International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-2 Issue-5, June 2013

A Survey: Access Patterns Mining Techniques and ACO

Abhishek Mathur, Trapti Agrawal

Abstract-In this paper we define Web Mining as Data Mining on the web. Further we define Web Usage Mining along with its applications and tools. Overall the focus of the paper will be to present a survey of the recent developments of the past and current work in Web Usage Mining and looks at Ant Colony optimization algorithm as a clustering technique for web usage patterns. In addition, there is an attempt to provide an overview of the state of the art on ACO in WUM.

Index Terms- Data Mining, Web Mining, Web Usage Mining, Site Customization, Ant Colony Optimization, Server Logs Data, Clustering

I. INTRODUCTION

According to [1], basically data mining techniques are used in web mining. Web mining is extended version of data mining. Data mining is work upon Off-Line whereas Web mining is work upon On-Line. In data mining data stored in (database) data warehouse and in web mining data stored in server database & web log. The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. [1] New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area. Web mining is a new and rapidly developing research and application area. With more collaborative research across different disciplines like database, artificial intelligence, statistics and marketing, we will be able to develop web mining applications that are very useful to the web based information systems. It is an import part of the online Knowledge discovery Process where data mining techniques are harvesting knowledge over the data collected from World Wide Web. Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure (hyperlink) or usage (Web log) data is used in the mining process (with or without other types of Web data)[11].

Manuscript published on 30 June 2013.
* Correspondence Author (s)
Prof..Abhishek Mathur, Asstt. Prof., SATI, Vidisha,M.P., India.
Trapti Agrawal, Research Scholar, SATI Vidisha,M.P., India.

Taxonomy of Web Mining can be understood using the below fig.

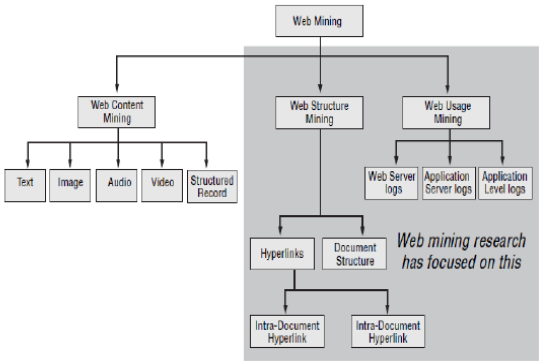

Fig 1: Web Mining Taxonomy

Web Mining is the area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely, Web Content Mining is that part of Web Mining which focuses on the raw information available in web pages; source data mainly consist of textual data in web pages (e.g., words, but also tags); typical applications are content-based categorization and content-based ranking of web pages. Web Structure Mining is that part of Web Mining which focuses on the structure of web sites; source data mainly consist of the structural information in web pages (e.g., links to other pages); typical 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. Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs, that are collected when users access web servers and might be represented in standard formats; typical applications are those based on user modeling techniques, such as web personalization, adaptive web sites, and user modeling.

II. WEB USAGE MINING

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the Web. It tries to make sense of the data generated by the Web surfer's sessions or behaviors. Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. It's an important technology for understanding user's behaviors on the Web and is one of the favorite areas of many researchers in the recent time. Markov models have been extensively used to model Web users' navigation behaviors on Web sites.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Web usage mining consists of three phases, namely preprocessing (data preparation), pattern discovery, and pattern analysis. The web data is typically unlabelled, distributed, heterogeneous, semi-structured, time varying and high dimensional so in the first phase, Web log data are preprocessed in order to identify users, sessions, page views; and so on this maps the usage data of the Web server into relational tables before an adapted data mining technique is performed. In the second phase, statistical methods, as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied in order to detect interesting patterns. These patterns are stored so that they can be further analyzed in the third phase of the Web usage mining process. The applications of Web usage mining could be classified into two main categories: On one hand, learning a user profile or user modeling in adaptive interfaces (personalized) and learning user navigation patterns (impersonalized) [3]. Web users would be interested in, among others, techniques that could learn their information needs and preferences, which is user modeling possibly combined with Web content mining. On the other hand, Information providers would be interested in, among others, techniques that could improve the effectiveness of the information on their Web sites by adapting the Web site design or by biasing the user's behavior towards satisfying the goals of the site. In other words, they are interested in learning user navigation patterns. Then the learned knowledge could be used for applications such as personalization (at a Web site level), system improvement, site modification, business intelligence, and usage characterization (Srivastava et al., 2000).

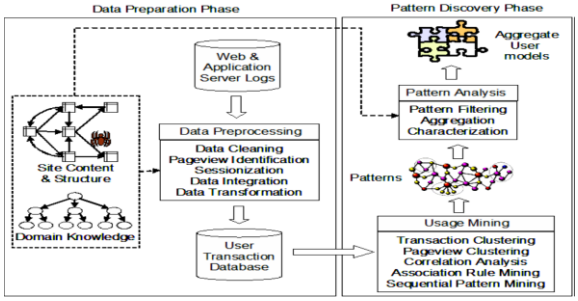

Fig2: Web Usage Mining Process

III. APPLICATION AND TOOLS FOR WUM [10]

Web mining for usage pattern is the key to discover marketing intelligence in e-commerce. It helps tracking of general access pattern, personalization of web link or web content and customizing adaptive sites. It can disclose the properties and inter-relationship between potential customers, users and markets, so as to improve Web performance, on-line promotion and personalization activities. Web Log Mining uses KDD techniques to understand general access patterns and trends to shed light on better structure and grouping of resource providers. For e.g., Web miner discovers association rules and sequential patterns automatically from server access logs. Commercial software Web Analyst by Megaputer learns the interests of the visitors, based on their interaction with the website. Clementine and DB2 Intelligent Miner for Data are two general-purpose data mining tools, which can be used for web usage mining with suitable data preprocessing.

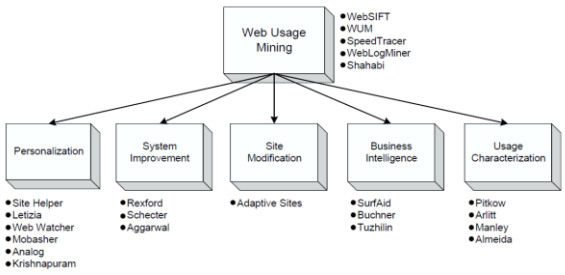

Fig3: Major Application Areas of Web Usage Mining

Web analysis tools provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools, e.g., 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. There are several commercial software tools (see Table 1) that could provide Web usage statistics. These stats could be useful for Web administrators to get a sense of the actual load on the server. For small web servers, the usage statistics provided by conventional Web site trackers may be adequate to analyze the usage pattern and trends. However as the size and complexity of the data increases, the statistics provided by existing Web log file analysis tools may prove inadequate and more intelligent knowledge mining techniques will be necessary.

Sr. No.	Tool	Researcher/Company	Feature	Function
1	ArchCollect	Esmin et al., 2008.	It uses fast semantic interaction acquisition algorithms and form a dimensional cube of information directly which serve as input to an IR approach [41].	To monitor users' interactions in web media.
2	i-Miner	Abraham et al., 2003.	To optimize the concurrent architecture of a fuzzy clustering algorithm (to discover data clusters) and a fuzzy inference system to analyze the trends [42].	Pattern Discovery and trend analysis from web usage data.
3	AWUSA	Tiedtke et al., 2002.	A framework based on combination of information architecture, automated usability evaluation and web mining techniques for data-gathering and analysis [43].	Automated website usability evaluation.
4	i-JADE Web-Miner	Lee and Liu, 2001.	To use Agent technology, together with Web mining technology, to automate a series of product search and selection activities. It is based on multiagent development platform iJADE [19].	E-commerce applications.
5	Web Quilt	Hong et al., 2001.	Web logging and visualization system that helps web design teams capture usage traces which can be aggregated and visualized in a zooming interface that shows the web pages people viewed [14].	To run usability tests and analyze the collected data from web logs.
6	KOINOTITES	Pierrakos et al., 2000.	A system which uses data mining techniques for the construction of user communities on the Web [44].	Personalisation
7	INSITE	Shahabi et al., 2000.	To generate user profiles in real time through the use of a unique Connectivity Matrix Model (CM-model). And show the efficacy and scalability[45].	INSITE acquisition Extracts and stores the essence of the captured information in real time and visualize the result.
8	STRATDYN	Berendt,B. 2000.	Developed as add on module to the WUM, extends its capability by exploiting site semantics [46].	Visualization of navigation patterns.
9	SEWeP	Eirinaki M. et al. 2003	Developed as a system that makes use of both the usage logs and the semantics of a Web site's content in order to personalize it [47].	Personalisation
9	WebTool	Masseglia et al., 2000.	It uses sequential pattern mining which relies on PSP an algorithm developed by the authors [48].	Usage profiling.
10	Web SIFT (Based on WEBMINER prototype 1997)	Cooley et al., 1999.	System uses the content and structure information from a website in order to identify interesting results from mining usage data [49].	To mine interesting frequent item sets automatically from real web data.
11	M/DAS (Mining Internet Data for Associative Sequences)	Büchner et al., 2000.	It introduces a new algorithm called M/DAS that extends traditional sequence discovery with a wide range of web-specific features [10].	Pattern discovery.
12	Web usage miner (WUM)	Spiliopoulou and Faulstich 1998.	It exploits an innovative aggregated storage representation for the information in the web server log. It discovers patterns comprised of not necessarily adjacent events [50].	Mining interesting navigation patterns in the form of graphs.
13	Web Mate	Chen & Sycara, 1998.	The user profile is inferred from training examples [13].	As Proxy agent provides effective browsing and searching Help.
14	SpeedTracer	Wu, Yu and Ballman, IBM, 1998.	Reconstructs the user transversal paths for session identification by using the referrer page and the URL of the requested page as a traversal step [51].	Mining web server log files.
15	WebLogMiner	Zaiane, 1998	Use data mining and OLAP on treated and transformed web access files [17].	Mining web server log files.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

16	DB2 Intelligent Miner for Data	IBM cradle	Provides a single framework for database mining using proven, parallel mining techniques.	User database miner
17	Poly Analyst version 6.0	Megaputer, (1997-2007)	Integrates the data and text mining capabilities of analytical software directly.	Profiles the website resources and dynamically identifies the most appropriate resources to serve each visitor.
18	Clementine	SPSS, Apache Software Foundation, 1997.	To browse data using interactive graphics to find important features and relationships.	CRM
19	WEBMINER	Cooley, Srivastava and Mobasher., 1997.	A general and flexible framework for Web usage mining, the application of data mining techniques, such as the discovery of association rules and sequential patterns, to extract relationships from data collected in large Web data repositories [16].	Restructure a Web site, and in analyzing user access patterns to dynamically present information tailored to specific groups of users.
20	WEBVIZ	Pitkow & Bharat, 1994	By incorporating Web-Path paradigm into interactive software, users can see not only the documents but also the hyperlinks traveled and have lot more options [47].	Provides graphical view of their local database and access patterns.

Table 1: Web Usage Mining Tools

IV. A REVIEW OF CHALLENGES IN WEB USAGE MINING

The web usage mining algorithms are applied on the preprocessed web log data. The log files are collected from web server. But there are certain reasons due to which the actual logs are not collected.

1) Due to the cache present on client browser, most of the request, if it is resent in the cache is not sent to web server.
2) Most of the time user does not visit the home page of a website. They directly navigate to a particular page, by getting the URL from search engines. So it reduces the hit count of index page.
3) Generally in web pages designed by server side scripting like PHP, JSP or ASP.NET they use inner page. That is, one page consisting of more than one page. In that case the request for main page records two entries in access log. It is difficult to identify an inner page.
4) Some web pages take query string as argument to the URL. E.g. dept.php? dept=CSE, dept.php? dept=IT like this. In this case the same page i.e. dept.php is accessed but with different arguments. It is difficult to count the page access of the web page without the argument.

In web usage mining the pattern extraction algorithms are applied on the log data after they are processed. So preprocessing is very much important and must be carried out with proper care. While preprocessing the web access log the above points should be taken into consideration so that it will produce a good set of access logs for pattern extraction.

V. TECHNIQUES IN WEB USAGE MINING

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. However, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of web usage data. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Various data mining techniques have been investigated for mining web usage logs. They are as following:

STATISTICAL ANALYSIS

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

The useful statistical information discovered from web logs is listed in Table2.Many web traffic analysis tools, such as WebTrends and WebMiner, are available for generating web usage statistics.

IMPORTANT STATISTICAL INFORMATION DISCOVERED FROM WEB LOGS

Statistics	Detailed Information
Website Activity Statistics	Total number of visits Mean number of hits Successful/failed/redirected/ hits Average view time Average length of a path through a site
Troubleshooting/Diagnostic Statistics	Server errors Page not found errors
Server Statistics	Top pages visited Top entry/exit pages

Table 2

PATH ANALYSIS

There are many different types of graphs that can be formed for performing path analysis. Graph may be representing the physical layout of a Web site, with Web pages as nodes and hypertext links between pages as directed edges. Graphs may be formed based on the types of Web pages with edges representing similarity between pages, or creating edges that give the number of users that go from one page to another. Path analysis could be used to determine important information e.g. 80% of clients left the site after four or less page references. This example indicates that many users don't browse more than four pages into the site, it can be concluded that important information is contained within four pages of the common site entry points.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

ASSOCIATION RULES

Association Rules are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining for finding frequent patterns, associations, and correlations among sets of items. When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users' sessions. The typical result has the form "A.html, B.html, C.html" which states that if a user has visited page A.html and page B.html, it is very likely that in the same session, the same user has also visited page C.html. This correlation might suggest that this information should be moved to a higher level to increase access to C.html. Discovery of such rules for organizations engaged in electronic commerce can help in the development of effective marketing strategies. [5] proposes and evaluates some interestingness measures to evaluate the association rules mined from web usage data.[6] exploits a mixed technique of association rules and fuzzy logic to extract fuzzy association rules from web logs.

Since usually such transaction databases contain extremely large amounts of data, current association rule discovery techniques try to prune the search space according to support for items under consideration. Support is a measure based on the number of occurrences of user transactions within transaction logs. Aside from being exploited for business applications, such observations also can be used as a guide for Web site restructuring, e.g., by adding links that interconnect pages often viewed together, or as a way to improve the system's performance through prefetching Web data.

SEQUENTIAL PATTERNS

In web usage mining, sequential patterns are exploited to find sequential navigation patterns that appear in users' sessions frequently. Sequential pattern discovery is an extension of association rules mining in that it reveals patterns of co occurrence incorporating the notion of time sequence. The typical sequential pattern has the form [7]: the 70% of users who first visited A.html and then visited B.html afterwards, in the same session, have also accessed page C.html. Sequential patterns might appear syntactically similar to association rules; in fact there are essentially two classes of algorithms that are used to extract sequential patterns: one includes methods based on association rule mining; one includes methods based on the use of tree-structures, data projection techniques, and Markov chains to mine navigation patterns. Some well-known algorithms for mining association rules have been modified to extract sequential patterns. [8] Presents a comparison of different sequential pattern algorithms applied to Web Usage Mining. The comparison includes PSP+, FreeSpan, and PrefixSpan. While PSP+ is an evolution of GSP, based on candidate generation and test heuristic, FreeSpan and the newly proposed PrefixSpan use a data projection based approach. PrefixSpan outperforms the other two algorithms and offers very good performance even on long sequences. [9] Proposes an hybrid method: data are stored in a database according to a so-called Click Fact Schema; an Hypertext Probabilistic Grammar (HPG) is generated by querying the databases; HPGs represent transitions among web pages through a model which resembles many similarities with Markov chains. The frequent sequential patterns are mined through a breadth first search over the hypertext probabilistic grammar. HPGs were first proposed in [7], and later improved in [9] where some scalability issues of the original proposal have been solved.

The discovery of sequential patterns in Web server access logs allows Web-based organizations to predict user visit patterns and helps in targeting advertising aimed at groups of users based on these patterns. By analyzing this information, the Web mining system can determine temporal relationships among data items such as the following:

1) 30% of clients who visited /company/products/, had done a search in Yahoo, within the past week on keyword data mining; or
2) 60% of clients, who placed an online order in /computer/products/webminer.html, also placed an online order in computer/products/iis.html within 10 days.

From these relationships, vendors can develop strategies and expand business. Using this approach, useful users' trends can be discovered, and predictions concerning visit patterns can be made.

CLUSTERING AND CLASSIFICATION

Clustering has been widely used in Web Usage Mining to group together similar sessions among large amount of data based on a general idea of distance function which computes the similarity between groups. [10] Was the first to suggest that the focus of web usage mining should be shifted from single user sessions to group of user sessions; [20] was also the first to apply clustering for identifying such cluster of similar sessions. [11] Proposes similarity graph in conjunction with the time spent on web pages to estimate group similarity in concept-based clustering. [12] Uses sequence alignment to measure similarity, while [10] exploits belief functions. [13] Uses Genetic Algorithms to improve the results of clustering through user feedback. [14] couples Fuzzy Artificial Immune System and clustering techniques to improve the users' profiles obtained through clustering.[15] applies multi-modal clustering, a technique which build clusters by using multiple information data features. [16] Presents an application of matrix clustering to web usage data.

In Web mining, classification techniques allow one to develop a profile for clients who access particular server files based on demographic information available on those clients, or based on their access patterns. For example classification on WWW access logs may lead to the discovery of relationships such as the following:

1) Clients from state or government agencies who visit the site tend to be interested in the page /company/lic.html or .
2) 60% of clients, who placed an online order in/company/products /product2, were in the 35-45 age groups and lived in Chandigarh.

For web usage mining, clustering techniques are mainly used to discover two kinds of useful clusters, namely user clusters and page clusters. User clustering attempts to find groups of users with similar browsing preference and habit, whereas web page clustering aims to discover groups of pages that seem to be conceptually related according to the users' perception.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Such knowledge is useful for performing market segmentation in ecommerce and web personalization applications.

The desirable features of clustering algorithms are scalability, ability to deal with different data types, discovery of clusters with arbitrary shape, able to deal with noise and outliers, insensitive to order of input records, incorporation of user- specified constraints, interpretability and usability, minimal requirements for domain knowledge to determine input parameters. There exist a large number of clustering algorithms in the literature. No single algorithm is suitable for all types of objects, nor all algorithms appropriate for all problems. Unfortunately, many of the traditional clustering algorithms share a number of drawbacks. Generally clustering algorithms can be categorized into hierarchical methods, partitioning methods, density-based methods, grid-based methods, and model-based methods [17].

Recently, algorithms inspired by nature used for clustering. These algorithms have advantages in many aspects, such as self-organization, flexibility, robustness, no need of prior information, and decentralization [18].

VI. A SURVEY ON RESEARCH IN WUM TECHNIQUES

[19] Makes use of a rough set based learning program for predicting web usage. In our approach, web usage patterns are represented as rules generated by the inductive learning program, BLEM2.Inputs to BLEM2 are clusters generated by a hierarchical clustering algorithm applied to preprocessed web log records. Empirical results show that the prediction accuracy of rules induced by the learning program is better than a centroid based method. In addition, the use of a learning program can generate shorter cluster descriptions.

[20] Makes use of Rough Set Clustering to cluster the most probable sessions. Rough Set Clustering (RST) is an approach to aid decision making in the presence of uncertainty. It classifies imprecise, uncertain or incomplete information expressed in terms of data acquired from experience. Any union of elementary sets is called a crisp set and other sets are referred to as rough set. As a result of this definition, each rough set has boundary-line elements. Indiscernibility is core concept of RST and is defined as equivalence between objects. Objects in the information system about which we have the same knowledge form an equivalence relation.

[21] Proposed a complete preprocessing methodology, which covers all the above stated phases. In preprocessing phase, proposed data cleaning and data filtering algorithms are applied on a sample web server log file. Algorithms are also proposed to identify the users on the basis of IP Address (an attribute of web log). By applying proposed session identification algorithm, the user sessions from web log were obtained and transformed into session vectors. The "Angular Separation", "Canberra Distance" and "Spearman Distance" similarity measures to compute the similarity among the session vectors. In the last phase, the proposed algorithm based on Swarm and Agglomerative algorithms were applied to obtain the hierarchical sessionization of user sessions.

In [22] Fuzzy c-means clustering incorporates fuzzy set theoretic concept of partial membership and may result in the formation of overlapping clusters. The algorithm calculates the cluster centers and assigns a membership value to each data item corresponding to every cluster within a range of 0 to 1. The algorithm utilizes a fuzziness index parameter q where q subset [1, infinity] which determines the degree of fuzziness in the clusters. As the value of q reaches to 1, the algorithm works like a crisp partitioning algorithm. Increase in the value of q results in more overlapping of the clusters.

FCM algorithm is hard on data sets too, so the data sets must be quite regular. In order to solve the problem [23] use information entropy to initialize the cluster centers to determine the number of cluster centers. It can be reduce some errors. [24] adopted a CLIQUE (CLUstering in QUEst) algorithm for clustering web sessions for web personalization. Then adopted various similarity measures like ED, projected Euclidean distance Jaccard, cosine and fuzzy dissimilarity measures to measure the similarity of web sessions using sequence alignment to determine learning behaviors. Clique automatically finds subspaces of the highest dimensionality such that high density clusters exist in those subspaces. The clique clustering technique integrates density based and grid based clustering. It is useful for clustering high dimensional data in large databases.

We compared different web usage mining techniques on the basis of several parameters and complexity as below:

TECHNIQUES	BASIS	COMPLEXITY	SIMILARITY MEASURES	UNCERTAIN DATA	ITERATIONS	FEATURE
RST	Maximum pages visited	reduced	N/A	yes	Any	Mine the vital sessions
Hierarchical Clustering	PSO based session clustering of session vectors	controlled	AS,CD,SD	no	Up to 100	Enhance the web log visualization and structured information for the next phases
CLIQUE	Density of web usage data	Time and space complexity	ED,PED, Jaccard, cosine and fuzzy dissimilarity measures	No	Any	Cope with high dimensional /large databases
Fuzzy Set Theoretic	Fuzzy c-partitions and sessions identified	reduced	Euclidian distance	yes	Any	Cope with growth and complexity of www and can handle outliers.
Improved FCM	Entropy of initial cluster centers	reduced	Euclidian distance	yes	until a stopping criterion is achieved	Can solve irregular datasets
Ant Colony Clustering	Continuous Learning Strategy	reduced	Depends on the clustering algo	yes	Until ant learns	Provide optimization & Reduces a significant time

Table3: Comparisons of Web Usage Mining Techniques

VII. A SURVEY ON RESEARCH IN ACO IN WUM

In [25], the hybrid framework uses an ant colony optimization algorithm to cluster Web usage patterns. This paper proposed an ant clustering algorithm (ACLUSTER) to segregate visitors or find the web usage patterns (data clusters) and a linear genetic programming approach to analyze the visitor trends.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The results are compared with the earlier works using self organizing map and evolutionary -fuzzy c means algorithm to segregate the user access records and several soft computing paradigms to analyze the user access trends.

[26] Propose an accelerated ant based clustering algorithm (ACCANTCLUST) which is based on chemical recognition system of ants and this algorithm finds the number of clusters automatically. In ANTCLUST algorithm, when meeting between two ants is simulated, if the meeting is between two ants with no nest, and if they accept each other, these two ants are placed in a new nest. If they do not accept each other, no nest is created. In the proposed algorithm, if the ants do not accept each other, two new nests are created and the ants are placed in two different nests. The advantage of the proposed algorithm generates optimum clusters even if the last two steps of ANTCLUST are removed and thereby reducing the time taken to form clusters. The experimental results of ACCANTCLUST show the fast convergence of clusters compared with ANTCLUST.

In [27], ant-based clustering is applied to pre-processed logs to extract frequent patterns for pattern discovery and then it is displayed in an interpretable format. In this paper, a new method is proposed for extracting patterns from web logs based on ant clustering algorithm. We apply ant-based clustering for pattern discovery, other similar methods applied ant colony clustering to segregate visitors. Some methods applied Markov models for modeling user web navigation behavior. But the proposed method has the similarity and speed of ant-based clustering algorithm rather than other clustering algorithms.

[28] Presents how to mine the secondary data (web logs) derived from the users' interaction with the web pages during certain period of Web sessions. At first Ant-based clustering algorithm is applied to pre-processed log files to extract frequent patterns, then it is displayed in an interpretable format and secondly decision tree method is used to find and predict user's navigation behavior. Decision trees are used in classification and prediction. It is simple yet a powerful way of knowledge representation. Two type of approaches are used were the offline phase is based on Ant based clustering and the online phase is based on decision trees. The experimental results represent that the approach can improve the quality of clustering for user navigation pattern in web usage mining systems. These results can be used for predicting user's next request in the huge web sites.

[29] Proposes that as the size of the cluster goes on increasing due to increase in users or growth of interest of users it will become inevitable need to optimize the clusters. This paper proposes a cluster optimizing methodology based on ants nestmate recognition ability and is used for eliminating the data redundancies that may occur after the clustering done by the web usage mining methods. This paper proposes an AntClusterTrack algorithm; a cluster optimization algorithm which takes its input from a neural network based training process, ART1. The clusters obtained by L2 layer of ART1 is feed into an ant based clustering approach that checks for the similarity of the pheromone values of the artificial ants. This is done on the fact that ants belonging to the same nest will have similar odor. In this algorithm clusters are considered as the ants nest and the url combinations in each cluster is considered as the artificial ants.

[30] Suggests a novel methodology for analyzing Web user behavior based on session simulation by using an Ant Colony Optimization algorithm In the first place, artificial ants learn from a clustered Web user session set through the modification of a text preference vector. Then, trained ants are released through a web graph and the generated artificial sessions are compared with real usage. The main result is that the proposed model explains approximately 80% of real usage in terms of a predefined similarity measure. Here it is proposed to perform a clustering process of the Web user sessions in order to reduce the number of subsequent comparisons. In short the model is based on a continuous learning strategy based on previous usage in which artificial ants try to fit their sessions with real usage through the modification of a text preference vector. An interesting result is related with the correlation between the most important keywords present in the pages that receive the majority of visitors.

Here we compared these variants in below table:

TECHNIQUE	CONVERGENCE/OPTIMALITY OF CLUSTERS	PRECISION	COVERAGE	CLUSTERING ALGOTITHM USED	OPTIMIZATION ALGORITHM	PREDICTION	FEATURE
Antclust	slow	good	Not Bad	N/A	--	good	Speed
Accantclust	fast	good	good	N/A	--	good	Reduced time
Antnestmate	---	Low/high	High/low	ART-1N N	ClusterTr-ack	good	precision/accuracy and can cope with increasing size of the web
Session Simulation Based on Ant	Same as k-means and SOM	normal	81%	Hierarchical clustering	---	vague	Restrained due to high pinch of proximity

Table4: Comparisons among Variants of Ant in WUM

VIII. CONCLUSION AND FUTURE WORK

As per the survey, it can be reasoned that it's the database properties which can adjudicate about the mining technique to be used for web data. Hierarchical clustering cluster the session vectors (structured) but its complexity increases when it's iterated more than 100 times. Improved FCM come out to be the best algorithm for usage logs as it can handle irregular as well as uncertain data and a stopping criterion is used for no. of iterations. Similarly Ant based algorithm and its variants can be used for optimization of the clusters formed or the usage data. Here the ACCANTCLUST appear to be dearest in terms of all the parameters. However future work needs to be done on session simulation based on ant which can predict the true sessions for future visitors. This will assist in automated user personalization of the websites. More new parameters and ant's features can be used to increase the efficiency of the ant algorithm. Moreover, work needs to be done to automate the whole process of the WUM.A complete WUM methodology, covering data preprocessing, pattern analysis and pattern discovery phases in one will be more useful.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Web Usage Mining plays an important role in realizing enhancing the usability of the website design, the improvement of customers' relations and improving the requirement of system performance and so on. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc

IX. ACKNOWLEDGEMENT

I acknowledge the work of all those researchers/scholars who have contributed to the field of WUM, which I have taken support of and could not mention due to limitation of space in this paper. I also acknowledge the work of all those authors which I have surveyed, my college faculty and friends.

REFERENCES

[1] Kavita sharma, gulshan shrivastava, and vikas kumar, "Web mining: today and tomorrow", **I**EEE International conference on Electronics Computer technology 2011
[2] M. eirinaki and m. vazirgiannis , web mining for web personalization, ACM transactions on internet technology, 3(1), 2000, 1-27.
[3] hussain, t., s. asghar, et al. (2010). Web usage mining: a survey on preprocessing of web log file. IEEE, international conference on (iciet) 2010, karachi pakistan
[4] chhavi rana "a study of web usage mining research tools" int. j. advanced networking and applications volume:03 issue:06 pages:1422-1429 (2012)
[5] xiangji huang, nick cercone, and aijun an. comparison of interestingness functions for learning web usage patterns, proceedings of the eleventh international conference on information and knowledge management, pages 617–620. ACM press, 2002.
[6] s. shiu c. wong and s. pal. mining fuzzy association rules for web access case adaptation. in case-based reasoning research and development: proceedings of the fourth international conference on case-based reasoning, 2001.
[7] eleni stroulia nan niu and mohammad el-ramly. understanding web usage for dynamic web-site adaptation: a case study. in proceedings of the fourth international workshop on web site evolution (WSE'02), pages 53–64. ieee, 2002.
[8] behzad mortazavi-asl. discovering and mining user web-page traversal patterns. master's thesis, simon fraser university, 2001.
[9] t. b. pedersen s. jespersen and j. thorhauge. a hybrid approach to web usage mining. technical report r02-5002, department of computer science aalborg university, 2002.
[10] 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.
[11] a. banerjee, j. ghosh, clickstream clustering using weighted longest common subsequences, in: proceedings of the web mining workshop at the 1st siam conference on data mining, 2001.
[12] b. hay, g. wets, k. vanhoof, clustering navigation patterns on a website using a sequence alignment method in: intelligent techniques for web personalization: ijcai 2001, 17th int. joint conf. on artificial intelligence,august 4, 2001, seattle, wa, usa, pp. 1–6.
[13] j.h. holland, adaptation in natural and artificial systems, university of michigan press, ann arbor, 1975,republished by the MIT press, 1992.
[14] o. nasraoui, f. gonzalez, d. dasgupta, the fuzzy artificial immune system: motivations, basic concepts, and application to clustering and web profiling, in: proceedings of the world congress on computational intelligence (wcci) and ieee international conference on fuzzy systems, 2002, pp. 711–716.
[15] a. ypma, t. heskes, clustering web surfers with mixtures of hidden markov models, in: proceedings of the 14th belgian–dutch conference on ai (bnaic_02), 2002.
[16] s. oyanagi, k. kubota, a. nakase, application of matrix clustering to web log analysis and access prediction,.
[17] p. berkhin, "survey clustering data mining techniques", technical report, accrue software, san jose, california, 2002.
[18] w. bin and s. zhongzhi, " a clustering algorithm based on swarm intelligence", proc. of the int. conf. on info-tech. and info-net, beijing,china, 2001, pp. 58-66. sigkdd explorations newsletter, 1(2), 2000, 12-23.
[19] natheer khasawneh, chien-chung chan web usage mining using rough sets.
[20] ms. jyoti dr. a. k. sharma dr. amit goel ms. payal gulati "a novel approach for clustering web user sessions using rst"
[21] tasawar hussain, dr. sohail asghar, simon fong "a hierarchical cluster based preprocessing methodology for web usage mining"
[22] zahid ansari_, a. vinaya babuy, waseem ahmed_ and mohammad fazle azeem "a fuzzy set theoretic approach to discover user sessions from web navigational data"
[23] k.suresh, r.madanamohana, a.ramamohanreddy, a.subrmanyam "improved fcm algorithm for clustering on web usage mining"
[24] ms k.santhisree, dr. a damodaram"clique: clustering based on density on web usage data: experiments and test.
[25] ajith abraham, vitorino ramos "web usage mining using artificial ant colony clustering and linear genetic programming"@IEEE 2010
[26] h. hannah inbarani, k. thangavel "clickstream intelligent clustering using accelerated ant colony algorith" ©2006 IEEE.
[27] kobra etminani 1 mohammad-r. akbarzadeh-t. 2 noorali raeeji anehsarikobra etminani 1 mohammad-r. akbarzadeh-t. 2 noorali raeeji yanehsari "web usage mining: users' navigational patterns extraction from web logs using ant-based clustering method" ifsa-eusflat 2009
[28] mrs. v. sujatha, dr. punithaval li " an approach to user navigation pattern based on ant based clustering and classification using decision trees" @ 2010
[29] anna alphy 1, s. prabakaran "cluster optimization for improved web usage mining using ant nestmate approach"@IEEE 2011.
[30] pablo loyola*, pablo e. rom´an* and juan d. vel´asquez* "clustering-based learning approach for ant colony optimization model to simulate web user behavior" 2011 IEEE.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)