American Journal of Systems and Software, 2015, Vol. 3, No. 2, 44-61
Available online at http://pubs.sciepub.com/ajss/3/2/3

DOI:10.12691/ajss-3-2-3

World towards Advance Web Mining: A Review

Shyam Nandan Kumar*

M.Tech-Computer Science and Engineering, Lakshmi Narain College of Technology-Indore (RGPV, Bhopal), MP, India
*Corresponding author: shyamnandan.mec@gmail.com

Received March 28, 2015; Revised April 05, 2015; Accepted April 16, 2015

Abstract With the advent of the World Wide Web and the emergence of e-commerce applications and social networks, organizations across the Web generate a large amount of data day-by-day. The abundant unstructured or semi-structured information on the Web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the Web. In this article, an attempt has been made to review the various web mining techniques to discover fruitful patterns from the Web, in detail. New concepts are also included in broad-sense for Optimal Web Mining. This paper also discusses the state of the art and survey on Web Mining that is used in knowledge discovery over the Web.

Keywords: *data mining, www, web mining, cloud mining, web usage mining, web content mining, web structure mining, semantic web mining, web mining algorithm, knowledge discovery, information retrieval*

Cite This Article: Shyam Nandan Kumar, "World towards Advance Web Mining: A Review." *American Journal of Systems and Software*, vol. 3, no. 2 (2015): 44-61. doi: 10.12691/ajss-3-2-3.

1. Introduction

Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – ***Users***, ***Web service providers***, ***Business analysts***. The users want to have the effective search tools to find relevant information easily and precisely. To find the relevant information, users either browse or use the search service when they want to find specific information on the Web. When a user uses search service he or she usually inputs a simple keyword query and the query response in the list of pages ranked based on their similarity to the query. But due to the problems [1] with browser like: Low precision, which is due to the irrelevance of many of search results, and Low recall, which is due to the inability to index all the information available on the Web, users feel difficulty to find the relevant information on the web. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs, like what the customer do and want. Mass customizing the information to the intended user or even to personalize it to individual customer is the big problem. Web mining is expecting tools or techniques to solve the above problems encountered on the Web. Sometimes, web mining techniques provide direct solution to above problems. On the other hand, web mining techniques can be used as a part of bigger applications that addresses the above problems. Other related techniques from different research areas, such as database, information retrieval, and natural language processing, can also be used. Therefore, Web mining becomes a very hot and popular research field.

Web mining combines two of the activated research areas: ***Data Mining*** and ***World Wide Web***. Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. It extracts the hidden predictive information from large database. With the widespread use of data and the explosive growth in their size, organizations are faced with the problem of information overload. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Web mining, when looked upon data mining terms, can be said to have three operations of interests: ***Clustering*** (e.g., finding natural grouping of users, pages, etc.), ***Association*** (e.g., which URLs tend to be requested together), ***Sequential Analysis*** (e.g., the order in which URLs tends to be accessed). As in most real world problems, the clusters and associations in web mining do not have clear-cut boundaries and often overlap considerably. The unstructured feature of Web data triggers more complexity of Web mining. In the present time, it is not easy task to retrieve the desired information because of more and more pages have been indexed by search engines. So, this redundancy of resources has enhanced the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "Web Data mining" [3]. Etzioni [4] came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data

sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become interesting topic.

Web mining is also an important component of content pipeline for web portals. It is used in data confirmation and validity verification, data integrity and building taxonomies, content management, content generation and opinion mining [2]. Web mining - is the application of data mining techniques to discover patterns from the Web. It is also related to text mining because much of the web contents are texts. 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***.

In this paper sections are organized as follows: Section 2 gives the idea about web mining and its types. Section 3 discusses the comparison of web mining with data mining and text mining. Description of ways of web mining techniques is explained in the section 4, 5 and 6. These sections focus on types of web mining approaches in detail. Section 7 describes Semantic Web Mining. Various web mining algorithms are given in section 8. Section 9 outlines the issue and challenges that are associated with web mining. Application areas of web mining are classified in section 10. Section 11 concludes the paper and presents avenues for future work. References for this paper are given in section 12.

2. Web Mining

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity of Web mining. Web mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence, and also psychology and statistics as well.

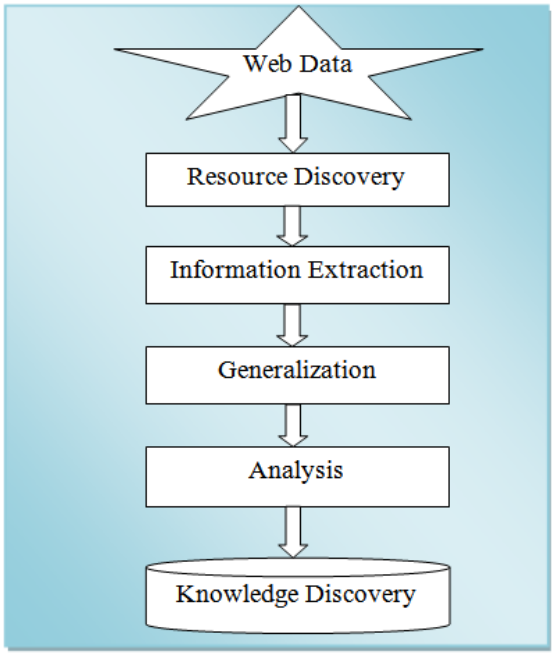

Figure 1. Steps of Web Mining

Web mining involves the analysis of Web server logs of a Web site. The Web server logs contain the entire collection of requests made by a potential or current customer through their browser and responses by the Web server. The information in the logs varies depending on the log file format and option selected on the Web server. Analysis of the Web logs can be insightful for managing the corporate e- business on a short-term basis; the real value of this knowledge is obtained through integration of this resource with other customer touch point information. Common applications include Web site usability, path to purchase, dynamic content marketing, user profiling through behavior analysis and product affinities.

Similar to [4], as shown in Figure 1, decomposition of web mining can be suggested into the following sub-tasks:

- **Resource Discovery**: the task of retrieving the intended information from Web.
- **Information Extraction**: automatically selecting and pre-processing specific information from the retrieved Web resources.
- **Generalization**: automatically discovers general patters at the both individual Web sites and across multiple sites.
- **Analysis**: analyzing the mined pattern.

Based on the main kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining as shown in Figure 2. Summary of Web Mining and its types are given in Table 4.

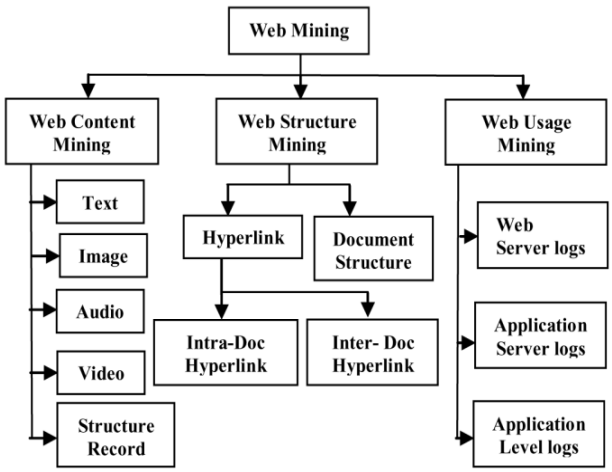

Figure 2. Types of Web Mining

2.1. Cloud Mining

Cloud computing [5,6] has become a viable mainstream solution for data processing, storage and distribution. It promises on demand, scalable, pay-as-you-go compute and storage capacity. To analyze “Big Data” [7] on clouds, it is very important to research data mining strategies based on cloud computing paradigm from both theoretical and practical views. Association rules, Clustering, Anomaly detection, Regression, and Classification are frequently used to mine the data over the cloud by supporting MapReduce [7] parallel computing platform. The implementation of data mining techniques through Cloud computing can allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and

storage. In brief, a MapReduce computation executes as follows:

- Some numbers of Map tasks each are given one or more chunks from a distributed file system. These Map tasks turn the chunk into a sequence of key-value pairs. The way key-value pairs are produced from the input data is determined by the code written by the user for the Map function.
- The key-value pairs from each Map task are collected by a master controller and sorted by key. The keys are divided among all the Reduce tasks, so all key-value pairs with the same key wind up at the same Reduce task.
- The Reduce tasks work on one key at a time, and combine all the values associated with that key in some way. The manner of combination of values is determined by the code written by the user for the Reduce function.

Cloud Mining can be classified as: Service Mining, Deployment Mining, Architecture Mining and Workflow Mining, as shown in Figure 3.

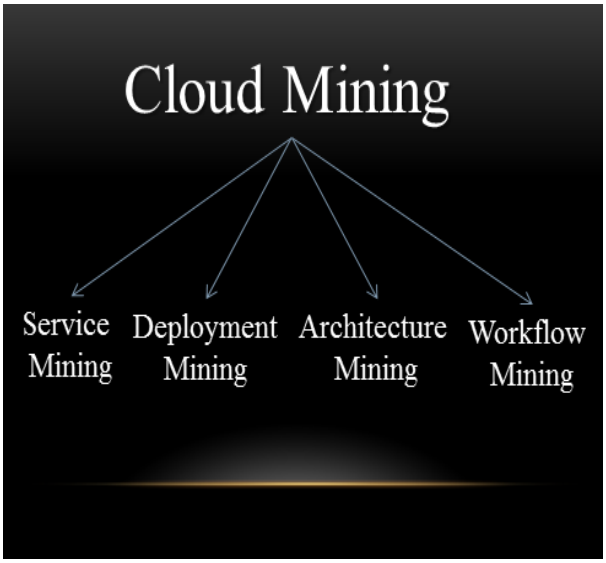

Figure 3. Cloud Mining Types

2.1.1. Service Mining

To quantitatively measure quality of service, several related aspects of the network service are often considered, such as error rates, bandwidth, throughput, transmission delay, availability, jitter, etc. Quality of service is particularly important for the transport of traffic with special requirements. Various cloud clients are interacted with cloud based services. Some known services are: Infrastructure as a service (***IaaS***), Platform as a service (***PaaS***) and Software as a service (***SaaS***). IaaS clouds often offer additional resources such as a virtual-machine disk image library, raw block storage, and file or object storage, firewalls, load balancers, IP addresses, virtual local area networks (VLANs), and software bundles. PaaS includes operating system, programming language execution environment, database, and web server. In the SaaS model, cloud providers install and operate application software in the cloud and cloud users access the software from cloud clients. Within these services there is a need of advance methodology. New Cloud Services can be mined for providing optimal services. Quality management is the most important issue in cloud mining. End-to-end quality of service can require a method of coordinating resource allocation between one autonomous system and another. Quality of service guarantees are important if the cloud capacity is insufficient, especially for real-time streaming multimedia applications such as voice over IP, online games and IP-TV, since these often require fixed bit rate and are delay sensitive, and in networks where the capacity is a limited resource, for example in cellular data communication.

2.1.2. Deployment Mining

In deployment mining, new cloud patterns can be discovered. Discovery of new types of cloud computing mechanism is required for satisfying the customer requirements. Based on cloud type application and services can be mined. In this case, knowledge discovery depends upon cloud types. Cloud platforms provide scalable processing and data storage and access services that can be exploited for implementing high-performance knowledge discovery systems and applications.

2.1.3. Architecture Mining

Cloud architecture typically involves multiple cloud components communication. Methods of efficient organization of these components are required. Cloud Computing Architectures and Cloud Solution Design Patterns can be mined under architecture mining. Cloud computing offers an effective support for addressing both the computational and data storage needs of Big Data mining and parallel analytics applications. In fact, complex data mining tasks involve data- and compute-intensive algorithms that require large storage facilities together with high performance processors to get results in acceptable times.

Cloud engineering brings a systematic approach to the high-level concerns of commercialization, standardization, and governance in conceiving, developing, operating and maintaining cloud computing systems. It is a multidisciplinary method encompassing contributions from diverse areas such as systems, software, web, performance, information, security, platform, risk, and quality engineering.

2.1.4. Workflow Mining

It involves mining the various techniques to minimize the workload over the cloud. Traffic Handling is one of the important issues over the cloud. To provide different priority to different applications, users, or data flows, or to guarantee a certain level of performance to a data flow, workflow mining is required. For example, a required bit rate, delay, jitter, packet dropping probability and/or bit error rate may be guaranteed. Autonomic Business Process and Workflow Management in Clouds should be efficiently managed.

Mining of cloud based framework leads development of distributed data analytics applications as workflows of services.

3. Data Mining vs. Web Mining

Data mining is the process of non-trivial discovery from implied, previously unknown, and potentially useful

information from data in large databases. Hence it is a core element in knowledge discovery, often used synonymously. Data mining involves using techniques to find underlying structure and relationships in large amounts of data. Common data mining applications discover patterns in a structured data such as database (i.e. DBMS). The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages: (1) Selection, (2) Pre-processing, (3) Transformation, (4) Data Mining, and (5) Interpretation/Evaluation.

Web mining describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of web data. The analyzed web resources contain (1) the actual web site (2) the hyperlinks connecting these sites and (3) the path that online users take on the web to reach a particular site. Web usage mining then refers to the derivation of useful knowledge from these data inputs. Web mining discovers patterns in a less structured data such as Internet (***WWW***). In other words, we can say that Web Mining is Data Mining techniques applied to the ***WWW***.

Text mining is different from what we're familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. In text mining, the goal is to discover heretofore unknown information, something that no one yet knows and so could not have yet written down. In text mining the patterns are extracted from natural language text rather than from structured databases of facts. Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

Table 1 shows the comparison between data mining and web mining while Table 2 shows the comparison between text mining and web mining.

Table 1. Data Mining vs. Web Mining

Data Mining	Web Mining
Data mining involves using techniques to find underlying structure and relationships in large amounts of data.	Web mining involves the analysis of Web server logs of a Web site.
Common data mining applications discover patterns in a structured data such as database.	Web mining; likewise discover patterns in a semi-structured data such as Internet (WWW). In other words, we can say that Web Mining is Data Mining techniques applied to the WWW.
It can handle large amount of Data**.**	It can handle big data compare than traditional data mining.
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.
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 markup.

3.1. Text Mining vs. Web Mining

Table 2. Text Mining vs. Web Mining

Text Mining	Web Mining
Sub-domain of Information Retrieval(IR) and Natural Language Processing	Sub-domain of IR and multimedia
Text Data: free-form, unstructured & semi-structured data	Semi-structured data: hyper-links and html tags Multimedia data type: Text, image, audio, video.
Content management & information organization.	Content management/mining as well as usage/traffic mining.
Patterns are extracted from natural language text rather than from structured database.	Patterns are extracted from Web rather than from structured database.

4. Web Usage Mining

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. It tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Usage data captures the identity or origin of Web users along with their browsing behavior at a web site. It deals with studying the data generated by web surfer's sessions or behaviors. Since the web content and structure mining utilize the real or primary data on the web. On the contrary, web usage mining mines the secondary data derived from the interactions of the users with the web. The secondary data includes the data from the proxy server logs, browser logs, web server access logs, user profiles, user sessions, user queries, registration data, bookmark data, mouse clicks and scrolls, cookies and any other data which are the results of these interactions. Log file pros and cons are given in Table 3. High level architecture of different web logs is shown in Figure 4.

Table 3. Log File Pros and Cons

File Type	Advantage	Disadvantage	Mapping
Client Log File	Authentic and Accurate	Modification, Collaboration	One to Many
Server Log File	Reliable and Accurate	Incomplete	Many to One
Proxy Log File	Control Efficiency of Corporate Access to the Internet, Log Traffic	Complex, Unsecure	Many to Many