WEB DATA MINING RESEARCH: A SURVEY

Brijendra Singh[1], Hemant Kumar Singh[2],
[1]Department of computer Science, University of Lucknow, LUCKNOW, INDIA.
[2]Department of computer Applications, AzadIET, Lucknow, INDIA.
(drbri_singh@hotmail.com, hemantbib@gmail.com)

ABSTRACT

Web Data Mining is an important area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web, It can be classified into three different types i.e. web content mining, web structure mining and web usages mining. The aim of this paper is to provide past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining and also outlines key future research directions. This paper also reports the comparisons and summary of various methods of web data mining with applications, which gives the overview of development in research and some important research issues.

Keywords*:* Web mining; web content mining; web usage mining; web structure mining; semantic web

I. INTRODUCTION

The amount of data kept in computer files and data bases is growing at a phenomenal rate. At the same time users of these data are expecting more sophisticated information from them .A marketing manager is no longer satisfied with the simple listing of marketing contacts but wants detailed information about customers' past purchases as well as prediction of future purchases. Simple structured / query language queries are not adequate to support increased demands for information. Data mining steps is to solve these needs. Data mining is defined as finding hidden information in a database alternatively it has been called exploratory data analysis, data driven discovery, and deductive learning [1].

In the data mining communities, there are three types of mining: data mining, web mining, and text mining [2]. There are many challenging problems [3] in data/web/text mining research like no unifying theory of

data mining, also opportunity and need for data mining researchers to solve some long standing problems in statistical research such as age-old problem of avoiding spurious correlations. This is sometimes related to the problem of mining for "deep knowledge," which is the hidden cause for many observations. For example: can we discover Newton's laws from observing the movements of objects [3].

The mining data may vary from structured to unstructured. Data mining mainly deals with structured data organized in a database while text mining mainly handles unstructured data/text. Web mining lies in between and copes with semi-structured data and/or unstructured data. Web mining calls for creative use of data mining and/or text mining techniques and its distinctive approaches. Mining the web data is one of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web and we can easily get overwhelmed with data [2].

There is no agreed definition of Web Data Mining but we present one simple definition:

"Web Data Mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process."

Word Wide web is the interactive and popular medium to distribute information today. Data on the web is rapidly increasing day by day and Web data is huge, diverse and dynamic so information users could encounter the following problems while interacting with the web [4].

1. *Finding Relevant Information*- People either browse or use the search service when they want to find specific information on the web. However today's search tools have problems like low precision which is due to irrelevance of many of the search results. This results in a difficulty in finding the relevant information. Another problem is low recall which is due to inability to index all the information available on the web.
2. *Creating new knowledge out of the information available on the web*-This problem is basically sub problem of the above problem. Above problem is query triggered process(retrieval oriented) but this problem is data triggered process that presumes that we already have collection of web data and we want to extract potentially use full knowledge out of it.
3. *Personalization of information*- When people interact with the web they differ in the contents and presentations they prefer.
4. *Learning about Consumers or individual users*-This problem is about what the customer do and want. Inside this problem there are sub problem such as customizing the information to the intended consumers or even to personalize it to individual user, problem related to web site design and management and marketing etc.

There are many tools like Database (DB), Information Retrieval (IR), and Natural Language Processing (NLP) etc. available to solve the above stated problem. Web mining techniques could be more efficiently used to solve the information overload problem directly or indirectly.

The purpose of the paper is to provide past, current evaluation and update in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining and also outlines key future research directions.

The literature in this paper is classified into the three types of web mining: web content mining, web usage mining, and web structure mining. We put the literature into following sections: Section 2.0 presents the web mining , section 3.1 presents the

978-1-4244-5967-4/10/$26.00 ©2010 IEEE

literature review for web content mining in which we described the enhancement made in image retrieval system year by year, section 3.2 presents the literature review for web structure mining which describes the improvements in the URL mining over the years, section 3.3 presents literature review for web usages mining which describes that how much important weblog data is; and we also review current topic on semantic web as Section. 3.4. Finally Section 4.0 concludes the paper and outlines some promising areas of future research.

II. WEB MINING

2.1 Overview-

In 1996 it's Etzioni [5] who first coined the term web mining. Etzioni starts by making a hypothesis that information on web is sufficiently structured and outlines the subtasks of web mining. According to Oren Etzioni Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and service. Similar to Kosala and Blockeel [4], Qingyu Zhang and Richard s. Segall[2] suggest decomposing Web mining into the following sub tasks:

Resource Discovery*:* locating unfamiliar documents and services on the Web.

Information Selection and Pre-Processing**:** automatically extracting and pre-processing specific information from newly discovered Web resources.

Generalization: uncovering general patterns at individual Web sites and across multiple Sites.

Analysis: Validation and interpretation of mined patterns.

Visualization- Presenting the results of an interactive analysis in a visual, easy to understand fashion.

Kosala and Blockeel [4] who perform research in the area of web mining and suggest the three web mining categories depending on which kind of data to be mined that is mining for information or mining the web link structure or mining for user navigation patterns . Mining for information focuses on the development of techniques for assisting a user in finding documents that meet a certain criterion that is web content mining. Web content mining refers to the discovery of useful information from web contents, including text, image; audio, video, etc mining the link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining. Web structure mining tries to discover the model underlying the link structures of the web. Model is based on the topology of hyperlinks with or without description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites. Finally, mining for user navigation patterns focuses on techniques which study the user behavior when navigating the web that is web usages mining. Web usage mining refers discovery of user access patterns from Web servers. Web usages data include data from web server access logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries, bookmark data, mouse clicks and scrolls or any other data as result of interaction.

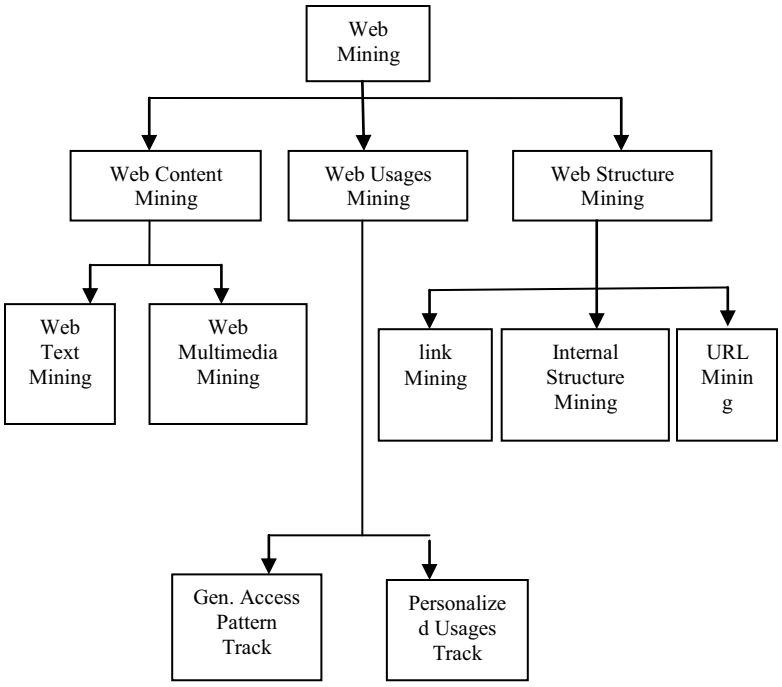

Figure-1: Taxonomy of Web Mining

Since searching, comprehending, and using the semi structured information stored on the Web poses a significant challenge because this data is more sophisticated and dynamic than the information that commercial database systems store. Figure 1, shows the Taxonomy of web mining.

Han and Chang[6] claimed that incorporating data mining to Web-page ranking helps Web search engines to find high-quality Web pages and enhances Web click stream analysis , data semantics could substantially enhance the quality of keyword-based searches and indicate research problems (tremendous number of documents have not been indexed, which makes searching the data contains extremely difficult) to use data mining effectively in developing web intelligence. It latter includes mining web search-engine data and analyzing web's link structure, classifying web documents automatically, mining web page semantic structures and page contents, and mining web dynamics. Web dynamics is the study of how the web changes in the context of its contents, structure, and access patterns.

Number of survey papers [7-10] had been published on web mining research -content, structure and usages

III. LITERATURE REVIEW

3.1 Web Content Mining- It deals with discovering useful information or knowledge from web page contents. Margaret H. Dunham [1] stated Web Content Mining can be thought of the extending the work performed by basic search engines. Web content mining analyzes the content of Web resources. Recent advances in multimedia data mining promise to widen access also to image, sound, video, etc. content of Web resources. The primary Web resources that are mined in Web content mining are individual pages. Information Retrieval is one of the research areas that provide a range of popular and effective, mostly statistical methods for Web content mining. They can be used to group, categorize, analyze, and retrieve documents [11].

Since it is not possible to annotate images on internet manually. So Web images are usually not well annotated using semantic descriptors. Due to multiplicity of contents in a single image and the subjectivity of human perception, it is hard to make exactly the same understanding to the similar image by different users. These problems have restricted the application of the keyword based image retrieval tools and also traditional text based methods could not handle the explosive load of images and hence the concept of Content Based Image Retrieval was born. Content-Based Image Retrieval (CBIR) attempts to automate the process of indexing or annotating image in image databases. Content Based Image Retrieval uses the visual contents of an image such as color, shape, texture, and spatial layout to represent and index an image. These visual contents are extracted using multi-dimensions feature vectors and then indexed in a database. When an image is given as input query for retrieval, its feature vectors are extracted and then images with similar feature vectors are retrieved from the database by comparison. The indexing scheme provides an efficient way to retrieve images from the database [12] .

A major bottleneck in content-based image retrieval (CBIR) systems or search engines is the large gap between low-level image features like color histograms used to index images and high-level semantic contents of images human subjectivity.

To reduce the gap between low-level image features used to index images and high-level semantic contents of images in content-based image retrieval (CBIR) systems or search engines, Zhang et al. [13] suggest applying relevance feedback technique to refine the query or similarity measures in image search process. Due to learning and searching nature of each relevance feedback algorithm in CBIR is a machine learning problem that is a computer program that automatically improves with experience. In which a user provides feedback examples from the retrieval results of a query and system learns from such examples to refine retrieval results. In CBIR, relevance feedback is a task to improve the retrieval performance and the experience here is feedback examples provided by the users. They presented a framework of relevance feedback and semantic learning where low-level features and keyword explanation are integrated in image retrieval and in feedback processes to improve the retrieval performance and effectiveness of information system. At the beginning most approaches performed relevance feedback at the low level features basically replacing keywords with features for document retrieval. Using only low level features may not be efficient in representing users' feedbacks and telling their intensions. The user can interact with image retrieval system by two ways. In first one, the user types in a list of keywords representing the semantic contents of the preferred images. In second one, the user provides a set of examples images as an input and the retrieval system will retrieve other analogous images. They took the advantage of combining these two approaches to improve both retrieval accuracy and ease of use of the system. A CBIR framework with integrated relevance feedback and query expansion consist of the following steps-

1. It consists of a semantic network which links images to semantic annotations in a database, a similarity measure that integrating both semantic features and image features, and a machine learning algorithm to iteratively update the semantic network and to improve the system's performance over time.
2. System supports both query by keyword and query by image example through semantic network and low-level feature indexing.

Figure 2, shows the framework of integrated relevance feedback and query expansion. As well as Figure 3, shows the semantic network.

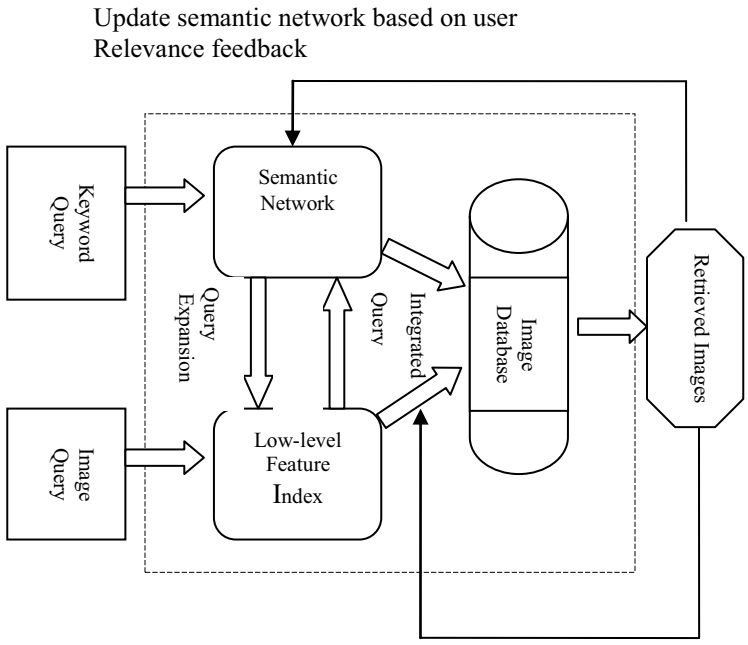

Figure 2: Framework of integrated relevance feedback and query expansions

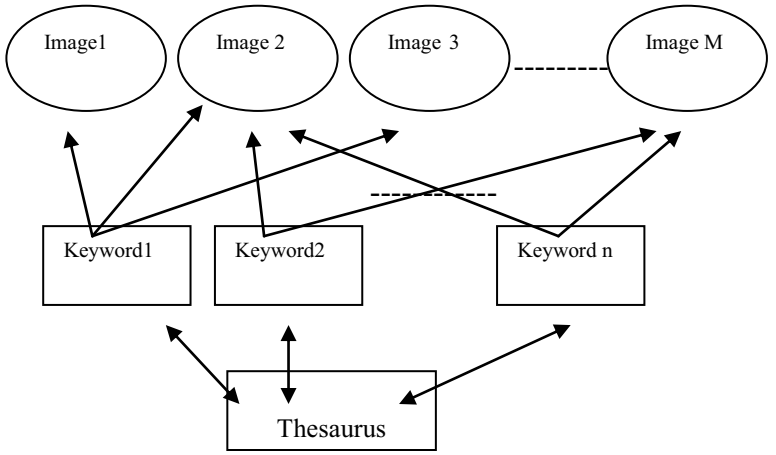

Figure 3: Semantic network

To further improve the retrieval performance of the proposed framework, a cross modality query expansion method is supported. That is, once a query is submitted in the form of keywords, the retrieved images based on keyword search are considered as the positive examples, based on which the query is expanded by features of these images. This is done by first searching the semantic network for the keywords .Then, the visual features of these images that contains these keywords

(referred as training images) are incorporated into the expanded query. The more images are annotated correctly, better the system retrieval performance will be . However, the reality is human labeling of images is tedious and expensive, hence not a feasible solution, To address this issue, a probabilistic progressive keyword propagation scheme is proposed by H. J. Zhang in the framework to automatically annotate images in the databases in the relevance feedback process utilizing based a small percentage of annotated images. Assume that initially only a few images in a database have been manual labeled with keywords and the retrieval is performed mainly based on low-level features. The initial keywords annotation can be from web through the crawler when the images are from the Web, or labeled by humans. While the user is interacting with the system by providing feedbacks in a query session, a progressive learning process is activated to propagate the keyword annotation from the labeled images to un-labeled images so that more and more images are implicitly labeled by keywords. In this way, the semantic network is updated in which the keywords with a majority of user consensus will emerge as the dominant representation of the semantic content of their associated images. They developed a prototype system (iFind- an Image Search Engine) performing better than traditional approaches.
The search options that iFind support include Keyword based search, Query by example, Relevance feedback, log mining.
The dynamic nature and size of the Internet can result in difficulty finding relevant information. Most users typically express their information need via short queries to search engines and they often have to physically sift through the search results based on relevance ranking set by the search engines, making the process of relevance judgment time-consuming. Chen et al[14] describe a novel representation technique which makes use of the Web structure together with summarization techniques to better represent knowledge in actual Web Documents. They named the proposed technique as Semantic Virtual Document (SVD). The SVD can be used together with a suitable clustering algorithm to achieve an automatic content-based categorization of similar Web Documents. This technique allows an automatic content-based classification of web documents as well as a tree-like graphical user interface for browsing post retrieval document browsing enhances the relevance judgment process for Internet users. They also introduce cluster-biased automatic query expansion technique to interpret short queries accurately. They present a prototype of Intelligent Search and Review of Cluster Hierarchy (iSEARCH) for web content mining.
Typically, search engines are low precision in response to a query, retrieving lots of useless web pages, and missing some other important ones. Ricardo Campos et al[15] study the problem of the hierarchical clustering of web and proposed an architecture of a meta-search engine called WISE that automatically builds clusters of related web pages embodying one meaning of the query. These clusters are then hierarchically organized and labeled with a phrase representing the key concept of the cluster and the corresponding web documents.
Mining search engine query log is a new method for evaluating web site link structure and information architecture. Mehdi Hosseini , Hassan Abol hassani [16] propose a new query-URL co-clustering for a web site useful to evaluate information architecture and link structure. Firstly, all queries and clicked URLs corresponding to particular web site are collected from a query log as bipartite graph, one side for queries and the other side for URLs. Then a new content free clustering is applied to cluster queries and URLs concurrently. Afterwards, based on information entropy, clusters of URLs and queries will be used for evaluating link structure and information architecture respectively.
Rouhollah Rahmani, Hui Zhang et al defined a Localized Content-Based Image Retrieval as a CBIR task where the user is only interested in a portion of the image, and the rest of the image is irrelevant and presented a localized CBIR system, ACCIO!, that uses labeled images in conjunction with a multiple-instance learning algorithm to first identify the desired object and weight the features accordingly and then to rank images in the database using a similarity measure that is based upon only the relevant portions of the image.

Table 1: Summary Table for Web Content mining

Author	Method	Application	Publication Year
Dr. Fuhui Long et al.[12]	Visual content description	Content based image retrieval	1999
H. Zhang et al[13]	Relevance feedback algorithm	Content based image retrieval (iFind)	2003
Chen et al[14]	Web structure together with summarization techniques	Semantic virtual document (iSearch)	2005
Ricardo Campos et al [15]	Graph-based overlapping, Clustering algorithm	Meta-search engine called WISE	2006
Mehdi Hosseini[16]	Query-URL co-clustering	Categorize queries and URLs related to special web site	2007
Hui Zhang et al[17]	SPARSE technique	Localized CBIR system,	2008
G. Poonkuzhali[18]	Signed approach and full word matching	Retrieval of documents takes less time and less space	2009

Most of the existing web mining algorithms have concentrated on finding frequent patterns while neglecting the less frequent ones that are likely to contain outlying data such as noise,

Irrelevant and redundant data. Retrieving relevant content from the web is a very common task. Thus there is need to develop more user friendly and automated tools for providing relevant information quickly and improving the results of web content mining by detecting and eliminating redundant links. It is a major challenge in web mining research today [18]. We have summarized various methods of web content mining with application in Table 1.

3.2 Web Structure Mining- It deals with discovering and modeling the link structure of web. Web information retrieval tools make use of only the text available on web pages but ignoring valuable information contained in web links. Web structure mining aims to generate structural summary about web sites and web pages. The main focus of web structure mining is on link information. Web bags may be used to discover visible web documents, luminous web documents (number of outgoing links) and luminous paths (set of inter-linked nodes) that most of the search engines fails to discover for eg. if a new business enterprise wants to do some analysis of their web sites which display products for buying? By finding the visibility of its web site with respect to other web sites selling. For such related products, the company can find ways to redesign (including changes in product's price etc.) its web site to improve visibility. For example, if a web site sells PC monitors, they must be providing links to web sites which sell CPU. Thus, if a web site finds that its visibility is lower in comparison to other web sites selling CPUs then the web site needs to improve in terms of design, products, etc. [19].Through an original algorithm for hyperlink analysis called HITS (Hypertext Induced Topic Search), Kleinberg [20] introduced the concepts of hubs (pages that refer to many pages) and authorities (pages that are referred by many pages). He developed a set of algorithmic tools for extracting information from the link structures of such environments and reported on experiments that demonstrate their effectiveness in a variety of contexts on the World Wide Web. The main issue they addressed within their framework is the refinement of broad search topics, through the discovery of "authoritative" information sources on such topics.

Furnkranz[21] described the Web may be viewed as a directed graph whose nodes are the documents and the edges are the hyperlinks between them and exploited the graph structure of the World Wide Web for improved retrieval performance and classification accuracy. Many search engines use graph properties in ranking their query results.

The continuous growth in the size and use of the Internet is creating difficulties in the search for information. To help users search for information and organize information layout, Smith and Ng [22] suggested using a self organizing map(SOM) to mine web data and provided a visual tool to assist user navigation. Based on the users' navigation behavior, they developed LOGSOM, a system that utilizes SOM to organize web pages into a two-dimensional map. The map provides a meaningful navigation tool and serves as a visual tool to better understand the structure of the web site and navigation behaviors of web users.

As the size and complexity of websites expands dramatically, it has become more and more challenging to design websites on which web surfers can easily find the information they seek. Fang and Sheng [23] address the design of the portal page of a web site. They try to maximize the efficiency, effectiveness, and usage of a web site's portal page by selecting a limited number of hyperlinks from a large set for the inclusion in a portal page. Based on relationships among hyperlinks (i.e. structural relationships that can be extracted from a web site and access relationship that can be discovered from a web log), they proposed a heuristic approach to hyperlink selection called Link Selector.Instead of clustering user navigation patterns by means of a Euclidean distance measure, Hay et al.[24] use the Sequence Alignment Method (SAM) to partition users into clusters, according to the order in which web pages are requested and

the different lengths of clustering sequences. They validate SAM by means of user traffic data of two different web sites and results show that SAM identifies sequences with similar behavioral patterns.

To meet the need for an evolving and organized method to store references to web objects, Guan and McMullen[25] design a new bookmark structure that allows individuals or groups to access the bookmark from anywhere on the Internet using a Java-enabled web browser. They proposed a prototype to include more features such as URL, the document type, the document title, keywords, date added, date last visited, and date last modified as they share bookmarks among groups of users.

Song and Shepperd [26] view the topology of a web site as a directed graph and mine web browsing patterns for e-commerce. They use vector analysis and fuzzy set theory to cluster users and URLs. Their frequent access path identification algorithm is not based on sequence mining.

HITS is one of the most influential algorithms for WSM. Nacim Fateh Chikhi, Bernard Rothenburger et al [27] showed the equivalence between HITS and Principal Component Analysis, a well-known technique for dimensionality reduction. They investigated the use of various dimensionality reduction techniques (DRTs) to extract the implicit structures hidden in the web hyperlink connectivity. And compared fours dimensionality reduction techniques that are Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF), Independent Component Analysis (ICA) and Random Projection (RP) for the task of web structure mining. They perform experiments on datasets and showed that Nonnegative matrix factorization to be a promising approach for web structure analysis because of its superiority over the other methods.

Table 2: Summary Table for Web Structure mining

Author	Method	Application	Publication Year
Sanjay Kumar Madria et al[19]	Warehouse of Web Data (WHOWEDA project)	To design the tools and techniques for web data mining	1999
Kleinberg et al [20]	HITS	Discovering authoritative sources in a hyperlinked environment	1999

Johannes F¨urnkranz et al[21]	Data mining and machine learning	Exploiting the graph structure of the Web	2002
Smith and Ng[22]	Clustering, self-organized map	Mapping user navigation patterns (LOGSOM)	2003
Fang and Sheng[23]	Heuristic approach	Hyperlink selection for portal page	2004
Hay et al.[24]	Sequence Alignment Method (SAM)	Mining navigation Patterns	2004
Guan and McMullen[25]	Design bookmark structure	Bookmark	2005
Song and Sheppard[26]	Frequent access path identification algorithm, fuzzy set theory	Mining Web Browsing patterns for e-commerce	2006
Nacim Fateh Chikhi et al[27]	Various dimensionality reduction techniques (DRTs)	To extract the implicit structures hidden in the web hyperlink connectivity	2007
Lefteris Moussiades et al[28]	Graph clustering algorithm	Mining the community structure of a graph	2009

Apart from search ranking, hyperlinks are also useful for finding Web communities. A web community is a collection of web pages that are focused on a particular topic or theme. Most community mining approaches are based on the assumption that each member of a community has more hyperlinks within than outside its community. In this context, many graph clustering algorithms may be used for mining the community structure of a graph as they adopt the same assumption, i.e. they assume that a cluster is a vertex subset such that for all of its vertices, the number of links connecting a vertex to its cluster is higher than the number of links connecting the vertex outside its cluster [28]. We have summarized various methods of web structure mining with application in Table 2.

Web Structure Mining plays a vital role with various benefits including quick response to the web users, reducing lot of HTTP transactions between users and server. We hope that the above survey on web structure mining could provide useful information related to knowledge issues on web.

3.3 Web Usages Mining- It deals with understanding user behavior in interacting with the web or with a website. One of the aims is to obtain information that may assist web site reorganization or assist site adaptation to better suit the user. Web usage mining model is a kind of mining to server logs and its aim is getting useful users' access information in logs to make sites can perfect themselves with appropriate users' requirements, serve users better and get more economy benefits. Several surveys on Web usage mining exist in [4, 29, 30, 42].

There are many web log analysis tools available to mine data from log record on web page. Log record contains plenty of useful information such as URL, IP address and time and so on. Analyzing and discovering Log could help organizations to find more potential customers, pages popularity (number of times a page has been visited) etc that can help in reorganizing the web site for fast and easy customer access, improving links and navigation, attracting more advertisement capital by intelligent adverts, turning viewers into customers by better site architecture, and monitoring the efficiency of the web site

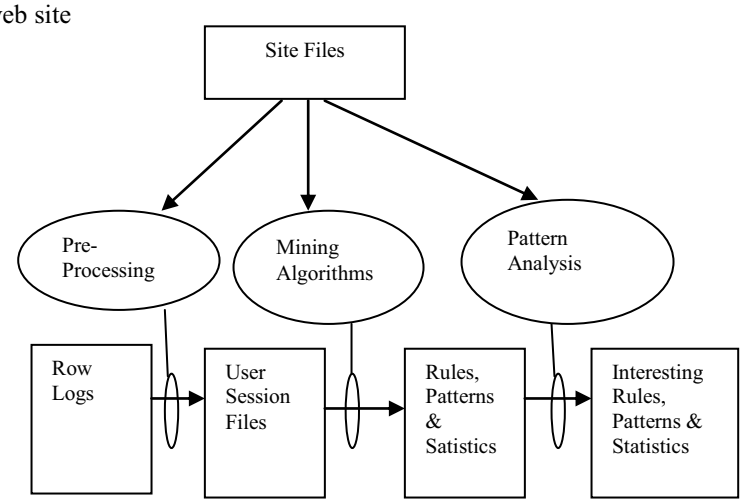

Figure-4: High Level Web Usage Mining Process

Most data used for mining [29] is collected from Web servers, clients, proxy servers, or server databases, all of them produce noisy data. Because Web mining is sensitive to noise, data cleaning methods are necessary. Jaideep Srivastava and R. Cooley categorize data preprocessing into subtasks and noted that the final outcome of preprocessing should be data that allows identification of a particular user's browsing pattern in the form of page views, sessions, and click streams. Click streams are of particular interest because they allow reconstruction of user navigational patterns.

Markov models have been extensively used to model Web users' navigation behaviors on Web sites. Jianhan Zhu, Jun Hong et al[31] proposed a clustering algorithm called CitationCluster to cluster conceptually related pages. The clustering results are used to construct a conceptual hierarchy of the Web site. Markov model based link prediction is integrated with the hierarchy to assist users' navigation on the Web site.

In the previous six years collection of user navigation session were presented in form of many models such as Hyper Text Probabilistic Grammar (HPG), N-Gram Model, Dynamic clustering based morkov model etc [32].

Web Access Pattern Tree (WAP-tree) stores the highly compressed access sequences, and mining frequent access sequences based on WAP-tree needs to scan transaction database only twice. However, producing conditional WAP-tree repeatedly in the algorithm influences the efficiency in a certain degree. Considering the shortage of WAP-tree, combined with the need of mining maximal access sequences, TAN Xiaoqiu, YAO Min et al [33] improves WAP-tree and introduces restrained sub tree structure to solve the problem that a mass of conditional WAP-tree is built in the traditional algorithm.

Many researches have developed Web usage mining (WUM) algorithms utilizing Web log records in order to discover useful knowledge to be used in supporting business applications and decision making. The quality of WUM in knowledge discovery, however, depends on the algorithm as well as on the data. This research by Yu-Hui Tao , Tzung-Pei Hong et al [34]explores a new data source called intentional browsing data (IBD) for potentially improving the effectiveness of WUM applications. IBD is a category of online browsing actions, such as ''copy'', ''scroll'', or ''save as,'' and is not recorded in Web log files. Consequently, the research aims to build a basic understanding of IBD which will lead to its easy adoption in WUM research and practice.

Recently, a number of Web Usage Mining algorithms [33, 34, 35] have been proposed to mining user navigation behavior. Partitioning method was one of the earliest clustering methods to be used in Web usage mining [34].

Web based recommender systems are very helpful in directing the users to the target pages in particular web sites. Web usage mining recommender systems have been proposed to predict user's intention and their navigation behaviors. We can take into account the semantic knowledge [explained in later section] about underlying domain to improve the quality of the recommendation. Integrating semantic web and web usage mining can achieve best recommendations in the dynamic huge web sites [16].

Prediction of user future movements and intentions based on the users' clickstream data. Mehrdad Jalali, Norwati Mustapha et al [36] develop a model for online predicting through web usage mining system and propose an approach for classifying user navigation patterns to predict users' future intentions. The approach is based on the using longest common subsequence algorithm to classify current user activities to predict user next movement.

The quality of recommendations in the current systems to predict user future requests in a particular Web site is below satisfaction. To effectively provide online prediction, M. Jalali, N. Mustapha et al have developed a recommendation system called WebPUM, an online prediction using Web usage mining system and propose a novel approach for classifying user navigation patterns to predict users' future intentions. The approach is based on the new graph partitioning algorithm to model user navigation patterns for the navigation patterns mining phase. Furthermore, longest common subsequence algorithm is used for classifying current user activities to predict user next movement. We have summarized various methods of web usage mining with application in Table 3.

Web in how to quickly and accurately finds the user-needed information and then provides personalized service to users

The increasing competitive E-commerce gradually enhances the attention paid to E-marking. Web log mining is of great importance to E-commerce, distance education platforms today.

Table 3: Summary Table for Web usage mining

Author	Method	Application	Publication Year
Jaideep Srivastava, R. Cooley[29]	Statistical Analysis Association Rules	Personalization Site Modification etc	2000
Jianhan Zhu et al[31]	Clustering algorithm called Citation Cluster	Construct a conceptual hierarchy of the Web site	2002
Borges and M. Levene[32]	Dynamic clustering-based method	Representing a collection of user web navigation sessions	2004
TAN Xiaoqiu, YAO Min et al[33]	Improved WAP tree	Sequential pattern mining	2006
Yu-Hui Tao , Tzung-Pei Hong et al[34]	Taxonomy of browsing data	Decision support	2007
Mehdi Hosseini et al[16]	Web based recommender systems	predict user's intention and their navigation behaviors	2008
Mehrdad Jalali et al [36]	Longest common subsequences algorithm	Predict user near future movement.	2009
M. Jalali, et al[37]	WebPUM	Predict user near future Movement	2010

3.4 Semantic Web Mining- The Semantic Web is based on a vision of Tim Berners-Lee who is known as the inventor of the WWW. The enormous success of the current WWW leads to a new challenge- A huge amount of data is interpretable by humans only and machine support is limited. Berners-Lee suggests to improve the Web by machine-processable information which supports the user in his tasks. For instance, today's search engines are already quite powerful, but still too often return excessively large or inadequate lists of hits. Machine processable information can point the search engine to the relevant pages and can thus improve both precision and recall [11].The Semantic Web [16, 44] is a web that is able to describe things in a way that computers can understand. Statements are built with syntax rules. The syntax of a

language defines the rules for building the language statements. But how can syntax become semantic? This is what the Semantic Web is all about. Describing things in a way that computers applications can understand it. The Semantic Web is not about links between web pages. The Semantic Web describes the relationships between things (like A is a part of B and Y is a member of Z) and the properties of things (like size, weight, age, and price). Semantic Web Mining aims at combining the two fast-developing research areas Semantic Web and Web Mining. More and more researchers are working on improving the results of Web Mining by exploiting semantic structures in the Web, and they make use of Web Mining techniques for building the Semantic Web. Last but not least, these techniques can be used for mining the Semantic Web itself [38]. The Semantic Web is a recent initiative, inspired by Tim Berners-Lee [39], to take the World-Wide Web much further and develop it into a distributed system for knowledge representation and computing. The aim of the Semantic Web is to not only support access to information "on the Web" by direct links or by search engines but also to support its use. Instead of searching for a document that matches keywords, it should be possible to combine information to answer questions. Instead of retrieving a plan for a trip to Hawaii, it should be possible to automatically construct a travel plan that satisfies certain goals and uses opportunities that arise dynamically. This gives rise to a wide range of challenges. Some of them concern the infrastructure, including the interoperability of systems and the languages for the exchange of information rather than data. Many challenges are in the area of knowledge representation, discovery and engineering. They include the extraction of knowledge from data and its representation in a form understandable by arbitrary parties, the intelligent questioning and the delivery of answers to problems as opposed to conventional queries and the exploitation of formerly extracted knowledge in this process.

The main areas of research in this domain are Web log data preprocessing and identification of useful patterns from this preprocessed data using mining techniques

IV. CONCLUSION AND FUTURE DIRECTIONS

This paper has provided a more current evaluation and update of web mining research available. Extensive literature has been reviewed based on three types of web mining, namely web content mining, web usage mining, and web structure mining. Year wise Summary of improvements in each type of mining is given in respective tables. Web data mining is a fast rising research area today. As the web data and its usage will rise in future. It will prolong to generate more content, structure and usage data. So the importance of web data continues increasing. Web data is mainly semi-structured to unstructured. Due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research Problems.

Most of the knowledge represented in HTML Web Documents, there are numerous other file formats that are publicly accessible on the Internet. Also, if both the actual Web Documents and corresponding Back Link Documents were mainly composed of multimedia information (e.g. graphics, audio, etc.), SVD will not be particularly effective in revealing more textual information. It would be worthwhile to research new techniques to include these file formats and multimedia information for knowledge representation. Web Data Mining is perhaps still in its infancy and much research is being carried out in the area.

V. ACKNOWLEDGEMENT

Second author of the paper offers the sincere gratitude to the management, Director, Head and faculty members in dept. of Computer Applications of Azad Institute of Engineering and Technology for the constant encouragement and unparallel support provided through out the period of this research work.

REFERENCES

[1] Margaret H. Dunham, "Data Mining Introductory & Advanced Topics", Pearson Education

.[2] Qingyu Zhang and Richard s. Segall," Web mining: a survey of current research,Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 (2008) 683–720

[3] Q. Yang and X. Wu, 10 challenging problems in data mining research, Int. J Inform.Technol. Decision Making 5(4) (2006) 597–604

[4] Kosala and Blockeel, "Web mining research: A survey," SIGKDD:SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000

[5] O.etzioni. The world wield web: Quagmire or Gold Mining. Communicate of the ACM, (39)11:65-68, 1996;

[6] J. Han and C. Chang, Data mining for web intelligence, Computer (November 2002), pp. 54–60, http://www-faculty.cs.uiuc.edu/~hanj/pdf/computer02.pdf.

[7] N. Barsagade, Web usage mining and pattern discovery: A survey paper, Computer Science and Engineering Dept., CSE Tech Report 8331 (Southern Methodist University,Dallas, Texas, USA, 2003).

[8] R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004), pp. 155–163

[9] P. Kolari and A. Joshi, Web mining: Research and practice, Comput. Sci. Eng.July/August (2004) 42–53

[10]B. Liu and K. Chang, Editorial: Special issue on web content mining, SIGKDD Explorations 6(2) (2004) 1–4

[11] Semantic Web Mining:State of the art and future directions" Web Semantics: Science, Services and Agents on the World Wide Web, Volume 4, Issue 2, June 2006, Pages 124-143

[12] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng."Fundamentals of content based image retrieval"www.cse.iitd.ernet.in/~pkalra/siv864/Projects/ch01_Long_v40-proof.pdf

[13] H. Zhang, Z. Chen, M. Li and Z. Su, Relevance feedback and learning in content-based image search, World Wide Web 6(2) (2003) 131–155.

[14] L. Chen, W. Lian and W. Chue, Using web structure and summarization techniques for web content mining, Inform. Process. Management: Int. J. 41(5) (2005) 1225–1242

[15] Ricardo Campos, Gael Dias, Celia Nunes, "WISE: Hierarchical Soft Clustering of Web Page Search Results Based on Web Content Mining Techniques," wi, pp.301-

304, 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), 2006.

[16] Mehdi Hosseini , Hassan Abol hassani ,"Mining Search Engine Query Log for Evaluating Content and Structure of a Web Site" in Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence.

[17] Rahmani, R.; Goldman, S.A.; Hui Zhang; Cholleti, S.R.; Fritts, J.E.; , "Localized Content-Based Image Retrieval," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.30, no.11, pp.1902-1912, Nov.2008

[18] G. Poonkuzhali, K.Thiagarajan et al "Signed Approach for Mining Web Content Outliers", World Academy of Science, Engineering and Technology 56 2009

[19] Sanjay Kumar Madria , Sourav S. Bhowmick , Wee Keong Ng , Ee-Peng Lim, Research Issues in Web Data Mining, Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery, p.303-312, September 01, 1999

[20] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.

[21] J. Furnkranz, Web structure mining — Exploiting the graph structure of the worldwide web, ¨OGAI-J. 21(2) (2002) 17–26

[22] K. A. Smith and A. Ng, Web page clustering using a self-organizing map of user navigation patterns, Decision Support Syst. 35(2) (2003) 245–256

[23] X. Fang and O. Sheng, LinkSelector: A web mining approach to hyperlink selection for web portals, ACM Trans. Internet Tech. 4(2) (2004) 209–237

[24] B. Hay, G. Wets and K. Vanhoof, Mining navigation patterns using a sequence alignment method, Knowledge Inform. Syst. 6(2) (2004) 150–163

[25] S. Guan and P. McMullen, Organizing information on the next generation web —design and implementation of a new bookmark structure, Int. J. Inform. Technol. Decision Making 4(1) (2005) 97–115

[26] Q. Song and M. Shepperd, Mining web browsing patterns for e-commerce, Comput. Indus. 57(7) (2006) 622–630

[27] Nacim Fateh Chikhi, Bernard Rothenburger, Nathalie Aussenac-Gilles "A Comparison of Dimensionality Reduction Techniques for Web Structure Mining", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence,P.116-119 ,2007

[28] Lefteris Moussiades, Athena Vakali, "Mining the Community Structure of a Web Site," bci, pp.239-244, 2009 Fourth Balkan Conference in Informatics, 2009

[29] Jaideep Srivastava, R. Cooley, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD, VOL.7 No. 2 Jan 2000

[30] Subhash K.Shinde, Dr.U.V.Kulkarni, "A New Approach For On Line Recommender System in Web Usage Mining",Proceedings of the 2008 International Conference on Advanced Computer Theory and Engineering Pages: 973-977

[31] Jianhan Zhu, Jun Hong et al," Using Markov Models for Web Site Link Prediction" College Park, Maryland, USA ACM June 11-15, 2002

[32] Borges and M. Levene,"A dynamic clustering-based markov model for web usage Mining", cs.IR/0406032, 2004

[33] TAN Xiaoqiu, YAO Min et al," Mining Maximal Frequent Access Sequences Based on Improved WAP-tree" Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)

[34] Yu-Hui Tao, Tzung-Pei Hong, Yu-Ming Su," Web usage mining with intentional browsing data" in international journal of Expert Systems with Applications 34 (2007) 1893–1904

[35] M. Jalali, N. Mustapha, M. N. Sulaiman, and A. Mamat, "Web User Navigation Pattern mining approach based on Graph Partitioning algorithm " Journal of Theoretical and Applied Information Technology vol. 4, pp. 1125-1130, 2008

[36] Mehrdad Jalali, Norwati Mustapha et al," A Recommender System Approach for Classifying User Navigation Patterns Using Longest Common Subsequence Algorithm", American Journal of Scientific Research ISSN 1450-223X Issue 4 (2009), pp 17-27

[37] M. Jalali, N. Mustapha et al ," WebPUM: A Web-based recommendation system to predict user future movements", in international journal Expert Systems with Applications 37 (2010) 6201–6212

[38] Berners-Lee, T., Fischetti, M.: Weaving the Web. Harper, San Francisco (1999

[39] Bettina Berendt, Andreas Hotho, Dunja Mladenic, Maarten van Someren, Myra Spiliopoulou and Gerd Stumme "A Roadmap for Web Mining:From Web to Semantic Web"DOI: 10.1007/978-3-540-30123-3_1

[40] J. Han and C. Chang, Data mining for web intelligence, Computer (November 2002),pp. 54–60, http://www-faculty.cs.uiuc.edu/~hanj/pdf/computer02.pdf

[41] W3Schools, Semantic web tutorial (2008) http://www.w3schools.com/semweb/default.asp

[42] Ramakrishna, Gowdar et al "Web Mining: Key Accomplishments, Applications and Future Directions", in the International Conference on Data Storage and Data Engineering 2010

[43] Feng Zhang Hhui-You Chang "Research and development in web usage mining system--key issues and proposed solutions: a survey", Proceedmgs of the First International Conference on Machine Learning and Cybernetics, Beijing, 4-5 November 2002

[44] K. A. Smith and A. Ng, Web page clustering using a self-organizing map of user navigation patterns, Decision Support Syst. 35(2) (2003) 245–256

ABOUT THE AUTHORS:

Dr. Brijendra Singh received his D.Phil degree in the field of Computer Science from J.K.Institute of applied Physis and technology, University of Allahabad, Allahabad. He has taught and researched at various reputed institutions including IIT Kharagpur, Vikram Sarabhai Space Centre (VSSC) Trivandrum and National Institute of Technology Kurukshetra, Prior to Join University of Lucknow as Professor of Computer Science, He worked as Professor and Head

Department of Information Technology, at Vellore Institute of Technology, Vellore.. At present Prof. Singh is working as Professor of Computer Science at University of Lucknow.

Prof. Singh is author of three books including "Data Communication and Network " and Network Security and Management" published by PHI Delhi. Besides He has published fifty research papers related to Computer Science and Reliability.

Prof. Singh is a recipient of the Gowari Memorial Awards in 1995 from Institution of Electronics and Telecommunications Engineers, Delhi. He has received the honor of International Man of the Year (1999-2000) from International Biographical Centre, Cambridge, England.

Prof. Singh has organized a number of conferences, Seminar and workshops in the areas of Network Security and Management, Software Quality and Reliability. He has also chaired several technical sessions and delivered invited talks in national and international conferences.

Hemant Kumar Singh received MCA degree from U.P. technical University Lucknow, India. Presently he is Assistant. Professor in department of computer applications, AzadIET Lucknow and pursuing M.Tech. degree in computer science from U.P. technical University Lucknow, India His research interests include data mining and web mining.