their features;

-*Libraries*: book ordering;

-*Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;

-*City-planning*: identifying groups of houses according to their house type, value and geographical location;

-*Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
WWW: document classification; clustering weblog data to discover groups of similar access patterns.
The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- interpretability and usability.

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of "distance" (for distance-based clustering);
- if an *obvious* distance measure doesn't exist we must "define" it, which is not always easy, especially in multidimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

D. Classification

Classification is the task of mapping a data item into one of several predefined classes [38]. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

Classification can be done by using supervised inductive learning algorithms such as: Decision trees classifiers, Rule-based induction, Neural networks, Memory(case) based reasoning,Genetic Algorithms, Naïve Bayesian classifiers, k-nearest neighbor classifiers, Support Vector Machines For example, classification on server logs may lead to the discovery of interesting rules such as : 30% of users who placed an online order in / Product/Music are in the 18-25 age group and live on the West Coast.

E. Sequential Patterns

The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this approach, Web marketers can predict future visit patterns which will be helpful in placing advertisements aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns includes trend analysis, change point detection, or similarity analysis.

In Web Usage Mining, sequential patterns are exploited to find sequential navigation patterns that appear in users_ sessions frequently. The typical sequential pattern has the following form : the 70% of users who first visited A.html and then visited B.html afterwards, have also accessed page C.html in the same session. Sequential patterns might appear syntactically similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining. However, sequential patterns include the notion of time, i.e., at which point of the sequence a certain event happened. In the above example, pages A, B, and C appears sequentially, one after another, in the user sessions; in the previous example on association rules, information about the event sequence is not considered. There are essentially two class of algorithms that are used to extract sequential patterns:

- One includes methods based on association rule mining;
- the other one includes methods based on the use of tree structures and Markov chains to represent navigation patterns.

Some well-known algorithms for mining association rules have been modified to extract sequential patterns. For instance, [32] used AprioriAll and GSP, two extensions of the Apriori algorithm for association rules mining [29]. Ref. [31] argues that algorithms for association rule mining (e.g., Apriori) are not efficient when applied to long sequential patterns, which is an important drawback when working with Web logs.

Accordingly, [31] proposes an alternative algorithm in which tree structures (WAP-tree) are used to represent navigation patterns. The algorithm (WAP-mine) [31] and the data structure (WAP-tree), specifically tailored for mining Web access patterns, WAP-mine outperforms other Apriori-like algorithms [31] like GSP.

Ref. [32] provides a comparison of different three sequential pattern algorithms applied to Web Usage Mining. The comparison includes (i) PSP+, an evolution of GSP, based on candidate generation and test heuristics, (ii) FreeSpan, based on the integration of frequent sequence mining and frequent pattern mining, and the newly proposed (iii) PrefixSpan that uses an approach based on data projection. The results of the comparison [32] show that PrefixSpan outperforms the other two algorithms and offers very good performance even on long sequences.

F. Dependency Modeling

Dependency modeling is another useful pattern discovery task in Web Mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the Web domain. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (ie. from a casual visitor to a serious potential buyer). There are several probabilistic learning techniques that can be employed to model the browsing behavior of users. Such techniques include Hidden Markov Models and Bayesian Belief Networks. Modeling of Web usage patterns will not only provide a theoretical framework for analyzing the behavior of users but is potentially useful for predicting future Web resource consumption. Such information may help develop strategies to increase the sales of products ordered by the Web site or improve the navigational convenience of users.

VI. Applications

As shown in Figures 3, usage patterns extracted from Web data have been applied to a wide range of applications. In Section we will brief overview of some of those applications.

A. Letizia

Letizia is an application that assists a user browsing the Internet. As the user operates a conventional Web browser such as Mozilla, the application tracks usage patterns and attempts to predict items of interest by performing concurrent and autonomous exploration of links from the user's current position. The application uses a best-first search augmented by heuristics inferring user interest from browsing behavior.

B. WebSift[3]

The WebSIFT (Web Site Information Filter) system is another application which performs Web Usage Mining from server logs recorded in the extended NSCA format (includes referrer and agent fields), which is quite similar to the combined log format which used in case of DSpace log files. The preprocessing algorithms include identifying users, server sessions, and identifying cached page references through the use of the referrer field. It identifies interesting information and frequent item sets from mining usage data.

C. Adaptive Websites

An adaptive website adjusts the structure, content, or presentation of information in response to measured user interaction with the site, with the objective of optimizing future user interactions. Adaptive websites are web sites that automatically improve their organization and presentation by learning from their user access patterns. User interaction patterns may be collected directly on the website or may be mined from Web server logs. A model or models are created of user interaction using artificial intelligence and statistical methods. The models are used as the basis for tailoring the website for known and specific patterns of user interaction.

VII. Conclusion

This paper has presented the detail study of Webusage mining System, Techniques, Areas, Applications. This paper has attempted to provide survey of the rapidly growing area and techniques of Web Usage mining. With the growth of Web-based applications, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has led to a number of commercial offerings for doing such analysis.

Refrences

[1] O. Etzioni, The world-wide Web: quagmire or gold mine? Communications of the ACM 39 (11) (1996) 65–68.

[2] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1) (2000) 1–15.

[3] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explorations 1 (2) (2000) 12–23.

[4] D.M. Kristol, Http cookies: standards, privacy, and politics, ACM Transactions on Internet Technology (TOIT) 1 (2) (2001) 151–198.

[5] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, The impact of site structure and user environment on session reconstruction in web usage analysis, in: Proceedings of the 4th WebKDD 2002 Workshop, at the ACMSIGKDD Conference on Knowledge Discovery in Databases (KDD_2002), 2002.

[6] K.D. Fenstermacher, M. Ginsburg, Mining client-side activity for personalization, in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS_02), 2002, pp. 205–212..

[7] Pilot Software, Web site analysis, Going Beyond Traffic Analysis http://www.marketwave.com/productssolutions/hitlist.html (2002).

[8] B. Mobasher, N. Jain, E. Han, and J. Srivastava. Web Mining: Pattern discovery from World Wide Web transactions. Technical Report TR 96-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.

[9] Data mining: Crossing the chasm, 1999. Invited talk at the 5th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining(KDD99).

[10] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Creating adaptive web sites through usage-based clustering of urls. In Knowledge and Data Engineering Workshop, 1999.

[11] Olfa Nasraoui, Raghu Krishnapuram, and Anupam Joshi. Mining web access logs using a fuzzy relational clustering algorithm based on a robust estimator. In Eighth International World Wide Web Conference, Toronto, Canada, 1999

[12] Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira. Characterizing reference locality in the www. Technical Report TR-96-11, Boston University, 1996.

[13] Charu C Aggarwal and Philip S Yu. On disk caching of web objects in proxy servers. In CIKM 97, pages 238{245, Las Vegas, Nevada, 1997.

[14] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy _lters. In Proc. ACM SIGCOMM,pages 241{253, 1998.

[15] S. Schechter, M. Krishnan, and M. D. Smith. Using path pro_les to predict http requests. In 7th International World Wide Web Conference, Brisbane, Australia, 1998.

[16] T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 53{62, San Diego, CA,1999. ACM.

[17] Martin F Arlitt and Carey L Williamson. Internet web servers: Workload characterization and performance implications. IEEE/ACM Transactions on Networking, 5(5):631{645, 1997.

[18] Alex Buchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4):54{61,1998.

[19] L. Catledge and J. Pitkow. Characterizing browsing behaviors on the world wide web. Computer Networks and ISDN Systems, 27(6), 1995.

[20] Chi E. H., Pitkow J., Mackinlay J., Pirolli P., Gossweiler, and Card S. K. Visualizing the evolution of web ecologies. In CHI '98, Los Angeles, California, 1998.

[21] Stephen Lee Manley. An Analysis of Issues Facing World Wide Web Servers. Undergraduate, Harvard, 1997.

[22] Mike Perkowitz and Oren Etzioni. Adaptive web sites:Conceptual cluster mining. In Sixteenth International Joint Conference on Arti_cial Intelligence, Stockholm,Sweden, 1999.

[23] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In Proc. of the 20th VLDB Conference, pages 487{499, Santiago, Chile, 1994.

[24] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In Proc. ACM KDD, 1994.

[25] R. Agarwal and R. Srikant. Fast algorithm for mining assosation rules. In proc. Of the 20th Int'l Conf. on Verry Large Databases (VLDB '94), Santiago, Chile June 1994.

[26] M. Houtsma and A. Swami. Set-oriented mining for aasosiation rules in relational databases. Technical Report RJ 9567, IBM Almaden Research Center, San Jose, California, October 1993.

[27] A. Savasere, E. Omiecinski and S. Navathe. An efficient algorithm for mining association rules in large databases. In Proc. Of the 21st Conf on Very Large Databases(VLDB '95), Zurich, Switzerland, September 1995.

[28] M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. New algorithms for fast discovery of association rules. In Proc. Of the 3rd Int'l Conf. on KDD and Data Mining(KDD'97),Newport Beach, California,August 1997.

[29] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001.

[30] Y. Xie, V.V. Phoha, Web user clustering from access log using belief function, in: Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), ACM Press, 2001, pp. 202–208

[31] J. Pei, J. Han, B. Mortazavi-asl, H. Zhu, Mining access patterns efficiently from web logs, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000, pp. 396–407.

[32] B. Mortazavi-Asl, Discovering and mining user web-page traversal patterns, Master_s thesis, Simon Fraser University, 2001.

* * * * *