© 2019 IJSRST | Volume 6 | Issue 3 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : 10.32628/IJSRST196325

Web Mining Algorithms

Neha Doomra[1], Asst. Prof. Rashmi Verma[2]
[1]Master of Engineering, Department of Computer Science and Engineering DPGITM, Gurugram, India
[2]H.O.D, Department of Computer Science and Engineering, DPGITM, Gurugram, India

ABSTRACT

Web is collection of inter related files on one or more web servers. Mining literally means the operations involved in digging for hidden treasures. Similarly data mining is used for the operations involved in digging out critical information from within an organization stored data for better decision support. It is a nontrivial process of extracting implicit, previously unknown and potentially useful patterns from large database. Web mining can be generally defined as the application of data mining techniques to extract useful knowledge from the Web Data. Web mining can be further categorized as web content that includes text, images, record etc, web structure which includes hyperlinks, tags etc, and web usage including http logs, app server logs.

Keywords : Web Mining, Link Mining, Web Usage Mining, Contents, Patterns, Algorithms.

I. INTRODUCTION

The Internet is a worldwide, publicly accessible series of interconnected computer networks that transmit data. Internet today is called as Information Super Highway. The www is a large, distributed hypertext repository of information, where people navigate through browsers on their terminals. The long-term success of www depends on the fast response time. The WWW has become a huge, diverse, and dynamic information reservoir accessed by people with different backgrounds and interests. On the Web, access information is generally collected by Web servers and recorded in the access logs. Web mining and user modeling are the techniques that make use of these access data, discover the surfer s browsing patterns, and improve the efficiency of Web surfing . Web mining can be generally defined as the use of data mining techniques to automatically discover useful knowledge from the Web. Data mining techniques are being used to extract user navigation patterns from web logs [1]. The techniques that have been used to find such patterns are association rules , sequential patterns , classification, clustering etc. The data mining techniques can be clubbed with prefetching. The discovered patterns can be used to predict which pages are likely to be clicked by a user given a sequence of pages that the user already visited. With this prediction, server can be made to automatically prefetch the predicted pages to the client cache before the pages are actually requested and thus, this reduces the network latency. The objective of this paper is to describe the algorithms of web mining to improve the efficiency in predicting the next page.

II. LITERATURE REVIEW

Data mining is the process of analyzing large amounts of data in order to discover patterns and other information. It is typically performed on databases, which store data in a structured format. By "mining" large amounts of data, hidden information can be discovered and used for other purposes. Data mining

is used to extract usable data from a larger set of any raw data. It implies analyzing data patterns in large batches of data using one or more software. The data is integrated and cleaned so that the relevant data is retrieved.. Web mining can be generally defined as the use of data mining techniques to automatically discover useful knowledge from the Web. It is a converging area from several research communities such as Information Retrieval, Database, Machine Learning, and Natural Language Processing. According to the data type, Web mining is divided into three categories: Web content mining, Web structure mining, and Web usage mining. Figure 1 shows the taxonomy of the Web mining. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization□s database. Depending on the location of the source, the type of collected data differs .It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are Unlabeled, Distributed, Heterogeneous (mixed media) ,Semi structured, Time varying, High dimensional. Therefore, web mining basically deals with mining large and hyper-linked information base having the aforesaid characteristics. Also, being an interactive medium, human interface is a key component of most web applications. Some of the issues which have come to light, as a result, concern

1) Need for handling context sensitive and imprecise queries;
2) Need for summarization and deduction;
3) Need for personalization and learning.

Web mining can be viewed as consisting tasks

1) Information Retrieval (IR) (Resource Discovery):.
2) Information Selection/Extraction and Preprocessing:
3) Generalization
4) Analysis

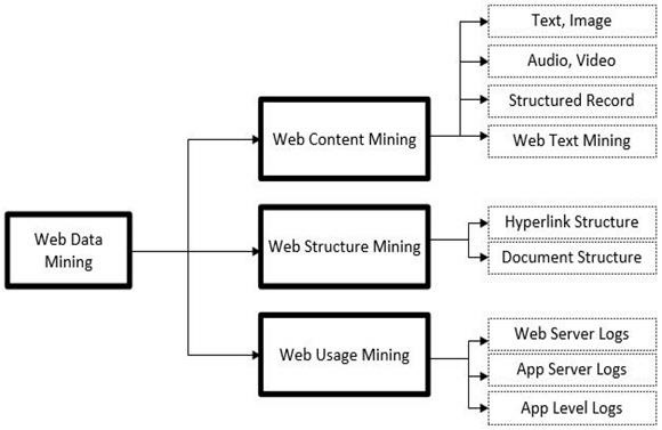


Fig1. Taxonomy of the Web mining

III. WEB MINING CATEGORIES

Web Mining can be further categorized into three types as shown in Fig. 1:

- *Web Content Mining*
- *1Web Structure Mining*
- *Web Usage Mining*

Web Mining consists of massive, dynamic, diverse and mostly unstructured data that provides big amount of data. Explosive growth of web leads to some problems like finding relevant data over the internet, observing user behavior. To solve such kind of problem efforts were made to provide relevant data in structure form (table) that is easy to understand and useful for organizations to predict customer□s need.

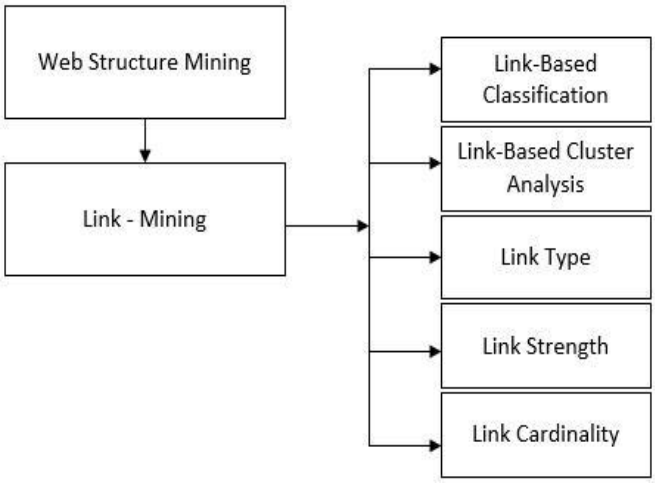
A. Web Content Mining:

It describes the discovery of useful information from the web contents/data /documents. However, what consist of web content could encompass a very broad range of data. Previously the internet consists of different type of services and the data sources such as Gopher, FTP and Usenet. Now most of those data are either ported to or accessible from the web. Basically the web content consists of several types of data such as textual image, audio, video, meta data as well as hyperlinks. The web content data consist of unstructured data such as free text, semi-structured data such as html documents, and a more structured data such as data in the tables or database generated HTML pages. However much of the web content data is unstructured text data. The research around applying data mining techniques to unstructured text is termed Knowledge discovery in text, or text data mining, or text mining. Web Mining uses many techniques to extract data and information from large amount of databases. Many algorithms are used to fetch useful information.

B. Web Structure Mining

Web structure mining is the process of discovering structure information from the web. It tries to discover the model underlying the link structure of the web. The model is based on the topology of hyperlinks with or with out the description of the link. A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page, In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. This model can be used to categories web pages and is useful to generate information such as similarity and relationship between different websites.Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

Link analysis is an old but very useful method that is way its value increases in the research area of web mining Structure analysis is also called as Link-mining. Fig 2 shows types of Web structure mining



The tasks of link-mining:

Link-based Classification: It is an upgrade classification version of classic data mining and its task is to link domains. Main focus is to predict webpage categories based on text, HTML tags, link between web pages and other attributes .

Link-based Cluster Analysis: Primary focus is on data segmentation. In cluster analysis data is categorized or grouped together [16]. Similar objects are grouped in a single group and dissimilar data objects are grouped separately. To dig hidden patterns from datasets link-based cluster analysis can be used [15].

Link Type: It helps to guess link type between entities (two or more) [16].

Link Strength: Link strength shows that links might be related to weights [16].

Link Cardinality: Main focus of link cardinality is to find duplicated website, finding comparison between

them, predicts links between objects, also page categorization [15].

C. Web Usage Mining:

Web usage mining tries to make sense of the data generated by the web surfer's sessions or behaviors. While the web content and structure mining utilize the real or primary data on the web .web usage mining mines the secondary data derived from the interaction of the users while interacting with web. The web usage data include the data from the web server logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries and any other data as the result of interaction. Typically, the usage mining is defined as a three-phase process or tecniques: Fig 3 shows web usage techniques.

1. **Data preprocessing**
2. **Pattern discovery**
3. **Pattern analysis**

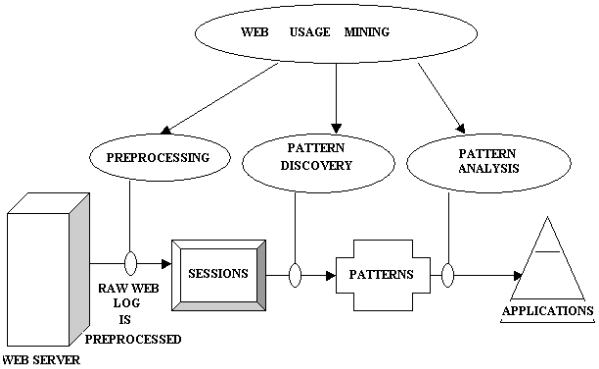
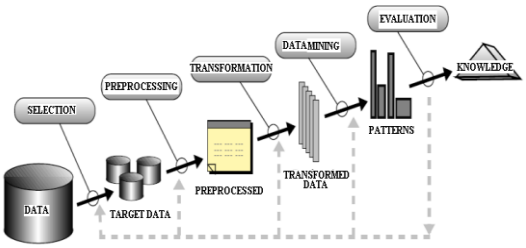


Fig 3 .WEB USAGE MINING TECNIQUES

The above process implicitly covers the standard process of knowledge discovery in databases (KDD).Fig 4shows the KDD process. Table 1 shows the comparison between techniques



DATA PREPROCESSING: It retrieves raw data from the Web resources, and automatically selects and preprocesses the retrieved data. It includes any kind of transformation of the original raw data. It mainly consists of data cleaning, user identification, and session identification

PATTERN DISCOVERY: Once the original Web data have been transformed into desirable formats, several types of techniques can be performed to study the Web surfer's access patterns. Statistical analysis is a powerful technique used for extracting knowledge about webpage visitors. Association rule is one of the basic rules of data mining and is mostly use in web usage mining. Association rule helps to find correlations between web pages that appears in a user session repeatedly. Clustering is a method of grouping items (users and pages) with similar features together. Usage mining consist of two types of clusters i.e. users and pages cluster.

PATTERN ANALYSIS: Pattern analysis is the last step in the overall Web Usage mining process [25]. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase.

The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations.

Table I shows the comparisons between the techniques

IV. WEB STRUCTURE MINING ALGORITHMS

There are various web structure mining algorithms as mentioned in Table I, the paper describes two of them i.e. Page rank algorithm and HITS algorithms. Both of them focuses on link structure of web and how it gives importance to web pages.

Page rank algorithm was developed in 1998 [16] by two famous authors L. Page and S. Brain. The idea was proposed in their PhD research. Both the authors suggested that well known search engine Google was formed by page rank algorithm. It is an algorithm that is frequently used to rank pages. Page rank approach leads to number of pages linking to a specific web page indicates, calculates or describes the importance of that page. Above calculated links are known as backlinks. If backlink is produced from key page or an important page then weightage of this link will be higher than those whose links are coming from non-important pages. Link from page A to page D is considered as a vote (Shown in Fig. 5: Back link Structure). More the vote receives by the page more the importance of that specific page will be. If vote produced from a high weightage page then the importance of linking page will become higher.

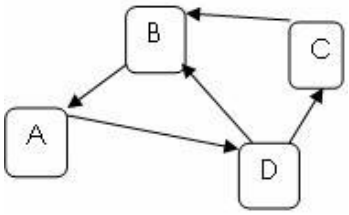


Fig. 5. Back link structure [16]

Following is the formula [14] to find page rank of page A:

$$PR(A) = (1-d) + \frac{d(PR(T1))}{C(T1)} + \ldots\ldots\ldots + \frac{PR(Tn))}{C(Tn)(1)}$$

Where:

PR (Ti) = Rank of Pages

Ti = links to A

C (Ti) = No. of outbound links

d = damping factor (0 to 1)

HITS is an algorithm that stands for Hyperlink Induced Topic Search and is use for web structure (hyperlink analysis) mining. HITS concept was developed by Jon Kleinberg [16] to rank pages. Two terminologies are used in HITS algorithm i.e. authorities and hubs. Good authority is a page that is pointed by high hub weights and good hubs are pages that points to many authority pages with high weights (Shown in Fig. 6 HITS). It is not easy to differentiate in between these two attributes as some sites can be hubs as well as authorities at the same time.

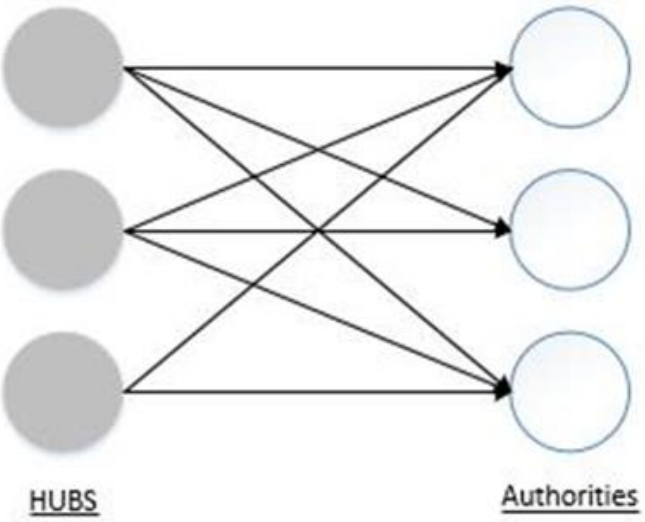


Fig. 6. HITS (Hubs and Authorities) [17].

HITS algorithm includes two steps. First is sampling in which related pages are collected for certain queries. In iterative step authorities and hubs are found with the help of sampling output. Because of the equal weights of pages HITS don't find the relevant pages requested by user queries.

Weighted PageRank algorithm (WPR): It is an extension to the standard PageRank algorithm, is introduced in this paper. WPR takes into account the importance of both the inlinks and the outlinks of the pages and distributes rank scores based on the popularity of the pages. The results of our simulation studies show that WPR performs better than thePage Rank algorithm. It is a research model and query independent. Table II shows the comparison between algorithms of web structure mining. Table II shows the comparison between these three algorithms.

V. WEB USAGE MINING ALGORITHMS

Web usage mining there are numbers of algorithms that can be used as few of them are listed in above Table I. This section will describe three important algorithms i.e. Apriori Algorithm, FP Growth Algorithm and Fuzzy c-means algorithm.

Apriori algorithm is an important and supervised algorithm mostly use for association rule (describe above) to find frequent sets of items during transaction. At first apriori algorithm observe initial database and captures those data sets which are large, then uses result of first captured data sets as a base or model to discover other data sets (large). In apriori algorithm there is a pre-defined support level, if the support level is greater than minimum then item sets are called large or frequent and if support level is below then item sets are known as small. Before AIS algorithm was used for mining regular item sets and association rules but after some time algorithm was modified and given a name Apriori Algorithm . Example: Suppose we have two transaction A1 ={1,2,3} and A2 = {2,3,4} where 1,2,3,4 are item set and 2 ,3 are frequent items in both transaction because of repetition.

FP growth is another efficient algorithm use for association rule. FP-Growth discovers frequent sets of data from FP tree without candidate generation and use bottom-up approach. FP tree is complete data structure, contains one root node "null" and sub tree nodes (prefix) as children. FP growth search FP tree and fetch frequent sets of data.

Fuzzy cMean is an algorithm use in usage mining using clustering approach. It was developed by Bezdek [29]. Fuzzy in an unsupervised algorithm that is applied to a wide range to connected data. FCM task is to group n number objects n number of clusters. In every cluster there is center point which describes features and importance of that cluster [30]. Objects close to the center of cluster become member of the cluster.

FCM Algorithm formula:

$$j(U,c_1,\ldots c_c) = \sum_{i=1}^{c}{}_{ji} = \sum_{i=1}^{c} \sum_{i=1}^{c} u_{ij}^{m} d_{ij}^{2}$$

Where: Ci = Cluster Center

Uij = Numerical Value [0, 1]

i = Euclidian Distance = $d_{ij} = \| c_i - x_i \|$

ith , jth = Cluster Center , data points

III. CONCLUSION

Web mining , which is the actually the application of data mining techniques to discover patterns from the World Wide Web including Web documents, hyperlinks between documents, usage logs of web sites, etc. Now in today advanced world, web becomes an important part of many of all organizations, businesspersons and daily individuals. As a web data is of very much different formats, we have studied the characteristics of web data. As it is very much important to mine particular data from web. Each type has different algorithms and techniques that are used for data retrieval. Various algorithms and techniques

for each type are described. all types and Table I shows comparison for web usage mining techniques. Web content mining is useful in terms of exploring data from text, table, images etc. Web structure mining classifies relationships between linked web pages. Web usage mining is also an important type that stores user access data and get information about specific user from logs. All techniques may have some advantages and disadvantages but drawbacks can be improved by further studies.

IV. REFERENCES

[1]. Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.

[2]. Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," International Journal of Novel Research in Computer Science and Software Engineering, vol. 2, no. 1, pp. 36-42, January - April 2015.

[3]. Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69 No.8, pp. 39-43, May 2013.

[4]. Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.

[5]. R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940-2945, Augest 2013.

[6]. Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[7]. Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 888), vol. Volume 47 No.11, pp. 44-50, June 2012.

[8]. Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, June 2013.

[9]. Claus Pahl and Dave Donnellan, "Data Mining Technology for the Evaluation of Web-based Teaching and Learning Systems," 7th Int. Conference on E-Learning in Business, Government and Higher Education, October 2002.

[10]. Anurag kumar and Kumar Ravi Singh, "A Study on Web Content Mining," International Journal Of Engineering And Computer Science, vol. 6, no. 1, pp. 20003-20006, January 2017.

[11]. Dr. S. Vijayarani and Ms. A. Sakila, "MULTIMEDIA MINING RESEARCH AN OVERVIEW," International Journal of Computer Graphics & Animation (IJCGA), vol. 5, pp. 69-77, January 2015.

[12]. Tina R. Patil and Mrs. S. S. Sherekar, "16. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal Of Computer Science And Applications, vol. 6, pp. 256-261, April 2013.

[13]. M. Bilal, P. M. L. Chan, and W. Khan, "Cooperative Network for Vehicular Communications: Game Theoretic Distribution of Reward among Contributing Vehicles," Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in

Telecommunications (JSAT), vol. 3, no. 8, pp. 11-25, Augest 2013.

[14]. Miguel Gomes da Costa Júnior and Zhiguo Gong, "Web Structure Mining: An Introduction," International Conference on Information Acquisition, pp. 590-595, June 27 - July 3 2005.

[15]. Anurag Kumar and Kumar Ravi Singh, "A Study on Web Structure Mining," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 1, pp. 715-720, January 2017.

[16]. B. L. Shivakumar and T. Mylsami, "SURVEY ON WEB STRUCTURE MINING," ARPN Journal of Engineering and Applied Sciences, vol. 9, 1914-1923, October 2014.

[17]. Monica Sehgal, "Analysis of Link Algorithms for Web Mining," International Journal of Scientific and Research Publications, vol. 4, no. 5, May 2014.

[18]. Pranit Bari and P.M. Chawan, "Web Usage Mining," Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, pp. 34-38, June 2013.

[19]. Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: A Survey," International Journal of Scientific & Engineering Research, vol. 4, no. 6, pp. 1762-1768, June 2013.

[20]. Sasa Bosnjak, Mirjana Marić, and Zita Bosnjak, "The Role of Web Usage Mining in Web Applications Evaluation," Management Information Systems, vol. 5, October 2009.

[21]. Prabha.K and Suganya.T, "A Guesstimate on Web Usage Mining Algorithms and Techniques," International Journals of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 6, 518-521, June 2017.

[22]. Liupu Wang et al., "Using Internet Search Engines to Obtain Medical Information: A Comparative Study," Journal of Medical Internet Research, May 2012.

[23]. Saeko Nomura, Satoshi Oyama, Tetsuo Hayamizu, and Toru Ishida, "Analysis and Improvement of HITS Algorithm for DetectingWeb Communities," Applications and the Internet, February 2002.

[24]. Yan Wang, Web Mining and Knowledge Discovery of Usage Patterns., February 2000.

[25]. Parth Suthar and Prof. Bhavesh Oza, "A Survey of Web Usage Mining Techniques," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6, pp. 5073-5076, 2015.

[26]. Nasrin JOKAR, Reza Ali HONARVAR, Shima AgHAMIRZADEH, and Khadijeh ESFANDIARI, "Web mining and Web usage mining techniques," Bulletin de la Société des Sciences de Liège, vol. 85, pp. 321 - 328, 2016.

[27]. Surajit Chaudhuri and Umeshwar Dayal, "An Overview of Data Warehousing and OLAP Technology," ACM SIGMOD, vol. 26, no. 1, 65-74, March 1997.

[28]. Ayse Yasemin SEYDIM, INTELLIGENT AGENTS: A DATA MINING PERSPECTIVE. Dallas, May 1999.

[29]. Ajith Abraham, "BUSINESS INTELLIGENCE FROM WEB USAGE MINING," Journal of Information & Knowledge Management, vol. 2, no. 4, December 2003.

[30]. M.SANTHANAKUMAR and C.CHRISTOPHER COLUMBUS, "Web Usage Based Analysis of Web Pages Using RapidMiner," WSEAS TRANSACTIONS on COMPUTERS, vol. 14, pp. 455-464, 2015.

[31]. Aanum Shaikh, "Web Usage Mining Using Apriori and FP Growth Algorithm," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6, pp. 354-357, 2015.

[32]. James E. Pitkow and Krishna A. Bharat, "WEBVIZ: A TOOL FOR WORLD-WIDE WEB ACCESS LOG ANALYSIS," In Proceedings of the First International WWW Conference, January 1994.

TABLE I. USAGE MINING TECHNIQUES COMPARISON

Usage Mining Techniques	Methods	Data Gathering	Data Store	Advantages	Important Algorithms
Data Preprocessing	- Web status codes	- Data logs - Website - Users login information - Web access logs - Cache - Cookies etc.	- Web logs	- Convert raw data to understandable Common LF and Extended CLF for recording	- Apriori algorithm - FP Growth
Pattern Discovery	- Frequency, median, mode used to show length, recently accessed, view time of pages	- Filtered data from preprocessing section	- Session logs	- Extract useful information from discovered patterns correlations	- K-means with Genetic algorithms - Fuzzy c-mean Algorithm
Pattern Analysis	- Roll-up - Drill Down/Up	- Pattern discovery	- Data cube (multi-dimensional database)	- Irrelevant rules and patterns are separated	- SQL Language - OLAP

Table II. Structure Mining Algorithms Comparison

Algorithm	Page Rank	HITS	Weighted Page Rank
Main Technique	Web Structure mining	Web Structure mining, Web Content mining	Web Structure mining
Methodology	It score for pages at the time of indexing	It uses hubs and authorities of the relevant pages	It performs on the basis of input and output links
Input parameter	Back links	Content, Back links and forward links	Back links and forward links
Quality	Medium	Low	Higher than PR
Search Engine	Google	IBM Search Engine	Research Model
Limitation	Query Independent	Efficiency Problem	Query Independent

Cite this article as :

Neha Doomra, Asst.Prof. Rashmi Verma, "Web Mining Algorithms", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 3, pp. 192-200, May-June 2019.

Journal URL : http://ijsrst.com/IJSRST196325