American Journal of Systems and Software, 2015, Vol. 3, No. 2, 44-61
Available online at http://pubs.sciepub.com/ajss/3/2/3

DOI:10.12691/ajss-3-2-3

World towards Advance Web Mining: A Review

Shyam Nandan Kumar*

M.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology-Indore (RGPV, Bhopal), MP, India
*Corresponding author: shyamnandan.mec@gmail.com

Received March 28, 2015; Revised April 05, 2015; Accepted April 16, 2015

Abstract With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the Web generate a large amount of data day-by-day. The abundant unstructured or semi-structured information on the Web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the Web. In this article, an attempt has been made to review the various web mining techniques to discover fruitful patterns from the Web, in detail. New concepts are also included in broad-sense for Optimal Web Mining. This paper also discusses the state of the art and survey on Web Mining that is used in knowledge discovery over the Web.

Keywords: *data mining, www, web mining, cloud mining, web usage mining, web content mining, web structure mining, semantic web mining, web mining algorithm, knowledge discovery, information retrieval*

Cite This Article: Shyam Nandan Kumar, "World towards Advance Web Mining: A Review." *American Journal of Systems and Software*, vol. 3, no. 2 (2015): 44-61. doi: 10.12691/ajss-3-2-3.

1. Introduction

Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – ***Users***, ***Web service providers***, ***Business analysts***. The users want to have the effective search tools to find relevant information easily and precisely. To find the relevant information, users either browse or use the search service when they want to find specific information on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response in the list of pages ranked based on their similarity to the query. But due to the problems [1] with browser like: Low precision, which is due to the irrelevance of many of search results, and Low recall, which is due to the inability to index all the information available on the Web, users feel difficulty to find the relevant information on the web. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs, like what the customer do and want. Mass customizing the information to the intended user or even to personalize it to individual customer is the big problem. Web mining is expecting tools or techniques to solve the above problems encountered on the Web. Sometimes, web mining techniques provide direct solution to above problems. On the other hand, web mining techniques can be used as a part of bigger applications that addresses the above problems. Other related techniques from different research areas, such as database, information retrieval, and natural language processing, can also be used. Therefore, Web mining becomes a very hot and popular research field.

Web mining combines two of the activated research areas: ***Data Mining*** and ***World Wide Web***. Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. It extracts the hidden predictive information from large database. With the widespread use of data and the explosive growth in their size, organizations are faced with the problem of information overload. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Web mining, when looked upon data mining terms, can be said to have three operations of interests: ***Clustering*** (e.g., finding natural grouping of users, pages, etc.), ***Association*** (e.g., which URLs tend to be requested together), ***Sequential Analysis*** (e.g., the order in which URLs tends to be accessed). As in most real world problems, the clusters and associations in web mining do not have clear-cut boundaries and often overlap considerably. The unstructured feature of Web data triggers more complexity of Web mining. In the present time, it is not easy task to retrieve the desired information because of more and more pages have been indexed by search engines. So, this redundancy of resources has enhanced the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Data mining" [3]. Etzioni [4] came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data

sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become interesting topic.

Web mining is also an important component of content pipeline for web portals. It is used in data confirmation and validity verification, data integrity and building taxonomies, content management, content generation and opinion mining [2]. Web mining - is the application of data mining techniques to discover patterns from the Web. It is also related to text mining because much of the web contents are texts. According to analysis targets, web mining can be divided into three different types, which are ***Web usage mining***, ***Web content mining*** and ***Web structure mining***.

In this paper sections are organized as follows: Section 2 gives the idea about web mining and its types. Section 3 discusses the comparison of web mining with data mining and text mining. Description of ways of web mining techniques is explained in the section 4, 5 and 6. These sections focus on types of web mining approaches in detail. Section 7 describes Semantic Web Mining. Various web mining algorithms are given in section 8. Section 9 outlines the issue and challenges that are associated with web mining. Application areas of web mining are classified in section 10. Section 11 concludes the paper and presents avenues for future work. References for this paper are given in section 12.

2. Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity of Web mining. Web mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence, and also psychology and statistics as well.

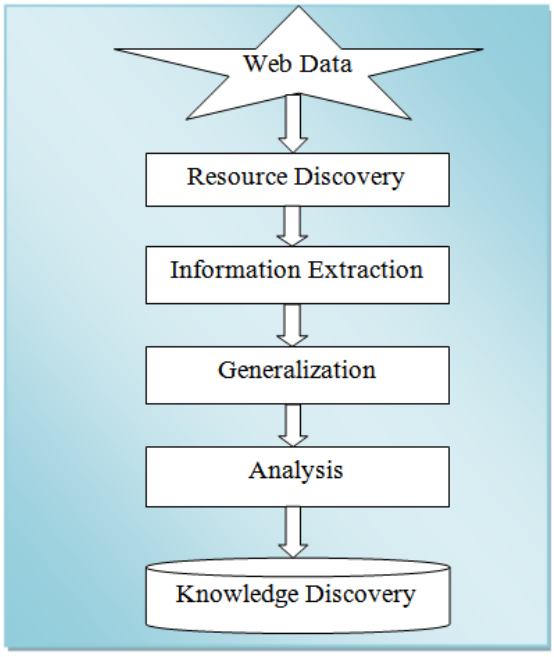

Figure 1. Steps of Web Mining

Web mining involves the analysis of Web server logs of a Web site. The Web server logs contain the entire collection of requests made by a potential or current customer through their browser and responses by the Web server. The information in the logs varies depending on the log file format and option selected on the Web server. Analysis of the Web logs can be insightful for managing the corporate e- business on a short-term basis; the real value of this knowledge is obtained through integration of this resource with other customer touch point information. Common applications include Web site usability, path to purchase, dynamic content marketing, user profiling through behavior analysis and product affinities.

Similar to [4], as shown in Figure 1, decomposition of web mining can be suggested into the following sub-tasks:

- **Resource Discovery**: the task of retrieving the intended information from Web.
- **Information Extraction**: automatically selecting and pre-processing specific information from the retrieved Web resources.
- **Generalization**: automatically discovers general patters at the both individual Web sites and across multiple sites.
- **Analysis**: analyzing the mined pattern.

Based on the main kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining as shown in Figure 2. Summary of Web Mining and its types are given in Table 4.

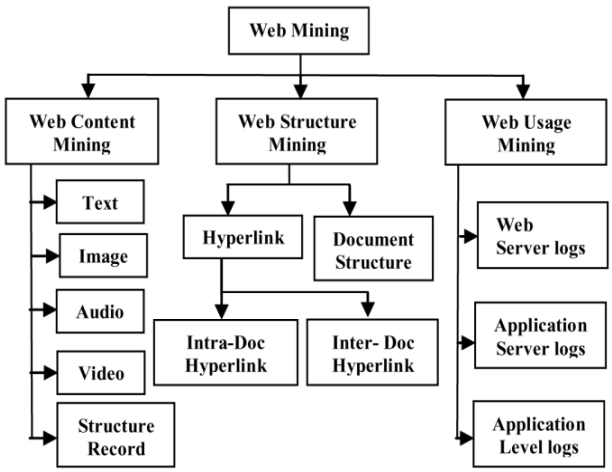

Figure 2. Types of Web Mining

2.1. Cloud Mining

Cloud computing [5,6] has become a viable mainstream solution for data processing, storage and distribution. It promises on demand, scalable, pay-as-you-go compute and storage capacity. To analyze “Big Data” [7] on clouds, it is very important to research data mining strategies based on cloud computing paradigm from both theoretical and practical views. Association rules, Clustering, Anomaly detection, Regression, and Classification are frequently used to mine the data over the cloud by supporting MapReduce [7] parallel computing platform. The implementation of data mining techniques through Cloud computing can allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and

storage. In brief, a MapReduce computation executes as follows:

- Some numbers of Map tasks each are given one or more chunks from a distributed file system. These Map tasks turn the chunk into a sequence of key-value pairs. The way key-value pairs are produced from the input data is determined by the code written by the user for the Map function.
- The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.
- The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by the user for the Reduce function.

Cloud Mining can be classified as: Service Mining, Deployment Mining, Architecture Mining and Workflow Mining, as shown in Figure 3.

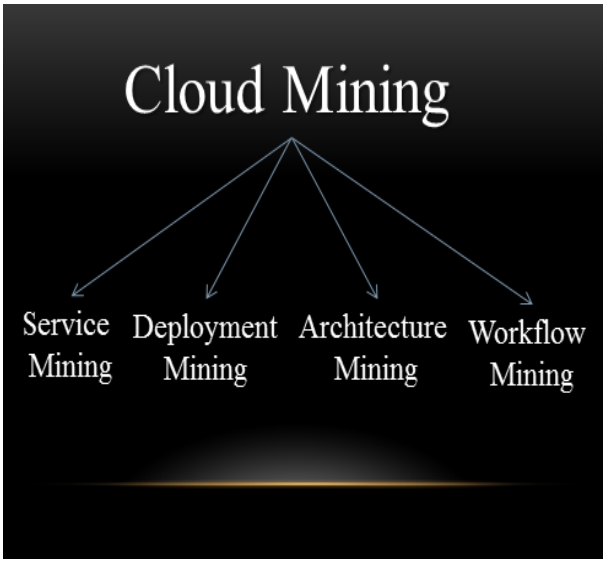

Figure 3. Cloud Mining Types

2.1.1. Service Mining

To quantitatively measure quality of service, several related aspects of the network service are often considered, such as error rates, bandwidth, throughput, transmission delay, availability, jitter, etc. Quality of service is particularly important for the transport of traffic with special requirements. Various cloud clients are interacted with cloud based services. Some known services are: Infrastructure as a service (***IaaS***), Platform as a service (***PaaS***) and Software as a service (***SaaS***). IaaS clouds often offer additional resources such as a virtual-machine disk image library, raw block storage, and file or object storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles. PaaS includes operating system, programming language execution environment, database, and web server. In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Within these services there is a need of advance methodology. New Cloud Services can be mined for providing optimal services. Quality management is the most important issue in cloud mining. End-to-end quality of service can require a method of coordinating resource allocation between one autonomous system and another. Quality of service guarantees are important if the cloud capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication.

2.1.2. Deployment Mining

In deployment mining, new cloud patterns can be discovered. Discovery of new types of cloud computing mechanism is required for satisfying the customer requirements. Based on cloud type application and services can be mined. In this case, knowledge discovery depends upon cloud types. Cloud platforms provide scalable processing and data storage and access services that can be exploited for implementing high-performance knowledge discovery systems and applications.

2.1.3. Architecture Mining

Cloud architecture typically involves multiple cloud components communication. Methods of efficient organization of these components are required. Cloud Computing Architectures and Cloud Solution Design Patterns can be mined under architecture mining. Cloud computing offers an effective support for addressing both the computational and data storage needs of Big Data mining and parallel analytics applications. In fact, complex data mining tasks involve data- and compute-intensive algorithms that require large storage facilities together with high performance processors to get results in acceptable times.

Cloud engineering brings a systematic approach to the high-level concerns of commercialization, standardization, and governance in conceiving, developing, operating and maintaining cloud computing systems. It is a multidisciplinary method encompassing contributions from diverse areas such as systems, software, web, performance, information, security, platform, risk, and quality engineering.

2.1.4. Workflow Mining

It involves mining the various techniques to minimize the workload over the cloud. Traffic Handling is one of the important issues over the cloud. To provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow, workflow mining is required. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Autonomic Business Process and Workflow Management in Clouds should be efficiently managed.

Mining of cloud based framework leads development of distributed data analytics applications as workflows of services.

3. Data Mining vs. Web Mining

Data mining is the process of non-trivial discovery from implied, previously unknown, and potentially useful

information from data in large databases. Hence it is a core element in knowledge discovery, often used synonymously. Data mining involves using techniques to find underlying structure and relationships in large amounts of data. Common data mining applications discover patterns in a structured data such as database (i.e. DBMS). The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages: (1) Selection, (2) Pre-processing, (3) Transformation, (4) Data Mining, and (5) Interpretation/Evaluation.

Web mining describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of web data. The analyzed web resources contain (1) the actual web site (2) the hyperlinks connecting these sites and (3) the path that online users take on the web to reach a particular site. Web usage mining then refers to the derivation of useful knowledge from these data inputs. Web mining discovers patterns in a less structured data such as Internet (***WWW***). In other words, we can say that Web Mining is Data Mining techniques applied to the ***WWW***.

Text mining is different from what we're familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down. In text mining the patterns are extracted from natural language text rather than from structured databases of facts. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Table 1 shows the comparison between data mining and web mining while Table 2 shows the comparison between text mining and web mining.

Table 1. Data Mining vs. Web Mining

Data Mining	Web Mining
Data mining involves using techniques to find underlying structure and relationships in large amounts of data.	Web mining involves the analysis of Web server logs of a Web site.
Common data mining applications discover patterns in a structured data such as database.	Web mining; likewise discover patterns in a semi-structured data such as Internet (WWW). In other words, we can say that Web Mining is Data Mining techniques applied to the WWW.
It can handle large amount of Data**.**	It can handle big data compare than traditional data mining.
When doing data mining of corporate information, the data is private and often requires access rights to read.	For web mining, the data is public and rarely requires access rights.
A traditional data mining task gets information from a database, which provides some level of explicit structure.	A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

3.1. Text Mining vs. Web Mining

Table 2. Text Mining vs. Web Mining

Text Mining	Web Mining
Sub-domain of Information Retrieval(IR) and Natural Language Processing	Sub-domain of IR and multimedia
Text Data: free-form, unstructured & semi-structured data	Semi-structured data: hyper-links and html tags Multimedia data type: Text, image, audio, video.
Content management & information organization.	Content management/mining as well as usage/traffic mining.
Patterns are extracted from natural language text rather than from structured database.	Patterns are extracted from Web rather than from structured database.

4. Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. It tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Usage data captures the identity or origin of Web users along with their browsing behavior at a web site. It deals with studying the data generated by web surfer's sessions or behaviors. Since the web content and structure mining utilize the real or primary data on the web. On the contrary, web usage mining mines the secondary data derived from the interactions of the users with the web. The secondary data includes the data from the proxy server logs, browser logs, web server access logs, user profiles, user sessions, user queries, registration data, bookmark data, mouse clicks and scrolls, cookies and any other data which are the results of these interactions. Log file pros and cons are given in Table 3. High level architecture of different web logs is shown in Figure 4.

Table 3. Log File Pros and Cons

File Type	Advantage	Disadvantage	Mapping
Client Log File	Authentic and Accurate	Modification, Collaboration	One to Many
Server Log File	Reliable and Accurate	Incomplete	Many to One
Proxy Log File	Control Efficiency of Corporate Access to the Internet, Log Traffic	Complex, Unsecure	Many to Many

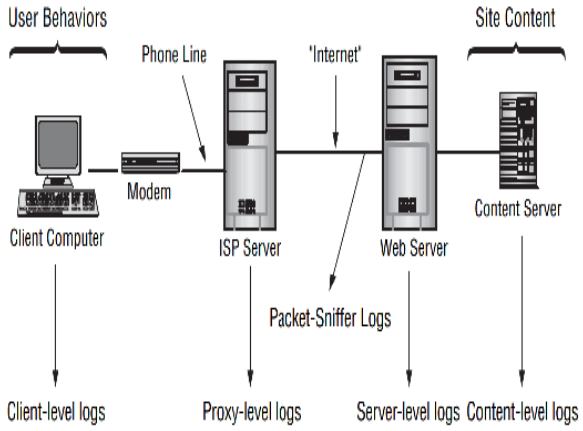

Figure 4. High Level Architecture of Different Web Logs

4.1. Phase of Web Usage Mining

There are generally three distinctive phases in web usage mining: Data collection and preprocessing, Knowledge discovery, and pattern analysis as shown in Figure 5.

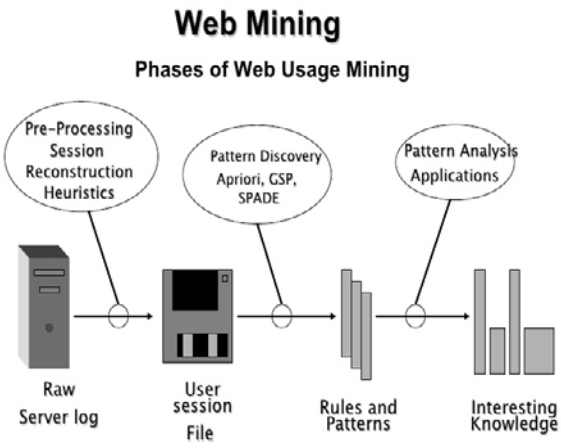

Figure 5. Phases of Web Usage Mining

4.1.1. Data Collection and Pre-processing Phase

It deals with generating and cleaning of web data and transforming it to a set of user transactions representing activities of each user during his/her website visit. This step will influence the quality and result of the pattern discovery and analysis. Therefore, it needs to be done very carefully. Details explanation of preprocessing phase of web usage mining is shown in Figure 6.

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

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

User and Session Identification: The task of user and session identification finds out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study. The rules adopted to distinguish user sessions can be described as follows:

- The different IP addresses distinguish different users;
- If the IP addresses are same, the different browsers and operation systems indicate different users;
- If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified in the URL in the Refer URI field hasn't been accessed previously, or there is a large interval usually (more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty;
- The session identified by rule 3 may contains more than one visit by the same user at different time, the time oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

Page view: Visual rendering of a Web page in a specific client environment at a specific point in time.

Path completion: In order to reliably identify unique user session, it should determine if there are important accesses that are not recorded in the access log. Reason causes such matter is mainly due to the presence of Cache. If user clicks backward to visit a page that has had a copy stored in Cache, browser will get the page directly from the cache. Such a page view will never be trailed in access log, thus causing the problem of incomplete path, which need mending. To accomplish this task needs to refer to referrer log and site topology. If the referred URL of a requesting page doesn't exactly match the last direct page requested, it means that the requested path is not complete. Furthermore, if the referred page URL is in the user's recent request history, we can assume that the user has clicked the "backward" button to visit page. But if the referred page is not in the history, it means that a new user session begins, just as we have stated above. We can mend the incomplete path using heuristics provided by referrer and site topology.

Episode Identification: It is the final step of preprocess. An episode is a subset of related user clicks that occur within a user session. In web usage mining [8] for example, user session is believed equivalent to transaction. It regards user session mentioned above as user session based on duration and transaction here as user session based on structure. As Web usage mining aims at

analyzing navigation patterns of users, so it believes it is reasonable to take session as transaction.

In order to divide user session into transaction, Web Miner classifies pages in a session into ***auxiliary pages*** and ***content pages***. Auxiliary pages are those that are just to facilitate the browsing of a user while searching for information. Content pages are those that user are of interest and that they really want to reach. Using the concept of auxiliary and content pages references, there are two ways to define a transactions. The first would be to define a transaction as all of the auxiliary references up to and including each content reference for a given user, which is a so-called ***auxiliary-content transaction***. The second method would be to define a transaction as all of the content references for a given user, which is ***content-only transaction***. Based on the two definitions, Web Miner employs two methods to identify transaction: one is ***reference length***; the other is ***maximal forward reference***. It also uses time window as a benchmark to evaluate the two methods.

A text episode can be define as a pair $E = (V, \leq)$, where V is the collection of feature vector and $\leq$ is a partial order in V. Feature vector is an ordered set of features, where a feature can be any of the textual feature. For a given text sequence S, a text episode $E = (V, \leq)$ occurs within S if there is a way of satisfying the feature vector in V, using the feature vectors in S so that the partial order $\leq$ is represented. In other words, the feature vectors of V can be found within S in an order that satisfies $\leq$.

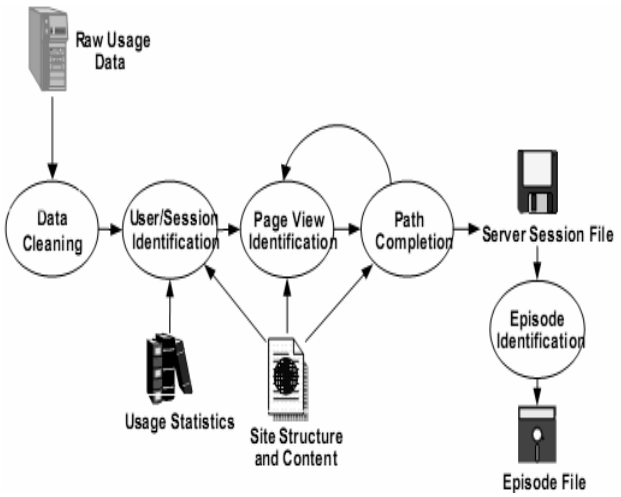

Figure 6. Steps of Preprocessing Phase during Web Usage Mining

4.1.2. Pattern Discovery Phase

Knowledge or pattern discovery is the key component of the Web mining, which uses the algorithms and techniques from data mining, machine learning, statistics and pattern recognition etc. research categories. Use statistical method to carry on the analysis and mine the pretreated data. We may discover the user or the user community's interests then construct interest model. At present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

Statistical Analysis: The analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file; the statistical techniques are the most powerful tools in extracting knowledge about visitors to a Web site. By analyzing the statistical information contained in the periodic Web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions.

Association Rules: It refers to sets of pages that are accessed together with a support value exceeding some specified threshold. This technique can be used to discover unordered correlation between items found in a database of transactions. When loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency.

Dependency Modeling: The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the Web domain. This technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future Web resource consumption.

Clustering: It is a technique to group users or data items (pages) together, with the similar characteristics. It can facilitate the development and execution of future marketing strategies. Clustering of users helps to discover the group of users, who have similar navigation pattern. It's very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The clustering of pages is useful for Internet search engines and Web service providers, since it can be used to discover the groups of pages having related content.

Sequential Pattern: It concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. The task of discovering frequent sequences is challenging, because the algorithm needs to process a combinatorial explosive number of possible sequences. Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min-support threshold, sequential pattern mining is to find all frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than min-support.

Classification: The technique to map a data item into one of several predefined classes, which help to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe te properties of a given class or category. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc.

4.1.3. Pattern Analysis Phase

Pattern Analysis is the final stage of the Web usage mining. Challenges of Pattern Analysis are to filter uninteresting information and to visualize and interpret the interesting patterns to the user. Analysis methodologies and tools used for this are Query mechanism like SQL, OLAP, and Visualization etc. This is a much fertilized research area. First delete the less significance rules or models from the interested model storehouse; Next use

technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce website.

4.2. Classification of Web Usage Mining

Web Usage data can be accumulated by the web server. Analysis of the web access logs of different websites can facilitate an understanding of the user behavior and web structure, thereby improving the design of this colossal collection of information. Web Usage Mining can be classified into following categories: Web Server Data, Application Server Data and Application Level data, as shown in Figure 2.

Web Server Data: User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data: Commercial application servers such as Weblogic and StoryServer have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above the categories.

4.3. Application of Web Usage Mining

The application of web usage mining could be classified into two main categories: Learning a user profile or User modeling in adaptive interface (personalized) and Learning user navigation patterns (Impersonalized) [9].

5. Web Content Mining

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables as shown in Figure 2. There may be also metadata as well as hyperlinks. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP).

Some of the web content data are hidden data, and some are generated dynamically as a result of queries and reside in Database (DB). These data are generally not indexed. The textual parts of web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and more structured data such as data in the tables or database-generated HTML pages. Undoubtedly, much of the web content data is unstructured, free text data. Broad view of Web content mining approaches are shown in Figure 7.

5.1. Unstructured Web Content Mining Approaches

5.1.1. Information Extraction

To extract information from unstructured data, pattern matching is used. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. This technique is very useful when there is large volume of text. Information Extraction (IE) is the basis of many other techniques used for unstructured mining [10].

5.1.2. Topic Tracking

Topic Tracking is a technique in which it checks the documents viewed by the user and studies the user profiles. According to each user it predicts the other documents related to users interest. In Topic Tracking applied by Yahoo, user can give a keyword and if anything related to the keyword pops up then it will be informed to the user. Same can be applied in the case of mining unstructured data.

5.1.3. Summarization

Summarization is used to reduce the length of the document by maintaining the main points. It helps the user to decide whether they should read this topic or not. The time taken by the technique to summarize the document is less than the time taken by the user to read the first paragraph. The challenge in summarization is to teach software to analyze semantics and to interpret the meaning. This software statistically weighs the sentence and then extracts important sentences from the document. To understand the key points summarization tool search for headings and sub headings to find out the important points of that document. This tool also give the freedom to the user to select how much percentage of the total text they want extracted as summary.

5.1.4. Categorization

Categorization is the technique of identifying main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document It decides the main topic from the counts. It ranks the document according to the topics. Documents having majority content on a particular topic are ranked first. Categorization can be used in business and industries to provide customer support [10].

5.1.5. Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. During pattern discovery phase of Web Usage Mining, Clustering is used. In same fashion, clustering is implemented for web content mining for

unstructured data. It helps the user to easily select the topic of interest.

5.1.6. Information Visualization

Visualization utilizes feature extraction and key term indexing to build a graphical representation. Through visualization, documents having similarity are found out [11]. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is useful to find out related topic from a very large amount of documents [10].

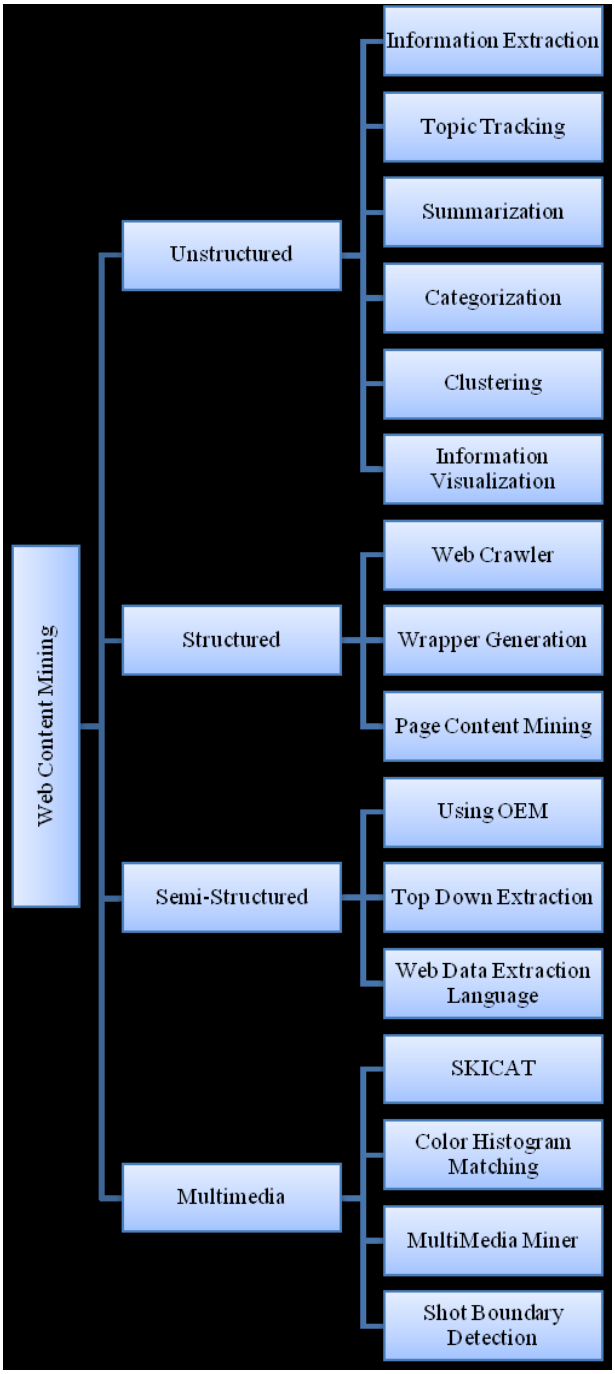

Figure 7. Web Content Mining Approaches

5.2. Structured Web Content Mining Approaches

5.2.1. Web Crawler

A Web crawler [12] is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing. It can validate ***hyperlinks*** and ***HTML code***. They can also be used for web scraping. Web search engines and some other sites use Web crawling software to update their web content or indexes of others site's web content. Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly.

A Web crawler starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies. If the crawler is performing archiving of websites it copies and saves the information as it goes. Such archives are usually stored such that they can be viewed, read and navigated as they were on the live web, but are preserved as 'snapshots'. High-level architecture of a standard Web crawler is shown in Figure 8.

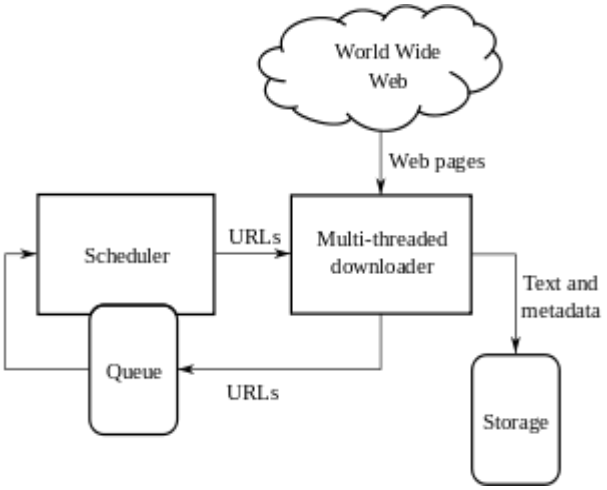

Figure 8. Standard Architecture of Web Crawler

5.2.2. Wrapper Generation

Wrapper in data mining is a program that extracts content of a particular information source and translates it into a relational form [13,14]. There are two main approaches to wrapper generation: ***wrapper induction*** and ***automated data extraction***. Wrapper induction uses supervised learning to learn data extraction rules from manually labeled training examples. Wrapper generation on the Web is an important problem with a wide range of applications. Extraction of such data enables one to integrate data/information from multiple Web sites to provide value-added services, e.g., comparative shopping, object search, and information integration.

5.2.3. Web Page Content Mining

Web page content mining aims to extract/mine useful information or knowledge from web page contents. By comparing page Content rank it classifies the pages. It identifies information within web page and distinguishes home page from other pages. Web page content mining uses two types of approaches: Database approach and Agent based approach. The first approach aims on

modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The second approach aims on improving the information finding and filtering.

5.3. Semi-Structured Web Content Mining Approaches

5.3.1. Object Exchange Model (OEM)

OEM [15] is a model for exchanging semi-structured data between object-oriented databases. It represents semi-structured data by a labeled graph. The data in the OEM is viewed as a graph, with objects as the vertices and labels on the edges. Each object is identified by an object identifier and a value that is either atomic, such as integer, string, etc. or complex in the form of a set of object references, denoted as a set of pair. The database view on Web content mining mainly uses OEM.

5.3.2. Top down Extraction

The page segmentation algorithm relies on the ***DOM tree*** [17] representation on the web page and traverses it in a top-down fashion in order to segment the content of page, which lies on the leaf nodes. A segment can be defined as a contiguous set of leaf nodes within a web page. A web page usually contains several pieces of information and it is necessary to partition a web page into several segments or information blocks before organizing the content into hierarchical groups. Meta-data extracted from the attribute the rich segments of web pages can be used to extract information from text with the help of Natural Language Processing (NLP) techniques. Top down extraction starts with a general rule and then aims to specialize it.

The Document Object Model (DOM) is a cross-platform and language-independent convention for representing and interacting with objects in HTML, XHTML, and XML documents. The nodes of every document are organized in a tree structure, called the DOM tree as shown in Figure 9. Objects in the DOM tree may be addressed and manipulated by using methods on the objects. The public interface of a DOM is specified in its application programming interface (API).

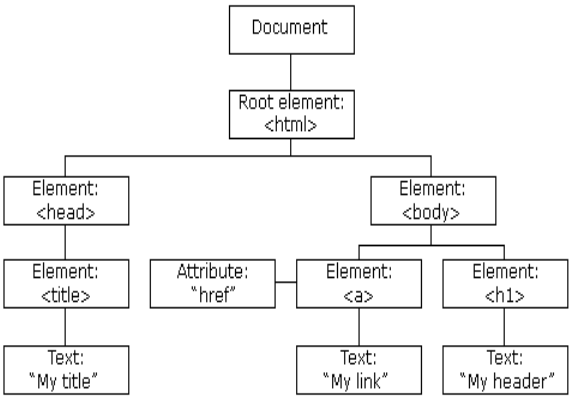

Figure 9. DOM Tree Representation

5.3.3. Web Data Extraction Language

Web data extraction languages are used to convert the other data formats to suitable web data format for supporting optimal services on the web. A Data Extraction Language (DEL) script specifies how to locate and extract fragments from input data and where to insert them in the resulting XML format. Data extraction process retrieves data out of (semi-structured) data sources for further data processing. This growing process of data extraction from the web is referred to as ***Web Scraping*** [16]. Uses of web scraping include online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration.

Web scraping (web harvesting or web data extraction) is a computer software technique of extracting information from websites. Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox. Web scraping is closely related to web indexing, which indexes information on the web using a bot or web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured or semi-structured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software.

5.4. Multimedia Mining Approaches

5.4.1. SKICAT

SKICAT stands for Sky Image Cataloging and Analysis Tool. Since sky's data are scalable and distributed, converting these large amounts of raw data into useful scientific results is a challenging task, which requires an entirely new generation of computing analysis tools. As collaboration between ***Caltech*** and the ***JPL*** Artificial Intelligence group, we have developed a powerful new software system, called ***SK***y ***I***mage ***C***ataloging and ***A***nalysis ***T***ool, or SKICAT. The system incorporates the latest in the AI technology, including machine learning, expert systems, machine-assisted discovery, etc., in order to automatically catalog and measure sources detected in the sky survey images, to classify them as stars or galaxies, and to assist an astronomer in performing scientific analyses of the resulting object catalogs.

5.4.2. Color Histogram Matching

Histogram matching is a method in image processing of color adjustment of two images using the image histograms. An example of histogram matching is shown in Figure 10. It is possible to use histogram matching to balance detector responses as a relative detector calibration technique. It can be used to normalize two images, when the images were acquired at the same local illumination (such as shadows) over the same location, but by different sensors, atmospheric conditions or global illumination.

For given two images, ***the reference*** and the ***adjusted images***, their histograms can be computed. Consider $F_1(\ )$ and $F_2(\ )$ as cumulative distribution functions for the reference image and for the target image respectively. Then for each gray level $G_1 \in [0,255]$, gray level G_2 can

be calculated such that $F_1(G_1) = F_2(G_2)$ and this is the result of histogram matching function: $M(G_1) = G_2$. Now the function M ( ) can be applied on each pixel of the reference image.

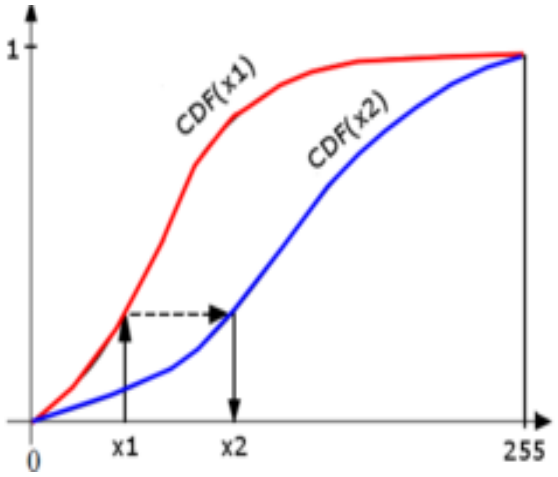

Figure 10. Example of Histogram Matching

The Histogram matching Algorithm can be extended to find a monotonic mapping between two sets of histograms [18]. Given two sets of histograms $P = \{p_i\}_{i = 1 \text{to } k}$ and $Q = \{q_i\}_{i = 1 \text{ to } k}$, the optimal monotonic color mapping M is calculated to minimize the distance between the two sets simultaneously, namely $min_M \sum_k d(M(p_k), q_k)$ where d( ) is a distance metric between two histograms. The optimal solution is calculated using dynamic programming.

5.4.3. Multimedia Miner

Multimedia Miner is a prototype of a data mining system for mining high-level multimedia information and knowledge from large multimedia databases. It includes the construction of multimedia data cubes which facilitate multiple dimensional analysis of multimedia data, and the mining of multiple kinds of knowledge, including summarization, classification, and association, in image and video databases.

In text mining there are two open problems: polysemy, synonymy. Polysemy refers to the fact that a word can have multiple meanings. Synonymy means that different words can have the same/similar meaning.

Image mining is used in variety of fields like medical diagnosis, space research, remote sensing, agriculture, industries, and also handling hyper spectral images. Images include maps, geological structures, and biological structures and even in the educational field. The fundamental challenge in image mining is to reveal out how low-level pixel representation enclosed in a raw image or image sequence can be processed to recognize high-level image objects and relationships.

Mining video data is even more complicated than mining image data. One can regard video to be a collection of moving images, much like animation. The important areas include developing query and retrieval techniques for video databases, including video indexing, query languages, and optimization strategies. In video mining, there are three types of videos: a) the produced (e.g. movies, news videos, and dramas), b) the raw (e.g. traffic videos, surveillance videos etc.), and c) the medical video (e.g. ultra sound videos including echocardiogram). Higher-level information from video includes: i) detecting trigger events (e.g. any vehicles entering a particular area, people exiting or entering a particular building), ii) determining typical and anomalous patterns of activity, generating person-centric or object-centric views of an activity, and iii) classifying activities into named categories (e.g. walking, riding a bicycle), clustering and determining interactions between entities.

In general, audio mining (as opposed to mining transcribed speech) is even more primitive than video mining. Since audio is a continuous media type like video, the techniques for audio information processing and mining are similar to video information retrieval and mining. Audio data could be in the form of radio, speech, or spoken language. To mine audio data, one could convert it into text using speech transcription techniques. Audio data could also be mined directly by using audio information processing techniques and then mining selected audio data.

5.4.4. Shot Boundary Detection

Shot transition detection (or shot boundary detection) also called ***cut detection*** or ***shot detection***, is a field of research of video processing. Its subject is the automated detection of transitions between shots in digital video with the purpose of temporal segmentation of videos. It is used to split up a film into basic temporal units called ***shots***; a shot is a series of interrelated consecutive pictures taken contiguously by a single camera and representing a continuous action in time and space. Shot detection methods can be classified into many categories: pixel based, statistics based, transform based, feature based and histogram based. Basic idea of cut detection is shown in Figure 11, where (1) represents ***Hit***: a detected hard cut, (2) represents ***Missed hit***: a soft cut (dissolve), that was not detected, and (3) represents ***False Hit***: one single soft cut that is falsely interpreted as two different hard cuts.

A digital video consists of frames that are presented to the viewer's eye in rapid succession to create the impression of movement. "***Digital***" in this context means both that a single frame consists of pixels and the data is present as binary data, such that it can be processed with a computer. Each frame within a digital video can be uniquely identified by its frame index, a serial number.

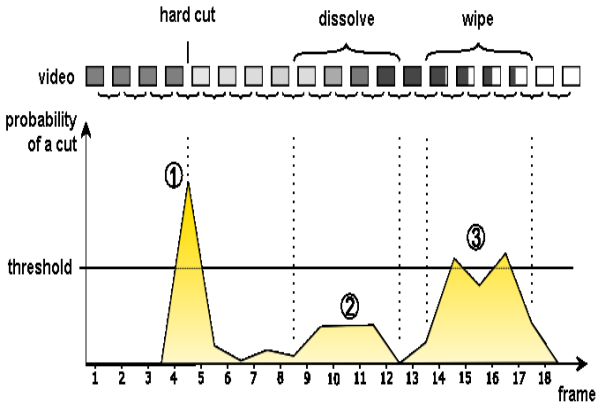

Figure 11. Cut Detection

6. Web Structure Mining

Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship

between Web pages linked by information or direct link connection. It is used to study the topology of ***hyperlinks*** with or without the description of the links. This structure data is discoverable by the provision of web structure schema through database techniques for Web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, linking the information through reference links to bring forth the specific page containing the desired information. The main purpose for structure mining is to extract previously unknown relationships between Web pages. This structure data mining provides use for a business to link the information of its own Web site to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining.

According to the type of web structural data, web structure mining can be divided into two kinds: ***Hyperlinks*** and ***Document Structure*** as shown in Figure 2.

6.1. Hyperlinks

Web Structure Mining extracts patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location, location, either within the same Web page or on a different web page. A hyperlink that connects to a different part of the same page is called an ***Intra-Document Hyperlink***, and a hyperlink that connects two different pages is called an ***Inter-Document Hyperlink***. Hyperlink hierarchy is also determined to path the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links. This enables clustering of connected Web pages to establish the relationship of these pages. A hyperlink points to a whole document or to a specific element within a document. ***Hypertext*** is text with hyperlinks. Link analysis algorithms are given in Section 8.

6.2. Document Structure

Web Structure Mining also mines the document structure. It includes analysis of the tree-like structure of page structures to describe HTML or XML tag usage. It supports ***DOM tree,*** which is described in ***Section 5.3.2.***

Web Structure Mining discovers structure information from the Web. The structure of a typical Web graph consists of Web pages as ***nodes***, and hyperlinks as ***edges*** connecting related pages. Any collection, V, of hyperlinked pages can be viewed as a directed graph $G = (V, E)$: the node corresponding to the pages and a directed edge $(p, q) \in E$ indicates the presence of the link from p to q. The out-degree of a node p is the number of nodes to which it has links, and the in-degree of p is the number of nodes that have links to it. If $W \subseteq V$ is the subset of the pages, $G[W]$ can be used to denote the graph induced on W, whose nodes are the pages in W, and its edges corresponds to all the links between the pages in W.

Table 4. Summary of Web Mining

	Web mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	Db View		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	-Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Serves Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing -User Modeling

The Relevancy of a page to a given query depends on its group and its location and position in the page list. The larger the relevancy value, the better is the result.

$$K = \sum_{i \in R(P)} (n - i) * W_i \quad (1)$$

Where i denote the i_{th} page in the result page-list $R(p)$, n represents the first n pages chosen from the list $R(p)$, and W_i is the weight of i_{th} page.

7. Semantic Web Mining

The research area of Semantic Web Mining is aimed at combining two fast developing fields of research: the ***Semantic Web*** and ***Web Mining***. The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C) [32]. The Semantic Web offers to add structure to the Web, while Web Mining can learn implicit structures. Semantic search seeks to improve search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data-space, whether on the Web or within a closed system, to generate more relevant results. Rather than using ranking algorithms such as Google's PageRank to predict relevancy, semantic search uses semantics or the science of meaning in language, to produce highly relevant search results. In most cases, the goal is to deliver the information queried by a user rather than have a user sort through a list of loosely related keyword results.

7.1. Semantic Web

The current WWW has a huge amount of data that is often unstructured and usually only human understandable. The Semantic Web aims to address this problem by

providing machine interpretable semantics to provide greater machine support for the user. ***Microformats*** extend HTML syntax to create machine-readable semantic markup about objects including people, organizations, events and products. The Semantic Web addresses the second part of this challenge by trying to make the data machine understandable, while Web Mining addresses the first part by automatically extracting the useful knowledge hidden in these data, and making it available as an aggregation of manageable proportions. The Semantic Web has a layer structure that defines the levels of abstraction applied to the Web. At the lowest level is the familiar World Wide Web, then progressing to XML, RDF, Ontology, Logic, Proof and Trust [19] as shown in Figure 12. The main tools that are currently being used in the Semantic Web are ontologies based on ***OWL*** (Web Ontology Language). The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Ontology is an agreed vocabulary that provides a set of well-founded constructs to build meaningful higher level knowledge for specifying the semantics of terminology systems in a well-defined and unambiguous manner. For a particular domain, ontology represents a richer language for providing more complex constraints on the types of resources and their properties. Compared to taxonomy, ontologies enhance the semantics of terms by providing richer relationships between the terms of a vocabulary.

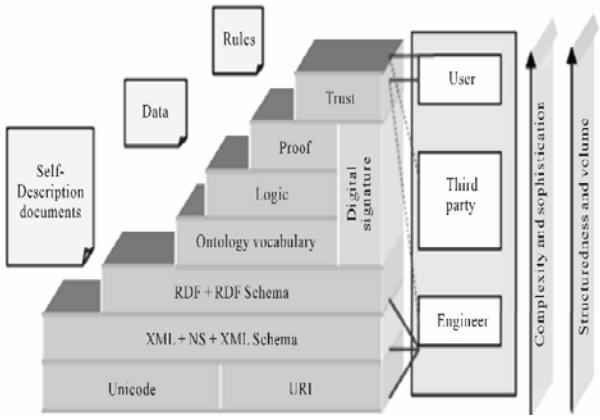

Figure 12. Layers of Semantic Web

The idea of the Semantic Web is introduced by the inventor of the World Wide Web, Tim Berners-Lee. The perspective of semantic web, its technologies and applications is depicted shortly in Figure 13. [20]. Brief explanation is given below.

Uniform Resource Identifier (URI): A universal resource identifier is a formatted string that serves as a means of identifying abstract or physical resource. Such identification enables interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols. Schemes specifying a concrete syntax and associated protocols define each URI. The most common form of URI is the uniform resource locator (URL), frequently referred to informally as a web address. A URI can be further classified as a locator, a name, or both. One of the advantages of using URIs is that they can be dereferenced using the HTTP protocol.

Extensible Markup Language (XML): It has been established as a generic technique to store, organize, and retrieve data on/from the web. By enabling users to create their own tags, it allows them to define their con-tent easily. Therefore, the data and its semantic relation-ships can be represented.

Resource Description Framework (RDF): The Resource Description Framework (RDF) is a common language that enables the facility to store resources' information that are available in the World Wide Web using their own domain vocabularies [19, 22]. Three types of elements contented in the RDF: ***resources*** (entities identified by Uniform Resource Identifiers URIs), ***literals*** (atomics values like strings and numbers), and ***properties*** (binary relationships identified by URIs). This is a very effective way to represent any kind of data that could be defined on the web [22].

Metadata: Metadata is data that describes other data. They serve to index Web pages and Web sites in the Semantic Web, allowing other computers to acknowledge what the Web page is about. In addition to document files, metadata is used for images, videos, spreadsheets and web pages. The use of metadata on web pages can be very important. Metadata for web pages contain descriptions of the page's contents, as well as keywords linked to the content. These are usually expressed in the form of meta-tags. The metadata containing the web page's description and summary is often displayed in search results by search engines, making its accuracy.

Ontology Web Language (OWL): The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies. Ontologies are a formal way to describe taxonomies and classification networks, essentially defining the structure of knowledge. It is considered as more complex language with better machine-interpret- ability than RDF. It precisely identifies the resources' nature and their relationships [21]. To represent the Se-mantic Web information, this language uses ontology, a shared machine-readable representation of formal explicit description of common conceptualization and the fundamental key of Semantic Web Mining [19,21]. The OWL languages are characterized by formal semantics. They are built upon a W3C XML standard for objects called the Resource Description Framework (RDF). OWL and RDF have attracted significant academic, medical and commercial interest.

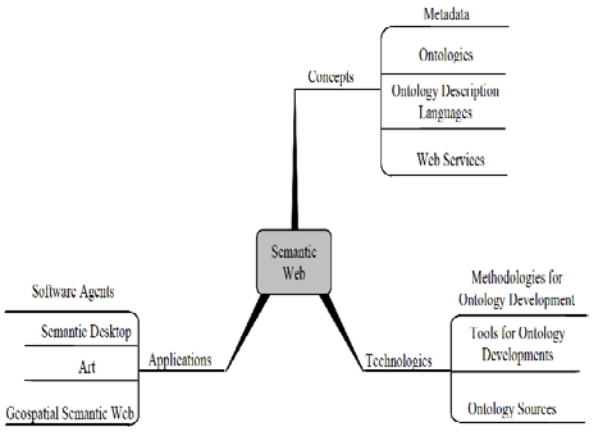

Figure 13. Semantic web perspectives

7.2. Semantic Web Functions

Various functions of semantic web are given in the Figure 14 [23]. Based on these functions, the process of

semantic web mining can be classified as: ***Ontology Mining***, and ***Web Mining***.

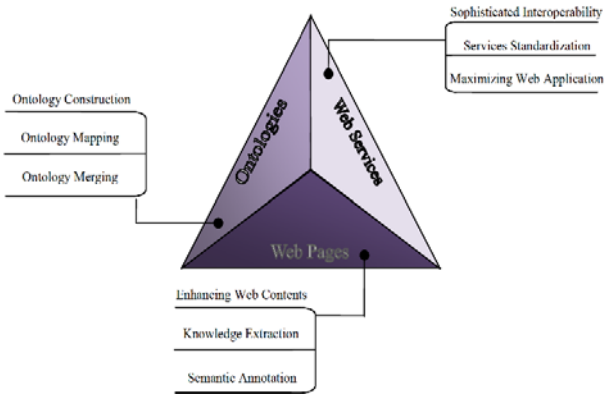

Figure 14. Functions of Semantic Web

8. Web Mining Algorithms

Ranking is an integral component of any information retrieval system. In the case of Web search, because of the size of the Web and the special nature of the Web users, the role of ranking becomes critical. It is common for Web search queries to have thousands or millions of results. Therefore, it is important for the ranking function to output the desired results within the top few pages; otherwise the search engine is rendered useless. Web offers a rich context of information which is expressed through the hyperlinks. The hyperlinks define the ***context*** in which a Web page appears. Intuitively, a link from page *p* to page *q* denotes an endorsement for the quality of page *q*.

A link analysis ranking algorithm starts with a set of Web pages; depending on how this set of pages is obtained. There are several algorithms proposed based on link analysis. Five important algorithms PageRank[24], Weighted PageRank[25], HITS (Hyper-link Induced Topic Search)[26], TSPR (Topic-Sensitive PageRank)[27], and SALSA (Stochastic Approach for Link-Structure Analysis)[33] are discussed below.

8.1. PageRank

Brin and Page [24] developed PageRank algorithm at Stanford University based on the mention analysis. PageRank algorithm is used by the famous search engine, Google. It is the most frequently used algorithm for ranking the various pages. Functioning of the Page Rank algorithm depends upon link structure of the web pages. The PageRank algorithm is based on the concepts that if a page surrounds important links towards it then the links of this page near the other page are also to be believed as imperative pages. The Page Rank reflects on the back link in deciding the rank score. Thus, a page gets hold of a high rank if the addition of the ranks of its back links is high. A simplified version of PageRank is given in Equation 2.

$$PR(u) = c\sum_{v\in B(u)} PR(v)/N_v \tag{2}$$

where u represents a web page, $B(u)$ is the set of pages that point to u, $PR(u)$ and $PR(v)$ are rank achieves of page u and v respectively, N_v indicates the number of outgoing links of page v, c is a factor applied for normalization.

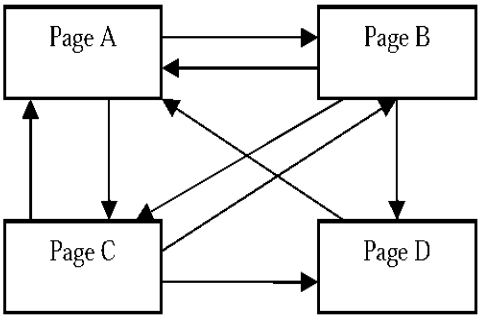

Figure 15. Hyperlink Structure for four pages PageRank Algorithm

Consider the example of hyperlink structure of four pages A, B, C and D as shown in Figure 15. It is observed that not all users follow the direct links on ***WWW***. Using modified version of Equation 2, the PageRank for pages A, B, C and D can be calculated as:

$$PR\,u = (1-d) + d\left(\sum_{v\in B(u)} PR(v)/N_v\right) \tag{3}$$

where d is a damping factor that is frequently set to 0.85 and $(1 - d)$ is the page rank distribution from non-directly linked pages. The PageRanks form a probability distribution over the Web pages. PageRank can be intended using a simple iterative algorithm, and keeps up a correspondence to the principal ***eigen-vector*** of the normalized link matrix of the Web.

8.2. Weighted PageRank

Wenpu Xing and Ali Ghorbani [25] projected a Weighted PageRank (WPR) algorithm which is an addition of the PageRank algorithm. This algorithm assigns a larger rank values to the more important pages rather than separating the rank value of a page evenly among its outgoing linked pages.

Each outgoing link gets a value proportional to its consequence. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}_{(m,\,n)}$ and $W^{out}_{(m,\,n)}$ respectively. $W^{in}_{(m,\,n),}$ as shown in Equation 4, is the weight of link (m, n) calculated based on the number of incoming links of page n and the number of incoming links of all orientations pages of page m.

$$W^{in}_{(m,n)} = I_n / \sum_{p\in R(m)} I_p \tag{4}$$

Where I_n and I_p are the number of incoming links of page n and page p respectively. $R(m)$ denotes the allusion page list of page m.

$$W^{out}_{(m,n)} = O_n / \sum_{p\in R(m)} O_p \tag{5}$$

$W^{in}_{(m,\,n)}$ is the weight of link (m, n) as shown is Equation 5. It calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m. O_n and O_p are the number of outgoing links of page n and p correspondingly. Now using Equation 4 and Equation 5, Weighted PageRank(WPR) formula is given as:

$$WPR(u) = (1-d) + d\sum_{m\in H(n)} WPR(n) W^{in}_{(m,n)} W^{out}_{(m,n)} \tag{6}$$

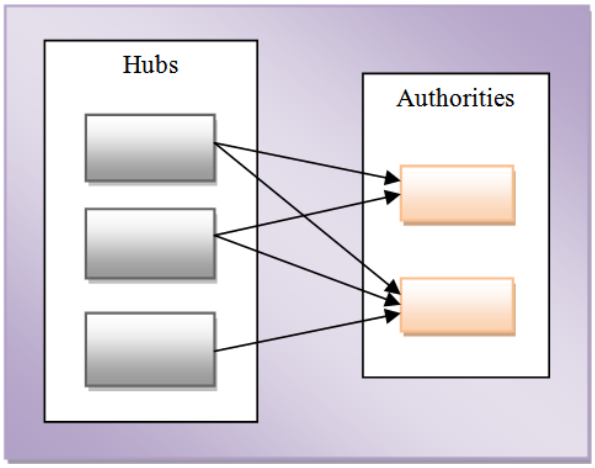

Figure 16. Hubs and Authorities

8.3. HITS

Kleinberg [26] developed a WSM based algorithm named Hyperlink-Induced Topic Search (HITS) which presumes that for every query given by the user, there is a set of ***authority*** pages that are relevant and accepted focusing on the query and a set of ***hub*** pages that contain useful links to relevant pages/sites including links to many authorities. Thus, fine hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many fine hub pages on the same subject. Hubs and Authorities are shown in Figure 16. The HITS algorithm treats ***WWW*** as a directed graph $G(V,E)$, where V is a set of vertices representing pages and E is a set of edges that match up to links

According to the Kleinberg, a page may be a good hub and a good authority at the same time. This spherical relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search).

There are two major steps in the HITS algorithm. The first step is the ***Sampling Step*** and the second step is the ***Iterative Step***. In the Sampling step, a set of relevant pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in influence pages. This algorithm starts with a root set R, a set of S is obtained, keeping in mind that S is comparatively small, rich in relevant pages about the query and contains most of the good authorities. The second step, Iterative step, finds hubs and authorities using the output of the sampling step that is given in Equation 7 and Equation 8.

$$H_p = \sum_{q \in I(p)} A_q \tag{7}$$

$$A_q = \sum_{q \in B(p)} H_q \tag{8}$$

Where H_p is the hub weight, A_p is the Authority weight, $I(p)$ and $B(p)$ denotes the set of reference and referrer pages of page p. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it; similarly, a page's hub weight is proportional to the sum of the influence weights of pages that it links to.

8.4. Topic Sensitive PageRank

Topic-Sensitive PageRank (commonly referred to as TSPR) is a context-sensitive ranking algorithm for web search developed by Taher Haveliwala [27]. It includes two steps: 1) set of biased PageRank vectors generation and 2) performance at query time.

First step computes multiple importance scores for each page, which includes computation of a set of scores of the importance of a page with respect to various topics. This step is performed once, offline, during the preprocessing of the Web crawl. During the offline processing of the Web crawl, algorithm generates 16 topic-sensitive PageRank vectors; each biased using URLs from a top-level category from the Open Directory Project (***ODP***) [28].

Let T_j be the set of URLs in the ODP category c_j. Then when computing the PageRank vector for topic c_j, in place of the uniform damping vector $\vec{p} = [1/N]_{Nx1}$, non-uniform vector $\vec{p} = \vec{v}$ is used, where

$$\begin{aligned} v_p &= 1/\left|T_j\right|, i \in T_j \\ &= 0, i \text{ not belong to } T_j \end{aligned} \tag{9}$$

Second step is performed at query time. At query time, the similarity of the query is compared to each of these vectors or topics; and subsequently, instead of using a single global ranking vector, the linear combination of the topic-sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query. For each web document query sensitive importance score. The results are ranked according to this composite score. It provides a scalable approach for search rankings using Link analysis.

Given a query q, let q' is the context of q. In other words, if the query was issued by highlighting the term q in some Web page u, then q' consists of the terms in u. For ordinary queries not done in context, let $q' = q$. Using a unigram language model, with parameters set to their maximum-likelihood estimates, the class probabilities for each of the 16 top-level ODP classes is computed, conditioned on q'. Let q_i' be the ith term in the query (or query context) q'. Then given the query q, we compute for each c_j the following:

$$\begin{aligned} &P(c_j \mid q') \\ &= \left(\left(P(c_j) \cdot P(q' \mid c_j)\right) / P(q')\right) \propto \left(P(c_j) \cdot \Pi_i P(q' \mid c_j)\right. \end{aligned} \tag{10}$$

8.5. SALSA Algorithm

Stochastic Approach for Link-Structure Analysis (SALSA) [33] is a web page ranking algorithm designed by R. Lempel and S. Moran to assign high scores to hub and authority web pages based on the quantity of hyperlinks among them.

SALSA is inspired by two other link-based ranking algorithms, namely HITS and PageRank, in the following ways:

- Like HITS, the algorithm assigns two scores to each web page: a hub score and an authority score. An authority is a page which is significantly more relevant to a given topic than other pages whereas a hub is a page which contains many links to authorities;
- Like HITS, SALSA also works on a focused subgraph which is topic-dependent. This focused subgraph is obtained by first finding a set of pages

most relevant to a given topic (e.g. take the top-n pages returned by a text-based search algorithm) and then augmenting this set with web pages that link directly to it and with pages that are linked directly from it. Because of this selection process, the hub and authority scores are topic-dependent;

- Like PageRank, the algorithm computes the scores by simulating a random walk through a Markov chain that represents the graph of web pages. SALSA however works with two different Markov chains: a chain of hubs and a chain of authorities. This is a departure from HITS's notions of hubs and authorities based on a mutually reinforcing relationship.

The authority weights are defined to be the stationary distribution of this random walk. Formally, the Markov Chain of the random walk has transition probabilities

$$P_a(i\,j) = \sum_{k \in B(i) \cap B(j)} (1/|B(i)|).(1/|F(k)|) \quad (11)$$

Consider $G_a = (A, E_a)$ denotes the authority graph where there is an (undirected) edge between two authorities if they share a hub. This Markov Chain corresponds to a random walk on the authority graph G_a where we move from authority i to authority j with probability $P_a(i, j)$. Consider W_r denote the matrix derived from matrix W by normalizing the entries such that, for each row, the sum of the entries is 1. Also consider W_c denote the matrix derived from matrix W by normalizing the entries such that, for each column, the sum of the entries is 1. Then the stationary distribution of the SALSA algorithm is the principal left eigenvector of the matrix

$$M_s = Wc^T W_r \quad (12)$$

If authority graph G_a has more than one component, then the SALSA algorithm selects a starting point uniformly at random and performs a random walk within the connected component that contains that node. Consider j be a component that contains node i, and A_j denote the set of authorities in the component j, and E_j the set of links in component j. Then the weight of authority i in component j is

$$a_i = \left(|A_j| / |A|\right).\left(|B(i)| / |E_j|\right) \quad (13)$$

Using the concepts of HITS, in SALSA algorithm each authority divides its weight equally among the hubs that point to it. Therefore,

$$a_i = \sum_{j \in B(i)} h_j / |F(j)| \quad (14)$$

$$h_i = \sum_{j \in F(i)} a_j / |B(j)| \quad (15)$$

9. Issues and Challenges in Web Mining

With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the world generate a large amount of data daily. Data security is the utmost critical issue in ensuring safe transmission of information through the internet. Also network security issues are now becoming important as society is moving towards digital information age. As more and more users connect to the internet it attracts a lot of cyber-criminals [30].

The multitude of rich information sources available on the Web today provides wonderful opportunities and challenges to Web mining for a diverse range of applications: constructing large knowledge bases, predicting the future, finding trends and epidemic in a population, marketing and recommendation, as well as filtering and cleaning Web content to improve the experience of users consuming it. This track covers data analysis for a wide variety of Web data including tweets, tags, links, logs, images, videos, and other multimodal data.

There are various issues and challenges with the web. Some challenges include:

- The Web pages are dynamic that is the information is changes constantly. Copping the changes and monitoring them is an important issue for many applications.
- Noise elimination on the web is another issue. A user feels noisy environment during searching the content, if the information comes from different sources. Typical Web page involves many pieces of information for instance the navigation links, main content of the page, copyright notices, advertisements, and privacy policies. Only part of the information is useful for a particular application but the rest is considered noise.
- The diversity of the information on the multiple pages show similar information in different words or formats, based on the diverse authorship of Web pages that make the integration of information from multiple pages as a challenging problem.
- Handing Big Data on the web is most important challenge, which is scalable in term of volume, velocity, variety, variability, and complexity.
- To maintain security and privacy of web data is not an easy task. Advanced cryptographic algorithm is required for optimal service on the web.
- Discovery of advance hyperlink topology and its management is the other mining issue on the web.
- Searching the web involves two main steps: Extracting the pages relevant to a query and ranking them according to their quality. To provide fruitful search to the user is a need of time.

The main challenge is to come up with guidelines and rules. With the rules and guidelines, site administrator may perform various analyses on the usage data without compromising the identity of an individual user. W3C [32] has initiated a project called Platform for Privacy Preferences (P3P), which provides a protocol try to solve the conflict between Web users and the site administrators. The Platform for Privacy Preferences Project (P3P) is a protocol allowing websites to declare their intended use of information they collect about web browser users. Designed to give users more control of their personal information when browsing, P3P was developed by the World Wide Web Consortium (W3C). P3P is also in proceeding to provide guidelines for independent organization which can ensure that sites comply with the policy statement they are publishing. It manages information through privacy policies.

Security of multimedia using neural network [31] can provides better service of web contents. The key formed by neural network is in the form of weights and neuronal functions which is difficult to break. Here, content data would be used as an input data for cryptography so that data become unreadable for attackers and remains secure from them. The ideas of mutual learning, self-learning, and stochastic behavior of neural networks and similar algorithms can be used for different aspects of cryptography. Web usage mining focuses on privacy concerns.

10. Web Mining Application Areas

Web mining is an important tool to gather knowledge of the behavior of Websites visitors and thereby to allow for appropriate adjustments and decisions with respect to Websites‘ actual users and traffic patterns. Along with a description of the processes involved in Web mining states that Website Design, Web Traffic Handling, e-Business and Web Personalization are four major application areas for Web mining. These are briefly described in the following sections.

10.1. Website Design

The content and structure of the Website is important to the user experience/impression of the site and the site‘s usability. The problem is that different types of users have different preferences, background, knowledge etc. making it difficult (if not impossible) to find a design that is optimal for all users. Web usage mining can then be used to detect which types of users are accessing the website, and their behavior, knowledge which can then be used to manually design/re-design the website, or to automatically change the structure and content based on the profile of the user visiting it.

10.2. Web Traffic Handling

The performance and service of Websites can be improved using knowledge of the Web traffic in order to predict the navigation path of the current user. This may be used for cashing, load balancing or data distribution to improve the performance. The path prediction can also be used to detect fraud, break-ins, intrusion etc.

10.3. e-Business

For Web based companies, Web mining is a powerful tool to collect business intelligence by using electronic business to get competitive advantages. Patterns of the customer’s activities on the Website can be used as important knowledge in the decision-making process, e.g. predicting customer’s future behavior; recruiting new customers and developing new products are beneficial choices. There are many companies providing (among other things) services in the field of Web Mining and Web traffic analysis for extracting business intelligence.

In business intelligence web usage mining can be mainly used by the website administrators who are all involved in ***e-commerce*** and marketing. Through this, the marketing advertisers suggest some more related recommendations to the web user who is involved in the online shopping transaction to increase their profit and to advertise their products.

E-commerce or eCommerce, is trading in products or services using computer networks, such as the Internet. Electronic commerce draws on technologies such as mobile commerce, electronic funds transfer, supply chain management, Internet marketing, online transaction processing, electronic data interchange (EDI), inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web for at least one part of the transaction's life cycle, although it may also use other technologies such as e-mail. E-commerce businesses may employ some or all of the following:

- Online shopping web sites for retail sales direct to consumers
- Providing or participating in online marketplaces, which process third-party business-to-consumer or consumer-to-consumer sales
- Business-to-business buying and selling
- Gathering and using demographic data through web contacts and social media
- Business-to-business electronic data interchange
- Marketing to prospective and established customers by e-mail or fax (for example, with newsletters)
- Engaging in pretail for launching new products and services.

10.4. Web Personalization

Based on Web Mining Techniques, websites are designed to have the look-and-feel and contents are personalized to the needs of an individual end-user. Web Personalization or customization is an attractive application area for Web based companies, allowing for recommendations, marketing campaigns etc. to be specifically customized for different categories of users, and more importantly to do this in real-time, automatically, as the user accesses the Website.

Personalizing web means, for a given query, the web search engine produces different ***SERPs*** or reorganizes the SERPs differently for different users. For this, the intuition of the user is captured by the usage patterns. The ***search engine results page*** (SERP) is the actual result returned by a search engine in response to a keyword query. The SERP consists of a list of links to web pages with associated text snippets. The SERP rank of a web page refers to the placement of the corresponding link on the SERP, where higher placement means higher SERP rank. Web Personalization Process contains:

1. Define an Audience based on the Visitor’s
2. Deliver Personalized Content
3. Optimize and Test to Perfection

10.4.1. Web Personalization Methods

Web pages are personalized based on the characteristics (interests, social category, context, etc.) of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed. The term customization is used instead when the site only uses explicit data such as ratings or preferences. Three methods are used to personalize or customize the web pages.

1. **Explicit**: Web pages are changed by the user using the features provided by the system.
2. **Implicit**: Personalization is customized by the web page based on the different categories such as: ***Collaboration based, Behavior based,*** and ***Profile / Group based***.
3. **Hybrid**: It combines explicit and implicit approaches.

10.5. E-Learning and Digital Library

Web mining can be used for improving the performance of electronic learning. Applications of web mining towards e-learning are usually web usage based. Machine learning and web usage mining improve web based learning.

An electronic library or digital library is a type of information retrieval system. It collects digital objects that can include text, visual material, audio material, video material, stored as electronic media formats.

10.6. Security and Crime Investigation

Along with the rapid popularity of the Internet, crime information on the web is becoming increasingly rampant, and the majority of them are in the form of text. Because a lot of crime information in documents is described through events, event-based semantic technology can be used to study the patterns and trends of web-oriented crimes [29,30].

11. Conclusion and Future Work

Web mining is a rapid growing research area. As the Web has become a major source of information, techniques and methodologies to extract quality information is of paramount importance for many Web applications and users. Web mining and knowledge discovery play key roles in many of today's prominent Web applications such as e-commerce and computer security.

In this paper, various concepts are outlined to extract the useful information from the web. The relationship between web mining and its related paradigm are explored. The paper focuses on basic concepts of data mining and knowledge discovery from the web in discipline and fruitful manner. Web design patterns are useful tools for web data mining. Web pages are analyzed to find out which useful information is included in web page. To find out fruitful information two methods were used. One is Information Retrieval and second one is Information Extraction. Information Retrieval is used to extract useful information from large collection of web pages while Information Extraction is used to find structure information. Clustering of hyperlinked documents can rely on a combination of textual and link-based information. Study about hyperlink topology and web structure would greatly enhance the capacity of web usage mining. This paper also focuses on cloud mining and semantic web mining approaches. Various issues and challenges which are associated with web mining are found out. Applications areas of web mining are also outlined in this article.

In the future, work can be done on various web mining algorithms for supporting Big Data on the web and user friendly services.

References

[1] S. Chakrabarti, *"Data mining for hypertext: A tutorial Survey,"* ACM, SIGKDD, Explorations, 1(2), 1-11, 2000.

[2] Galitsky B, Dobrocsi G, de la Rosa JL, Kuznetsov SO. "Using generalization of syntactic parse trees for taxonomy capture on the web", ICCS. 2011; 8323.

[3] Sankar K. Pal, Varun Talwar, and Pabitra Mitra, "*Web Mining in Soft Computing Framework: Relevance, State of the Art and Future Directions*", IEEE transactions on neural network, Vol. 13, No. 5, September 2002, pp. 1163-1177.

[4] Oren Etzioni, "*The world wide Web: Quagmire or gold mine*", Communications of the ACM, 39(11):65-68, 1996.

[5] Shyam Nandan Kumar, *"Cryptography during Data Sharing and Accessing Over Cloud*." International Transaction of Electrical and Computer Engineers System, vol. 3, no. 1 (2015): 12-18.

[6] Shyam Nandan Kumar and Shyam Sunder Kumar, *"Advancement of Human Resource Management with Cloud Computing,"* International Journal of Research in Engineering Technology and Management, Paper Id: IJRETM-2014-SP-048, Issue: Special, June-2014, pp. 1-6.

[7] Shyam Nandan Kumar, "Advanced Mechanism to Handle Big Data of HD Video File for Mobile Devices," International Journal of Research in Engineering Technology and Management, Paper Id: IJRETM-2014-02-06-006, Vol: 02, Issue: 06, Nov-2014, pp. 1-7.

[8] Berendt, B.spiliopoulou M., "Analysing navigation behaviour in web sites integrating multiple information system", VLDB Journal, Special issue on databases and the web 9, I(2000),56-75.

[9] M. Spiliopoulou, *"Data Mining for Web. In Principals of data mining and knowledge discovery"*, Second European Symposium, PKDD-1999, pp. 588-589.

[10] Fan, W., Wallace, L., Rich, S. and Zhang, Z., *"Tapping into the Power of Text Mining*", Communications of the ACM – Privacy and Security in highly dynamic systems. Vol. 49, Issue-9, 2005.

[11] Gupta, V. and Lehal, G. S., "*A Survey of Text Mining Techniques and Applications*", Journal of Emerging Technologies in Web Intelligence. Vol. 1. pp 60-76, 2009.

[12] "*Web crawler*", http://en.wikipedia.org/wiki/Web_crawler.

[13] "*Wrapper (data mining)*", http://en.wikipedia.org/wiki/Wrapper_(data_mining)

[14] Nicholas Kushmerick, Daniel S. Weld, Robert Doorenbos, "*Wrapper Induction for Information Extraction*", Proceedings of the International Joint Conference on Artificial Intelligence, 1997.

[15] Papakonstantinou, Y. and Garcia-Molina, H. and Widom, J. (1995). "*Object exchange across heterogeneous information sources*". Proceedings of the Eleventh International Conference on Data Engineering: 251-260.

[16] "Web scraping", http://en.wikipedia.org/wiki/Web_scraping.

[17] "Document Object Model (DOM)", http://en.wikipedia.org/wiki/Document_Object_Model.

[18] Shapira D., Avidan S., Hel-Or Y., "Multiple Histogram Matching", 20th IEEE International Conference on Image Processing (ICIP), 2013, pp. 2269-2273.

[19] V. Sugumaran and J. A. Gulla, *"Applied Semantic Web Technologies,"* Taylor & Francis Group, Boca Raton, 2012.

[20] K. K. Breitman, M. A. Casanova, and W.Truszkowski, "*Semantic Web: Concepts, Technology and Applications*", Springer, 2007.

[21] A. Jain, I. Khan and B. Verma, *"Secure and Intelligent Decision Making in Semantic Web Mining,"* Interna-tional Journal of Computer Applications, Vol. 15, No. 7, 2011, pp. 14-18.

[22] D. Jeon and W. Kim, *"Development of Semantic Deci- sion Tree,"* Proceedings of the 3rd International Confer- ence on Data Mining and Intelligent Information Tech- nology Applications, Macau, 24-26 October 2011, pp. 28-34.

[23] H. Hassanzadeh and M. R. Keyvanpour, *"A machine Learning Based Analytical Framework for Semantic Annotation Requirements*", International Journal of Web and Semantic Technology (IJWeST), vol. 2, no. 2, pp. 27-38, 2011.

[24] S. Brin, and L. Page, "*The Anatomy of a Large Scale Hypertextual Web Search Engine*", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[25] Wenpu Xing and Ali Ghorbani, "*Weighted PageRank Algorithm*", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[26] J. Kleinberg, "*Authoritative Sources in a Hyper-Linked Environment*", Journal of the ACM 46(5), pp. 604-632, 1999.

[27] Taher H. Haveliwala, "*Topic-Sensitive PageRank*", Eleventh International World Wide Web Conference (Honolulu, Hawaii), USA, May-2002, ACM 1-58113-449-5/02/0005.

[28] "*The Open Directory Project: Web directory for over 2.5 million URLs*", http://www.dmoz.org/.

[29] J. Hosseinkhani, M. Koochakzaei, S. Keikhaee and Y. Amin, "*Detecting Suspicion Information on the Web Using Crime Data Mining Techniques*", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 3, No. 1, 2014, Page: 32-41.

[30] Shyam Nandan Kumar, "*Review on Network Security and Cryptography*." International Transaction of Electrical and Computer Engineers System, vol. 3, no. 1 (2015): 1-11.

[31] Shyam Nandan Kumar, "*Technique for Security of Multimedia using Neural Network*," Paper id-IJRETM-2014-02-05-020, IJRETM, Vol: 02, Issue: 05, pp.1-7. Sep-2014.

[32] "W3C- World Wide Web Consortium (W3C)", http://www.w3.org/.

[33] "SALSA Algorithm", http://en.wikipedia.org/wiki/SALSA_algorithm.