Global Journal of Computational Intelligence Research.
ISSN 2249-0000 Volume 2, Number 1 (2012), pp. 1-8

http://www.ripublication.com/gjcir.htm

Web Mining: A Basic Key to Enrich the Business on Web

Dr. D. Suresh Babu, B. Srinivas, G. Sridhar and B. Raju

Dept. of CSE, KITS, Warangal, India
E-mail: sureshd123@gmail.com, srinu1032@gmail.com
raju.nestham@gmail.com

Abstract

This paper takes an overview of the web mining concept and how it can be useful and beneficial to the business improvement by facilitating its applications in various areas over the internet. The contribution of this paper is towards the various areas containing web sites on internet, which can make best use of different web mining techniques to improve their business decisions based on the user behavior analysis which can ultimately help in improving the relevance of their web site to suit their user needs and adding value to their business growth. It also contributes about the factors responsible and governing the usage of web mining for the web sites to improve business intelligence. The need for techniques which would be able to classify, categories, cluster the web pages in such a way that the web page retrieval can be done in a optimum way and to reduce the burden on the user to keep on searching the required web page from the sea of the information. Web mining is helpful in making business decisions for further trends and patterns of user access of content of the web pages and customer behavior in an effective way. This paper discusses the major business areas which can be benefited by applying web mining techniques.

Keywords: Web Mining, Advantages, Benefits, Business applications of web mining

Status of Web Information

The information on the internet is in the form of static and dynamic web pages of various areas from education, industry to every walk of life including blogs.

As per the web sites' survey more than 170,000,000 web sites are having inter, intra linked web pages. The speed of increase of web information is rapid. The hidden

knowledge discovery, patterns and trends of user access can be found from the way the web sites and web pages are accessed and it is useful from the business perspective giving future directions for decision making.

The Data Mining techniques help in identifying the patterns implying the future trends in the studied data. The Web Mining is an application of the data mining techniques to find interesting and potentially useful knowledge from web data.

Reasons for Web Mining

Infinite web pages are either used or unused by users adding to large volume of space and their occurrence in web searches. 30- 40 % web pages are having duplication of the content approx. Best estimate of unique static HTML pages is in billions from widely used search engines such as Yahoo, Google and increase continually. The following table shows the facts of web sites increase from 2000 till February 2010

Table 1: Increase in the web sites from November 2000 to February 2010

Web Site Survey Month & Year	**Number of total web sites across all domains**	**Observations**
November 2000 till May 2010	**1.GROWTH - HOSTNAME AND ACTIVE WEB SITES**	
May 2005 till May 2009 Hostname Active Web Sites	**2. GROWTH IN HOSTNAME AND ACTIVE WEB SITES** **2.1 There was an increase from 19,000,000 to 68,000,000** **2.2 The growth was observed from 0 to 34,000,000**	Rapid Increase in websites Active websites indicating the Presence of inactive web sites or web pages too.

May 2008till Feb 2010	**3. GROWTH IN HOSTNAME AND ACTIVE WEB SITES**	Further increase in websites requiring study of user access and behavior, link analysis of hyperlinks accessed by user adding value to business
Hostname	**3.1 There was an increase from 68,000,000 to 170,000,000**	
Active Web Sites	**3.2 The growth was observed from 34,000,000 to appr 75,000,000**	

Categories of Web Sites

The web information is categorized as deep web and shallow web. The deep web includes information stored in searchable databases often inaccessible to search engines and it is accessed only by Web Site's interface; the shallow web information can be accessed by search engines without accessing the web databases.

It is necessary to study the relation of web pages among the same and other web sites, useful to improve the business decisions. This requires the web mining to be modeled and applied on the collected data. Thus it requires the basic need to represent the relations and linkages between the web pages hyperlinks, their traversals in a diagrammatic way using data structure techniques such as Graph. Directed graph can be represented as a set of nodes which correspond to pages on the web denoted by V and edges which correspond to links on the web denoted by E. Thus a graph of (V, E) where all edges are directed is ordered pair of nodes which they link. It considers linking of web pages with hyperlinks and their path of the traversal. Cookies generated at client and server side can support user access.

Web Mining Techniques Used and Web Mining Categories

Web Mining techniques make use of the web information and are based on web content mining, web structure mining, and web usage mining. They provide clustering analysis, web link analysis, pattern analysis, association analysis and correlation analysis. The web consists of a large number of unstructured text-based documents using information retrieval, where the user needs effective techniques like keyword based retrieval and indexing techniques. The study of web servers and web log analysis are helpful in applying the web mining techniques. It is normally expected that either the hyperlink structure of the web or the web log data or both have been used in the mining process.

Web Content Mining

It deals with discovering useful information or knowledge from web page contents

than hyperlinks and goes beyond using keywords in a search engine. Web content consists of information such as unstructured free text, image, audio, video, metadata, and hyperlink. Search engines, subject directories, intelligent agents, cluster analysis, and portals are used to find out what a user might be looking for.

Web Structure Mining

It deals with discovering and modeling the hyperlink structure of the web pages based on the topology of the hyperlinks. This gives similarity between sites or sites for a particular topic or web communities.

Web Usage Mining

It deals with understanding user behavior with a web site and to obtain information that may assist in web site reorganization to suit user needs. The mined data includes data logs of user's web interaction, having web server logs, proxy server logs and browser logs, having data about referring page, user identification, user spent time at site and sequence of pages visited. Also cookie files contain information.

While web structure mining shows that page A has a link to Page B, web usage mining shows who or how many people took that link, which site they came from and where they went when they left page B. The important factors considered here are hyperlinks, dynamic content generation as per user preferences, quality of the content in web pages, huge size of the data.

How Web Mining is going to Make the Users Life Easy?

Ultimate objective of the most data mining projects is to use the insights and the models to improve business, by sharing the results across departments by representing with visualization charts, maps effectively.

Solution for Business Decision Problems of E-Commerce for Retailer's Web Site Solved Using Web Mining

In an e -commerce web site a reduction in user's web site behavior analysis time and web site usage, trends will be a value addition provided by web mining. This e-commerce web site, required to develop a web data mining system for business users and data analysts as an end to end solution comprising of data gathering, cleansing, ETL operations, warehousing. The business intelligence systems created user friendly, flexible, dynamic, multidimensional factual reporting, supported by visualization, and web data mining techniques.

In the e-commerce web site the data gathering and data sources includes not only customer registration and demographic information but also web click-streams, response to direct-mail, email campaigns, and orders placed through a website, call center amongst the other sources. The quantity of data can vary above 100 million records. The E-commerce Web Site Architecture can collect additional click stream data besides the data in the web logs; web logs have sensitive information about customer's login, session information, IP addresses indicating the area, region they belong to and their age, frequency of using the web site etc.

The focus on Business to Customer(B2C) e-commerce for retailers helps in understanding and fulfilling the business needs to develop the required expertise and design out of the box reports and analysis of the domain's future trends and patterns of customer behavior understand in a better way to the business user. It can answer the business questions such as to identify heavy spenders at the web site, which are the customers who express willingness to receive emails from the web site are heavy spenders? Such answers reflect the customer's loyalty, based on these results promotion offers and discount offers can also be derived by the business decision maker and possibility to increase in customers can be increases for registrations to web site.

Solution to the Search Engine Problems and How Web Mining Can Help in Improving the Business Decisions

As the search engines use enormous information existing in the web sites, web pages, it is a challenging task to engineer, implement and to improvise the search engine. This specifies that indexing of web pages involves a huge task. Per day tens of millions of queries are given to search engine. It indicates the tremendous magnitude of data. The problems of scaling traditional search techniques to this magnitude data; new technical challenges are involved in using the additional information present in hypertext to produce better search results. The real question of how to build a practical large-scale system which can exploit the hypertext information can be answered by using web mining techniques and improving the capabilities of the search engines by giving better results to customers. It helps in problems of how to effectively deal with uncontrolled hypertext collection where anyone can publish anything they want. Web Mining Applications have been used by these web sites such as Web search e.g., Google and Yahoo , Web Vertical Search e.g.,FatLens and Become, Web Recommendations e.g., Amazon.com , Web Advertising e.g., Google and Yahoo, Web site design e.g., landing page optimization.

The Various Business Areas Where Web Mining has Helped in Improving the Business Decision Making

E-Business

Analysis of click-stream data i.e. web mining uncovers real-time e-business opportunities across geography. It provides ways to target right customers and understand their needs and to customize services and strategies in near-or-real time. The area of advertising is no exception for utilizing the opportunities provided by online customer analytics to promote right products in real time to the right customer. It also helps in effectiveness of a web site as a channel for marketing by quantifying the user's behavior while on the web site.

CRM

Analytical CRM utilizes business intelligence and reporting methodologies such as data mining and analytical processing to CRM applications. While the earlier CRM implementations focus on improving operational efficiencies in the sales and service functions through tailor-made solutions for call-center management, analytical CRM

solutions use intelligence solutions to analyze the data, identify the demographic profiles and measure the purchase frequency and other behavioral patterns of the customers.

With the amount of available online content, today organizations put premium on understanding, adopting and managing the same, convert them into appropriate knowledge suitable to serve their customers better, and thus improve the operations and accelerate the process of delivery of products to markets. The World Wide Web is a fertile area for web Mining and it can provide applications, methods, algorithms to be beneficial in various real-world applications with respect to the critical e-CRM function.

Customer Behavior

Web Mining helps in understanding the concerns such as current and future probability of every customer, relationship between behavior and the loyalty at the website The models based on customer-centric web behavior can be used not only for identifying improvements in the appeal of web site segmentation, which are based on web behavior providing a precise basis for personalization but also for predicting customer's future behavior that is essential for website content planning and design.

Web Usage Mining for Proxy Server

Web Usage Mining is an aspect of data mining that has received a lot of attention in recent years. Commercial companies as well as academic researchers have developed an extension array of tools that perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular website. Performing this kind of investigations on your website can provide information that can be used to better accommodate the user's needs. An area that has received much less attention is the investigation of user-behavior on proxy servers. Servers of Internet Service Providers (ISPs) log traffic from thousands of users of websites. This can give a general overview of user behavior on the Internet or an overview of behavior within a specific sector.

Cross Selling

Web Data mining usage which will allow to cross- sell into web store application with a minimal effort.

Web Site Service Quality Improvement

The World Wide Web is one of the most used interfaces to access remote data and commercial, non-commercial services and the number of actors involved in these transactions is growing very quickly. Everyone using the Web Experiences knows that how the connection to a popular website may be very slow during rush hours and it is well known that web users tend to leave a site if the wait time for a page to be served exceeds a given value. Therefore, performance and service quality attributes have gained enormous relevance in service design and deployment. This has led to the development of web benchmarking tools that are largely available in the market. One of the most common criticism to this approach is that synthetic workload produced by

web stressing tools is far from realistic. Moreover, websites need to be analyzed for discovering commercial rules and user profiles and models must be extracted from log files and monitored data.

Conclusion

In today's era where the entire world has become a global village and the driving force is internet having e-business to internet blogs to search engines, the major questions in front of the business users is while they would like to retain the existing customers and also would like to understand the patterns and trends of customer behavior so that their decisions can be supported with facts represented with visualizations and appropriate reporting made possible with web mining. The success of accuracy of deriving patterns is directly proportional to the amount of sample data used for the data mining techniques.

The advantages of using web mining in search engines and e-commerce, CRM, customer behavior analysis, cross selling; web site service quality improvement is noticeable. The recommendation of using web mining techniques can be applied successfully with a keen analysis of clearly understood business needs and requirements. Also one more governing factor is the amount of data, as the data is voluminous the results can be more towards the correct trends and patterns to be predicted from the given set of data.

But although the web mining techniques can be applied to even the small web sites with a few number of web pages and links within them, web mining may not be the answer for its improvement as it will not be the optimum solution as far as the cost factor in terms of parameters such as complexity of web mining techniques using algorithms may not be recommended.

Possible applications can be On-line social networking community software applications can use web mining techniques to explore the effectiveness of on -line networking, also areas such as knowledge management web sites and web mining can also be useful in bioinformatics, e-governance and e-learning.

References

[1] Berson Alex, et al, 2000. *Building Data Mining Applications for CRM.* Publishers, TATA McGRAW HILL, New Delhi, INDIA.

[2] G. K. Gupta, 2006. *Introduction to Data Mining with Case Studies.* Publishers, Prentice Hall, New Delhi, INDIA. N. Girija, 2006. *Web Mining.* Publishers, ICFAI University Press, Hyderabad, INDIA.

[3] J. Srivastava, et al, 2000. Web Usage Mining: Discovery and Applications of Usage patterns from Web Data, *ACM SIGKDD Explorations,* Vol 1, No 2, pp 12-23.

[4] Magdalini Eirinaki et al, 2003. Web Mining for web personalization. *In ACM Transactions on Internet Technology,* Vol. 3, No. 1, pp 1- 27.

[5] Kosala, et al, 2000. Web Mining Research: A Survey, *ACM SIGKDD Explorations*, Vol 2, No 1, pp 1-15.
[6] Chkrabarti, et al, 2000. Data Mining for Hypertext: A Tutorial Survey, *ACM SIGKDD Explorations*, Vol 1, No 2, pp 1-11.
[7] Joshi, A. et al, 2000. On mining web access logs. *In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp 63-69
[8] Magdalini Eirinki, et al, 2005. Web path recommendations based on page ranking and Markov models, *Proceedings of 7th annual ACM international workshop on web information and data management, Bremen, Germany.*
[9] Mobashir, B. et al, 2000. Discovery of aggregate usage profiles for web personisation. *Proceedings of Web Mining for E-Commerce Workshop(WEBKDD'2000), Sixth ACM SIGKDD International Conference on Knowledge Discovery andData Mining, Boston, USA*
[10] Mohsen Jafari Asbagh, et al, 2007. Web service usage mining: mining for executable sequences, *Proceedings of 7th conference on 7th WSEAS International Conference on Applied Computer Science*, pp 266-271, Venice, ITALY.

Author's Biography

Dr. D. Suresh Babu is currently working as a Professor at Vaagdevi College of Engineering, Warangal,A.P, INDIA. He has received his Ph.D Degree in Computer science & Engineering from Acharya Nagarjuna University, Guntur, A.P., INDIA. His main research interest includes Data Mining, neural networks, data retrieval process and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published more than 15 papers in International journals and conferences.