Web Data Mining: Extracting From 'Big Data'

Abdul Karim Siddiqui
Asst. Professor, Dept. of Computer Applications, Awadhoot Bhagwan Ram P.G. College Anpara- Sonebhadra (U.P.)

Abstract:

Being an interdisciplinary subject of study Data Mining has become new and curious subject among researchers. As our capabilities of both generating and collecting data have been increasing rapidly, it has become dynamic and fast expanding field with great strength. The thirsts of required data include the computerization of business, scientific, and government transactions; informative data search on different topics, digital images, e- purchasing of online products etc. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge.

Introduction

Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. This large amount of Data is sometimes called as 'Big Data'.

Web Data Mining

The term Web Data Mining is a technique used to crawl through various web contents to collect précised contents i terms of information. This technique enables us to promote our business, understanding marketing dynamics, new promotions supports on the Internet. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. 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. Web Mining produces patterns from the Web. We may say it as screen scraping, web scraping or data extraction using software and tools to extract data from sources that are not formatted to be used as automated data sources. Most of the our searches come from web pages, often in the form of HTML that is formatted to read, rather than a computer. This presents obstacles for the automated system, such as embedded images, multimedia, or formatting elements that are not a part of the desired text (which is to be analyzed).

The Big Data Extraction

Web data mining has grown out of the large mass of free data available on the websites. Prior to data mining becoming a stand-alone task, business analysts and statisticians extracted and analyzed datasets. However, the large mass and technical behaviour of data necessitated the development of data mining tools designed specifically for web data mining process. WWW provides us with huge amount of required data digitally available as hypertext Data. It may be webpages, images, information and other type. This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information.

The economic importance of web will enhance the academic interest. The Database Administrator, Management persons or others wishing to perform data mining on large number of web pages will require the services of web crawler or its based tools. For these reasons crawlers are normally multi threaded by which millions of WebPages may be extracted parallel by only one process. So Web Crawler for automatic Data and Web Mining is Useful to Us.

Web Crawler

Search engines move to see what are there on Web pages. These tasks are performed by a piece of software, called a *crawler* or a *spider* (or Googlebot). Spiders follow links from one page to another and index everything they find on their way. It is impossible for a spider to visit a site daily just to see if a new page has appeared or if an existing page has been modified in more than 20 billion pages over Web space. Sometimes crawlers may not end up visiting our site for a month or two.

Data Warehousing

Data Warehousing involves data definition, data analysis and data retrieval. It may be the place where required data should be kept. A data warehouse is a repository of information collected from diverse sources. Theses informative data are stored under a unified schema, normally residing at a single site. Data warehousing business have existed in one form or another since the invention of the computer though the data warehouse idea hadn't been fully formed. It is constituted of process of Data cleaning, Data integration, Data transformation, Data loading and periodic Data refreshing. In simple way a Data warehouse integrates dta originating from multiple sources and various timeframes. It may be modelled as multidimensional data structure, called as 'Data cube'. A Data cube provides a multidimensional view of data and allows the pre-computation and fast access of summarized data.

Applications Of Web Mining Technique

According to analysis targets, web mining can be divided into three different types – i. Web content mining , ii. Web usage mining, iii. Web structure mining.

I. Web Content Mining

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. It is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. This scanning is completed after the clustering of web pages through structure mining and provides the results based upon the level of relevance to the suggested query. With the large amount of information that is available on the WWW, it provides the results lists

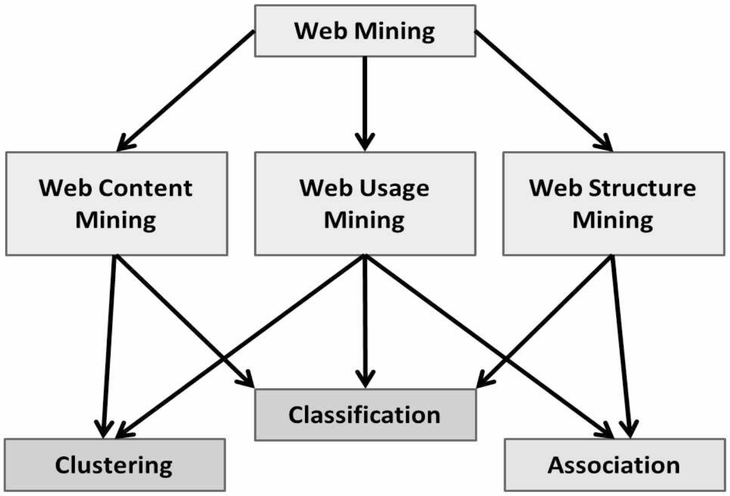

Fig. 1 Web Mining

to search engines in order of highest relevance to the keywords in the query. Web Content Mining also termed as Text mining is directed toward specific information provided by the person search information in search engines. The results are pages relayed to the search engines through the highest level of relevance to the lowest. Though, the search engines have the ability to provide links to Web pages by the thousands in relation to the search content, this type of web mining enables the reduction of irrelevant information.

Web text mining is very effective when used in relation to a content database dealing with specific topics. For example online universities use a library system to recall articles related to their general areas of study.

II. Web Usage Mining

Web usage mining is the process of finding out what users are looking for on the Internet. It is the process of extracting useful information from server logs. It works on usage of data as some users might be looking at only textual data, whereas some others might be interested in multimedia. 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 behaviour at a Web site. Based on usage of data it can be further divided into three form-

- 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.
- 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.

Usage mining is valuable not only to businesses using online marketing, but also to e-businesses whose business is based solely on the traffic provided through search engines. It helps to gather the important information from persons visiting the site. E-businesses depend on this information to direct the company to the most effective Web server for promotion of their product or service.

III. Web Structure Mining

Web structure mining 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 the following 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 structure mining is a tool used to identify the relationship between Web pages linked by information or direct link connection. This structure data is discoverable by the provision of web structure schema through database techniques for web pages. This connection allows a search engine to pull data relating to a search query directly to the linking Web page from the Web site the content rests upon. This completion takes place through use of spiders scanning the Web sites, retrieving the home page, then, and linking the information through reference links to bring forth the specific page containing the desired information.

The main purpose for structure mining is to extract previously unknown relationships between web pages. This structure data mining provides use for a business to link the information of its own website to enable navigation and cluster information into site maps. This allows its users the ability to access the desired information through keyword association and content mining. On the WWW, the use of structure mining enables the determination of similar structure of web pages by clustering through the identification of underlying structure. This information can be used to project the similarities of web content. The known similarities then provide ability to maintain or improve the information of a site to enable access of web spiders in a higher ratio. The larger the amount of Web crawlers, the more beneficial to the site because of related content to searches.

Ranking Metrics

Searching the web involves two main steps: Extracting the pages relevant to a query and ranking them according to their quality. Ranking is important as it helps the user look for "quality" pages that are relevant to the query. Different metrics have been proposed to rank web pages according to their quality.

PageRank- PageRank is a metric for ranking hypertext documents based on their quality. Page, Brin, Motwani, and Winograd (1998) developed this metric for the popular_search engine Google. The key idea is that a page has a high rank if it is pointed to by many highly ranked pages. So, the rank of_a page depends upon the ranks of the pages pointing to it. This process is done_iteratively until the rank of all pages are determined.

The rank of a page p can_be written as -

$$PR(p) = d/n + (1-d) \sum_{(q,p)\in G} (\frac{PR(q)}{Outdegree(q)})$$

Here, n is the number of nodes in the graph and Out Degree(q) is the number of hyperlinks on page q. Here d is the probability that the surfer chooses a URL directly and 1−d is the probability that a person arrives at a page by traversing a link.

Hubs and Authorities-

Hyperlink-Induced Topic Search (**HITS**; also known as **hubs and authorities**) is a link analysis algorithm that rates Web pages, developed by Jon Kleinberg. The idea behind Hubs and Authorities stemmed from a particular insight into the creation

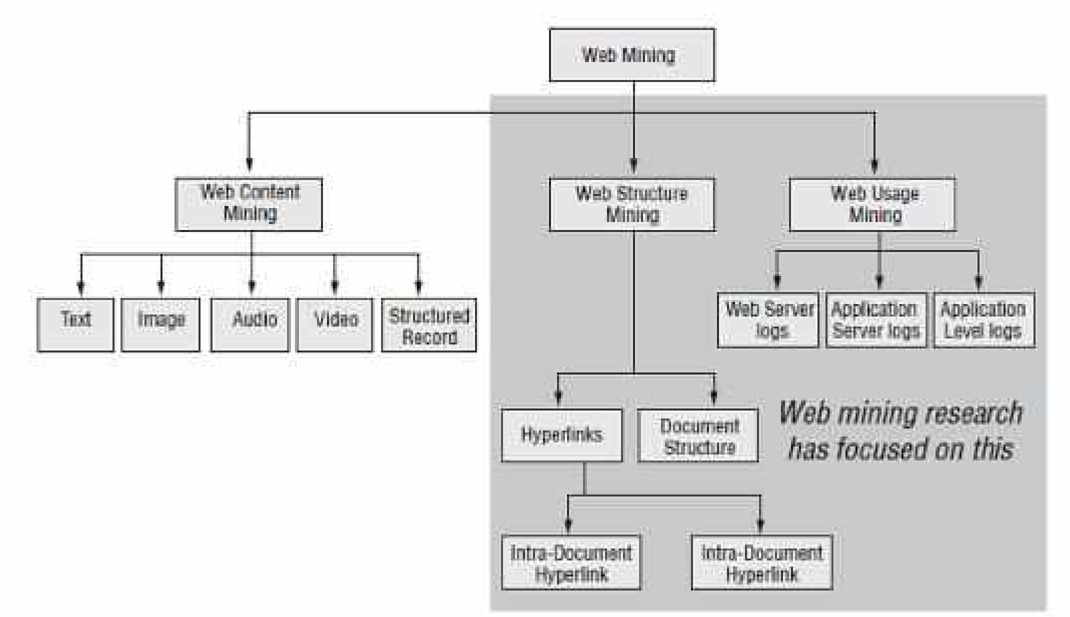

Fig. 2.2 Web Mining Taxonomy

of web pages when the internet was originally forming; that is, certain web pages, known as hubs, served as large directories that were not actually authoritative in the information that it held, but were used as compilations of a broad catalog of information that led users directly to other authoritative pages. In other words, a good hub represented a page that pointed to many other pages, and a good authority represented a page that was linked by many different hubs. This phenomenon also occurs in the Internet. Counts the number of links to a page can give us a general estimate of its prominence on the web, but a page with very few incoming links may also be prominent, if two of these links come from the home pages of Yahoo! or Google or MSN. Thus, because these sites are of very high importance but are also search engines, there can be very irrelevant results. The Twitter Social Network uses a HITS style algorithm to suggest user accounts to follow. Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to. Some implementations also consider the relevance of the linked pages.

Robot Detection and Filtering

Web robots are software programs that automatically traverse the hyperlink structure of the web to locate and retrieve information. First, e-commerce retailers are particularly concerned about the unauthorized deployment of robots for gathering business intelligence at their web sites. Second, web robots tend to consume considerable network bandwidth at the expense of other users. Sessions due to web robots also make it difficult to perform click-stream analysis effectively on the web data.

Web Search

Google is one of the most popular and widely used search engines. It provides users access to information from over 2 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. 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.

Personalized Portal for the Web

Yahoo was the first to introduce the concept of a “personalized portal,” i.e. a web site designed to have the look-and-feel and content personalized to the needs of an individual end-user. Mining MyYahoo usage logs provides Yahoo valuable insight into an individual's web usage habits, enabling Yahoo to provide personalized content, which in turn has led to the tremendous popularity of the Yahoo website.

Conclusion

Without Data mining, many businesses may not be able to perform effective market analysis, compare customer feedback on similar products, discover the strengths and weaknesses of their competitors, retain highly valuable customers and make smart business- decisions. Today, data warehousing and web data mining are much more sophisticated. A web data mining program scans the web continuously looking for targeted information. This information is sorted and stored according to our needs. If we are looking to launch a new product line, a web mining and data warehousing subscription can gather the information relevant to our line, analyze the data and sort it, then help us prepare a report to determine the best marketing plan. The savings, in terms of man hours, would be tremendous as Big Data.

REFERENCES

[1] Zdravko Markov, Daniel T. Larose "Data Mining the Web: Uncovering Patterns in Web Content,

Structure, and Usage", Wiley, 2007 ISBN: 978-0-471-66655-4

[2] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques" [Third Edition] ISBN:978-93-80931-91-3

[3] T.Sudha, M. Usha Rani "Application of Data Mining" ISBN: 978-81-8356-330-7

[4] Soumen Chakrabarti, "Mining the Web: Analysis of Hypertext and Semi Structured Data", Morgan Kaufmann, 2002

[5] Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents and Usage Data", Springer, 2007 ISBN-10 3-540-37881-2

[6] Palvi Arora, Tarun Bhalla "A Synonym Based Approach of Data Mining in Search Engine Optimization" (IJCTT) – volume 12 number 4 – Jun 2014

[7] Web Mining — Concepts, Applications, and Research Directions; University of Minnesota http://dmr.cs.umn.edu/

[8] http://www.web-datamining.net/

[9] www.statsoft.com What is Data Mining (Predictive Analytics, Big Data)

[10] http://en.wikipedia.org/wiki/Web_mining

Author's Biography:

The author is a lecturer (Computer Science) at Awadhoot Bhagwan Ram P.G. College Anpara, Sonebhadra (U.P.) – accredited 'B' by NAAC. He was the Steering Committee Co-ordinator for NAAC A & A and continually co-ordinating IQAC at college. He also, gained 2 years work experience in IT industries.