WEB MINING: A COMPARATIVE STUDY

Aishwarya Rastogi
Assistant Professor
Dept. of CS & IT
MIT Moradabad

Smita Gupta
Assistant Professor
Dept. of CS & IT
MIT Moradabad

Srishti Agarwal
Assistant Professor
Dept. of CS & IT
MIT Moradabad

Nimisha Agarwal
Assistant Professor
Dept. of CS & IT
MIT Moradabad

Abstract:
Currently, World-Wide Web has developed to a distributed information space with nearly 100 million workstations and several billion pages, which brings the people great trouble in finding needed information although huge amount of information available on webs. The search engine is a very important tool for people to obtain information on Internet, but the low-precision and low-recall exist widely in current search engines. With the rapid development of Internet, the effective and accurate intelligent search engine based on the Web mining technology has become the most important research issue.

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. Web content mining, web structure mining and web usages mining. Through this paper we presents a view about how to extract the useful and relevant information on the web using web mining and also give the superficial knowledge and brief comparison about data mining. This paper discusses 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- Data mining: Web Mining; Web Content Mining; Web Structure Mining; Web Usage Mining; Semantic Web; Cloud Mining.

1. Introduction

Because of the rapid development and wide application of the Internet, World Wide Web has become a pool, exchange, sharing of information and effective tool for collaborative work. People's attention and frequent use of the Web not only promotes the development of various technologies, but also make the Web information resources on the rapid growth. The wide adoption of the Internet has fundamentally changed the ways in which we communicate, gather information, conduct businesses and make purchases. Therefore, resulting in flood of information resources distributed on the Web that provides various facilities to satisfy the user's need. Earlier people used to communicate through postal services, purchase the products from nearby markets, and gather information from news papers and magazines. Even people do business and banking transactions manually through paper work. But in today era we have a vast ocean of data which we called as internet or web. This huge library of data originates as a result of modernization and globalization of data over internet. All the activities discussed above, which we used to do manually earlier, now becomes the part of internet and hence the result of increasing data over web.

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. For example a science student is concerned about the text or videos related to physics, chemistry or any other subject of his concern while a commerce student is in need of a statistical or economical knowledge. Similarly engineers, doctors, scientist, homemakers, search the web according to their own requirements. If a person is only concerned about a small information of Web, and not interested in the rest of the information contained in Web, because of the two reasons he may become unsatisfactory: first, the desired search results will be submerged by the traditional search engines which are based on the keywords; second, since the majority of Web data is unstructured, which lead to the traditional data mining results will be unsatisfactory. Therefore, to avoid all these problems, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Various data mining methods are used to discover the hidden and useful 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. 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. This paper is organized as follows: Section 2 describes background and origin of web mining. Section 3 describes about the web mining and all its categories Section 4 describes the comparative study between the web mining and data mining. Section 5 contains some prominent applications that can be used as future directions of web mining. Section 6 finally concluded with all the necessary aspects of web mining.

2. Origin Of Web Mining

Web mining techniques are the result of long process of research and product development. This evolution began when the amount of data kept in computer files and databases 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. 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. Data mining can be best applied by using a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates architecture for advanced analysis in a large data warehouse.

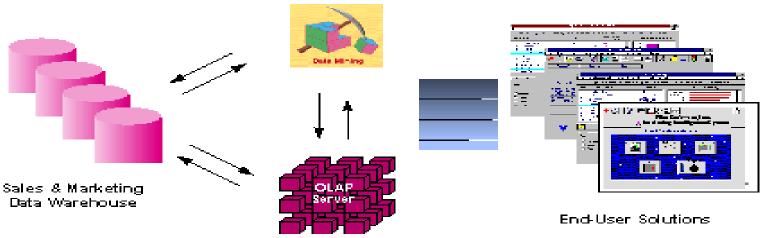

Figure 1 - Integrated Data Mining Architecture

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.

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.

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in Mumbai last March?"	Relational databases (RDBMS), Structured 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 Mumbai last March? Drill down to Delhi."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (2000s)	"What's likely to happen to Delhi 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 Delhi 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

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.

Drawbacks in the existing approaches

1. The explosive growth of the Web has imposed a heavy demand on networking
2. Resources and Web servers.
3. 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.
4. 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 bottleneck.
5. Main drawback of systems which have enhanced prefetching policies is that some prefetched objects may not be eventually requested by the users.

The purpose of the paper is to provide past, current evaluation and future direction in each of the three different types of web mining i.e. web content mining, web structure mining and web usages mining.

3. Web Mining

OVERVIEW: Web mining means employing the technique of data mining into the documents on the net. Web mining can be used for studying varied aspects of a site can recognize the patterns and relationships in the user behavior so as to get the insight in crucial information**.**

Web mining is an extension of data mining. Web Mining is based on knowledge discovery from web. It is the extraction of the knowledge framework represents in a proper way. 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 and relevant information. Web mining is based on data mining technique by using data mining technique discover the hidden data in web log.

The main component of Web Mining Technology has 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].

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 extracts 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. 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.

	Web Mining			
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR View	DB View		
View Of Data	- Unstructured -Semi Structured	-Semi Structured - Web Site as DB	-Links Structure	- Interactivity
Main Data	-Text Documents -Hypertext Documents	-Hypertext Documents	-Links Structure	- Server Logs - Browser Logs
Representation	- Bag of Words, n-grams - Terms, phrases - Concepts or Ontology - Relational	- Edge-Labeled Graph - Relational	- Graph	- Relational Table - Graph
Method	- TFIDF and Variants -Machine Learning -Statistical (including NLP)	-Proprietary algorithms - ILP -(Modified) association rules	- Proprietary algorithms	- Machine Learning - Statistical -(Modified) association rules
Application Categories	- Categorization - Clustering -Finding Extraction Rules - Finding Pattern in text - User Modeling	-Finding Frequent Sub-Structures -Web site schema discovery	- Categorization - Clustering	- Site Construction, Adaptation and Management -Marketing -User Modeling

4. Comparison Between Data Mining And Web 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 fromweb 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.
Data	Data mining is work upon Off-Line data	Web mining is work upon On-Line Data
Data Storage	In data mining data stored in (database) data warehouse	In web mining data stored in server database & web log.

5. Future Directions

This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the Web and the recent interest in e-commerce. This phenomenon partly creates confusion when we ask what constitutes Web mining and when comparing research in this area. This trend is likely to continue as Web services continue to flourish. As the Web and its usage grow, it will continue to generate evermore content, structure, and usage data, and the value of Web mining will keep increasing. Outlined here are some research directions that must be pursued to ensure that we continue to develop Web mining technologies that will enable this value to be realized.

i) Web mining using 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. 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 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. With the significant advances in Information and Communications Technology over the last half century, there is an increasingly perceived vision that computing will one day be the 5th utility (after water, electricity, gas, and telephony). This computing utility, like all other four existing utilities, will provide the basic level of computing service that is considered essential to meet the everyday needs of the general community. To deliver this vision, a number of computing paradigms have been proposed, of which the latest one is known as Cloud computing.

ii) Web metrics and measurements

From an experimental human behaviorist’s viewpoint, the Web is the perfect experimental apparatus. Not only does it provide the ability of measuring human behavior at a micro level, it (i) eliminates the bias of the subjects knowing that they are participating in an experiment, and (ii) allows the number of participants to be many orders of magnitude larger. Research needs to be done in developing the right set of Web metrics, and their measurement procedures, so that various Web phenomena can be studied.

iii) Process mining

Mining of ’market basket’ data, collected at the point-of-sale in any store, has been one of the visible successes of data mining. However, this data provides only the end result of the process, and that too decisions that ended up in product purchase. Research needs to be carried out in (i) extracting process models from usage data, (ii) understanding how different parts of the process model impact various Web metrics of interest, and (iii) how the process models change in response to various changes that are made - changing stimuli to the user.

iv) Temporal evolution of the Web

Society’s interaction with the Web is changing the Web as well as the way the society interacts. While storing the history of all of this interaction in one place is clearly too staggering a task, at least the changes to the Web are being recorded by the pioneering Internet Archive project. Research needs to be carried out in extracting temporal models of how Web content, Web structures, Web communities, authorities, hubs, etc. are evolving.

v) Web services optimization

As services over the Web continue to grow, there will be a need to make them robust, scalable, efficient, etc. Web mining can be applied to better understand the behavior of these services, and the knowledge extracted can be useful for various kinds of optimizations.

vi) Fraud and threat analysis

The anonymity provided by the Web has led to a significant increase in attempted fraud, from unauthorized use of individual credit cards to hacking into credit card databases for blackmail purposes.

6. Conclusions

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper we have briefly described the key computer science contributions made by the field, the prominent successful applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

7. 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 datamining research, Int. J Inform.Technol. Decision Making5(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]. 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.

[6]. http://www.cs.bu.edu/groups/oceans/papers/ Home.html.

[7]. http://www.thearling.com/text/dmwhite/dmwhite.htm

[8]. http://www.engr.sjsu.edu/~fayad/workshops/UDME07/cfp.php

[9]. http://www.data-mining-software.com/data_mining_history.htm

[10]. Raymond Kosala, Hendrik Blockeel,” Web Mining Research: A

Survey”, In ACM SIGKDD, July 2000.

[11]. http://www.expertstown.com/web-mining/

[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.