Web Mining: Knowledge Discovery on the Web

Anjali B. Raut,
Department of Computer Science & Engg.,
HVPM's COET, PRMITR, Badnera
Amravati, India

G. R. Bamnote, PhD.
Department of Computer Science & Engg.
HVPM's COET, PRMITR, Badnera
Amravati, India

ABSTRACT

Web mining is the use of data mining techniques to automatically discover and extract information from web documents This paper summarizes the different types of web mining, and their current states of the art.

Keywords—Web Mining, World Wide Web, Web Content Mining, Web Structure Mining, Web Usage Mining.

1. INTRODUCTION

Over the last decade there is tremendous growth of information on World Wide Web (WWW). It has become a major source of information. Web creates the new challenges of information retrieval as the amount of information on the web and number of users using web growing rapidly. It is practically impossible to search through this extremely large database for the information needed by user. Hence the need for Search Engine arises. Search Engines uses crawlers to gather information and stores it in database maintained at search engine side. For a given user's query the search engine searches in the local database and very quickly displays the results.

2. RELATED WORK

At present most users commonly use searching engines to find their favorite information. Each searching engines having its own characteristics and employing different algorithms to index, rank, and present web documents. But because all these search engines are build based on exact key words matching and it's query language belongs to some artificial kind, with restricted syntax and vocabulary other than natural language, there are some limitations that searching engines cannot overcome such as narrowly searching scope. Finding the relevant information on www is not an easy task. The user can encounter the following problems when interacting with the web[6].

1. Low precision: Today's search tools have the low precision problem, which is due to the irrelevance of many search results. This results in a difficulty finding the relevant information.
2. Low recall: It is due to the inability to index all the information available on the web. This results in a difficulty finding the unindexed information that is relevant.
3. Difficulty in discovering new knowledge out of the information available on the web: As the web is huge, diverse and dynamic and thus it raises the scalability, multimedia data, and temporal issues respectively.
4. Personalization of the information: This problem is often associated with the type & presentation of information , since it is likely that people differ in the contents & presentations they prefer while interacting with the web.
5. Learning about users: It is problem related to knowing what the individual user's interests is.
6. Unable to Search Multimedia Data

As the web is huge, diverse, and dynamic, and raises the issues of scalability, heterogeneity, and dynamism, respectively. Due to these characteristics, we are currently drowning in information, but starving for knowledge; there by making the web a fertile area of data mining research with the huge amount of information available online. Today data mining has emerged as a new discipline in world of increasingly massive datasets. Data Mining is the process of extracting or mining knowledge from data. Data Mining is becoming an increasingly important tool to transform data into information. Knowledge Discovery from Data i.e. KDD is synonym for Data Mining.

2.1 *Web Mining*

World Wide Web is a major source of information and it creates new challenges of information retrieval as the amount of information on the web increasing exponentially. Web Mining is use of Data Mining techniques to automatically discover and extract information from web documents and services [1].

Oren Etzioni was the person who coined the term Web Mining first time. Initially two different approaches were taken for defining Web Mining. First was a "process-centric view", which defined Web Mining as a sequence of different processes [1] whereas, second was a "data-centric view" , which defined Web Mining in terms of the type of data that was being used in the mining process [2]. The second definition has become more acceptable, as is evident from the approach adopted in most research papers[3][5]. Web Mining is also a cross point of database, information retrieval and artificial intelligence [4].

2.2 *Web Mining Process*

Web mining may be decomposed into the following subtasks:

1. Resource Discovery: process of retrieving the web resources.
2. Information Pre-processing : is the transform process of the result of resource discovery
3. Information Extraction: automatically extracting specific information from newly discovered Web resources.
4. Generalization: uncovering general patterns at individual Web sites and across multiple sites[3].

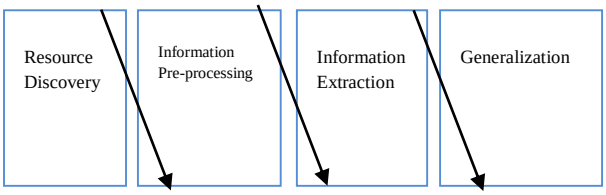

Fig 2.1: Web mining Process

2.3 Web Mining Taxonomy

Web has different facets that yield different approaches for the mining process:

1) Web pages consist of text.
2) Web pages are linked via hyperlinks
3) User activity can be monitored via Web server logs.

This three facets leads to the distinction into three categories i.e. Web content mining, Web structure mining and Web usage mining [4-7].

Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, etc. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main applications of web usage mining is to learn user profiles. The Taxonomy of web mining is as Figure 2.2 and Table 2.1[6].

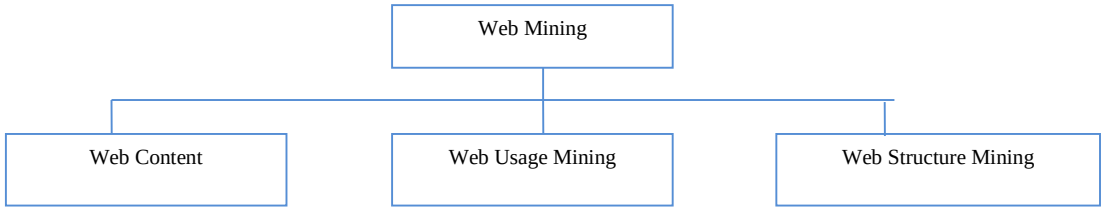

Fig 2.2: Web mining Taxonomy

TABLE 2.1 WEB MINING CATEGORIES

	Web Mining			
	Web Content Mining		**Web Structure Mining**	**Web usage Mining**
	IR View	**DB View**		
View of Data	Unstructured Semistructured	Semistructured Web Site	Link Structure	Interactivity
Main Data	Text Documents Hypertext Documents	Hypertext Documents	Graph	Server logs Browser logs
Representation	Bag of Words, n-grams Terms, phrases Concepts or ontology Relational	Edge labeled graph Relational	Link Structure	Relational table Graph
Method	TFIDF Machine Learning Statistical	Proprietary algorithms ILP Association Rules	Proprietary algorithms	Machine Learning Statistical Association Rules
Application Categories	Categorization Clustering Finding extraction rules Finding patterns in text User Modeling	Finding frequent substructures Web site schema discovery	Categorization Clustering	Site construction Adaption and management Marketing User modeling

2.4 Web Content Mining (WCM)

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 such as lists and tables. Application of text mining to web content has been the most widely researched.

Traditional search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information. The web content mining is differentiated from two different points of view: Information Retrieval View and Database View. [6] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the

web, the mining always tries to infer the structure of the web site to transform a web site to become a database. There are several ways to represent documents.

Vector Space Model:

In Web Text Mining clustering is of the data mining tool used for grouping web documents into clusters such that the objects from the same cluster are similar and objects from different cluster are dissimilar .Vector space model is the most Widely used model used to represent document. In the statistically based vector-space model, a document is conceptually represented by a vector of keywords extracted from the document, with associated weights representing the importance of the keywords in the document and within the whole document collection; likewise, a query is modelled as a list of keywords with associated weights representing the importance of the keywords in the query. The weight of a term in a document vector can be determined in many ways. A common approach uses the so called *tf * idf* method, in which the weight of a term is determined by two factors: how often the term *j* occurs in the document *i* (the term frequency $tf_{i,j}$) and how often it occurs in the whole document collection (the document frequency df_j). Precisely, the weight of a term *j* in document *i* is

$$w_{i,j} = tf_{i,j} * idf_j = tf_{i,j} * \log N/df_j$$

Where *N* is the number of documents in the document collection and *idf* stands for the inverse document frequency. This method assigns high weights to terms that appear frequently in a small number of documents in the document set. Once the term weights are determined, we need to measure similarity between the document vectors.

A common similarity measure, known as the *cosine measure*, determines the angle between the document vectors and when they are represented in a *V*-dimensional Euclidean space, where *V* is the vocabulary size. The similarity between a document D_i and D_q is defined as

$$sim(D_q, D_i) = \frac{\sum_{j=1}^{V} w_{q,j} * w_{i,j}}{\sqrt{\sum_{j=1}^{V} w_{q,j}^2 * \sum_{j=1}^{V} w_{i,j}^2}}$$

Where $w_{q,j}$ is the weight of term *j* in the query, and is defined in a similar way as $w_{i,j}$. The denominator in this equation, called the normalization factor, discards the effect of document lengths on document scores. Thus, a document containing {*x, y, z*} will have exactly the same score as another document containing {*x, x, y, y, z, z*} because these two document vectors have the same unit vector. Because the exact vector-space model is expensive to implement, Here we have given some family of successively simpler approximations.

2.5 Web Structure Mining (WSM):

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.

Link structures enable web pages express more information than normal documents. The link numbers point to a page suggest the popularity of the page, while links point out a page hint the topics of the page or content richness. A page frequently cited may be an important page. Thus we can mine the web through link structures. The examples are Page Rank and CLEVER. HITS (Hyperlink Induced Topic Search) is the foundational algorithm for web structure mining. HITS is divided into sub steps: constructing a sub-graph of WWW and computing hubs and authorities.

2.6 Web Usage Mining(WUM):

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data. Web usage data includes data from web server logs, browser logs, user profiles, registration data, cookies etc.

Web usage mining is the type of web mining activity that involves the automatic discovery of user access patterns from one or more web servers. As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by web servers and collected in server access logs. Other sources of user information include *referrer logs*, which contain information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts. Analyzing such data can help these organizations to determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. Analysis of server access logs and user registration data can also provide valuable information on how to structure a web site in order to create a more effective presence for the organization. Using intranet technologies in organizations, such analysis can shed light on more effective management of workgroup communication and organizational infrastructure. Finally, for organizations that sell advertising on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users.

WCM and WSM uses real or primary data on the web whereas WUM mines the secondary data derived from the interaction of the users while interacting with the web.

2.7 Data Preprocessing

It is necessary to perform a preprocessing on data to convert the raw data for further processing in data mining. It has separate subsections as follows.

Content Preprocessing: Content preprocessing is the process of converting text, image, scripts and other contents into the forms that can be used by the usage mining. It has different processes like stemming and removing non relevant words like a, an, the etc from web content.

Structure Preprocessing: The structure of a web site is formed by the hyperlinks between page views. The structure preprocessing can be treated similar as the content preprocessing. However, each server session may have to construct a different site structure than others.

Usage Preprocessing: The inputs of the preprocessing phase may include the web server logs, referral logs, registration files, index server logs, and optionally usage statistics from a previous analysis. The outputs are the user session file, transaction file, site topology, and page classifications.

2.8 Pattern Discovery

This is the key component of the web mining. Pattern discovery covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. It has separate subsections as follows.

Statistical Analysis: Statistical analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions

Association Rules: In the web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions.
Clustering: Clustering analysis is a technique to group together users or data items with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies.
Classification: Classification is the technique to map a data item into one of several predefined classes. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbour classifier, Support Vector Machines etc.
Sequential Pattern: This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis.
Dependency Modelling: The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the web domain. The modelling technique provides a theoretical framework for analyzing the behaviour of users, and is potentially useful for predicting future web resource consumption.

Pattern Analysis

Pattern Analysis is a last stage in web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations. All these methods assume the output of the previous phase has been structured.

3. MAJOR APPLICATIONS

In the past few years the web applications has been developed at a much faster rate in the industry than research in web related technologies. Many of these are based on the use of web mining concepts. Some of the most successful applications described here.

Web Search

Google is one of the most popular and widely used search engines. It provides users access to information from over 10 billion web pages that it has indexed on its server. The quality and quickness of the search facility makes it the most successful search engine. Earlier search engines concentrated on web content alone to return the relevant pages to a query. Google was the first to introduce the importance of the link structure in mining information from the web. PageRank, which measures the importance of a page, is the underlying technology in all Google search products, and uses structural information of the web graph to return high quality results [1].

Understanding Users Behavior

Knowledge gained from web mining is the key intelligence behind Amazon online book stores features such as he instant recommendation, purchase circles, wish list etc .Amazon gain the knowledge by observing users browsing pattern and interest.

Understanding Auction Behavior

As individuals in a society where we have many more things than we need, the allure of exchanging our useless stuff for some cash, no matter how small, is quite powerful. This is evident from the success of flea markets, garage sales and estate sales. The genius of eBay's founders was to create an infrastructure that gave this urge a global reach, with the convenience of doing it from one's home PC. In addition, it popularized auctions as a product selling and buying mechanism and provides the thrill of gambling without the trouble of having to go to Las Vegas. All of this has made eBay as one of the most successful businesses of the internet era. Unfortunately, the anonymity of the web has also created a significant problem for eBay auctions, as it is impossible to distinguish real bids from fake ones. eBay is now using web mining techniques to analyze bidding behavior to determine if a bid is fraudulent. Recent efforts are geared towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

Understanding Web Communities

One of the biggest successes of America Online(AOL) has been its sizable and loyal customer base. A large portion of this customer base participates in various AOL communities, which are the collection of users with similar interests. In addition to provide providing forum for each communities to interact among themselves, AOL provides them useful information and services. Applying web mining to the data collected from community interactions provide AOL with a very good understanding of its communities, which it has used for targeted marketing through advertisement and e-mail solicitation.

Personalized Portal for the Web

The concept of personalized portal was firstly introduced by Yahoo .In this a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. This becomes a an extremely popular concept and has led to creation of other personalized portals.

Digital Library and Autonomous Citation Indexing

CiteSeer is one of the most popular online bibliographic indices related to computer science. The key contribution of CiteSeer repository is its autonomous citation indexing. By using data mining techniques indexing makes it possible to extract information about related articles. Automating such a process reduce a lot of human efforts and makes it more effective and faster.

4. CONCLUSION

Web Mining is a new and promising research issue to help users in extracting useful information from web. In this paper we present a preliminary discussion about Web Mining including its definition, Taxonomy, and applications.

5. REFERENCES

[1] Jaideep Shrivastava, Prasanna Desikan, Vipin Kumar. Web mining–Concepts, Application, Research Direction.

[2] O. etzioni. The world wield web: Quagmire or Gold Mining. Communicate of the ACM, (39)11:65-68, 1996;

[3] Http://news.netcraft.com

[4] WangBin, LiuZhijing. Web mining Research. In Proceeding of the fifth International conference on computational intelligence and multimedia application, 2003.

[5] Yan Li, Xin-Zhong Chen , Bing-Ru Yang .Research on Web mining based intelligent search engine. In Proceeding of the fifth International conference on machine lerning and cybermetics, 2002.

[6] Raymond Kosala, Hendrik Blockeel, Web Mining Research: A Survey, SIGKDD Expirations, ACM SIGKDD, July 2000.

[7] U. Fayyad, G.Piatetsky-Shapiro, P.Smyth. from Data mining to knowledge discovery in databases.AI Magazine, pages 37-54,1996.

[8] Andreas Hotho, Gerd Stumme. Mining the World Wide Web- Methods, Application and Perspecivites.

[9] Bettina Berendt, Andreas Hotho, Gred Stumme. Towards Semantic Web Mining. In First International semantic web conference, 2002.

[10] Bettina Berendt, Andreas Hotho, Gred Stumme. Semantic Web Mining and the Representation. Analysis and Evolution of Web Space

[11] Gred Stumme, Andreas Hotho, Bettina Berendt. Semantic Web Mining State of tne art and future directions, 2006

[12] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the World Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997