USAGE OF WEB MINING IN MANAGEMENT RESEARCH

Mrs Shivani Chaudhary
Lecturer
A.SM' Group's Institute of Computer Studies
Pimpri, Pune India

ABSTRACT

The World-Wide Web provides every internet citizen with access to an abundance of information. The challenge of extracting knowledge from data draws upon research in marketing, management, statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions.

Web Mining is the application of data mining to discover useful knowledge from the Web. Web mining focuses now on four main research directions related to the categories of Web data: Web content mining, Web usage mining, Web structure mining, and Web user profile mining. Web content mining discovers what Web pages are about and reveals new knowledge from them. Web usage mining concerns the identification of patterns in user navigation through Web pages and is performed for the reasons of service personalization, system improvement, and usage characterization. Web structure mining investigates how the Web documents are structured, and discovers the model underlying the link structures of WWW. Web user profile mining discovers user's profiles based on users' behavior on the Web.

This paper provides a brief overview of the accomplishments of the field, both in terms of technologies and applications, and outlines key future research direction.

Keywords: Web mining, information retrieval, Web pages, information extraction.

INTRODUCTION:

The World Wide Web is a popular and interactive medium to disseminate information today. The Web is very huge diverse, and dynamic and thus raise scalability, multimedia data, and temporal issues respectively. Due to these situations, we are currently drowning in information and facing information overload. The researcher and other information seeker faces the following problems when their interacting with the Web.

Finding right information:

People either browse or use the search engine to find out the specific information they need on the Web. While doing this they simply fire the keyword query and the search result in a list of related pages which are ranked on their similarity to the query. However today's search tool have the following problems. The first problem is low precision, which is due to the irrelevance of many of the search results. This result in a difficulty finding the relevant information. The second problem is low recall, which 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..

Creating new knowledge from the information available on the web:

This is a kind of data triggered process that presumes that we already have a collection of the Web data and we want to exact potentially use full knowledge out of it . Recent research mainly focuses on the using Web as knowledge base for information gathering and making use of it with great effort.

Making the information personal:

This problem is often associated with the type and presentation of information, since people needs and the type of data they require differ in content and presentations.

1. WEB MINING:

1.1 Overview:

Web mining techniques can be used to solve the information overloading problem. 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. It aims to present a full picture of the state-of-the-art research and development of actionable knowledge discovery (AKD) in real-world businesses and applications. This paper features the latest methodological, technical and practical progress on promoting the successful use of data mining in a collection of business domains. The huge information space spurs the development of data mining and information retrieval techniques. Web mining, which is moving the World Wide Web toward a more useful environment in which users can quickly and easily find information, can be regarded as the integration of techniques gathered by means of traditional data mining methodologies and its unique techniques.

The intended audience of this paper will mainly consist of researchers, research students and practitioners in data mining and knowledge discovery .Two different approaches were taken in initially defining Web mining. First was a 'process-centric view', which defined Web mining as a sequence of tasks .Second was a 'data-centric view', which defined Web mining in terms of the types of Web data that was being used in the mining process. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue. In this paper we follow the

data-centric view of Web mining which is defined as follows Web involves three types of data on the Web, Cooley classified the data type as . Web Content, Web Structure and Web Usage data.

2.2.1 Web Content Mining:

Web content mining describes the automatic search of information resource available online and involves particularly mining Web content data. It is a combination of novel methods from a wide range of fields including data mining, machine learning, natural language processing, statistics, databases, information retrieval and so on.

Unfortunately, much of the data is unstructured and semi-structured. The Web document usually contains different types of data, such as text, image, audio, video, metadata and hyperlinks. Providing a relational interface to all such databases may be complicated. This unstructured characteristic of Web data forces the Web content mining towards a more complicated approach.

In the following, some key technologies concerning Web mining are demonstrated: the way of constructing conceptual semantic space, multi-hierarchy text classification and clustering algorithm based on Swarm Intelligence and k-Means.

2.2.2 Web Structure Mining: Page Rank vs. HITS:

Web structure mining is essentially about mining the links on the Web. Web pages are actually instances of semi-structured data, and thus mining their structure is critical to extract information from them. 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. In the following, we would like to compare famous link analysis methods: Page Rank vs. HITS. Two most influential hyperlink based search algorithms Page Rank and HITS were reported during 1997-1998. Both algorithms exploit the hyperlinks of the Web to rank pages according to their levels of “prestige" or “authority". PageRank Algorithm is originally formulated by Sergey Brin and Larry Page, PhD students from Stanford University, at Seventh International World Wide Web Conference (WWW) in April, 1998 The algorithm is determined for each page individually according to their authoritativeness. More specifically, a hyperlink from a page to another page is an implicit conveyance of authority to the target page. The more in-links that a page i receives, the more prestige the page i has. Let the Web as a directed graph $G = (V, E)$ and let the total number of pages be n. The PageRank score of the page i (denoted by $P(i)$) is defined by

$$P(i) = \Sigma(i, j) \in EP(j)Oj$$

2.2.3 Web Usage Mining:

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 [68]. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

2.2.3.1 Web Server Data:

The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

2.2.3.2 Application Server Data:

Commercial application servers such asWeblogic StoryServer [72] have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

2.2.3.3 Application Level Data:

New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of the especially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

3 KEY CONCEPTS:

In this section we briefly describe the key new concepts introduced by the Web mining research community.

3.1 User profiles - Understanding how users behave:

The Web has taken user profiling to completely new levels. For example, in a 'brick and-Mortar' store, data collection happens only at the checkout counter, usually called the 'point-of-sale'. This provides information only about the final outcome of a complex human decision making process, with no direct information about the process itself. In an on-line store, the complete click-stream is recorded, which provides a detailed record of every single action taken by the user, providing a much more detailed insight into the decision making process. Adding such behavioral information to other kinds of information about users, e.g. demographic, psychographic, etc., allows a comprehensive user profile to be built, which can be used for many different applications While most organizations build profiles of user behavior limited to visits to their own sites, there are successful examples of building 'Web-wide' behavioral profiles, e.g. Alexa Research and DoubleClick These approaches require browser cookies of some sort, and can provide a fairly detailed view of a user's browsing behavior across the Web.

3.2 Pre-processing – making Web data suitable for mining:

In the panel discussion referred to earlier, pre-processing of Web data to make it suitable for mining was identified as one of the key issues for Web mining. A significant amount of work has been done in this area for Web usage data, including user identification, session creation, robot detection and filtering extracting usage path patterns etc.

3.3 ONLINE BIBILIOMETRICS:

With the Web having become the fastest growing and most up to date source of information, the research community has found it extremely useful to have online repository of publications. Lawrence et al. have observed that having articles online makes them more easily accessible and hence more often cited than articles that are offline. Such online repositories not only keep the researchers updated on work carried out at different centers, but also makes the interaction and exchange of information much easier. With such information stored in theWeb, it becomes easier to point to the most frequent papers that are cited for a topic and also related papers that have been published earlier or later than a given paper. This helps in understanding the 'state of the art'in a particular field, helping researchers to explore new areas.

Fundamental Web mining techniques are applied to improve the search and categorization of research papers, and citing related articles. Some of the prominent digital libraries are SCI, ACM portal, CiteSeer and DBLP.

3.4 Visualization of the World Wide Web:

Mining Web data provides a lot of information, which can be better understood with visualization tools. This makes concepts clearer than is possible with pure textual representation. Hence, there is a need to develop tools that provide a graphical interface that aids in visualizing results of Web mining. Analyzing the web log data with visualization tools has evoked a lot of interest in the research community.

2. PROMINENT APPLICATIONS:

An outcome of the excitement about the Web in the past few years has been that Web pplications have 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, even though the organizations that developed these applications, and invented the corresponding technologies, did not consider it as such. We describe some of the most successful applications in this section.

2.1 Wikipedia:

Technically it is not a search engine, the Wikipedia Encyclopedia is one of the web's most popular research tools. Here user can add material or edit articles at any time. In appropriate material and unreferenced statements are removed quickly. It is very useful for the researchers for getting relevant data for their research. It provides the following facilities.

- **Help desk** – Ask questions about using Wikipedia.
- **Reference desk** – Serving as virtual librarians, Wikipedia volunteers tackle your questions on a wide range of subjects.
- **Village pump** – For discussions about Wikipedia itself, including areas for technical issues and policies.
- **Community portal** – Bulletin board, projects, resources and activities covering a wide range of Wikipedia areas.
- **Site news** – Announcements, updates, articles and press releases on Wikipedia and the Wikimedia Foundation.

2.2 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.

2.3 Web Search – Google:

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. Earlier search engines concentrated on Web content alone to return the relevant pages to a query . PageRank, that measures the 2The truth of this fundamental insight has been borne out by the phenomenon of 'shopping cart abandonment', which happens frequently in on-line stores, but practically never in a brick-and-mortar one. 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. 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.

2.4 Understanding Web communities – AOL:

One of the biggest successes of America Online (AOL) has been its sizeable and loyal customer base A large portion of this customer base participates in various 'AOL communities', which are collections of users with similar interests. In addition to providing a forum for each such community to interact amongst themselves, AOL provides them with useful information and services. Over time these communities have grown to be well-visited 'waterholes' for AOL users with shared interests. Applying Web mining to the data collected from community interactions provides AOL with a very good understanding of its communities, which it has used for targeted marketing through ads and e-mail solicitation. Recently, it has started the concept of 'community sponsorship', whereby an organization, say Nike, may sponsor a community called 'Young Athletic Twenty Something'. In return, consumer survey and new product development experts of the sponsoring organization get to participate in the community, perhaps without the knowledge of other participants. The idea is to treat the community as a highly specialized focus group, understand its needs and opinions on new and existing products, and also test strategies for influencing opinions.

2.5 Understanding auction behavior – eBay:

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/buying mechanism, which 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 towards understanding participants' bidding behaviors/patterns to create a more efficient auction market.

2.6 Personalized Portal for the Web – MyYahoo:

The company is perhaps best known for its web portal, search engine (Yahoo! Search), Yahoo! Directory, Yahoo! Mail, Yahoo! News, advertising, online mapping (Yahoo! Maps), video sharing (Yahoo! Video), and social media websites and services. Yahoo! provides Internet communication services such as Yahoo! Messenger and Yahoo! Mail. In March 2007, Yahoo! announced that their e-mail service would offer unlimited storage beginning May 2007.

Yahoo! also offers social networking services and user-generated content in products such as My Web, Yahoo! Personals, Yahoo! 360°, Delicious, Flickr and Yahoo! Buzz. Yahoo! partners with numerous content providers in products such as Yahoo! Sports, Yahoo! Finance, Yahoo! Music, Yahoo! Movies, Yahoo! News, Yahoo! Answers and Yahoo! Games to provide media content and news. Yahoo! also provides a personalization service, My Yahoo!, which enables users to combine their favorite Yahoo! features, content feeds and information onto a single page.

2.7 CiteSeer - Digital Library and Autonomous Citation Indexing:

NEC Research Index, also known as CiteSeer is one of the most popular online bibiliographic indices related to Computer Science. The key contribution of the CiteSeer repository is the "Autonomous Citation Indexing" (ACI) .Citation indexing makes it possible to extract information about related articles. Automating such a process reduces a lot of human effort, and makes it more effective and faster. CiteSeer works by crawling the Web and downloading research related papers. Information about citations and the related context is stored for each of these documents. The entire text and information about the document is stored in different formats. Information about documents that are similar at a sentence level (percentage of sentences that match between the documents), at a text level or related due to co-citation is also given. Citation statistics for documents are computed that enable the user to look at the most cited or popular documents in the related field. They also a maintain a directory for computer science related papers , to make search based on categories easier. These documents are ordered by the number of citations.

CONCLUSIONS:

As the Web and its usage continues to grow, so 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 we have briefly described the key computer science contributions made by the field, a number of prominent applications, and outlined some promising areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.
\

REFERENCES:

Google News. http://news.google.com.

Google Inc. http://www.google.com.

SRIVASTAVA, DESIKAN AND KUMAR 69

T. Haveliwala. Topic-sensitive PageRank. In *In Proceedings of the Eleventh*

International World Wide Web Conference, Honolulu, Hawaii, May 2002., 2002.

N. Imafuji and M. Kitsuregawa. Effects of maximum flow algorithm on identifying

web community. In *Proceedings of the fourth international workshop on Web*

information and data management, pages 43–48. ACM Press, 2002.

The Internet Archive Project. http://www.archive.org/.

J.Ghosh and J. Srivastava. Proceedings of Workshop on Web Analytics. http:

//www.lans.ece.utexas.edu/workshop_index2.htm, 2001.

J.Ghosh and J. Srivastava. Proceedings of Workshop on Web Mining. http:

//www.lans.ece.utexas.edu/workshop_index.htm, 2001.

L.R. Ford Jr and D.R. Fulkerson. Maximal Flow through a network. *Canadian J.*

Math, 8:399–404, 1956.

R.H. Katz. Pervasive Computing: It's All About Network Services, 2002.J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of*

the ACM, 46(5):604–632, 1999.

R. Kohavi. Mining e-commerce data: The good, the bad, and the ugly. In Foster J. Srivastava, R. Cooley, M. Deshpande, and P.N. Tan. Web Usage Mining: Discovery

and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*,

1(2):12–23, 2000.