ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 3, pp: (455-460), Month: July - September 2014, Available at: www.researchpublish.com

WEB MINING AN APPLICATION OF DATA MINING

Sumit Dalal, Sumit Kumar, Vivek Dixit

Dept. of Computer Science Engineering, Dronacharya Collage of Engineering (Gurgoan) Haryana, India

Abstract: **Web mining is a very hot research topic which combines two of the activated research areas: Data Mining and World Wide Web. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Although there exists quite some confusion about the Web mining, the most recognized approach is to categorize Web mining into three areas: Web content mining, Web structure mining, and Web usage mining. Web content mining focuses on the discovery/retrieval of the useful information from the Web contents/data/documents, while the Web structure mining emphasizes to the discovery of how to model the underlying link structures of the Web. The distinction between these two categories isn't a very clear sometimes. Web usage mining is relative independent, but not isolated, category, which mainly describes the techniques that discover the user's usage pattern and try to predict the user's behaviors. This paper is a survey based on the recently published research papers. Besides providing an overall view of Web mining, this paper will focus on Web usage mining. Generally speaking, Web usage mining consists of three phases: Pre-processing, Pattern discovery and Pattern analysis. A detailed description will be given for each part of them, however, special attention will be paid to the user navigation patterns discovery and analysis. The user privacy is another important issue in this paper. An example of a prototypical Web usage mining system, Web SIFT, will be introduced to make it easier to understand the methodology of how to apply data mining techniques to large Web data repositories in order to extract usage patterns. Finally, along with some other interested research issues, a brief overview of the current research work in the area of Web usage mining is included.**

Keywords: **Data mining, Pattern Discovery, Web mining.**

I. INTRODUCTION

Web usage mining is the process of extracting useful information from server logs e.g. use Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. 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 behavior at a Web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

- **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. A key feature is the ability to track various kinds of business events and log them in application server logs.
- **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. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

Studies related to work [Weichbroth et al.] are concerned with two areas: constraint-based data mining algorithms applied in Web Usage Mining and developed software tools (systems). [Costa and Seco] demonstrated that web log mining can be used to extract semantic information (hyponymy relationships in particular) about the user and a given community.

II. WEB MINING TAXONOMY

Web mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. Figure1 shows the taxonomy.

III. WEB CONTENT MINING

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on this topic have drawn heavily on techniques developed in 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 been limited.

IV. WEB STRUCTURE MINING

The structure of a typical web graph consists of web pages as nodes, and hyper-links as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used.

V. HYPERLINKS

A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects 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..

VI. 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. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents

VII. WEB USAGE MINING

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000). 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:

VIII. WEB SERVER DATA

User logs are collected by the web server and typically include IP address, page reference and access time.

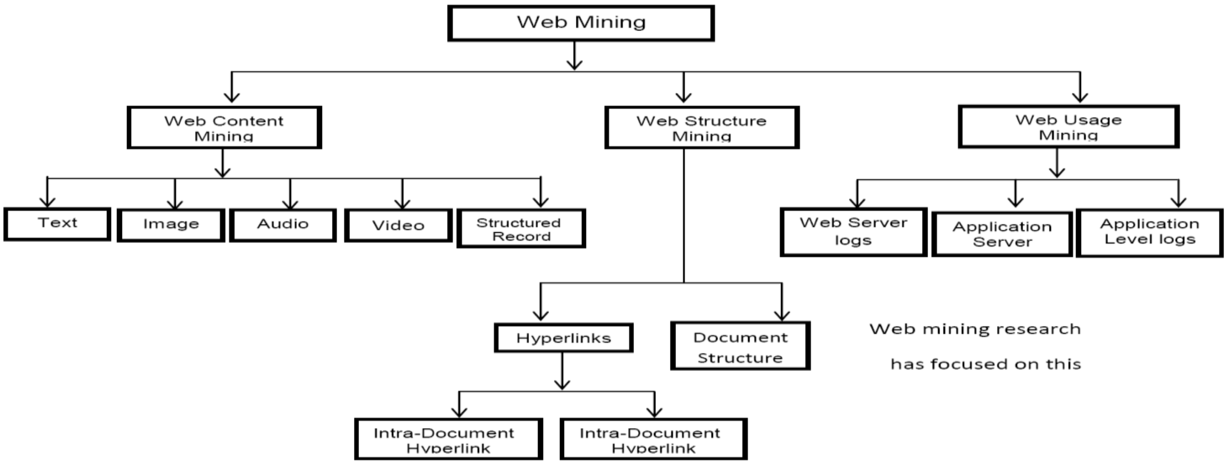

Figure 1: Web mining Taxonomy

IX. WEB STRUCTURE MINING

Most of the Web information retrieval tools only use the textual information, while ignore the link information that could be very valuable. 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 shemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlinks in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain, therefore the query processing will be easier and more efficient.

Web structure mining has a nature relation with the Web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the Web. It's quite often to combine these two mining tasks in an application.

X. WEB USAGE MINING

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. The heterogeneity and the lack of structure that permits much of the ever-expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web such as Lycos, Alta Vista, WebCrawler, ALIWEB [6], MetaCrawler, and others provide some comfort to users, but they do not generally provide structural information nor categorize, filter, or interpret documents. In recent years these factors have prompted researchers to develop more intelligent tools for information retrieval, such as intelligent web agents, as well as to extend database and data mining techniques to provide a higher level of organization for semi-structured data available on the web. The agent-based approach to web mining involves the development of sophisticated AI systems that can act autonomously or semi-autonomously on behalf of a particular user, to discover and organize web-based information.

Web content mining is differentiated from two different points of view:[1] Information Retrieval View and Database View. R. Kosala et al.[2] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site to transform a web site to become a database.

There are several ways to represent documents; vector space model is typically used. The documents constitute the whole vector space. If a term t occurs n(D, t) in document D, the t-th coordinate of D is n(D, t) . When the length of the words in a document goes to [corrupted text]. This representation does not realize the importance of words in a document. To resolve this, tf-idf (Term Frequency Times Inverse Document Frequency) is introduced.

By multi-scanning the document, we can implement feature selection. Under the condition that the category result is rarely affected, the extraction of feature subset is needed. The general algorithm is to construct an evaluating function to evaluate the features. As feature set, Information Gain, Cross Entropy, Mutual Information, and Odds Ratio are usually used. The classifier and pattern analysis methods of text data mining are very similar to traditional data mining techniques. The usual evaluative merits are Classification Accuracy, Precision, Recall and Information Score.

Web mining is an important component of content pipeline for web portals. It is used in data confirmation and validity verification, data integrity and building taxonomies, content management, content generation and opinion mining.

XI. THE USAGE MINING ON THE WEB

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications In the same paper, the Web usage mining is parsed into three distinctive phases: preprocessing, pattern discovery, and pattern analysis. I think it is an excellent approach to define the usage mining procedure. It also clarified the research sub direction of the Web usage mining, which facilitates the researchers to focus on each individual process with different applications and techniques. With the assistance of the diagram of the high-level

Web usage mining process shown in Figure 1, which is presented in [4, 5, 6], reader may understand the architecture of the Web Usage Mining easily.

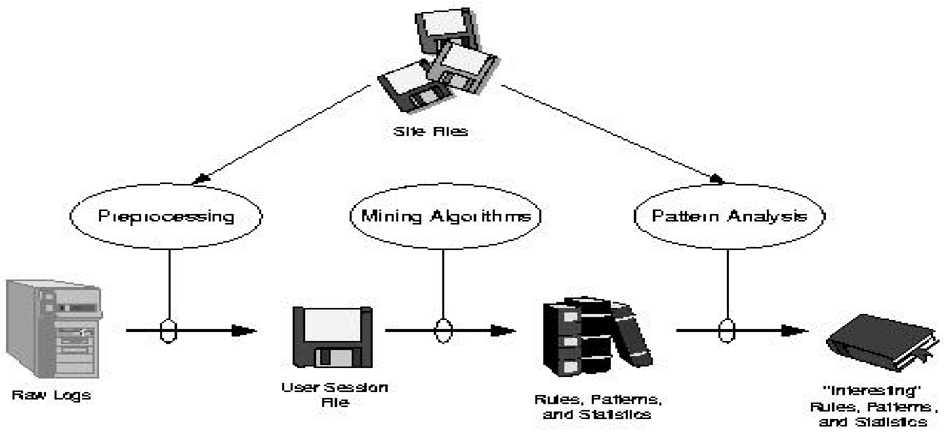

Figure 1: High Level *Web Usage Mining* Process

XII. PERSONALIZATION VS. USER NAVIGATION PATTERN

The applications of Web usage mining can be classified into two main streams: personalized vs. impersonalized. Personalized means learning a user profile of user modeling in adaptive interfaces, while impersonalized means learning user navigation pattern . With the technique of personalization, the Web user would prefer an intelligent Web server which capable to learn their information needs and preferences. On the other hand, with the technique of learning user navigation patterns, the information providers would be glad to view the improvement of the effectiveness on their Web sites, which results in adapting the Web site design or by biasing the user’s behavior towards satisfying the goals of the site.

1. Personalization

The Web provides a direct communication medium between the vendors of products and services, and their customer with very low cost. There come tremendous opportunities for ecommercedevelopment. The Web personalization is a very important, if not necessary, part of the e-commerce. Even outside of the e-commerce, Web personalization has many applications. In the context of Web mining, personalization is the provision to the individual of tailoredproducts, services, information or information relating to products or service. The goal of personalization systems is to provide users with what they need or want without explicit indication . B. Mabasher broadened the definition as the Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users.

Today, three of the major categories of existing personalization systems are manual decision rule systems, collaborative filtering system, and content-based filtering system. Mabasher compared these three kinds of system, and claimed that the new generation of Web personalization tools is attempting to incorporate techniques for pattern discovery from Web usage data.

2. User Navigation Pattern

The research of user navigation pattern focuses on the techniques to study the user behavior when navigating within a web site. While the World Wide Web turns to be the largest information resource available online, awareness of the user navigation preferences becomes an essential step. It is not only in the process of customizing and adapting the site's interface for individuals, but also in improving the site's static structure of the underlying hypertext system as well . Good knowledge on the way of visitors navigate in a web site could prevent disorientation and help the provider to place the information properly.

XIII. PRIVACY ON THE WEB

Due to the massive growth of the e-commerce, privacy becomes a sensitive topic and attracts more and more attention recently. The basic goal of Web mining is to extract information from data set for business needs, which determines its application is highly customer-related. As I mentioned in the above section, there exists unavoidable conflict between the Web user and the administrator in the view of privacy. From the administrators point of view, many of the uses of data mining are innocuous, such as the data analysis to detect hidden behavioral patterns to allow supermarkets to arrange items in ways that will encourage customers to buy more of certain products or to look for seasonal buying variations. However, from individual point of view, many users believe that some applications of Web mining, may raise privacy concern, such as junk mails stuck mail account or personal information divulged during online shopping. The privacy concern has become the most critical concern for the Web user, and e-commerce developer.

The lack of regulations in the use and deployment of Web mining systems and the widely spread privacy abuses reports related to data mining has made privacy a hot iron like never before. Privacy touches a central nerve with people and there are no easy solutions. To solve the problem, the privacy legislation is as important as the technique efforts.

XIV. WEB USAGE MINING PROS AND CONS

Pros

Web usage mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled e-commerce to do personalized marketing, which eventually results in higher trade volumes. Government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of mining applications can benefit society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

Cons

Web usage mining by itself does not create issues, but this technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web usage mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. The obtained data will be analyzed, and clustered to form profiles; the data will be made anonymous before clustering so that there are no personal profiles. Thus these applications de-individualize the users by judging them by their mouse clicks. De-individualization, can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits.

Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests.

The growing trend of selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded increasing the likeliness of one's privacy being invaded. The companies which buy the data are obliged make it anonymous and these companies are considered authors of any specific release of mining patterns. They are legally responsible for the contents of the release; any inaccuracies in the release will result in serious lawsuits, but there is no law preventing them from trading the data.

Some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation, right now this situation can be avoided by the high ethical standards maintained by the data mining company. The collected data is being made anonymous so that, the obtained data and the obtained patterns cannot be traced back to an individual. It might look as if this poses no threat to one's privacy, actually many extra information can be inferred by the application by combining two separate unscrupulous data from the user.

XV. RELATED WORKS

As far as we know, it was Etzioni who first coined the term Web mining. Etzioni starts by making a hypothesis that the information on the Web is sufficiently structured and outlines the subtasks of Web mining. His paper describes the Web mining processes. There have been some works around the survey of data mining on the Web. The first paper that we know that noticed the confusion in the Web mining research. It gives a Web mining taxonomy but restricted to Web content and Web usage mining, and gives a survey on Web usage mining. It divides the Web content mining into the agent based approach and the database approach. We use a similar division but divide it into the IR approach instead of the agent approach. Later, in they classify Web mining into three categories that are similar to our categories. Compared to their paper, our paper points out three confusions on the usage of the term Web mining, identifies additional user-centered Web mining processes, and provides new perspectives for the Web mining categories. We use the Web mining categories suggested in and proposes a new model for mining Web log data, while discusses the research issues of Web mining in the context of Web warehouse project.

XVI. CONCLUSIONS

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 ten 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 areas of future research. Our hope is that this overview provides a starting point for fruitful discussion.

ACKNOWLEDGEMENT

We thank our collage teachers for giving the advice on the legislation issue of privacy and the Interesting discussion.

REFERENCES

[1] B. Berendt. Web usage mining, site semantics, and the support of navigation.

[2] J. Borges and M. Levene. Data mining of user navigation patterns. In Proceedings of the WEBKDD'99 Workshop on Web Usage Analysis and User Profiling, August 15, 1999, San Diego, CA, USA, pages 31-39, 1999.

[3] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide Web. In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997

[4] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining world wide Web browsing patterns. Knowledge and Information Systems, 1(1), 1999.

[5] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.

[6] R. Cooley. WebSIFT: The Web Site Information Filter System.

[7] Oren Etzioni. The world wide Web: Quagmire or gold mine. Communications of the ACM, 39(11):65-68, 1996.