Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview

Article *in* International Journal of Advanced Computer Science and Applications · June 2018

DOI: 10.14569/IJACSA.2018.090630

1 author:

Muhammad Jawad Hamid Mughal
Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Dubai
4 PUBLICATIONS **0** CITATIONS

Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview

Muhammd Jawad Hamid Mughal
Department of Computer Science
SZABIST Dubai Campus
Dubai, United Arab Emirates

***Abstract*—Web data mining became an easy and important platform for retrieval of useful information. Users prefer World Wide Web more to upload and download data. As increasing growth of data over the internet, it is getting difficult and time consuming for discovering informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the web is not an easy task to perform. Different mining techniques are used to fetch relevant information from web (hyperlinks, contents, web usage logs). Web data mining is a sub discipline of data mining which mainly deals with web. Web data mining is divided into three different types: web structure, web content and web usage mining. All these types use different techniques, tools, approaches, algorithms for discover information from huge bulks of data over the web.**

Keywords*—Web data mining; hyperlinks; usage logs; contents; patterns

I. Introduction

Now a day's data over the internet is enormous and increasing frequently day by day. It is must to manage that massive information and display most related queried information on user's screen. Analyzing and fetching relevant data from large data bases is not possible manually, for this automated extraction tools are required through which user queried data can be fetch from billions of pages over the internet and discovers relevant information. Usually users find data from world wild web WWW by using different search engines like Yahoo, Bing, MSN, Google etc. Data mining is a process of analyzing usable information and extract data from large data warehouses, involving different patterns, intelligent methods, algorithms and tools. This process can help business to analyze data, user behavior and predict future trends. Data mining includes four strategies steps for relevant data extraction. Data source is a set on data in large data base which can have problem definition in it. Data exploration is a step of investigation true information from bulks of unfamiliar data. Third step is modeling, in this different models are designed and then evaluate. At the end tested models are deployed, that occurs in final step of data mining strategies. Organizations can use data mining techniques to change raw data into convenient information. It can also help business to improve their marketing strategies and increase the profit by learning more about customer's behavior.

Web mining is one of the types of techniques use in data mining. The main purpose of web mining is to automatically extract information from the web. For discovering useful data (videos, tables, audio, images etc.) from the web different techniques and tools are used. Information over the internet is huge and increasing with passage to time due to which size of data bases are also growing. Digging knowledgeable information and analyzing the data sets for relevant data is much difficult because data over the internet in not in plain text. It could be unstructured data, multimedia, table, tag.

Purpose of this paper is to describe web mining, its three different types, tools and techniques. All three types are explained in detailed and main focus is on web usage mining, its techniques. Summarization table is detailed for all three types.

II. Literature Review

Data mining is a process of discovering knowledge from data warehouse. This knowledge can be classified in different rules and patterns that can help user/organization to analyze collective data and predicted decision processes [9]. Centralized database of any organization is known as Data warehouse, where all data is stored in a single huge database. Data mining is a method that is used by organization to get useful information from raw data. Software's are implemented to look for needed patterns in huge amount of data (data warehouse) that can help business to learn about their customers, predict behavior and improve marketing strategies.

Web mining is actually an area of data mining related to the information available on internet. It is a concept of extracting informative data available on web pages over the internet [1]. Users use different search engines to fetch their required data from the internet, that informative and user needed data is discovered through mining technique called Web Mining. Different tools and algorithms are used for extraction of data from web pages that includes web documents, images etc. Web mining is rapidly becoming very important due to size of text documents increasing over the internet and finding relevant patterns, knowledge and informative data is very hard and time consuming if it is done manually. Structure (Hyperlinks), Usage (visited pages, data use), content (text document, pages) are included in information gathered through Web mining [2], [5]. Term World Wide Web is related to the combination of web documents, videos, audios etc. Some processes included in web mining are:

Information Retrieval is a process of retrieving relevant and useful information over the web. Information retrieval has more focuses on selection of relevant data from large collection of database and discovering new knowledge from large quantity of data to response user query.IR steps includes searching, filtering and matching [5], [6].

Information extraction is an automatic process of extracting analyzed data (structured). IE is a task that work same like information retrieval but more focuses on extracting relevant facts [5].

Machine Learning is support process that helps in mining data from web. Machine learning can improve the web search by knowing user behavior (interest). Different machine learning methods are used in search engine to provide intelligent web service. It is much more efficient than traditional approach i.e. information retrieval. It is a process that has ability to learn user behavior and enhance the performance on specific task.

III. Web Mining Categories

Web Mining is sub categorized in to three types as shown in Fig. 1:

A. Web Content Mining

B. Web Structure Mining

C. Web Usage Mining

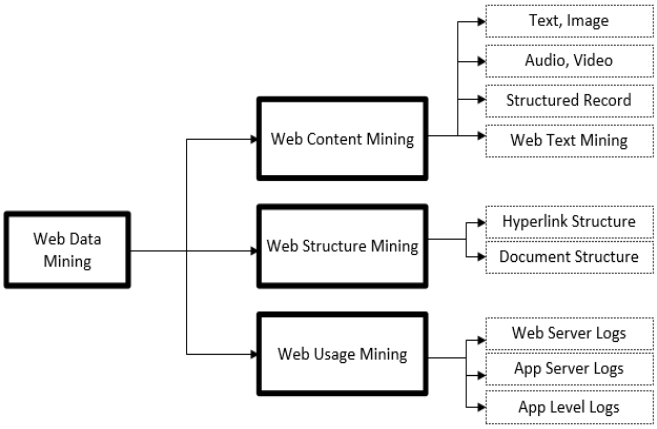

Fig. 1. Web mining taxonomy [8], [15].

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 needs [4].

A. Web Content Mining

Content Mining is a process of Web Mining in which needful informative data is extracted from web sites (WWW). Content includes audio, video, text documents, hyperlinks and structured record [1]. Web contents are designed to deliver data to users in the form of text, list, images, videos and tables. Over last few decades the amount of web pages (HTML) increases to billions and still continues to grow. Searching query into billions of web documents is very difficult and time consuming task, content mining extracts queried data by performing different mining techniques and narrow down the search data which become easy to find required user data [3].

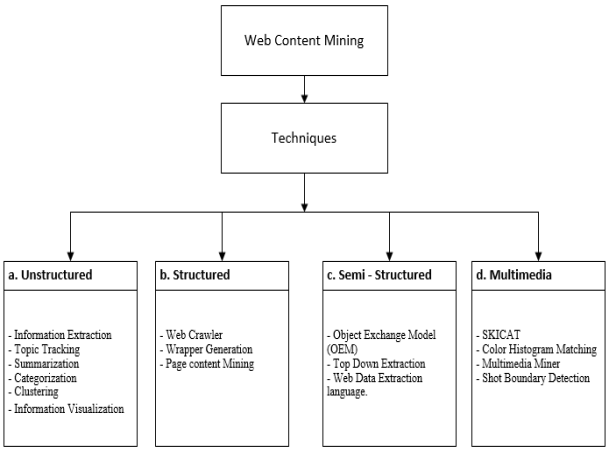

Fig. 2. Web content techniques [7].

1) Web Content Mining Techniques: Web content mining uses different techniques Fig. 2 to dig data. Following are four techniques described used by web content mining.

Mostly in web contents data is in unstructured text form. For extraction of unstructured data, web content mining requires text mining and data mining approaches [5]. Text documents are related to text mining, machine learning and natural language. Main purpose of text mining is to extract previous information from content source [7]. Text mining is a part of web content mining and hence different techniques are used for text data mining from web contents over the internet/website to provide unknown data, some of them are mentioned below:

- Information Extraction
- Summarization
- Information Visualization
- Topic Tracking
- Categorization
- Clustering

Structured is a technique that mines structured data on the web. Structure data mining is an important technique because it represents the host page on the web. Compare to unstructured, in structured data mining it is always easy to extract data [8]. Following are some techniques used for structured data mining:

- Web Crawler
- Page Content Mining
- Wrapper Generation

Semi-Structured is form of structured data but not full, text in semi-structured data is grammatical. Its structure is Hierarchical, not predefined. Representation of semi-structured data is in form of tags (such as HTML, XML). HTML is an intra-document structure case [4]. Techniques used to extract semi-structured data are:

- OEM - Object Exchange Model
- Web Data Extraction Language
- Top Down Extraction [5]

Traditionally computing data was consider as text and numbers but now a days there are different computing data types of multimedia data like videos, images, audios etc. This mining process is use for extracting interesting multimedia data sets and also converted data set types in to digital media [11]. Techniques uses for multimedia data mining are:

- Multimedia Miner
- Shot Boundary Detection
- SKICAT
- Color Histogram Matching [10]

2) Web Content Mining Algorithms: Multiple techniques are used by web mining to extract information from huge amount of data bases. There are different types of algorithms that are used to fetch knowledge information, below are some classification algorithms are described:

Decision tress is a classification and structured based approach which consist of root node, branches and leaf nodes. It is hierarchical process in which root node is split into sub branches and leaf node contains class label. Decision tress is a powerful technique [10].

Naïve Bayes is an easy, simple, powerful algorithm for classification and also known as Native Bayes classifier. It is based on Bayes' Theorem. From predefined dataset values, probabilities are calculated for each class by counting combinations on values. Most likely class is the one with highest probability [12].

Bayes' theorem: [13]

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

Support Vector Machine is a well-known and simple machine learning and classification algorithm. SVM is a method that can be used for linear and non-linear data sets [10]. Optimal separating hyper plane (decision boundary) is just a line that is used to draw to separate the two classes depends on the different classification features.

Neural network is another web content mining approach which use back propagation algorithm. The algorithm consist of multiple layers i.e. input layer, some hidden layers and then output layer, each feeds the next layer till last layer (output). Neuron is the basic unit of neural network. Inputs are fed simultaneously to units. From input layer, inputs are simultaneously feeding to hidden layers. Usually there is one hidden layer but numbers of hidden layers are arbitrary [10]. Last hidden layer fed the input and make up the output layer.

B. Web Structure Mining

Now a day's massive amount of data is increasing on web. World Wide Web is one of the most loved resources for information retrieval. Web mining techniques are very useful to discover knowledgeable data from web. Structure mining is one of the core techniques of web mining which deals with hyperlinks structure [14]. Structure mining basically shows the structured summary of the website. It identifies relationship between linked web pages of websites. Continues growth of data over the internet become a challenging task to find informative and required data [15]. Web mining is just a data mining which digs data from the web. Different algorithmic techniques are used to discover data from web. Structure mining analyzes hyperlinks of the website to collect informative data and sort out in categories like similarities and relationship. Intra-page is a type of mining that is performed at document level and at hyperlink level mining is known as inter-page mining. 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 [16]. Few of the tasks of link-mining Fig. 3 are summarized as:

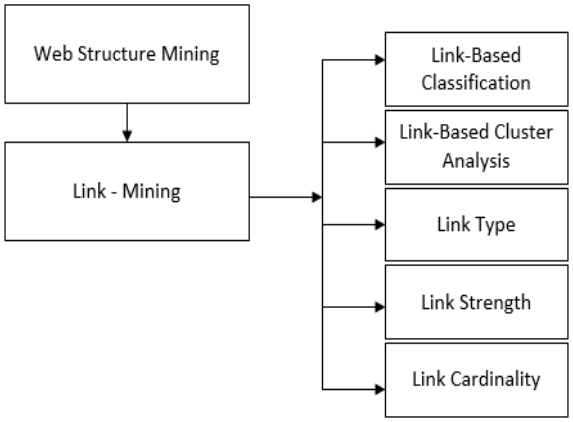

Fig. 3. Web structure mining.

- ***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 [15].
- ***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].

3) 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. 4: 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.

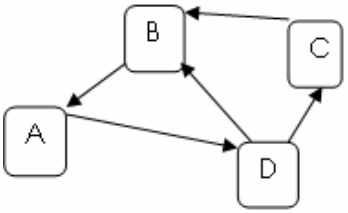

Fig. 4. Back link structure [16].

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

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

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 Fig. 5. 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.

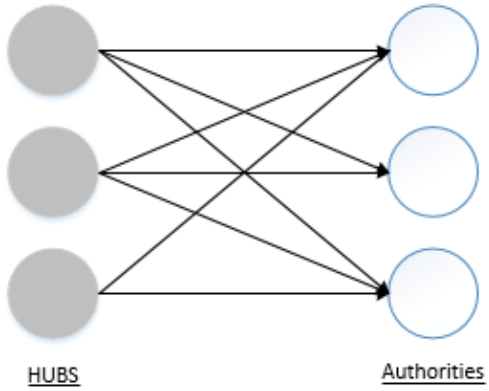

Fig. 5. HITS (Hubs and Authorities) [17].

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 [17].

4) Web Structure Mining Tools: Following mention tools are used for web structure mining. Google PR checker is a tool designed for page rank and is used for Page rank algorithm. It helps to rank of web pages in search engine result. It is simple to find page rank just by pasting website URL and click search – it will show rank of each page of website [22]. Link viewer is used for HITS to visualize analyzes links process [23].

- Google PR Checker (for PageRank)
- Lin Viewer (for HITS)
- Web Mining Categories Summarization

5) Web Usage Mining: Web usage mining also called log mining is a process of recording user access data on the web and collect data in form of logs. After visiting any website user leaves some information behind such as visiting time, IP address, visited pages etc. This information is collected, analyzed and store in logs. Which helps to understand user behavior and later can improves website structure [18]. Web usage mining is a technique that automatically archives access patterns of user and this information is mostly provided by web servers which are later collected in access logs. Logs stores much needed information like URL address, visiting time, Internet Protocol addresses etc. which can help an organization to understand their customer's behavior and insure good service quality. Web usage mining dig and analyze data present in log files which contains user access patterns. Main purpose of web usage mining is to observer user behavior at the time of his interacting with web. There are two types of pattern tracking i.e. general tracking and customized tracking. In general tracking information is collected from web page history. In customized tracking the information is gathered for specific user [19].

TABLE I. SUMMARIZATION TABLE FOR WEB DATA MINING CATEGORIES

Web Mining Categories	Techniques	Tools	Algorithms
Web Content Mining	- Unstructured Data Mining - Structured Data Mining - Semi – Structure Data Mining - Multimedia Data Mining	- Screen Scaper - Mozenda - Automation Anywhere7 - Web Content Extractor - Web Info Extractor - Rapid Miner	- Decision Tree - Naive Bayes - Support Vector Machine - Neural Network
Web Structure Mining	- Link-based Classification - Link-based Cluster Analysis - Link Type - Link Strength - Link Cardinality	- Google PR Checker - Link Viewer	- Page Rank Algorithm - HITS algorithms (Hyperlink Induced Topic Search) - Weighted Page Rank Algorithm - Distance Rank Algorithm - Weighted Page Content Rank Algorithm - Webpage Ranking Using Link Attributes - Eigen Rumor Algorithm - Time Rank Algorithm -Tag Rank Algorithm - Query Dependent Ranking Algorithm
Web Usage Mining	- **Data Preprocessing** • Data Cleaning • User & Session Identification - **Pattern Discovery** • Statistical Analysis • Association Rules • Clustering • Classification • Sequential Patterns - **Pattern Analysis** • Knowledge Query Mechanism • OLAP (Online Analytical processing) • Intelligent Agents	- **Data Preprocessing Tools** • Data Preparator • Sumatra TT • Lisp Miner • SpeedTracer - **Pattern Discovery Tools** • SEWEBAR-CMS • i-Miner • Argunaut • MiDas(Mining In-ternet Data for As-sociative Sequenc-es) - **Pattern Analysis Tools** • Webalizer • Naviz • WebViz • WebMiner • Stratdyn	- **Association Rules** • Apriori Algorithm • Maxi-mal Forward References • Markov Chains • FP Growth • Prefix Span - **Clustering** • Self-Organized Maps • Graph Partitioning • Ant Based Technique • K-means with Genetic algorithms • Fuzzy c-mean Algorithm - **Classification** • Decision Trees • Naïve Bayesian Classifiers • K-nearest Neighbor Classifiers • Support Vector Machine - **Sequential Patterns** • MIDAS (Mining Internet Data for Association Sequences) algorithm

6) Web Usage Mining Techniques: Following three techniques are described in detail with their sub approaches use in web usage mining. Each technique performs different tasks in a hierarchy.

- Data Preprocessing

Real world data and some data bases are incomplete, inconsistent and not understandable. Data preprocessing is a mining technique that integrate databases and make raw data understandable and consistent [18]. In data preprocessing information stored web logs are processed because of insufficient and noisy nature. Raw data cleaning is done is done in early step by removing redundant, useless, error, incomplete, inconsistent data [19]. Preprocessing task is to clean, correct the data and ready input data for mining. There are many e-sources in web usage mining from data can be collected and analyze such as data logs, website, users login information, web access logs, cache, cookies etc. The reliable source for usage mining is considered as web access logs because they use standard logs format (Common LF and Extended CLF) for recording [20]. Data preprocessing includes methods like Data cleaning, User and session identification are describe as follow.

Data cleaning is not only important for usage mining but important for other analysis techniques as well. Purpose is to remove irrelevant and no needed information from logs. Graphics and videos needs to be removed from web logs as they are unnecessary for usage mining [21]. When user requests for a web server for a particular web page, multiple entries are stored in log file. Those records that are not useful for usage mining must be removed.

User and Session identification technique is used to find user sessions from access log file. After data cleaning next step is to identify users. Different approaches are used for user identification like user login information, cookies to detect visitors with unique ID for specific webpage. Session identification is to know number of pages visited by a single user in a row on one visit to a website. Session is a set of

webpage visited by users, new IP mean new user. Difficult step is when proxy server is used, same IP addresses for different users in log file. Referrer method is suggested as a solution to this problem. As different IP indicates new users, if IP's are same then different browsers / OS can identify new users. If OS, IP and browsers are same then Referrer approach consider URL account information. If account in URL was not accessed before it will consider it as a new user [18].

- Pattern Discovery

Consider as key component in web data mining. After data cleaning and user identification, some web usage pattern discovery techniques are used to discover interesting patterns. Main and tough task is to discover patterns produced by preprocessing section and extract useful knowledge [19]. Pattern discovery techniques are describe as follow.

Statistical analysis is a powerful technique used for extracting knowledge about webpage visitors. Analysts perform to describe statistical analysis on session log while analyzing using different variables. Knowledge obtained by statistical analyzing result may help to improve performance and enhance the system security as well as marketing strategies [24]. Frequency, median and mode are three statistical analyses are used mostly on sessions to show length of page, recently accessed pages and view time [19].

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 webpages that appears in a user session repeatedly. The rule describes the relationship between pages visited one after another by user at the time of his visit session. The rule X => Y (where X and Y are pages) state that items (transaction) includes in page X also contain in page Y [26]. Rule format can be shown as:

X.html, Y.html => Z.html

It means that if user will observe page X and Y, most probable he will also observe page Z in the same session.

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. Users cluster provides information about set of users with a similar activities or browsing patterns [25]. Similar webpage content can be discovered from pages clusters. Different algorithms are used for clustering technique as shown in Table I.

Classification technique is use to classify data items and map them to different predefined classes. In usage mining, one with an interest of generating user profile will use this technique to establish user profile of user fitting to particular class [24]. Classification can be performed by use of different algorithms as mentioned in Table I.

Sequential sessions are discovered in sequential patterns. Many algorithms are used to find sequential patterns in usage mining, some of them are listed in Table I. MIDAS is commonly used algorithm for finding sequential sessions [19]. This technique catches patterns like one or multiple bulks of pages visited or accessed one after another in same time sequence. It is helpful for web admin/marketer to predict trends and prepare advertisements, place them to target group of users [25].

- Pattern Analysis

Pattern analysis is considered as a last and final step of usage mining. In this step all not interesting, irrelevant rules or patterns discovered in above phases are separated and interesting or relevant rules or patterns are extracted. This can help to improve system performance [24]. Following are the approaches uses for pattern analysis:

For query mechanism the most commonly language used is SQL. SQL stands for Structured Query Language and is use to extract useful information from patterns discovered [18].

After the pattern discovery data is receive into OLAP phase. In this phase data is store in data cube (multi-dimensional database) format and OLAP operation (roll up etc.) is performed. In OLAP measure term refers to dimensions (tables) [27].

An agent can be defining as an assistant that can help to perform some tasks on user's behalf. An intelligent agent can sense receiving element, recognizer them and determines which task should be performed. In usage mining agents analysis the pattern that are discovered at previously phases [18], [28].

7) Web Usage Mining Algorithms: For 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 [31]. Example: Suppose we have two transactions $A1 = \{1, 2, 3\}$ and $A2 = \{2, 3, 4\}$ where 1,2,3,4 are item sets and 2, 3 are frequent items in both transactions 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 [31].

Both Apriori and FP growth algorithms are suitable and scalable for association rule but FP growth is considered more efficient that Apriori algorithm but in Apriori full database needs to be scan for frequent sets of data where as in FP growth, FP tree is made and new sets are updated while transaction.

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: [29], [30]

$$J(U, c_1, \dots c_c) = \sum_{i=1}^{C} Ji = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}$$

Where: Ci = Cluster Center

Uij = Numerical Value [0, 1]

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

ith , jth = Cluster Center , data points

8) Web Usage Mining Tools: Speed Tracer is an analysis tool and use for usage mining. This tool help to discover users surfing behavior and analyze with entries stored in server logs by using different data mining techniques. Cookies are not required for identifying user session, speed tracer used different kind for information like: IP address, URL of page, agent etc. Collection of browsing patterns helps to understand user behavior in a better way [18]. Three types of understanding are generated by speed tracer: User based which refers to user access time duration. Path based relates to process of frequently visited path in web. Group based generates information of repeatedly visited groups of web pages.

Suggest 3.0 is a system that provides familiar information to user about web pages they might have interest in. Customers/user needs are successfully achieved by set of constant changes in page links. Suggest 3.0 uses graph partitioning algorithm for historical information to be maintained. Main purpose is to keep record of incremental change [18].

WebViz is a tool that is used for statistical analysis of web access logs. The main idea of developing this tool was to provide WWW database designers graphical outlook of their local db and access patterns. Relation between access logs and databases (local) is displayed by use of Web-Path paradigm [32]. It presents local database documents and association of documents in graph structure. Information about accessed documents is collected from access log. Number of visited paths by users is also collected for display.

9) Web Usage Mining Techinques Comparison

IV. Conclusion

Data mining is a concept that helps to find information which is needed from large data warehouses by using different techniques. It is also used to analyze past data and improve future strategies. Web data mining is considered as sub approach of data mining that focuses on gathering information from web. Web is a large domain that contains data in various forms i.e.: images, tables, text, videos, etc. As size of web is continuously increasing; it is becoming very challenging task to extract information. In this paper we described three important types of web data mining that can help in finding informative data. Each type has different algorithms, tools and techniques that are used for data retrieval. Various algorithms, tools and techniques for each type are described. Table I summarizes all types and Table II 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.

TABLE II. 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

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, pp. 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] Saša Bošnjak, Mirjana Marić, and Zita Bošnjak, "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, pp. 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, pp. 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.