IJARIIT INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

A Survey report for Data Mining based on web research

Gaurav Saini
gauravhpror@gmail.com

Abstract— Web Data Mining is an important area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. It defines the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc. Therefore, the process of extracting useful information from the contents of web documents and the Content data is the collection of facts a web page is designed to contain, it may consist of text, images, audio, video, or structured records such as lists and tables. The data used for web content mining includes both text and graphical data. Content mining is divided into two parts, one is webpage content mining and other is search result mining. Here, it defines the information retrieval and information extraction from web and making research for data mining.

I. INTRODUCTION

Web mining is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. The extraction of hidden predictive information from large databases is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [1]. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.
- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web pages and inside of documents.

At Scale Unlimited we focus on the last one – extracting value from web pages and other documents found on the web. Note that there's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results. A partial list includes:

- Business intelligence
- Competitive intelligence
- Pricing analysis
- Events
- Product data
- Popularity
- Reputation

Web usage mining
It is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

Web usage mining itself can be classified further depending on the kind of usage data considered:

- Web Server Data: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
- Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
- Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above [3].

Web structure mining

It is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.
2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Web content mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web.

Four Steps in Content Web Mining

When extracting Web content information using web mining, there are four typical steps.

1. Collect – fetch the content from the Web
2. Parse – extract usable data from formatted data (HTML, PDF, etc)
3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
4. Produce – turn the results of analysis into something useful (report, search index, etc)

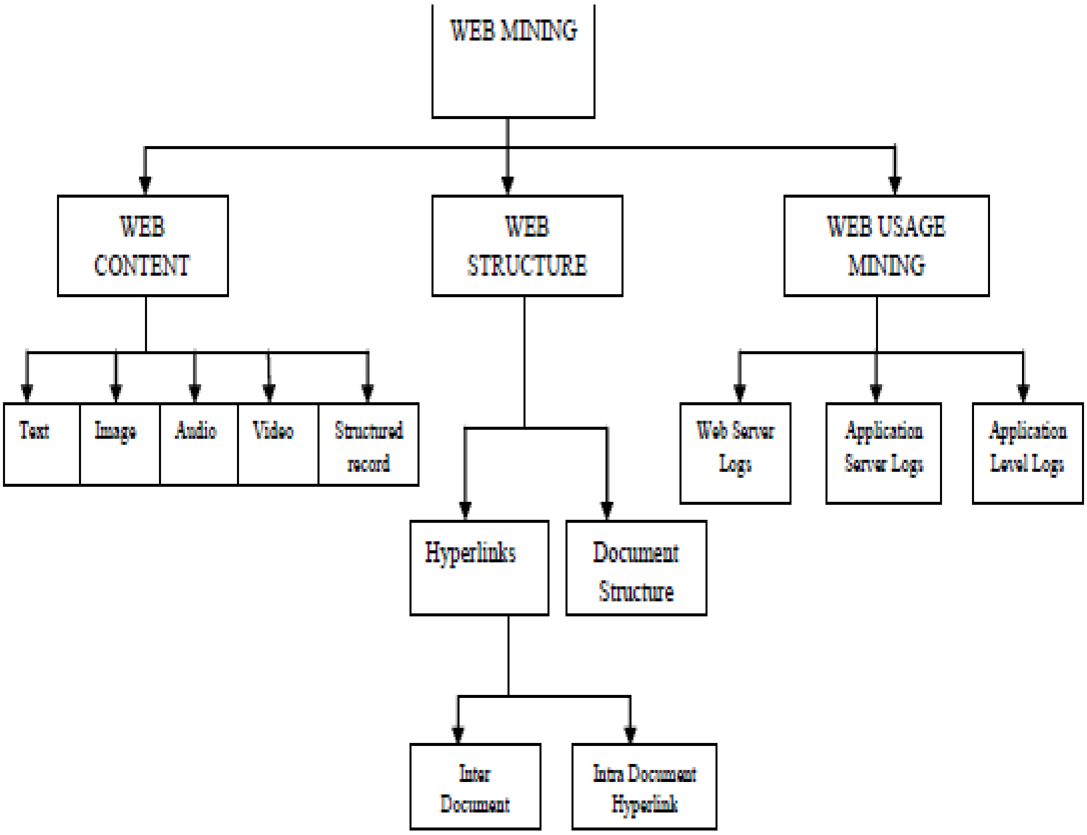

Figure 1: Web mining taxonomy

Web Mining versus Data Mining

When comparing web mining with traditional data mining, there are three main differences to consider:

1. **Scale** – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.
2. **Access** – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated (non-user) access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler (a) promises not to overload the site, and (b) has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful during the crawling process, to avoid causing any problems for the webmaster.
3. **Structure** – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML mark-up.

Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery [7]. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data mining derives its name from the similarities between searching for valuable business information in a large database. For example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides [5]. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.
- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together [2]. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

The most commonly used techniques in data mining are:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.

Many of these technologies have been in use for more than a decade in specialized analysis tools that work with relatively small volumes of data. These capabilities are now evolving to integrate directly with industry-standard data warehouse and OLAP platforms. The appendix to this white paper provides a glossary of data mining terms [8].

II. DATA MINING ARCHITECTURE

Data mining is described as a process of discover or extracting interesting knowledge from large amounts of data stored in multiple data sources such as file systems, databases, data warehouse etc. This knowledge contributes a lot of benefits to business strategies, scientific, medical research, governments and individual.

Business data is collected explosively every minute through business transactions and stored in relational database systems. In order to provide insight about the business processes, data warehouse systems have been built to provide analytical reports that help business users to make decisions [6].

Data is now stored in databases and/or data warehouse systems so should we design a data mining system that decouples or couples with databases and data warehouse systems? This question leads to four possible architectures of a data mining system as follows:

- **No-coupling**: in this architecture, data mining system does not utilize any functionality of a database or data warehouse system. A no-coupling data mining system retrieves data from a particular data sources such as file system, processes data using major data mining algorithms and stores results into file system. The no-coupling data mining architecture does not take any advantages of database or data warehouse that is already very efficient in organizing, storing, accessing and retrieving data. The no-coupling architecture is considered a poor architecture for data mining system however it is used for simple data mining processes.
- **Loose Coupling**: in this architecture, data mining system uses database or data warehouse for data retrieval. In loose coupling data mining architecture, data mining system retrieves data from database or data warehouse, processes data using data mining algorithms and stores the result in those systems. This architecture is mainly for memory-based data mining system that does not require high scalability and high performance.
- **Semi-tight Coupling**: in semi-tight coupling data mining architecture, beside linking to database or data warehouse system, data mining system uses several features of database or data warehouse systems to perform some data mining tasks including sorting, indexing, aggregation etc. In this architecture, some intermediate result can be stored in database or data warehouse system for better performance.
- **Tight Coupling**: in tight coupling data mining architecture, database or data warehouse is treated as an information retrieval component of data mining system using integration. All the features of database or data warehouse are used to perform data mining tasks. This architecture provides system scalability, high performance and integrated information [9].

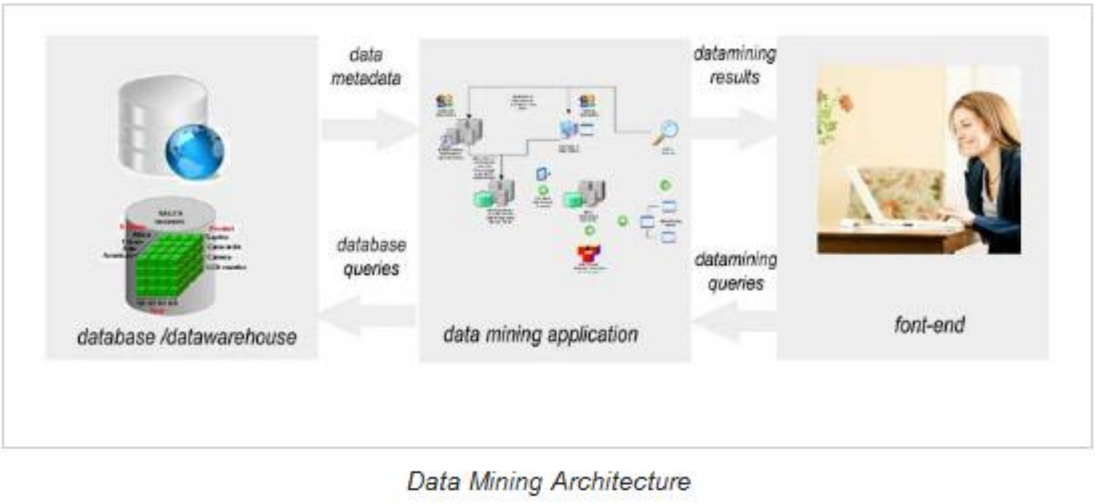

Figure 2: Architecture of data mining

There are three tiers in the tight-coupling data mining architecture:

1. **Data layer**: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end-user in form of reports or other kind of visualization.

2. **Data mining application layer** is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.
3. **Front-end layer** provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

III. WEB MINING ARCHITECTURE

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the Worldwide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the Worldwide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs [4].

- **Web Content Mining**

Web content mining is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine readable semantic, some approaches have suggested restructuring the document content in a representation that could be exploited by machines. The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. Techniques using lexicons for content interpretation are yet to come. There are two groups of web content mining strategies: Those that directly mine the content of documents and those that improve on the content search of other tools like search engines.

- **Web Structure Mining**

World Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. The Page Rank and CLEVER methods take advantage of this information conveyed by the links to find pertinent web pages. By means of counters, higher levels cumulate the number of artifacts subsumed by the concepts they hold. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized.

- **Web Usage Mining**

Web servers' record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analyses can shed light on better structure and grouping of resource providers. Many web analysis tools existed but they are limited and usually unsatisfactory. We have designed a web log data mining tool, Weblog Miner, and proposed techniques for using data mining and On-line Analytical Processing (OLAP) on treated and transformed web access files. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends [8,11].

Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed.

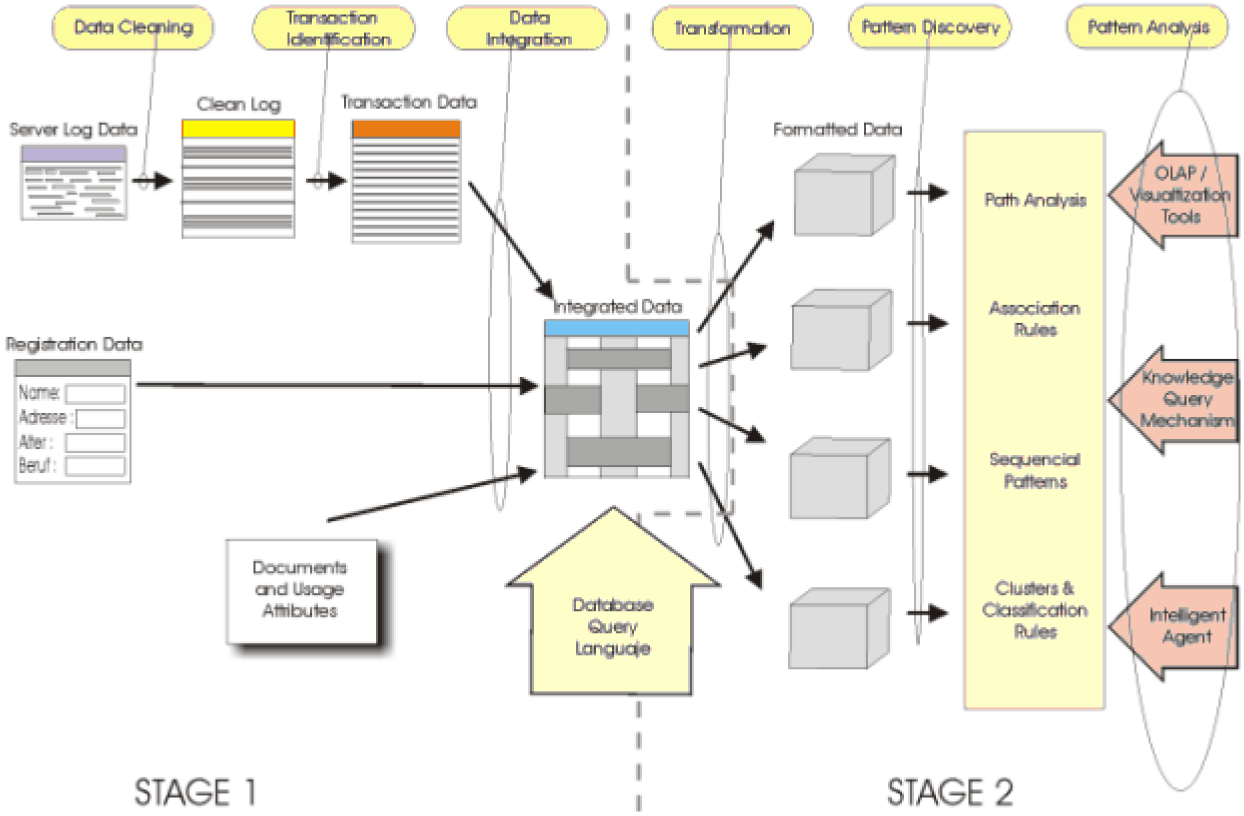

Figure 3: Architecture of web mining

Association Rules

This technique is generally applied to a database of transactions consisting of a set of items. This rule implies some kind of association between the transactions in the database. It is important to discover the associations and correlations between these set of transactions. In the web data set, the transaction consists of the number of URL visits by the client, to the web site. It is very important to define the parameter support, while performing the association rule technique on the transactions. This helps in reducing the unnecessary transactions from the database. Support defines the number of occurrences of user transactions within the transaction log. The discovery of such rules from the access log can be of tremendous help in reorganizing the structure of the web site. The frequently accessed web pages should be organized in their order of importance and be easily accessible to the users.

IV. CONCLUSION

The topic of this paper defining web mining and data researches, may be found, particularly in the commercial area as common data mining applications. However, there is space for research in the public scientific field and especially the functionality of Clear Research, which is literally determined for researchers, might be a source of inspiration. The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. Thus various Data mining techniques and web content mining tools are used to extract useful information or knowledge from web page contents. By these techniques we can make our search of contents over the web faster and exact.

ACKNOWLEDGMENT

Thanks to my Guide and family member who always support, help and guide me during my dissertation. Special thanks to my father who always support my innovative ideas.

REFERENCES

[1] Yong Shi, Yuqing Song and Aidong Zhang. A shrinking-based approach for multidimensional data analysis. In the 29th VLDB conference, September 2003.
[2] M Eirinaki and M Vazirgiannis, “Web Mining for Web Personalization,” ACM Trans. Internet Technology, vol. 3, no.1,2003.

[3] Diligenti M., Gori M., Maggini M.: “A Unified Probabilistic Framework for Web Page Scoring Systems” , IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 1, pp. 4 – 16, 2004.

[4] J.B. Schafer, J. Konstan, and J. Riedl, “Electronic Commerce Recommender Applications,” J. Data Mining and Knowledge Discovery, vol. 5, nos. 1/2, 2000.

[5] M. Spiliopoulou, “Web Usage Mining for Site Evaluation,” Comm. ACM, , vol. 43, no. 8, 2000.

[6] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[7] Kosala R., Blockeel H.:, “Web Mining Research: A Survey”, ACM SIGKDD Explorations Newsletter, Vol. 2, No. 1, pp. 1 – 15, 2000.

[8] Xiaoqing Zheng,Yiling Gu,Yinsheng Li,”Data Extraction from Web Pages Based on Structural Semantic Entropy”, International World Wide Web conference Committee (IW3C2),April 2012.

[9] Ananthi.J, “A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.

[10] Monika Yadav, Mr. Pradeep Mittal, “Web Mining: An Introduction”,IJARCSSE, March – 2013.

[11] Govind Murari Upadhyay, Kanika Dhingra, “Web Content Mining: Its Techniques and Uses”, IJARCSSE, November, 2013.