Web Mining: Today and Tomorrow

Kavita Sharma
M.Tech. (Information Security)
Ambedkar Institute of Technology
(G.G.S.I.P. University),
New Delhi, India
kavitasharma_06@yahoo.co.in

Gulshan Shrivastava
M.Tech. (Information Security)
Ambedkar Institute of Technology
(G.G.S.I.P. University),
New Delhi, India
gulshanstv@gmail.com

Vikas Kumar
M.Tech. (Computer Engineering)
PDM College of Engineering
(M.D. University),
Haryana, India
getforvikas@yahoo.in

***Abstract*—In this paper we presents study about how to extract the useful information on the web and also give the superficial knowledge and comparison about data mining. This paper describes the current, past and future of web mining. Here we introduce online resources for retrieval Information on the web i.e. web content mining, and the discovery of user access patterns from web servers, i.e. web usage mining that improve the data mining drawback. Furthermore, we also described web mining through cloud computing i.e. cloud mining. That can be seen as future of Web Mining.**

Keywords-Web Mining; Web Content Mining; Web Structure Mining; Web Usage Mining; Cloud Mining

I. INTRODUCTION

The wide adoption of the Internet has fundamentally altered the ways in which we communicate, gather information, conduct businesses and make purchases. As the use of the World Wide Web and email skyrocketed, computer scientists and physicists rushed to characterize this new phenomenon. While initially they were surprised by the tremendous variety the Internet demonstrated in the size of its features, they soon discovered a widespread pattern in their measurements: there are many small elements contained within the Web, but few large ones. A few sites consist of millions of pages, but millions of sites only contain a handful of pages. Few sites contain millions of links, but many sites have one or two. Millions of users flock to a few select sites, giving little attention to millions of others.

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. [1] [12] New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area.

A. Historical Evolution of Web Mining

Web mining techniques are the result of long process of research and product development. This evolution began when business data was first stored on computers & internet, continued with improvements in data access and more recently real-time technologies that users in the WWW (World Wide Web) allow us to navigate through our data. Data gathered from surveys, or input from several independent or networked locations via Computer & tapes are define Data collection. Use the data storage typically refers to software and related activities, the retrieving, or stored in a database or other data source acting on.

In the evolution from business data to business information each new step has built upon the previous one. For example, the ability to store large databases is critical to web mining. From the user point of view, the five step listed in Table 1 were revolutionary because they allowed new business question to be answered accurately and quickly.

Basically data mining technique are used in web mining. Web mining is extended version of data mining. Data mining is work upon Off-Line whereas Web mining is work upon On-Line. In data mining data stored in (database) data warehouse and in web mining data stored in server database & web log.

The main component of Web Mining Technology have been under development for decades, in research area such as internet, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make these technologies practical for current data warehouse environments. [10]

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue	Computer, Tapes, Disks	IBM,CDC	Retrospective, Static data delivery

978-1-4244-8679-3/11/$26.00 ©2011 IEEE

	in the last five years?"			
Data Access (1980s)	"What were unit sales in Delhi last March?"	Relational Database (RDBMS), Structure Query language (SQL),ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, Dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in Delhi last march? Drill down to Mumbai."	On-Line Analytic Processing (OLAP), Multidimensional Databases, Data warehouses	Pilot, Comshare, Arbor, Congnos, Microstrategy	Retrospective, Dynamic data delivery at multiple levels
Data Mining (2000s)	"What's likely to happen to Mumbai unit sales next month? Why?"	Advanced algorithms, Multiprocessor Computers, Massive Databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, Proactive information delivery
Web Mining (Emerging Today)	"What's likely to happen to Mumbai unit sales next/previous millions months? "	WWW, Internet, monumental scale Database	RockWare, Apteco Ltd., Simon Fraser University, IBM, Web Trends, SPSS, Flowerfire, Angoss, Net Genesis	Powerful, Affordable tool to mine large data warehouse and Relational databases fast and efficiently using multiple mining functions

Table 1 : (Steps in Evolution of Web Mining)

B. Drawbacks in the existing approaches

1. The response time perceived by the user is too long.
2. The explosive growth of the Web has imposed a heavy demand on networking
3. Resources and Web servers.
4. Hence, an obvious solution in order to improve the quality of Web services would be the increase of bandwidth, but such a choice involves increasing economic cost.
5. Web caching scheme has three significant drawbacks: If the proxy is not properly updated, a user might receive stale data, and, as the number of users grows, origin servers typically become bottlenecks.
6. The several factors diminish the ideal effectiveness of Web caching. The obvious factors are the limited system resources of cache servers (i.e., memory space, disk storage, I/O bandwidth, processing power, and networking resources). However, even if the cache space is unlimited, there are significant problems that cannot be avoided by such an approach. Specifically, large caches are not a solution because, the problem of updating such a huge collection of Web objects is unmanageable.
7. Main drawback of systems which have enhanced prefetching policies is that some prefetched objects may not be eventually requested by the users. In such a case, the prefetching scheme increases the network traffic as well as the Web servers' load.[9]

II. WEB MINING

Web Mining is based on knowledge discovery from web. It is extract the knowledge framework represents in a proper way. Web min g is like a graph & all pages are node & each connects with hyperlinks. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. By using web mining easily extract all features and information about multimedia before this web mining difficult to extract information in proper way from web. We search the any topic from web difficult to get accurate topic information but Now's day it is easy to get the proper

information about any things. [2][1]Web mining is based on data mining technique by using data mining technique discover the hidden data in web log. Thus, web mining, though considered to be a particular application of data mining, warrants a separate field of research.

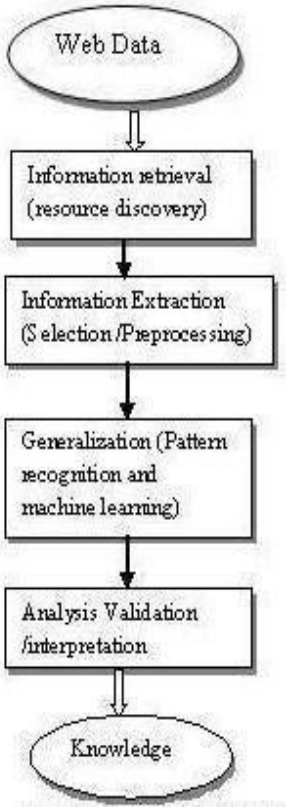

Figure 1: (Web Mining Subtask)

Based on the aforesaid four subtasks (Figure. 1), web mining can be viewed as the use of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. Here, evaluation includes both "generalization" and "analysis.

A. Web Mining Categories

Web mining can be categorized in to three area of interest based on which part of the web to mine:

a) Web Content Mining
b) Web Structure Mining
c) Web Usage Mining

i. Web Content Mining

Web Mining is basically extract the information on the web. Which process is happen to access the information on the web. It is web content mining. Many pages are open to access the information on the web. These pages are content of web. Searching the information and open search pages is also content of web. Last accurate result is defined the result pages content mining.

ii. Web Structure Mining

We can define web structure mining in terms of graph. The web pages are representing as nodes and Hyperlinks represent as edges. Basically it's shown the relationship between user & web. The motive of web structure mining is generating structured summaries about information on web pages/webs. It is shown the link one web page to another web page.

iii. Web Usage Mining

It is discovery of meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of inter related files on one or more web servers. It is automatically generated the data stored in server access logs, refers logs, agent logs, client sides cookies, user profile, meta data, page attribute, page content & site structure.[3]

Web mining usage aims at utilize data mining techniques to discover the usage patterns from web based application. It is technique to predict user behavior when it is interact with the web. [11] Web usage mining is categories in three phases:-

- Preprocessing
- Pattern Discovery
- Pattern Analysis

Preprocessing- According to client, server and proxy server it is first approach to retrieves the raw data from web resources and processed the data .it is automatically transformed the original raw data.

Pattern Discovery- According the data preprocessing discovered the knowledge and implements the techniques to discover the knowledge like as machine learning and data mining procedures are carried out at this stage.

Pattern Analysis- pattern analysis is the process after pattern discovery. Its check the pattern is correct on the web and how to implement on web to extract the information on your web search / extract knowledge from the web.

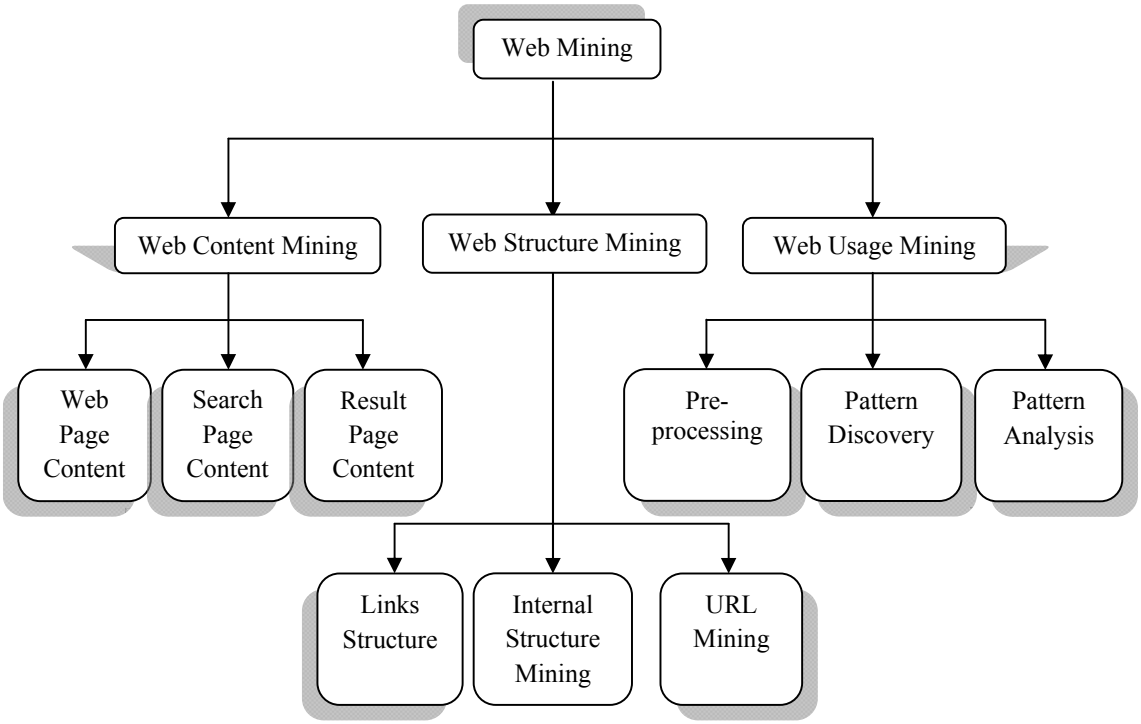

Figure 2: (Classification of Web Mining)

B. Web Mining v/s Data Mining

Comparison	Web Mining	Data Mining
Scale	In this the search processing is not a big, 10 million job in web server database	In this the search processing is large, a 1 million jobs in data base
Access	Web Mining is access data publicly. In this not hide the data which is access in web database. But take permission to web log master and access the data.	Data Mining is access data privately and only authorize user access data in the database**.**
Structure	In Web mining get the information from structured, unstructured and semi structured from web pages. web mining fetch the information from wide database	In Data mining get the information from explicit structure. Data mining is not fetch the information from wide database compares to web mining database.

Table 2: (Web Mining v/s Data Mining) [7]

III. WEB MINING THROUGH CLOUD COMPUTING

Cloud Computing is clearly one of today's most seductive technology areas due at least in part to its cost efficiency and flexibility. However, despite increased activity and interest, there are significant, persistent concerns about cloud computing that are impeding momentum and will eventually compromise the vision of cloud computing as a new IT procurement model [8]. The term 'cloud' is a symbol for the Internet, an abstraction of the Internet's underlying infrastructure, used to mark the point at which responsibility moves from the user to an external provider.

Basically Cloud Mining is new approach to faced search interface for your data. SaS (Software-as-a-Service) is used for reducing the cost of web mining and try to provide security that become with cloud mining technique. Now a day we are ready to modify the framework of web mining for demand cloud computing. [6] In terms of "mining" clouds, the Hadoop and MapReduce communities who have developed a powerful framework for doing predictive analytics against complex distributed information sources.

IV. RELATED WORK

Many researchers have looked for way of represent the web mining and future of web mining. Some of these are said that cloud mining is the future of web mining.

As we know, Etzioni is the first person who coined the term Web Mining. This paper describes the web mining subtask and process [3].

D. Sravan Kumar and B. Naveena Devi [4] described the web mining classification like Content mining, Structure mining & Usage mining. In this paper they also describe the process of web usage mining.

V. CONCLUSION & FUTURE WORK

We provide a survey about the research in the area of Web mining's today structure and tomorrow view. We point some confusion between data mining and web mining. Web data is growing at a significant rate. Web Mining is fertile area of research. Many Successful applications exist. We also suggest the subtask of web mining & future of web mining. Now we also work for the process mining and try to combine usage mining with structure mining. We also go for the mining from cloud. Whenever we work on mining over cloud computing that time we hesitate for the cost but that come very less by cloud mining. So, we can say that cloud mining can seen as future of web mining.

ACKNOWLEDGMENT

We are grateful to Dr. Vishal Bhatnagar (Associate Professor, Ambedkar Institute of Technology), Mr. C. M. Sharma (Research Scholar-PhD and Asstt. Prof., BPIT), & Mr. Parbal Partab (Scientist, DRDO) for taking time to read carefully drafts of this paper and provide us with valuable comments.

The authors are also grateful for thoughtful comments from reviewers who improved the content of the paper.

REFERENCES

[1]. Virgilio Almeida, Azer Bestavros, Mark Crovella, and Adriana de Oliveira, " Characterizing reference locality in the WWW " , In IEEE International Conference in Parallel and Distributed Information Systems, Miami Beach, Florida, USA, December 1996. http://www.cs.bu.edu/groups/oceans/papers/ Home.html

[2]. Pei, J. Han, J. Mortazavi, B. and Zhu, H. "Mining Access Patterns efficiently from Web Logs," Proc. Pacific- Asia Conf. Knowledge discovery and Data Mining (PAKDD'00) 2000.

[3]. Etzioni, O. "The World Wide Web: Quagmire or gold mine", Communication of the ACM, Vol. 39, No. 11, pp. 65-68, 1996.

[4]. Sravan Kumar, D. and Naveena Devi, B. "Learner's Centric Approach for Web Mining " et al. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1(2), 2010.

[5]. Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization" in ACM Transaction on Internet Technology, Vol. 3, No.1, Feb. 2003.

[6]. Ajay Ohri "Data mining through Cloud Computing". http://knol.google.com/k/data-mining-through-cloud-computing# See on Dec. 2010.

[7]. Michael Jennings," What are the major comparisons or differences between Web mining and data mining?" Information Management Online, June 25, 2002.

[8]. Deyi Li, Kaichang Di, Deren Li and Xuemei Shi, "Mining association rules with linguistic cloud models", Research and Development in Knowledge Discovery and Data Mining ,Lecture Notes in Computer Science, 1998.

[9]. Faten Khalil, Jiuyong Li and Hua Wang "A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses" ,Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184.

[10]. Raymond Kosala, Hendrik Blockeel," Web Mining Research: A Survey", In ACM SIGKDD, July 2000.

[11]. Wu, K.L. Yu, P. S. Ballman, A. "A Web usage mining and analysis tool", IBM Systems Journal, 2010.

[12]. Chen, M. S, Han, J. and Yu, P. S. "Data Mining: An overview from a database perspective", IEEE transaction on knowledge and data engineering, Vol. 08, No. 6, pp: 866-883, 1996.