RESEARCH ARTICLE OPEN ACCESS

A Brief Study and Analysis of Web Mining and It's Applications

Mrs.M. Porkizhi
Assistant Professor
Nadar Saraswathi College of Arts and Science, Theni
TamilNadu – India

ABSTRACT
Web Mining is moving the World Wide Web towards a more useful environment in which users can quicklyand easily find the information they need. Large amount of text documents, multimedia files and images are available in the web and it is still increasing.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 is the application of data mining techniques to discover patterns from the World Wide Web.Web mining consists of web content mining, web structure mining, web usage mining. This paper focuses the major study of Web mining, its categories and techniques, its applications and advantages/disadvantages.In this paper, we are trying to give a brief idea regarding web mining concerned with its techniques, andApplications.

Keywords:- Web content mining, Web structure mining, and Web usage mining.

I. INTRODUCTION

Web mining is the process of using data mining techniques and algorithms to extract information directly from the Web by extracting it from Web documents and services, Web content, hyperlinks and server logs. The goal of Web mining is to look for patterns in Web data by collecting and analyzing information in order to gain insight into trends, the industry and users in general.Web mining is the application of data mining techniques to discover patterns from the World Wide Web.

CATEGORIES OF WEB MINING

- web content mining,
- web structure mining,
- web usage mining

Web Content Mining

This is the process of mining useful information from the contents of Web pages and Web documents, which are mostly text, images and audio/video files. Techniques used in this discipline have been heavily drawn from natural language processing (NLP) and information retrieval.

Web structure mining

This is the process of analyzing the nodes and connection structure of a website through the use of graph theory. There are two things that can be obtained from this: the structure of a website in terms of how it is connected to other sites and the document structure of the website itself, as to how each page is connected.

Web usage mining

This is the process of extracting patterns and information from server logs to gain insight on user activity including where the users are from, how many clicked what item on the site and the types of activities being done on the site.

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.

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

II. 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/polite 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 markup.

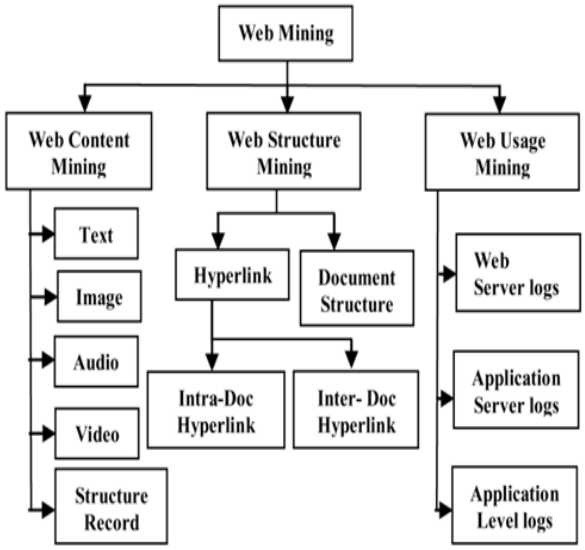
FOUR STEPS IN CONTENT WEB MINING

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)

Web mining is very useful to e-commerce websites and e-services.

III. THREE TYPES OF WEB MINING



1. Web Content Mining

Web Content Mining Problems/Challenges

Data/Information Extraction: Extraction of structured data from Web pages, such as products and search results is a difficult task. Extracting such data allows one to provide services. Two main types of

techniques, machine learning and automatic extraction are used to solve this problem.

Web Information Integration and Schema Matching: Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. Identifying or matching semantically similar data is a very important problem with many practical applications.

Opinion extraction from online sources: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking.

Knowledge synthesis: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.

Segmenting Web pages and detecting noise: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page to extract the main content of the pages is interesting problem.

- Web content mining can be used for mining of useful data, information and knowledge from web page content.
- Web structure mining helps to find useful knowledge or information pattern from the structure of hyperlinks.
- Due to heterogeneity and absence of structure in web data, automated discovery of new knowledge pattern can be challenging to some extent.
- Web content mining performs scanning and mining of the text, images and groups of web pages according to the content of the input (query), by displaying the list in search engines.

For example: If an user wants to search for a particular book, then search engine provides the list of suggestions.

2. Web Usage Mining

Web Usage Mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining, discover user navigation patterns from web data, tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. Web usage mining collects the data from Web log records to discover user access patterns of web pages. There are several available research projects and commercial tools that analyze those patterns for different purposes. The insight knowledge could be utilized in personalization, system improvement, site modification, business intelligence and usage characterization.

- Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
- Web server registers a web log entry for every web page.
- Analysis of similarities in web log records can be useful to identify the potential customers for e-commerce companies.

IV. SOME OF THE TECHNIQUES TO DISCOVER AND ANALYZE THE WEB USAGE PATTERN ARE

i) Session and visitor analysis

- The analysis of preprocessed data can be performed in session analysis ,which includes the record of visitors, days, sessions etc. This information can be used to analyze the behavior of visitors.
- Report is generated after this analysis, which contains the details of frequently visited web pages, common entry and exit.

ii) OLAP (Online Analytical Processing)

- OLAP performs Multidimensional analysis of complex data.
- OLAP can be performed on different parts of log related data in a certain interval of time.
- The OLAP tool can be used to derive the important business intelligence metrics.

3. Web Structure Mining

Web Structure Mining focuses on analysis of the link structure of the web and one of its purposes is to identify more preferable documents. The different objects are linked in some way. The intuition is that a hyperlink from document A to document B implies that the author of document. A thinks document B contains worthwhile information. Web structure mining helps in discovering similarities between web sites or discovering important sites for a particular topic or discipline or in discovering web communities.

Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models.

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

- The web structure mining can be used to discover the link structure of hyperlink.
- It is used to identify that the web pages are either linked by information or direct link connection.
- The purpose of structure mining is to produce the structural summary of website and similar web pages.

Example: Web structure mining can be very useful to companies to determine the connection between two commercial websites.

V. APPLICATIONS

There are many applications of Web Mining, some of them are like:

- E-commerce Customer Behavior Analysis
- E-commerce Transaction Analysis
- E-commerce Website Design
- E-Banking
- Web Advertisement
- Search Engine
- Online Auction
- E-Learning
- E-Governance

Personalized Customer Experience in B2C E-commerce—Amazon.com

Early on in the life of Amazon.com,12 its visionary CEO Jeff Bezos observed,

> "In a traditional (brick-and-mortar) store, the main effort is in getting a customer to the store. Once a customer is in the store they are likely to make a purchase - since the cost of going to another store is high - and thus the marketing budget (focused on getting the customer to the store) is in general much higher than the in-store customer

> experience budget (which keeps the customer in the store). In the case of an on-line store, getting in or out requires exactly one click, and thus the main focus must be on customer experience in the store."

This fundamental observation has been the driving force behind Amazon's comprehensive approach to personalized customer experience, based on the mantra "a personalized store for every customer" (Morphy 2001). A host of web mining techniques, such as associations between pages visited and click-path analysis are used to improve the customer's experience during a "store visit." Knowledge gained from web mining is the key intelligence behind Amazon's features such as "instant recommendations," "purchase circles," "wish-lists," etc.

Web-Wide Tracking—DoubleClick

"Web-wide tracking," i.e. tracking an individual across all sites he visits, is an intriguing and controversial technology. It can provide an understanding of an individual's lifestyle and habits to a level that is unprecedented, which is clearly of tremendous interest to marketers. A successful example of this is DoubleClick Inc.'s DART ad management technology. DoubleClick serves advertisements, which can be targeted on demographic or behavioral attributes, to the end-user on behalf of the client, i.e. the web site using DoubleClick's service. Sites that use DoubleClick's service are part of The DoubleClick Network and the browsing behavior of a user can be tracked across all sites in the network, using a cookie. This makes DoubleClick's ad targeting to be based on very sophisticated criteria. Alexa Research has recruited a panel of more than 500,000 users, who have voluntarily agreed to have their every click tracked, in return for some freebies. This is achieved through having a browser bar that can be downloaded by the panelist from Alexa's website, which gets attached to the browser and sends Alexa a complete click-stream of the panelist's web usage. Alexa was purchased by Amazon for its tracking technology. Clearly web-wide tracking is a very powerful idea. However, the invasion of privacy it causes has not gone unnoticed, and both Alexa/Amazon and DoubleClick have faced very visible lawsuits. Microsoft's Passport technology also falls into this category. The value of this technology in applications such as cyber-threat analysis and homeland defense is quite clear, and it might be only a matter of time before these organizations are asked to provide information to law enforcement agencies.

Personalized Portal for the Web—MyYahoo

Yahoo23 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. This has been an extremely popular concept and has led to the creation of other personalized portals such as Yodlee for private information like bank and brokerage accounts. 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 web site.

VI. COMMON MINING TECHNIQUES

The more popular and top basic mining techniquesthat can help you create optimal results.

CLASSIFICATION ANALYSIS

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes.Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster. So, in classification analysis you would apply algorithms to decide how new data should be classified.A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

ASSOCIATION RULE LEARNING

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset.Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

ANOMALY OR OUTLIER DETECTION

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention.This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy.

CLUSTERING ANALYSIS

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise.A result of this analysis can be used to create customer profiling.

REGRESSION ANALYSIS

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa.It is generally used for prediction and forecasting.

All of these techniques can help analyze different data from different perspectives. Now you have the knowledge to decide the best technique to summarize data into useful information – information that can be used to solve a variety of business problems to increase revenue, customer satisfaction, or decrease unwanted cost.To learn more about implementing data mining techniques, as well as various predictive models like machine learning, check out this data sheet.

VII. CONCLUSION

As the web and its usage continues to grow, so too grows the opportunity to analyze web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper briefly described the keycomputer science contributions made by the field, a number of prominent applications, and outlined some areas of future research. Thisoverview provides a starting point for fruitful discussion.

REFERENCES

[1] Ngu, Anne; Kitsuregawa, Masaru; Chung, Jen-Yao; Neuhold, Erich; Sheng, Quan (2005). Web Information Systems Engineering -

WISE 2005. Berlin: Springer. p. 15. ISBN 9783540300175.

[2] D. Sridevi, Dr. A. Pandurangan, Dr. S. Gunasekaran, "Survey on Latest Trends in Web Mining", International Journal of Research in Advent Technology, Vol. 2, No.3, March 2014.

[3] R. Agrawal, T. Imielinski and A. Swami, database mining: a performance perspective, IEEE Transactions on Knowledge and Data Engineering, 1993.

[4] KamikaChaudhary, Santosh Kumar Gupta, Web Usage Mining Tools & Techniques: A Survey in International Journal of Scientific & Engineering Research, Volume 4, Issue 6,June-2013 1762 ISSN 2229-5518.

[5] V. Bharanipriya& V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[6] Bodyan G.C, Shestakov T.V, "Web Mining in Technology Management", Engineering Universe for Scientific Research and Management, Vol 1 Issue 2, April 2009.

[7] Preeti Chopra, Md. Ataullah, a Survey on Improving the Efficiency of Different Web Structure Mining Algorithms in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

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

[9] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.

[10] Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).

[11] DarshnaNavadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.

[12] Mozenda, at: http://www.mozenda. com/web-mining-software Viewed 18 February 2013.

[13] Web Mining https://www. techopedia.com/definition/ 15634/ web-mining

[14] ChidanshAmitkumar Bhatt, Mohan S. Kankanhalli Multimedia data mining: state of the art and challenge. Journal Multimedia Tools and Applications archive Volume 51 Issue 1, January 2011.

[15] J. Srivastava, R. Cooley, M. Deshpande and P. Tan., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web data", Department of Computer Science and Engineering, University of Minnesota. SIGKDD Explorations, 1(2):12, January 1999.

[16] http://www.google.com/press/pressrel/b2b.html

[17] http://news.google.com

[18] http://www.doubleclick.com/dartinfo/

[19] http://www.alexa.com.

[20] http://www.wired.com/news/business/0,1367,36434,00.html.

[21] http://www.microsoft.com/netservices/passport/.

[22] http://www.aol.com.

[23] http://www.ebay.com