Web Mining Evolution & Comparative Study with Data Mining

Anu,
Assistant Professor (Resource Person)
University Institute of Engineering and Technology
Mahrishi Dayanand University
Rohtak-124001, India
E-Mail: anukadyan01@gmail.com

Abstract:- Web Technology is evolving very fast and Internet Users are growing much faster than estimated. The website users are using a wide range of websites leaving back a variety of information. This information must be used by the websites administrator to manipulate their websites according to the users of the websites. Aim of research in web mining is to develop a new technique for extracting and mining useful information or knowledge from web pages. Thus it's a challenging task for automated discovery of targeted or unexpected knowledge due to heterogeneity and lack of structure of web data. In this paper we will discuss about the evolution of web mining. This paper will contain detailed description about the other parts of web mining. Paper also analyse data mining and made a comparison between data mining and web mining on the basis of various parameters.

Keywords: *Web mining, Data mining, web content mining, web structure mining, web usage mining*.

*****

I. INTRODUCTION

Use of internet is growing continuously with rapid increase in the volume of information transactions and number of requests made by web users from all over the world. Therefore it become necessary for the web administrator to improve the quality of the web information service by discovering the hidden information about the user's access. Also from the business point of view, marketing and management of E-business, E-services, E-searching and E-education etc. could be directly affected by the knowledge obtained from the access patterns of web users.

Web mining is an application of data mining technique commonly used to extract information from web data, which may include web documents, hyper linking of documents, history of website usage. Initially when web mining was introduced, two different approaches were taken in to define web mining

- First approach was 'process-centric view' which defined web mining as a sequence of tasks.
- Second approach was 'data-centric view' which defined web mining in terms of type of web data used in the mining process.This approach is more suitable and accepted as it is evident from the recent papers that have addressed this issue.

Web mining can be used to discover three general classes of information:-

1. **Web activity**, from server logs and web browser activity tracking.
2. **Web graph**, from link between different pages, people and other data.
3. **Web content**, for the data found on web pages and inside of documents.

II. DATA MINING

Data mining is the process of extracting information from a data set and transforms it into an understandable structure for future use. Data stored in a Data Warehouse is of a wide range that may or may not contain the relevant information that the user desire to use. Searching some information from the wider data space involves analysing of large data which probably may result in degrading of its efficiency. Thus to make the best use of the data of the data warehouse ,a tool called the data mining is introduced so that it can provide the required information to its user from the data pool.Now a days, data mining is used in various fields just because of its efficiency in searching the required information needed by the user in a short period of time and with more accuracy. Data mining has broadened its area of implementation that is why it has become a topic of concern for the researchers.

Data mining comprises of several techniques and methods which help in analysing large data sets for the purpose of discovering and extracting unknown relations and structures from huge heap of details. Data mining does so with the use of algorithms and techniques drawn from the field of machine learning, data base management systems. In large data, data mining is popularly known as knowledge discovery database.

 ISSN: 2321-8169
1010 – 1014

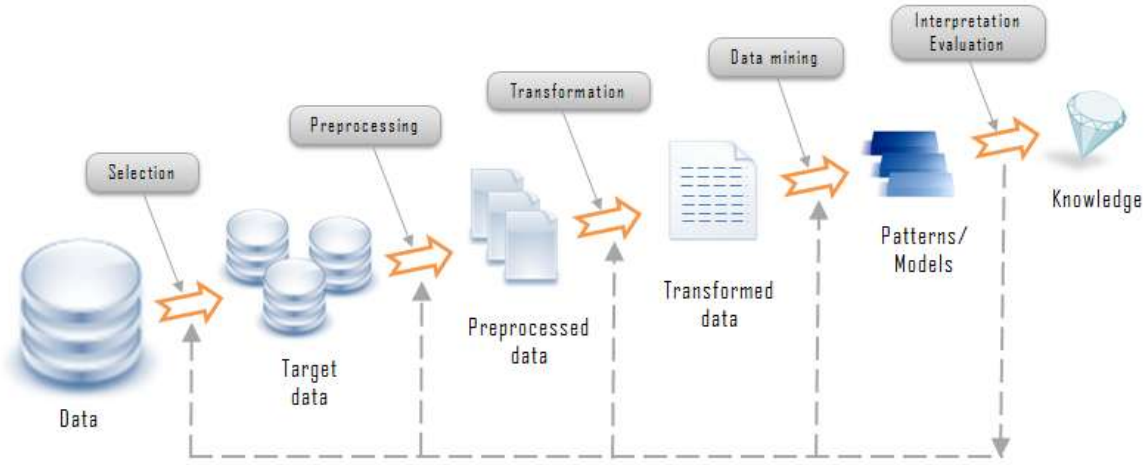

Figure 1:- Data Mining Process In KDD

III. WEB MINING

Web mining can be said as a part of or use of data mining technique for extracting and discovering information from web services and documents. Researcher's interest is growing in this area of research because of tremendous growth of sources of information available on the web and interest in e-commerce.

The whole process of web mining is divided into subtasks, namely:-

- **Resource finding:-** Itinvolves the task of retrieving the required document from the web. It is the same process by which we extract the data either from online or offline sources
- **Information selection and pre processing:** - It involve automatic selection and pre-processing of specific information retrieved from web sources.
- **Generalization:** -It automatically discovers general patterns at individual website as well as multiple sites. Data mining techniques are used in generalization.
- **Analysis:** - Involves validation and Interpretation of mind patterns. Human plays an important role in information or knowledge discovery process on web.

This achievement of web mining is not a one day or rapid process, it has absorbed many years for its evolution and this success. In this paper we will discuss about data mining evolution history starting from its very first introduction.

Years of Evolution	Technology Involved	Product Providers	Need in Business	Characteristics
Data Collection 1960_s	Computers, Disks, Tapes	CDC, IBM	Evaluation of total revenue in last year's.	Delivery of Static data.
Data Access 1980_s	SQL, RDBMS, ODBC	Oracle, IBM, Sybase, Microsoft	To know about unit sales	Delivery of Dynamic Data.
Data Warehousing and Decision Support 1990_s	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
Data Mining 2000_s	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

Table 1: - step by step evolution of web mining

 ISSN: 2321-8169
1010 – 1014

IV. CATAGORIES OF WEB MINING

According to the type of data to be mined, web mining is divided into three main categories. These are: -

1. **Web Content Mining**

Web content mining is the process of mining, extracting and integration of useful knowledge or information from the contents of web document. The information may comprises of text document, audio, video, pictures, graphs or structured records like list, tables. World Wide Web contains massive amount of information, content mining provides list of only those results to the search engine which are highly relevant to the keywords that are in the query. Web content mining is differentiated from two different point of views i.e. information retrieval view and database view.

The concept of web content mining is somehow related but different to data mining and text mining. Relation between web contact and data mining is that many data mining techniques can be applied in the web content mining. But the difference between the two is that in web content mining web data are mainly semi-structured or unstructured wherelse data mining deals with structured data. Relation between web content mining and text mining is only that the web content is text. And the difference between these two is that web content is semi-structured in nature but text mining focuses on unstructured texts.

Various contents of content mining are: -

Web page: - A web page contains a mixture of many kind of information for eg.

advertisements, navigation panels, copyright notice etc.

Search page: - A search page is used for searching a particular web page of a site, which is accessed several times in response to a query. Web content clustering and organization in a database is done for effective navigation of pages by users and search engine itself.

Result page: - A result page contains the result for the query.

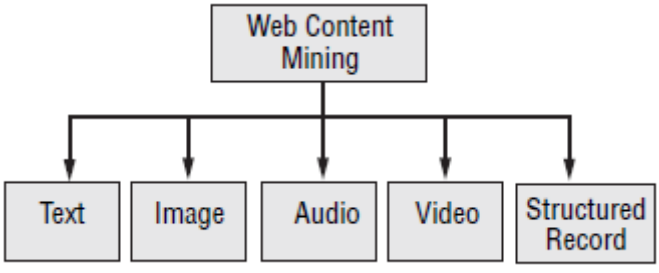

Figure 2: - types of web content mining

2. **Web structure mining**

A web graph structure contains of web nodes and hyperlinks as edges connecting related pages. It isthe process of discovering structure information from the web. There are two types of web structure mining: -

- **Hyperlinks: -** It is a unit that connects one location in a web page to the other location of a similar page or a different web page. Hyperlink connectivity to a different part within the same page is called intra-document hyperlink, and the hyperlink connectivity between two different pages is called inter-document hyperlink.
- **Document structure:** -Web page content can be organised in tree structure, based on HTML and XML tags within page.

Various content of Web structure mining are: -

Links structure mining: - Link structure mining contains Classification, Cluster Analysis, Type, Strength.Internal structure mining: - Provide information of page rank, authoritativeness and enhance search result through filtering.

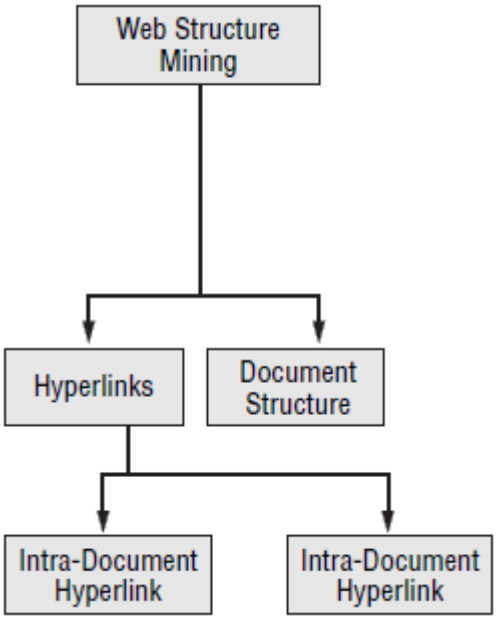

Figure 3: - types of web structure mining

3. Web Usage miningWeb usage mining is one of the data mining techniques for discovering interesting usage Pattern from web usage data to know and better serve the needs of web based applications.The use of usage data is to capture the identity of web users along with their browsingbehaviour to a particular website.

 ISSN: 2321-8169
1010 – 1014

Web usage mining is of three types: -

Web server data: - Web server data include IP address, page reference and access time.

Application server data: -E-commerce uses various features of application server datathat enable it to reach at the top with less efforts.

Application level data: - Much new kind of events can be defined in an application, and logging to them creates histories to these specially defined events.

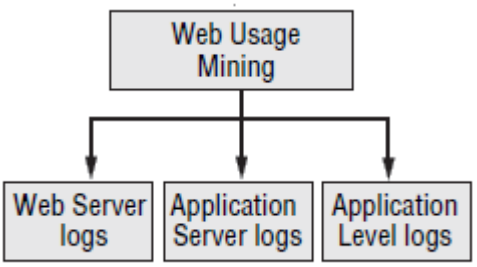

Figure 4: - Types of Web Usage Mining

V. Comparison between web mining and data mining

Comparison	Web Mining	Data Mining
Definition	It is a process of extracting information from the web documents.	It is a process used to extract hidden information from the database.
Scope	It contains 10 million jobs in serverdatabase, and therefore searchprocessing 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 un structured 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.
Accessibility	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 with on-line data.	It works with off-line data.
Storage of data	Data is stored in server logs and web server database.	Data is stored in datawarehouse.
Application Areas	E-learning, Digital Libraries, E-Government, Electronic Commerce, E-Politics, E-Democracy, Security & Crime Investigation, Electronic Business.	Banking, marketing, manufacturing & production, health-care, insurance, law, airlines, computer hardware & software, government & defence, etc.
Techniques	Web Content Mining, Graph Based Web Mining, Utilization in Web Mining, Text Mining and many others.	Artificial Neural Network, Decision Trees, Rule Induction, Nearest Neighbor Method and many others.
Challenges	Complexity of web pages, web is too huge, relevancy of information, web is dynamic information source, diversity of user communicates, etc.	Network settings, data quality, privacy preservation, scalability, complex and heterogeneous data, etc
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.

VI. Conclusion

At the end of this paper we conclude that need to access and manage web data is increasing vary rapidly. Web mining is basically used to ease the efforts of user to access information from World Wide Web and find web mining applications in fields like Clustering, in extraction rules and many other. To extract the specific data from web warehouse, the three categories of web mining plays a major role. When Web mining is compared with data mining we reach to a result that web mining is used to retrieve online data wherelse data mining is used to retrieve offline data. In web mining data is stored in server database but in data mining data is stored in data warehouse.

References

[1] Raymond Khosala and HardikBlockeel "Web Mining Research: A Survey" SIGKDD Exploration, Copyright 2000 ACM SIGKDD, july 2000, volume, issue-1, page 1

[2] R.Munilatha et.al. "A Study on issues and Techniques of Web Mining", international Journal of Computer Science & mobile Computing, volume 3 issue 5, May 2014, page 331-341, ISSN 2320-088X.

[3] Monika yadavet.al. "Web Mining: An Introduction", International Journal of Advanced Research in Computer Science & Software Engineering, volume 3, issue 3, march 2013 ISSN 2277-128X.

[4] Akshayet.al. "A Review on: Web Mining Techniques" International Journal of Engineering Trends and Technology, volume 10, number 3, april 2014.

[5] Dushyant B. Rathodet.al. "A Review on Emerging Trends of Web Mining & It's Applications", ISSN: 2321-9939.

[6] RachitAdhewaryu "A Review Paper on Web Usage Mining & Pattern Discovery", Journal of Information, Knowledge &Research in Computer Engineering, ISSN: 0975- 6760, 2013, volume 2, issue 2 page 279.

[7] Ankit rathiet.al. "Web Usage Mining- A Review" International Journal of Advanced Research in Computer and Communication Engineering, volume 5, issue 2, February 2016 N 2278-1021.

[8] Mohinder Singh et.al. "A Review on Various Web Mining Techniques with Proposed Algorithm of K-means Web Ranking" International Journal of Computer Science and Mobile Computing, volume 2, issue 4, April 2013, page 79-83.

[9] J.H Kroezeet.al. "Differentiating Between Data-Mining and Text-Mining terminology", South African Journal of Information Management, volume 6, issue 4, December 2014.

[10] Simranjeet Kaur et.al. "Web Mining and Data Mining: A Comparitive Approach", International Journal of Noval Research in Computer Science and Software Engineering, volume 2, issue 1, page 36-42, april 2015.