# A Review on Web Mining Technologies and Applications

Dr. Dushyantsinh B. Rathod
*Associate Professor & HOD, Computer Engineering Dept.,*
*D.A.DegreeEngg& Technology*
*Dushyantsinh.rathod@gmail.com*
*Ahmedabad, Gujarat, India*

***Abstract*—** The World Wide Web is a gigantic, data place for an assortment of uses. Web contains a dynamic and rich assortment of hyperlink data. It permits Web page get to, utilization of data and gives various sources to information mining. The objective of Web mining is to find the example of access and concealed data from gigantic assortments of reports. Right now are exhibiting the different developing web mining strategies that are viably productive in defeating the negative marks of existing innovations and furthermore give the shallow information and correlation about information mining. This paper portrays the past, current and fate of web mining. Web mining endeavors to decide helpful information from optional information got from the connections of the clients with the web. We have likewise portrayed the personalization on web which is utilized to control the data introduced to the clients through the different personalization systems.

*Key words* - ***Web Mining; Web Content Mining; Web Structure Mining; Web Usage Mining.***

## 1. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from Web data - including Web documents, hyperlinks between documents, usage logs of web sites, etc. Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks. Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process . The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue. In this paper we follow the data-centric view, and refine the definition of Web mining as, **Web mining** is the application of data mining techniques to extract knowledge from Web data, where **at least one of structure (hyperlink) or usage (Web log) data is used in the mining process** (with or without other types of Web data)[1].

The attention paid to Web mining, in research, software industry, and Web-based organizations, has led to the accumulation of a lot of experiences. It is our attempt in this paper to capture them in a systematic manner, and identify directions for future research.[2]

## 2. WEBMINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks [4]:

1. *Resource finding:* It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available onweb.
2. *Information selection and pre-processing:* It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in trainingcorpus.
3. *Generalization:* It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used ingeneralization
4. *Analysis:* It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process onweb[3].

## 3. WEB MININGCLASSIFICATION

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined:

### *A. Web ContentMining*

Web content mining is the process of extracting useful information from the contents of web documents. 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. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has beenlimited.

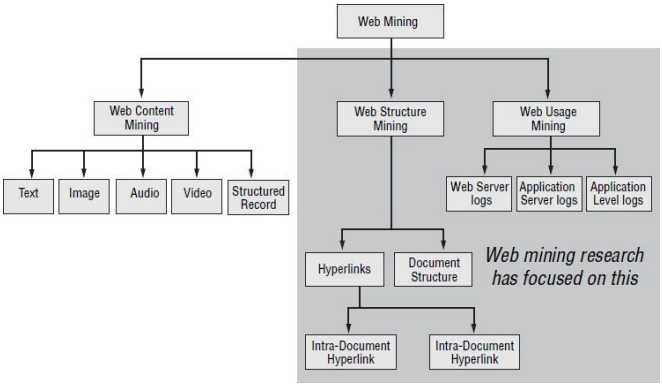


Fig. 1 . Web Mining Classification

Web content mining is related but different from data mining and text mining. It is related to data mining because many data mining techniques can be applied in Web content mining. It is related to text mining because much of the web contents are texts. However, it is also quite different from data mining because Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data. Web content mining is also different from text mining because of the semi-structure nature of the Web, while text mining focuses on unstructured texts. Web content mining thus requires creative applications of data mining and/or text mining techniques and also its own unique approaches.

The various contents of Web Content Mining are
- Web page
- Searchpage
- Resultpage

**Web Page:** A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only some part of the information is useful and the rest are noises.
**Search Page:** A search page is typically used to search a particular Web page of the site, to be accessed numerous times in relevance to search queries. The clustering and organization of Web content in a content database enables effective navigation of the pages by the customer and search engines.
**Result page:** A result page typically contains the results, the web pages visited and the definition of last accurate result in the result pages of content mining.

### *B. Web Structure Mining*

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure informationused.

***Hyperlinks***

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan, Srivastava, Kumar, and Tan (2002) provide an up-to-date survey.

***Document Structure***

In addition, the content within a Web page can also be organized in a treestructured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Liu 1998; Moh, Lim, and Ng 2000).

It derives information and knowledge mainly from the Web and the links between the organizational structures. Based on scientific citation analysis theory, the interconnection between the data in the document contains a wealth of useful information. The usual search engines consider only the Web as a flat collection of documents because of taking into account the complexity of the structure, ignoring the structure of information. Mining of structure and Web page structure, guides the classification and clustering of pages to find authoritative pages, center pages, to improve retrieval performance. Web page also can be used to guide the collection work to improve collectionefficiency.

The various contents of Web structure mining are
- Links StructureMining
- Internal StructureMining
- URLMining

**Links Structure:** Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts have resulted in a newly emerging research area called Link Mining. It consists Link-based Classification, Link-based Cluster Analysis, Link Type, Link Strength and Link Cardinality.
**Internal Structure Mining:** It can provide information about page ranking or authoritativeness and enhance search results through filtering i.e., tries to discover the model underlying the link structures of the web. This model is used to analyze the similarity and relationship between different web sites.
**URL Mining:** It gives a hyperlink which is a structural unit that connects a web page to different location, either within the same web page (intra_document hyperlink) or to a different web page (inter_document) hyperlink.

### *C. Web Usage Mining*

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). 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*
User logs are collected by the web server and typically include IP address, page reference and access time.

*Application Server Data*
Commercial application servers such as Weblogic,[1,2]StoryServer,[3] 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 serverlogs.

*Application LevelData*
New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the above thecategories.

### *D. TextMining*

Due to the continuous growth of the volumes of text data, automatic extraction of implicit previously unknown and potentially useful information becomes more necessary to properly utilize this vast source of knowledge. Text mining, therefore, corresponds to extension of the data mining approach to textual data and its concerned with various tasks, such as extraction of information implicitly contained in collection of documents or similarity- based structuring. Text collection in general, lacks the imposed structure of a traditional database. The text expresses the vast range of information, but encodes the information in a form that is difficult to decipher automatically [2].

It focuses on techniques that could predict user behavior while the user interacts with the web and also it discovers the meaningful pattern from data generated by client server transaction on one or more web localities. A web is a collection of interrelated files on one or more web servers. Web usage mining aims at utilization of data mining techniques to discover the usage patterns from web based application. It automatically generates the data for the server access logs, refers logs, agent logs, client sides cookies, user profiles, metadata, page attributes, page contents & site structures. It is a technique to predict user behavior when the user interacts with the web. Web usage mining is categorized into three phases:

- Preprocessing
- PatternDiscovery
- PatternAnalysis

**Preprocessing :**According to the client, server and proxy server, the preprocessing is the first approach to retrieve the raw data from web resources and process the data. It automatically transforms the original raw data to the next process.
**Pattern Discovery :**According to the data preprocessing, the raw data is used to discover the knowledge and to implement the techniques which will be used for machine learning. This makes use of data mining procedures.
**Pattern Analysis :**It is the process after pattern discovery. It checks whether the pattern is correct on the web and guides the process of extraction of the information/ knowledge from the web.

TABLE: 1 Web Mining Categories

| | Web Mining | | | |
|---|---|---|---|---|
| | Web Content Mining | | Web Structure Mining | Web Usage Mining |
| | IR view | DB View | | |
| View of Data | -Unstructured<br>-Structured | -Semi Structured<br>-Web Site as DB | -Link Structure | -Interactivity |
| Main Data | - Text documents<br>-Hypertext documents | -Hypertext documents | -Link Structure | -Server Logs<br>-Browser Logs |
| Representation | -Bag of words, n-gram Terms, -phrases, Concepts or ontology<br>-Relational | -Edge labeled Graph,<br>-Relational | -Graph | -Relational Table<br>-Graph |
| Method | -Machine Learning<br>-Statistical (including NLP) | -Proprietary algorithms<br>-Association rules | -Proprietary algorithms | -Machine Learning<br>-Statistical<br>-Association rules |
| Application Categories | -Categorization<br>-Clustering<br>-Finding extract rules<br>-Finding patterns in text | -Finding frequent sub structures<br>-Web site schema discovery | -Categorization<br>-Clustering | -Site Construction<br>-adaptation and management<br>-Marketing.<br>-User Modeling |

## 4. WEB MINING APPLICATIONS

There are many benefits that can be obtained through the applications of web mining technology.

**Application areas of Web Mining**

• E-Commerce
• SearchEngines
• Personalization
• WebsiteDesign

**How Web mining is different from classical data mining?**

***Web mining***

- The web is a collection of interrelated files, even though it is not arelation.
- Web mining is the discovery of knowledge from theweb.
- Usage data is huge and growing rapidly.
- Ability to react in real-time usage ofpatterns.

***Data mining***

- Textual information and linkagestructure.
- Google's usage logs are bigger than their webcrawl.
- Data generated per day is comparable to largest conventional datawarehouse.
- No human in theloop.

The attention paid to Web mining, in research, software industry and Web-based organizations, has led to the accumulation of lot of experiences. It is our attempt in this paper to capture them in a systematic manner and identify the directions for future research.

The various fields where web mining is applied are:

- E-Commerce
- Information filtering
- Fraud detection
- Education andresearch

***E-Commerce***

In e-commerce, web mining helps in generating user profiles by customizing the choice of users. For example, web mining enables a user to search for an advertisement and information regarding a product of his interest. Internet advertising is one of the major fields in e-commerce, where web mining is widely used. Advertising in a specific domain of an e-commerce web site or a general web site and is considered as one of the major application area of web mining.

***Information filtering***

Information filtering is the method to identify the most important results from a list of discovered frequent set of data items for which you can make use of web mining.

***Fraud detection***

Fraud detection can be performed using web mining by maintaining a list of signatures of all the users. Web mining is also applied for plagiarism detection and research works.

## 5. CONCLUSION AND FUTUREWORK

Web data is growing at a significant rate. Web mining is a fertile area of research with many successful applications. As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract useful knowledge from it. To extract the specific data from web warehouse, the three categories (Web Content Mining, Web Structure Mining and Web Usage Mining) of web mining play a major role. Web mining is one of the most important applications of data mining. It is having its own benefits and successful applications with which we can overcome the problems or difficulties faced in data mining. In this paper a clear picture of how web mining is efficiently different from data mining has been highlighted. As the usage of the internet in the present day is growing in faster rate, the personalization process of the web mining provides us a great opportunity of maximizing the efficient usage of theinternet.

Cloud mining is a new approach to apply data mining to the customer data by using web mining process. By a cloud, we mean an infrastructure that provides resources and/or services over the Internet. In fact among all the potential use of web mining in future, the growing online shopping activities, e-services industry and e-commerce are important domains. Hence the future of the web mining can be seen in the field of Cloud Mining architectures and algorithms that can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of

intelligent Web applications.

## REFERENCES


[1] F. Masseglia, P. Poncelet, and R. Cicchetti, "An Efficient Algorithm for Web Usage Mining", Networking and Information Systems Journal (NIS), vol.2, no. 5-6, pp. 571-603,1999.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, vol. I, no. 2, pp. 12-23,2000.

[3] Raymond Kosala, HendrikBlockeel, Web Mining Research: A Survey, SIGKDD Explorations, Copyright 2000 ACM SIGKDD, July2000.

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

[5] YingZhang,–ThestudyofwebdataminginEB,‖theexcellentcollectionofmagisterialanddoctoralthesis,May 2007.

[6] 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.

[7] KavitaSharma,GulshanShrivastava,VikasKumar,–WebMining:TodayandTomorrow‖.

*[8]* B.N.Laxmi,G.HRaghunandhan,–AConceptualOverviewofDataMining‖,ProceedingsoftheNationalConferenceon Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.17 & 18 February,2011.pp.27-*32.*

[9] JaideepSrivastava,PrasannaDesikan,VipinKumar,–WebMining-Concepts,Applications&ResearchDirectionsand Web Mining‖ – Accomplishments & FutureDirections

[10] A.JebarajRatnakumar,–AnImplementationOfWebPersonalizationUsingWebMiningTechniques‖,Journalof Theoretical and Applied Information Technology © 2005 - 2010 JATIT. All rightsreserved.

[11] ManojPandia,SubhenduKumarPani,SanjayKumarPadhi,LingarajPanigrahy,R.Ramakrishna,–AReviewOfTrends In Research On Web Mining‖, A Review Of Trends In Research On Web Minin, International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1,2011.

[12] Tan,Steinbach,Kumar,–IntroductiontoDataMining‖

[13] Introduction to Data Mining and Knowledge Discovery, Third Edition by Two CrowsCorporation.

[14] Michael J. A. Berry, Gordon S. Linoff, Mining the Web: Transforming Customer Data, J. Wiley,2002.

[15] Eirinaki, M., Vazirgiannis, M. (2003) "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol.3, No.1, February2003.