Novelty Journals

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 1, pp: (36-42), Month: January - April 2015, Available at: www.noveltyjournals.com

Web Mining and Data Mining: A Comparative Approach

[1]Simranjeet Kaur, [2]Kiranbir Kaur

[1,2] Department of Computer Science & Engineering, Guru Nanak Dev University, Amritsar, India

***Abstract*: Search engine is used to retrieve data from the web in response to user's queries. Web mining is a technique of data mining, which is used to discover data from the web documents. With the help of Web mining, the user obtains the required information accurately. Web mining is categorized into three types: web content mining, web structure mining and web usage mining. In this paper, difference of three categories of web mining and comparison of web mining and data mining is also presented. Data mining helps to extract data from database, whereas data is extracted from web in web mining.**

***Keywords*: Web Mining, Data Mining, Web Structure Mining, Web Content Mining, Web Usage Mining.**

I. INTRODUCTION

As the World Wide Web is very enormous and contain all sort of information. With the growing information needs and sources, it becomes necessary to manage the information and to display most relevant information to the user screen through search engine.

Users extract information from www with the use of search engine as information retrieval tool. Commonly used search engines are Google, Yahoo, Bing, etc. They index, download and store billions of web pages, and act as content aggregators because they keep every record of web pages available on the WWW [1].

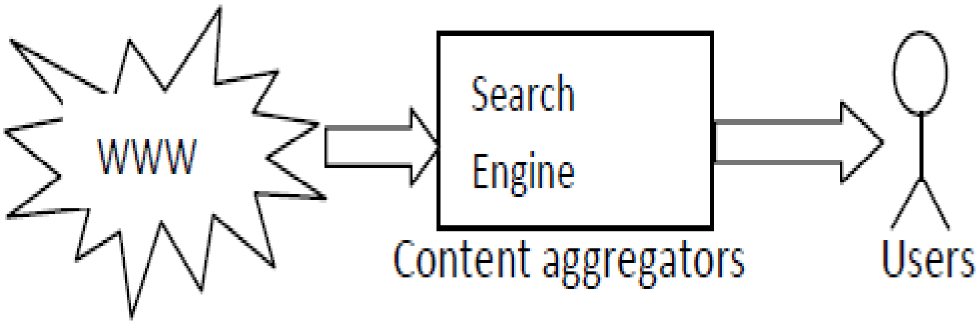

Figure 1: Concept of search engine [1]

There are primary two kinds of search services available on the internet [2]:

A. Organic Search Engine: To index websites into large database, automated programs known as "spiders" or "robots" are used. Spider index entire website when user submits page to organic search engine. Examples are Google, MSN, Yahoo, etc.

B. Search Directories: It include list of categories reviewed by human that rely on site owner's submission. It's all reviewer responsibility to describe the site in what text. Examples include Open Directory (Google Directory), Yahoo Directory, Look smart and others.

II. OVERVIEW OF SEARCH ENGINE OPTIMIZATION

Search Engine Optimization (SEO) technique is a part of Search Engine Marketing (SEM), to increase the traffic to websites. It helps the spider program to find the relevant webpage and provide it higher rank on the web, by analyzing the webpage content, keywords and webpage code [3].

III. EVOLUTION OF WEB MINING

Web Mining process began when the business data started storing on computers and internet and continued with improvements in data access, and with use of WWW. Data is collected from various sources like surveys, networked locations through the use of computers, and used as required [7].

Web mining is a revolutionary process, and the user gets the answer quickly and accurately. The evolution of web mining is shown below in table 1:

Table I: Evolution of web mining [7]

Evolutionary Steps	Technologies	Product Providers	Business Need	Characteristics
1960s **Data Collection**	Computers, disks and tapes	CDC, IBM	To know total revenue in last years	Delivery of static data
1980s **Data Access**	SQL, RDBMS, ODBC	Oracle, IBM , Sybase, Microsoft	To know about unit sales	Dynamic data delivery
1990 **Data Warehousing and Decision Support**	Multi-dimensional database, data warehouses, On-Line Analytic Processing	Microstrategy, Arbor, Congnos, Pilot	To know about unit sale and also to take decision on it	Dynamic data delivery at multiple levels
2000s **Data Mining**	Multiprocessor computers, advanced algorithms, Massive databases	IBM, SGI, Pilot, Lockheed	What to do with data and why	Prospective information delivery
Emerging Today **Web Mining**	WWW, Internet, monumental scale database, learning technology like supervised like rule generalization and unsupervised like pattern mining	IBM, Web Trends, Net Genesis, Rockware, Apecto Limited, Heckyl Technologies	To know about business in past or future	Affordable tool to mine large data warehouse and relational databases efficiently and fast using multiple mining functions, Powerful

Web Mining is a data mining technique that helps to extract or discovers information from the web documents. Web data can be in the form of web pages like text and images, inter-page structure i.e. linkage structure between web pages, intra page structure i.e. include XML or HTML tags and about the usage of data on the internet [5].

IV. WEB MINING PROCESS

The process of extracting knowledge from web is as follows [4]:

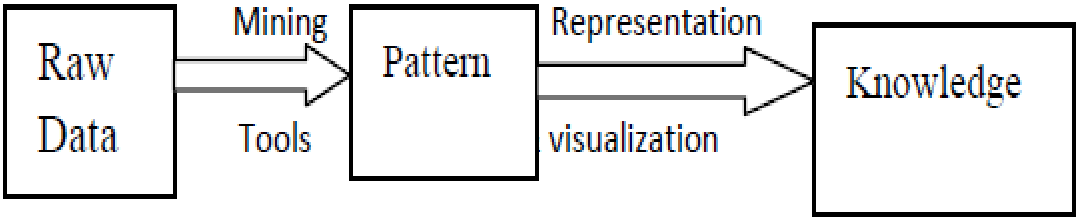

Figure 2: Web Mining Process [4]

In Web mining, data mining is used to extract the hidden data in web log (usage data). The subtasks of web mining are shown in figure 3:

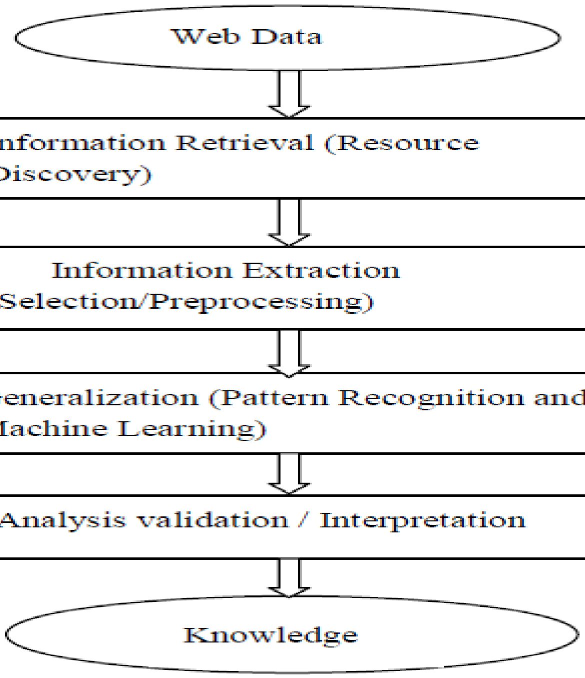
Web Data
Information Retrieval (Resource Discovery)
Information Extraction (Selection/Preprocessing)
Generalization (Pattern Recognition and Machine Learning)
Analysis validation / Interpretation
Knowledge

Figure 3: Subtasks of web mining [7]

V. WEB MINING CATEGORIES

Web Mining is categorized into three types: web content mining, web structure mining and web usage mining as shown in figure 4.

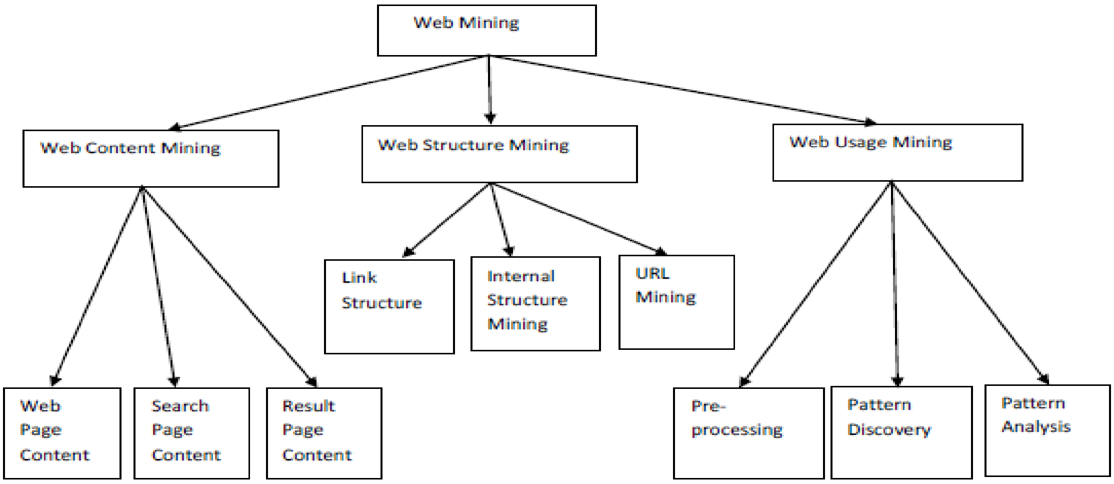

Figure 4: Categories of Web Mining [7,9]

A. Web Content Mining

It is the process to retrieve information from web into structured form and index the information, and extract it quickly. Its main focus is on the structure on the inner document and it consists of text, images, audio, video or structured records such as tables and lists [6].

B. Web Structure Mining

It is a process to model the linking structure of web pages. Its main motive is to generate structured summary of web pages and websites [6]. This is carried out at the hyperlink level (inter-page) and at document level (intra-page). This type of mining is useful for retrieving information [4].

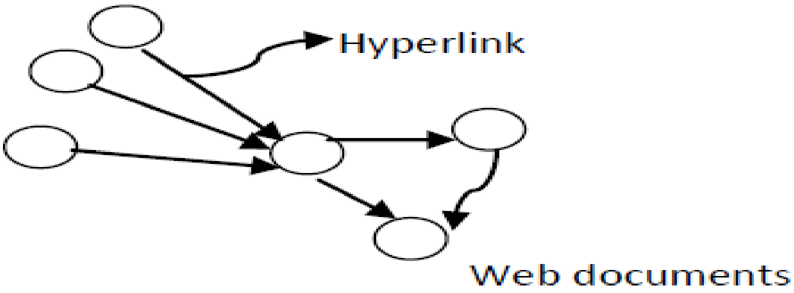

Figure 5: Web graph structure [8]

Web structure mining consists of three categories [9]:

- ***Link Structure:*** It include link based cluster analysis, link based classification, link type and link strength.
- ***Internal Structure Mining:*** It provides the information about ranking of pages, and to enhance the search results by discovering the model underlying the link structure. It is also helpful to analyze the relationship between different websites.
- ***URL Mining:*** In this category, hyperlink is used to connect a web page to different locations, either within same or different web page.

C. Web Usage Mining:

This type of mining is used to discover useful information and navigation patterns from the web present in server logs, agent logs, referrers log, client-side cookies, Meta data and user profile. It introduces privacy concern [8].

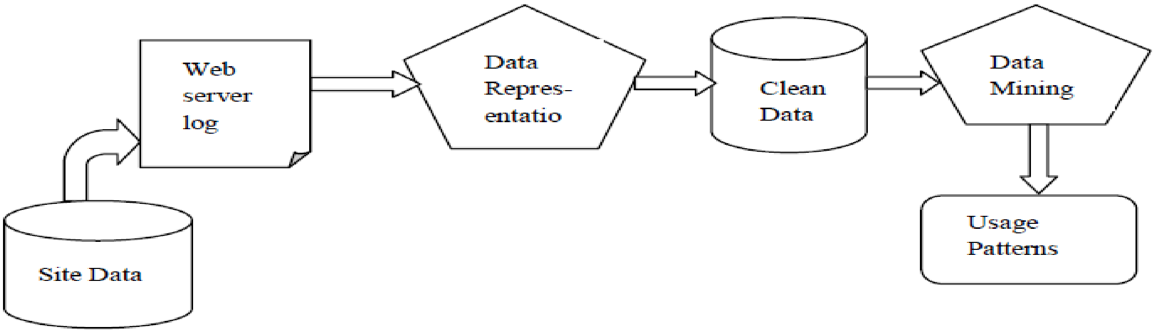

Figure 6: Web usage mining process [8]

When the user interacts with web, it is used to predict user behavior [7]. It consists of three phases:

- ***Preprocessing:*** Proxy server and server is the first way used to retrieve raw data from the web and process it, and it makes automatic transformation of original raw data [9].
- ***Pattern Discovery:*** After discovering the knowledge from preprocessing, techniques are implemented to discover the knowledge such as machine learning, data mining procedures and other related procedures [9].
- ***Pattern Analysis:*** After the stage of pattern discovery, pattern analysis work is to check the pattern on the web and its implementation to extract the knowledge from the web [9].

Table II: Categories of Web Mining [6, 9]

Web Mining Categories			
Specification	**Web Content Mining**	**Web Structure Mining**	**Web Usage Mining**
View of data	Structured Semi-structured Unstructured	Structure linkage	Interactive data
Type of data used	Primary	Primary	Secondary
Main data	Text document Hypertext document	Link structure	Browser logs Server logs
Representation	Concepts or ontology Relational Edge labeled graph n-grams Terms, Phrases	Graph	Relational table Graph
Method	Machine learning Association rules Proprietary algorithm Statistical method	Proprietary algorithm	Machine learning Statistical method
Tasks	It tells about the discovery of useful information from web documents/contents	It discover the model underlying web's link structure	Its work is to make sense of the data generated by web surfer's behavior
Scope	The scope of data is local in DB and is global in IR.	Global	Global
Goal	It mainly targets to discover knowledge	To generate structural summary about the web page and web site [10]	To analyze the behavioral pattern and users profile while interacting with web sites
Application areas	Clustering Categorization Finding extraction rules and patterns User modeling Finding frequent sub-schemas	Clustering Categorization	User modeling Site construction Marketing Adaptation and Management
Challenges	Data/Information extraction Opinion extraction from online sources Segmenting web pages and noise detection [10]	Not all pages have relevant meta information Entire text of predecessor page is not relevant	Preprocessing challenges about the users i.e. who will be user, how long it stay, where will it go and user view?

VI. COMPARISON OF WEB MINING AND DATA MINING

Data Mining (also called data or knowledge discovery) is the process to find hidden information or patterns in databases [11], and summarizing it into useful information. It is also the process to find co-relations among large number of fields in huge relational databases [17], whereas Web Mining is a technique that is used to retrieve the documents from the web.

Data Mining includes the five major steps: extract, store and manage data, provide data access, analyze the data and present it in a useful form, like in the form of table or graph [17].

Table III: Comparison of Web Mining and Data Mining [19]

Web Mining VS Data Mining		
Comparison	**Web Mining**	**Data Mining**
Definition	Process used to extract information from web documents.	Process used to extract hidden information from the database.
Scale	It contains 10 million jobs in server database, and therefore search processing is not big.	It contains 1 million jobs in database and search processing is large.
Structure	The information is obtained from structured, semi-structured and unstructured web forms. It gets the information from wide database.	It obtains the information from explicit structure. It is not able to get all the information from wide database as compared to web mining.
Access	Data is accessed publicly. In this, data is not hidden in web database and only permission is required to access the data from web log master.	Data is accessed privately and only authorized user can access the data.
Data	It works upon on-line data.	It works upon off-line data.
Data Storage	Data is stored in server logs and web server database.	Data is stored in data warehouses.
Application Areas	E-learning, Digital Libraries, E-Government, Electronic Commerce, E-Politics, E-Democracy, Security & Crime Investigation, Electronic Business [13].	Banking, marketing, manufacturing & production, health-care, insurance, law, airlines, computer hardware & software, government & defence, etc [12].
Disadvantages	url's can be tracked to access the data, multiplicity of events and url's, large amount of data remain unused	Privacy issues, security issues, misuse of information/ inaccurate information [12].
Challenges	Complexity of web pages, web is too huge, relevancy of information, web is dynamic information source, diversity of user communicates, etc [14].	Network settings, data quality, privacy preservation, scalability, complex and heterogeneous data, etc [12].
Techniques	Web Content Mining, Graph Based Web Mining, Utilization in Web Mining, Text Mining and many others [15].	Artificial Neural Network, Decision Trees, Rule Induction, Nearest Neighbor Method and many others [16].

VII. CONCLUSION

The need to access and manage data on the web is increasing day by day. Web mining enhances user's ability to access information from www and finds its applications in various fields like Clustering, Categorization, in extraction rules and patterns and many others. In the comparison of web mining with data mining, it is concluded that web mining is used to retrieve online data, and data mining retrieves offline data. Data is stored in server database in web mining and it can handle multiple transactions at the same time. Data can be discovered and extracted from multiple locations of the world by sitting at one location and is able to provide desired information at the time of requirement.

REFERENCES

[1] N. Duhan, A.K. Sharma and K. K. Bhatia, “Page Ranking Algorithms: A Survey”, IACC 2009

[2] Clifton and N. Rae, “How Search Engine Optimization Works”, SEO Whitepaper

[3] Z. Hui, Q. Shigang, L. Jinhua and C. Jianli, “Study on Website Search Engine Optimization”, International Conference on Computer Science and Service System, 2012

[4] N. Tyagi and S. Sharma, “Comparative study of various Page Ranking Algorithms in Web Structure Mining”, IJITEE, ISSN: 2278-3075, Volume-1, Issue-1, June 2012

[5] Jain, R. Sharma, G. Dixit and V. Tomar,” Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages”, International Conference on Communication Systems and Network Technologies, 2013

[6] R. Jain and Dr. G.N Purohit, “International Journal of Computer Applications (0975-8887) Volume 13-No.5, January 2011

[7] K. Sharma, G. Shrivastava and V. Kumar, “Web Mining: Today and Tomorrow”, IEEE, 2011

[8] M. T. Ramakrishna, L. K. Gowdar, M. S. Havanur and B. P. M. Swamy, ”Web Mining: Key Accomplishments, Applications and Future Directions”, International Conference on Data Storage and Data Engineering, 2010

[9] B. Rathod and Dr. S. Khanna, “A Review on Emerging Trends of Web Mining and It’s Applications”, IJEDR, ISSN: 2321-9939

[10] Web Content Mining Problems/Challenges, available at, “https://sites.google.com/site/assignmentssolved /mca/ semester6 /mc0088/12”

[11] N. Padhy, Dr. P. Mishra and R. Panigrahi, “The survey of Data Mining Applications And Future Scope”, International Journal of Computer Science, Engineering and Information Technology, Vol.2, No.3, June 2012

[12] S. H. Begum, “Data Mining Tools and Trends – An Overview”, International Journal of Emerging Research in Management & Technology, ISSN: 2278-9359, February 2013

[13] S. Yadav, K. Ahmad and J. Shekar, “Analysis of Web Mining Applications and Beneficial Areas”, IIUM Engineering Journal, Vol. 12, No. 2, 2011

[14] Data Mining – Mining World Wide Web available at “http://www.tutorialspoint.com/data _mining/dm_ mining_ www.htm”

[15] P. Bhisikar and A. Sahu, “Overview on Web Mining and Different Technique for Web Personalization”, In International Journal of Engineering Research and Applications, ISSN: 2248-9622, Vol. 3, Issue 2, March-April 2013

[16] An Introduction to Data Mining, available at, “http://www.thearling.com/text/dmwhite/dmwhite.htm”

[17] Data Mining: What is Data Mining? Available at,”http://www.anderson.ucla.edu/faculty/jason. frand/teacher/ techno -logies/palace/datamining.htm”

[18] Rastogi, S. Gupta, S. Agarwal, N. Agarwal, ”Web Mining: A Comparative Study”, International Journal of Computational Engineering Research, ISSN: 2250-3005, Vol. 2, Issue No. 2, Mar-Apr 2012

[19] R. Sharma, “A Framework to Compare Web Mining Types”, In International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277-128X, Volume 3, Issue 7, July 2013

Author Profile:

Simranjeet Kaur belongs to Amritsar, Punjab (India) and her Date of Birth is 15 December, 1989. She is B.Tech (IT) and pursuing M.Tech (Software System) from Guru Nanak Dev University in the Department of Computer Science and Engineering, Amritsar, India. Her interest area is search engine optimization and web information retrieval. She completed this paper under the guidance of Kiranbir Kaur, Assistant Professor in Guru Nanak Dev University, Department of Computer Science and Engineering, Amritsar, India.