Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 1, 2016
ISSN: 2455-7528

Web Mining: Prominent application

Neha Mittal

M.Tech, Computers Science, IV semester, SGVU, Jaipur

ABSTRACT

In recent years, we have witnessed the ever – interesting and upcoming publishing medium is the World Wide Web. Undoubtedly, much of the web content is unstructured, so gathering and making sense of such data is very tedious. An application of data mining techniques to the World Wide Web called Web Mining, makes it easier. Web mining is used to discover useful and interesting patterns from the web. Web mining can be further divided into three types. The first is, Web content mining - the process of discovering information or resource from millions of sources across the World Wide Web. The second is, Web Structure Mining – concerned with discovering the structural summary of the web page and the hyperlinks. The third is, Web usage mining – discovering the navigation patterns of the surfers from the web data..

The data available on the web is so voluminous and heterogeneous that it becomes an essential factor to mine this available data to make it presentable, useful, pertinent to a particular problem. Web mining deals with extracting these interesting patterns developing useful abstracts from diversified sources. The present paper deals with a preliminary discussion of WEB mining, few key computer science contributions in the field of web mining , the prominent successful applications and outlines some promising areas of future research.

Keywords – Content Mining Structure Mining, HITS, Page Rank, Usage Mining, Netcraft Survey

I. INTRODUCTION

WWW enriches us with enormous amount of widely dispersed, interconnected, beneficial and dynamic hypertext information. It has profoundly caters the different needs of us in various stages like communication, business, entertainment and so on.

The current World Wide Web has been reached the peak of its success with respect to:

- Valuable resources of information
- Enormous number of users
- Multiform and multitude of data
- Efficient digital commerce

The abundant unstructured or semi-structured information on the web leads a great challenge for both the users, who are seeking for effectively valuable information and for the business people, who needs to provide personalized service to the individual consumers, buried in the billions of web pages. To overcome these problems, data mining techniques must be applied on the www.

In this paper we present a preliminary discussion about Web mining, key accomplishments, applications and future directions

II. DEFINITION OF WEB MINING

Web mining is the application of data mining techniques to discover the patterns from the web. The main objective of web mining is to develop more intelligent tools for potentially help the user in finding, extracting, filtering and evaluating valuable information and resources.

Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 1, 2016
ISSN: 2455-7528

III. REASONS FOR WEB MINING

As like a coin, WWW has two sides, the user and the information provider. Both the sides face problems while dealing with the web data.

A. *The User Problems*

1) Finding relevant information*:* People, either browse or use the search service to find specific information on web.

Today's search tools have two problems. (1) Low precision due to the irrelevance of various search results. (2) Low recall due to the inability to index all the information available on the web as some of the relevant pages is not properly indexed. This is a Query – triggered process.

2) Extracting new knowledge form the web*:* This is a data triggered process. As it is hard to get relevant information, it's very hard to make sense out of it.

B. *The Information Providers' Problems*

Deficient in gathering information about –

- What do the customers do?
- What do the customers want?
- How to personalize the individual users?
- How effectively use the web data to market products and to service the customer?

IV. WEB MINING TAXONOMY

According to analysis targets, Web Mining techniques can be categorized into three areas. They are:

A. *Web Content Mining*

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Research activities in this field also involve using techniques from 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 not been very rapid.

B. *Web Structure Mining*

The structure of a typical Web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.Web Structure Mining can be regarded as the process of discovering structure information from the Web. This type of mining can be further divided into two kinds based on the kind of structural data used.

Hyperlinks: A Hyperlink is a structural unit that connects a Web page to different location, either within the same Web page or to a different Web page. A hyperlink that connect 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 an up-to-date survey. There has been a significant body of work on hyperlink analysis, of which provides an up-to-date survey.

Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

C*. Web Usage Mining*

It is the techniques to discover interesting usage patterns from Webdata, in order to understand and better serve the needs of Web-based applications It deals with the prediction of the surfer's behaviour and interaction with the web. The web content and structure mining uses the primary data web but, the web usage mining mines the secondary data, that is, the data from the web server, access logs, proxy server logs, browser logs, user profiles, registration data, user sessions or transactions, user queries and so on. The success of usage mining depends on what and how much valid and reliable knowledge one can discover from the huge, raw log data.

Real world application: Google Analytics – a service offered by Google that generates detailed statistics about the visits to a website.

Suresh Gyan Vihar University, Jaipur
International Journal of Converging Technologies and Management (IJCTM)
Volume 2, Issue 1, 2016
ISSN: 2455-7528

V. WEB MINING APPLICATIONS

Some of the web mining applications:

- E-Commerce
- Search Engines and Web search
- Website Design
- Recommendation engines
- Web communities and web market places

VI. PROMINENT APPLICATIONS

This section describes some of the most successful applications in this section. Clearly, realizing that these applications use Web mining is largely a retrospective exercise.

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

Early on in the life of Amazon.com, 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. A host of Web mining techniques, e.g. associations between pages visited, click-path analysis, etc., 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.

B .*Web Search—Google*

Google has successfully used the data available from the Web content (the actual text and the hyper-text) and the Web graph to enhance its search capabilities and provide best results to the users. Google has expanded its search technology to provide site-specific search to enable users to search for information within a specific website. The Google Toolbar' is another service provided by Google that seeks to make search easier and informative by providing additional features such as highlighting the query words on the returned web pages. The full version of the toolbar, if installed, also sends the click-stream information of the user to Google. The usage statistics thus obtained would be used by Google to enhance the quality of its results. Google also provides advanced search capabilities to search images and look for pages that have been updated within a specific date range. Built on top of Netscape's Open Directory project, Google's web directory provides a fast and easy way to search within a certain topic or related topics. The Advertising Programs introduced by Google targets users by providing advertisements that are relevant to search query. This does not bother users with irrelevant ads and has increased the clicks for the advertising companies by four or five times.

One of the latest services offered by Google is,' Google News'. It integrates news from the online versions of all newspapers and organizes them categorically to make it easier for users to read "the most relevant news". It seeks to provide information that is the latest by constantly retrieving pages that are being updated on a regular basis.

C. *Personalized Portal for the Web – My Yahoo*

Yahoo was the first to introduce the concept of a 'personalized portal', i.e. a Web site designed to have the look-and-feel as well as 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, e.g. Yodlee for private information. Mining My Yahoo usage logs provides Yahoo valuable insight into an individual's Web usage habits, enabling Yahoo to provide compelling personalized content, which in turn has led to the tremendous popularity of the Yahoo Web site.

VII. FUTURE RESEARCH DIRECTIONS

As the Web and its usage grows, it will continue to generate ever more content, structure, and usage data, andthe value of Web mining will keep increasing. Outlined hereare some research directions that must be pursued to ensure that we continue to develop Web mining technologies thatwill enable this value to be realized.

A. *Temporal evolution of the Web*

Large organizations generally archive (at least portions of) usage data from there Web sites. With these sources of data available, there is a large scope of research to develop techniques for analyzing of how the Web evolves over time.
The temporal behavior of the three kinds of Web data: Web Content, Web Structure and Web Usage. The methodology suggested for Hyperlink Analysis in [9] can be extended here and the research can be classified based on Knowledge Models, Metrics, Analysis Scope and Algorithms. For example, the analysis scope of the temporal behavior could be restricted to the behavior of a single document, multiple documents or the whole Web graph. The other factor that has to be studied is the effect of Web Content, Web Structure and Web Usage on each other over time.

B*. Fraud and threat analysis*

The anonymity provided by the Web has led to a
significant increase in attempted frauds, from unauthorized use of individual credit cards to hacking into credit card database for blackmail purposes. Yet another example is auction fraud, which has been increasing on popular sites like eBay. Since all these frauds are being perpetrated through the Internet, Web mining is the perfect analysis technique for detecting and preventing them. Research issues include developing techniques to recognize known frauds, and characterize and then recognize unknown or novel frauds, etc. The issues in cyber threat analysis and intrusion detection are quite similar in nature.

VIII. CONCLUSION

As the Web and its usage continues to grow, so does the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years has 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. In this paper we have briefly described the key computer science achievements made in this field, the prominent successful applications, and outlined some promising areas of future research in the area of web mining.

IX. REFERENCES

[1] T. Berners-Lee, R. Cailliau, A. Loutonen, H. Nielsen, and A.Secret. The World- Wide Web. *Communications of the ACM*, 37(8):76- 82, 1994
.

[2] R.Albert, H. Jeong, and A.-L. Barabasi.Diameter of the World-wide web. *Nature, 401:13Q-131,* September 1999.

[3] J. Borges and M. Levene. Mining Association Rules in Hypertex Databases. *In knowledge Discovery and Data Mining*.

[4] B. Berendt. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*,37-59,2002.

[5] B. Berendt, A. Hotho, and G. Stumme. Towards semantic web mining. In I. Horrocks and 1. Hendler, editors, *Proceedings of the1st International Semantic Web Conference (ISWC-02)*, pages 264-278. Springer-Verlag, 2002.

[6] J. Srivastava, B. Mobasher, Panel discussion on “Web Mining: Hype or Reality?” at *the 9th IEEE International Conference on CA,1997.*

AUTHORS INFORMATION