International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 11 | Nov - 2021 ISSN: 2582-3930

DATA MINING BASED ON THE DATA OF WEB A REVIEW

Animesh Singh[1], Gursimar Singh[2], Lakshay Mittal[3]

[123]Department of Computer Science and Engineering,

Chandigarh College of Engineering and Technology, Panjab University, INDIA.

I. INTRODUCTION

The data in today's Information technology Devices which growing at a rate which is Exponentials has a common issue when it comes to cleaning them in an Efficient manner.

Managers of their respective institutions are no longer satisfied with the Raw information about Their interacting clients. Simple Query languages like MySQL aren't no longer the standard to garner crucial information on which certain analysis can be done. 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].

Classifying the mining concept into two different types that is structured and unstructured. Text mining is mainly concerned with unstructured data/text. Web mining uses data creatively 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]

In Many regards we present a simple and unified 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. -
A typical user browses the internet by using various search engine services. But when looking in detail we find that due to inaccurate results which caused because of the low precision errors. 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 a user interacts with the web they differ in the contents and preferences

4. Learning about Consumers or individual users-This problem is about what the customer does and want. Inside this problem there are sub problem such as altering the information to the targeted consumers or even to personalize it to individual user. Other problems related management of the Web site Design, Structure 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 literature in this paper is divided into the three types of web mining:

Web content mining, Web usage mining, Web structure mining.
We classify the literature into following sections: Section 2.0 presents the web mining, section 3.1 presents the literature view review for web content mining in which we discuss 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 assembles the follow

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 Etzioni Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and service., Qingyuan Zhang and Richard s. Seagal [2] suggest decomposing Web mining into the following sub tasks:

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 Etzioni Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and service., Qingyuan Zhang and Richard s. Seagal [2] suggest decomposing Web mining into the following sub tasks:

Resource Discovery: locating unknown documents and services on the Web.

Information Selection and Pre-Processing: automatically extracting and refining specific information from New Web resources.'

Generalization: Recognizing general patterns at individual across multiple Sites.
Analysis: Validation of mined patterns.

Visualization- Representing the results in order to See through the outcomes in an understandable manner.

Kosala and Bloc keel [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. keyword-based searches and indicate research problems to use data mining effectively in developing web intelligence. It also 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.

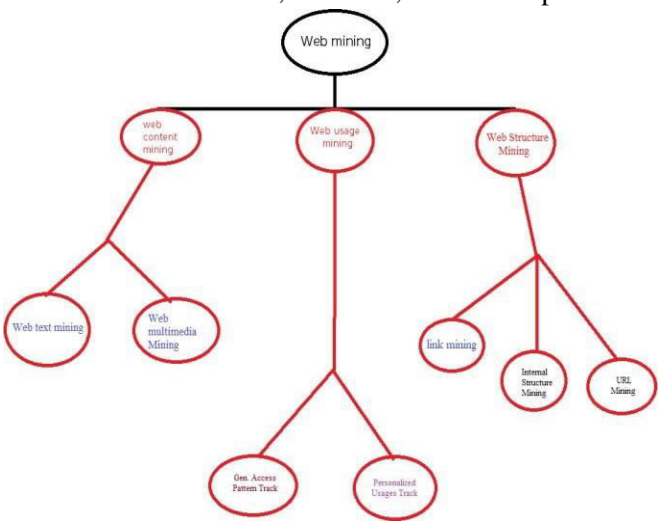


Figure-1: Taxonomy of Web Mining

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

III. LITERATURE REVIEW

3.1 Web Content Mining- It deals with finding useful knowledge from web page contents
Web content mining analyzes the content of Web resources. Which includes movies, images, multimedia etc. The primary Web resources that are mined in Web content mining are pages. Information Retrieval is one of the research areas that provide a range of popular and effective, mostly substantial methods for Web content mining. They can be used to group, categorize, analyze, and retrieve documents [11].

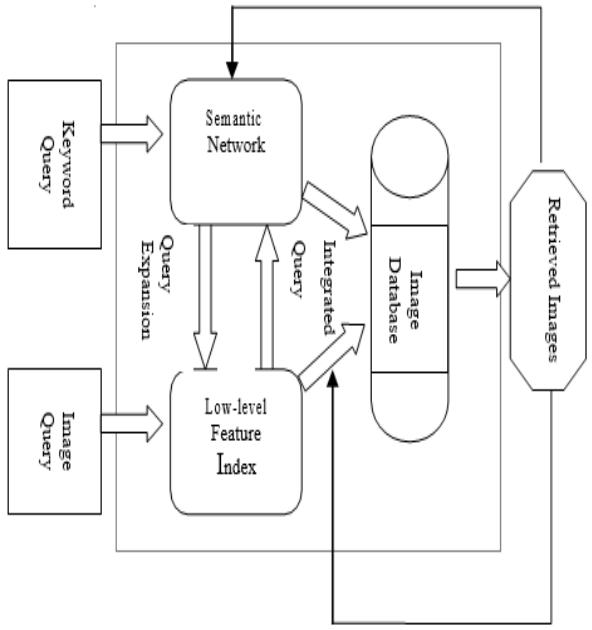
Since it is not possible to annotate images on internet manually. So, Web images are usually not well annotated using semantic descriptors. Due to multiplicity of contents in a single image and the subjectivity of human perception, it is hard to make exactly the same understanding to the similar image 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 drawback 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 of

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 generate corresponding results.

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.



Update query and similarity based on user relevance feedback

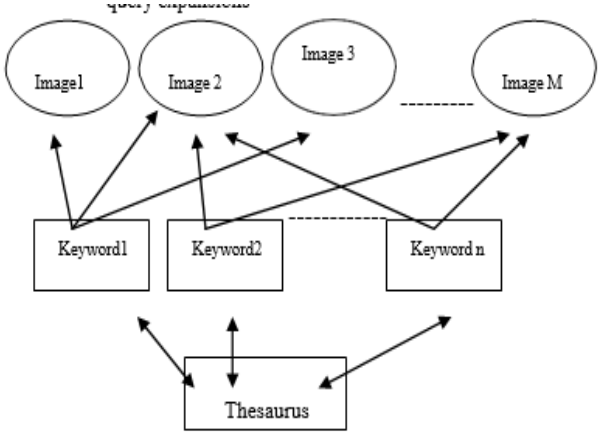


Figure 3: Semantic network

(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. Commonly, search engines when provided with a query, retrieve lots of useless web pages, and miss some important ones. The situation of hierarchical clustering of web was analysed by Ricardo Campos et al[15] and he came out with a model of meta-search engine called WISE that automatically creates clusters of inter linked web pages according to the meaning of the query. These linked pages are then hierarchically arranged and labeled according to the validity and

relevance of the information that they carry.

Mehdi Hosseini, Hassan Abul 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.

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 take 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,			

Table 1: Summary Table for Web Content mining

Irrelevant and redundant data to obtain useful information from the web is a day-to-day task. is an increasing need for the development of entities that automate the process of relevant web data mining. This is a challenge which is yet to be completed [18]. Various methods of Data mining are listed in Table 1.

3.2 Web Structure Mining- It uncovers the structure by which the web pages are linked together. Various retrieval tools use only the text that is available on the web pages and the actual relevant data is neglected. The structure mining aims at generating structural information about the web pages and websites. Web structure mining targets the linking data of the web pages. Klienberg [20] came out with hubs (reference to many other pages) and authorities (pages referred by many pages) buy using the Hypertext induced topic search as the base algorithm. He came out with a set of tools bases on algorithms for retrieving information from the link structures and reported on experiments that show their effectiveness on the World Wide Web. Authoritative information sources on certain topics helped to refine their related broad search topics. Furnkranz[21] used the concept of directed graphs to describe the web where he raised the similarities between the edges and the hyperlinks and also the similarities between the nodes and the documents thus embracing the graph structure of the World Wide Web.
To retrieve useful information is now becoming difficult with the growing size of the Internet. Smith and Ng {22] suggested using self-organizing maps to search and organize the data on the web which provided the tools for navigation. On checking the users behavior, LOGSOM was introduced which made use of a two dimensional mapping of the SOMs.

As the websites become more and more complex is has become a difficult task to retrieve useful information, such a scenario makes a troublesome situation for the people surfing the web. Fang and Sheng [23] address the design of the portal page of a website. They proposed a heuristic approach to the hyperlink selection (Link Selector) instead of using the Euclidean distance measure approach. The approach of dividing the users into groups according to which the web pages would be requested, by the use of SAM (Sequence Alignment Method) was adopted by Hay et al[24]. Guan and McMullen [25] developed a method to access a bookmark from anywhere using a Java - Enables Web browser. They came up with relating the information of the bookmarked page like its URL etc., and sharing it with all the group of users.

Various contributions in the idea of web structure mining are provided in the Table 2 which is based on dealing with the World Wide Web with a structural approach so as to be able to retrieve relevant information from the web with ease (reducing a lot of HTTP transactions).
Web structure mining was the topic of interest of many famous authors.

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 Fürnkranz 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[15]	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

1.3 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 Citation Cluster 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 Markov model etc [32]. Preprocessing Site Files Mining Algorithms Pattern Analysis Row Logs User Session Files Rules, Patterns & Satistics Interesting Rules, Patterns & Statistics 2010 IEEE International Conference on Computational Intelligence and Computing Research 666 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 WAPtree 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 Xiaoxia, 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 researchers 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 miningrecommender 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.

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 humansonly 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

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 .

I. CONCLUSION AND FUTURE DIRECTIONS

This paper aims to provide a brief overview of all the advancements made in the area of web data mining and its importance in the future . With increasing complexity and advancement in the different aspects of the web , the need for data mining to extract various valid information becomes absolutely essential. This paper also highlights the contributions made by various authors in the area of web mining – Majorly classified as Web Content Mining , Web Structure Mining , Web Usage Mining and Semantic Web Mining . Such an approach to data mining also highlights the

Model, Dynamic clustering based Moskov model etc. [32]

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.

II. ACKNOWLEDGEMENT

Second We would like to eagerly thank our HOD, Faculty and Management, who gave us a chance to research and review the concept of Web Data Mining . Their support was crucial and of utmost need during our review process.

REFERENCES

[1] Margaret H. Dunham, ―Data Mining Introductory & Advanced Topics‖, Pearson Education .[2] Qingyu Zhang and Richard s. Segal,‖ 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:
[8] A survey paper, Computer Science and Engineering
Dept., CSE Tech Report 8331 (Southern Methodist University,Dallas, Texas, USA, 2003).
[9] R. Chau, C. Yeh and K. Smith, Personalized multilingual web content mining, KES (2004), pp. 155–163
[10] 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, Servicesand 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/c h01_Long_v40-proof.pdf
[13] H. Zhang, Z. Chen, M. Li and Z. Su, Relevance feedback .

[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.
[11] 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.
[12] 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
[13] G. Poonkuzhali, K.Thiagarajan et al ―Signed Approach for Mining Web Content Outliers‖, World Academy of Science, Engineering and Technology 56 2009
[14] 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
[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604–632, 1999.
[16] J. Furnkranz, Web structure mining — Exploiting the graph structure of the worldwide web, ¨OGAI-J. 21(2) (2002) 17–26
[17] 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
[18] 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
[19] B. Hay, G. Wets and K. Vanhoof, Mining navigation patterns using a sequence alignment method, Knowledge Inform. Syst. 6(2) (2004) 150–163
[20] 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
[21] Q. Song and M. Shepperd, Mining web browsing patterns for e-commerce, Comput. Indus. 57(7) (2006) 622–630
[22] 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
[23] Lefteris Moussiades, Athena Vakali, "Mining the Community Structure of a Web Site," bci, pp.239-244, 2009 Fourth Balkan Conference in Informatics, 2009
[24] Jaideep Srivastava, R. Cooley, ―Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data‖, ACM SIGKDD, VOL.7 No. 2 Jan 2000
[25] 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
[26] Jinshan Zhu, Jun Hong et al,‖ Using Markov Models for Web Site Link Prediction‖ College Park, Maryland, USA ACM June 11- 15, 2002