Web Mining and Knowledge Discovery of Usage Patterns

CS 748T Project (Part I)

Yan Wang

February, 2000

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, WebSIFT, 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.

1. Introduction

It is not exaggerated to say the Web World Web is the most excited impacts to the human society in the last 10 years. It changes the ways of doing business, providing and receiving education, managing the organization etc. The most direct effect is the completed change of

information collection, conveying, and exchange. Today, Web has turned to be the largest information source available in this planet. The Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository, which supplies incredible amount of information, and also raises the complexity of how to deal with the information from the different perspectives of view – users, Web service providers, business analysts. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs. All of them are expecting tools or techniques to help them satisfy their demands and/or solve the problems encountered on the Web. Therefore, Web mining becomes an active and popular research field.

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services [7]. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity of Web mining. Web mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence [8], and also psychology and statistics as well.

As a forerunner of my term project specified in Web mining, the paper is organized as following:

Section 1 – Introduction

Section 2 – A general introduction of the Web data mining

Section 3 – Usage mining on the Web

Section 4 – A usage mining system: WebSIFT

Section 5 – Personalization vs. User navigation pattern

Section 6 – Privacy on the Web

Section 7 – Related Work

Section 8 – Conclusion

2. Web Data Mining

2.1 Overview

As many believe, it is Oren Etzioni first proposed the term of Web mining in his paper [7] 1996. In this paper, he claimed the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Many of the following researchers cited this explanation in their works. In the same paper, Etzioni came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a few research projects and papers. Some of the commercial consideration has presented on the schedule.

In both [7] and [8], they suggested a similar way to decompose Web mining into the following subtasks:

a. Resource Discovery: the task of retrieving the intended information from Web.
b. Information Extraction: automatically selecting and pre-processing specific information from the retrieved Web resources.
c. Generalization: automatically discovers general patters at the both individual Web sites and across multiple sites.
d. Analysis: analyzing the mined pattern.

In brief, Web mining is a technique to discover and analyze the useful information from the Web data. The authors of [10] claims the Web involves three types of data: data on the Web (content), Web log data (usage) and Web structure data. The authors of [5] classified the data type as content data, structure data, usage data, and user profile data. M. Spiliopoulou [14] categorized the Web mining into Web usage mining, Web text mining and user modeling mining; while today the most recognized categories of the Web data mining are Web content

mining, Web structure mining, and Web usage mining [2,8,10]. It is clear that the classification is based on what type of Web data to mine.

2.2 Web Content Mining

Web content mining describes the automatic search of information resource available online [10], and involves mining web data contents. In the Web mining domain, Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. The Web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents, or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of Web data force the Web content mining towards a more complicated appoach.

The Web content mining is differentiated from two different points of view [3]: Information Retrieval View and Database View. R. Kosala et al. [8] 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 of to transform a Web site to become a database.

S. Chakrabarti [19] provides a in-depth survey of the research on the application of the techniques from machine learning, statistical pattern recognition, and data mining to analyzing hypertext. It's a good resource to be aware of the recent advances in content mining research.

Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources. Multimedia data mining on the Web has gained many researchers' attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done. For the details about multimedia mining, please refer [8, 18] to find the related resource information.

2.3 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. The detailed works on it can be referred to [17].

What is on earth the structural information, and how to discover it? S. Madria et al. [17] gave a detailed description about how to discover interesting and informative facts describing the connectivity in the Web subset, based on the given collection of interconnected web documents. The structural information generated from the Web structure mining includes the

follows: the information measuring the frequency of the local links in the Web tuples in a Web table; the information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document; the information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites; the information measuring the frequency of identical Web tuples that appear in a Web table or among the Web tables.

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.

2.4 Web Usage Mining

Web usage mining tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web. M. Spiliopoulou [14] abstract the potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources,

which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.

There are lots of works have been done in the IR, Database, Intelligent Agents and Topology, which provide a sound foundation for the Web content mining, Web structure mining. Web usage mining is a relative new research area, and gains more and more attentions in recent years. I will have a detailed introduction in the next section about usage mining, based on some up-to-date research works.

3. 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 [5]. 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. I will give a detailed introduction as follows, encompassing these three-phase processing.

3.1 Data Pre-processing for Mining

From the technique point of view, Web usage mining is the application of data mining techniques to usage logs (secondary Web data) of large Web data repositories. The purpose of it is to produce results that can be used in the design tasks such as Web site design, Web server design and of navigating through a Web site [4]. However, before applying the data mining algorithm, we must perform a data preparation to convert the raw data into the data

abstraction necessary for the further process. The data can be collected at the server-side, client-side, proxy servers, or obtained from database. For each type of data collection, the difference is not only the location, but also the available data type, the segment of population from which the data was collected and the method of implementation [5]. The information sources available to mine include Web usage logs, Web page descriptions, Web site topology, user registries, and questionnaire [14]. It's natural to think that the preprocess has three different conversions: Usage converting, Content converting, and Structure converting.
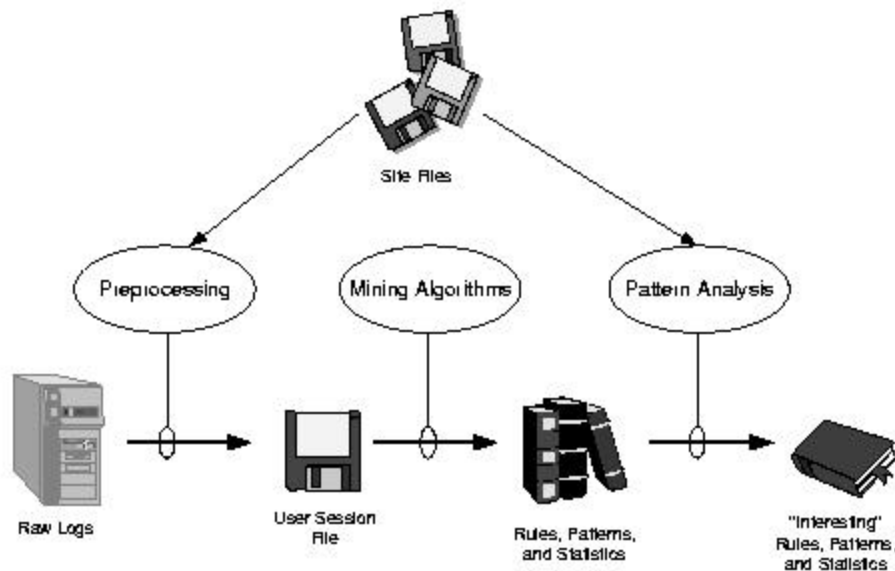

Figure 1: High Level *Web Usage Mining* Process

Since the data abstraction is very important in the data preprocess, it's necessary to clarify the definitions of the related data abstractions before the description of the different type of the data converting. The following definitions are from the Web characterization terminology & definition sheets drafty published by the World Wide Web Committee Web usage characterization activity (http://www.w3.org/WCA, http://www.w3.org/1999/05/WCA-terms/).

User –The principal using a client to interactively retrieve and render resources or resource manifestations.

Page view – Visual rendering of a Web page in a specific client environment at a specific point in time.

Click stream – A sequential series of page view request.

User session – A delimited set of user clicks (click stream) across one or more Web servers.

Server session (visit) – A collection of user clicks to a single Web server during a user session. Also called a visit.

Episode - A subset of related user clicks that occur within a user session.

3.1.1 Content Preprocessing

Content preprocessing is the process of converting text, image, scripts and other files into the forms that can be used by the usage mining. It's not hard to understand that the Web content can be used to filter the input to, or output from the pattern discovery algorithm [5]. R. Cooley also described how the page views play the important roles in the preprocessing. For the content of static page views, the preprocessing can be easily done by parsing the HTML and reformatting the information or running additional algorithm as desired. It would be much more complicated to the content of dynamic page views. To perform the preprocessing, the content of each page view must be "assembled", either by an HTTP request from a crawler, or a combination of template, script, and the database accesses. Please refer to the [5] for the detailed information.

3.1.2   Structure Preprocessing

The structure of a Web site is formed by the hyperlinks between page views. The structure preprocessing can be treated similar as the content preprocessing. However, each server session may have to construct a different site structure than others.

3.1.3   Usage Preprocessing

The inputs of the preprocessing phase may include the Web server logs, referral logs, registration files, index server logs, and optionally usage statistics from a previous analysis.

The outputs are the user session file, transaction file, site topology, and page classifications. It's always necessary to adopt a *data cleaning* techniques to eliminate the impact of the irrelevant items to the analysis result. The usage preprocessing probably is the most difficult task in the Web Usage Mining processing due to the incompleteness of the available data [5]. Without sufficient data, it is very difficult to *identify the users*. The easiest way to improve the data quality is to get user cooperation, but it's not easy at all. There exists a conflict between the analysis needs of the analysts (who want more detailed usage data collected), and the privacy needs of the individual users (who want as little data collected as possible) [3]. However, the heuristics and statistics methods can be used to improve the quality of the Web usage data [14]. We may find some approach to solve the problem, but it is impossible to avoid the misidentification completely, since the Web is so dynamic and versatile. For example, any page view accessed through a client or proxy-level cache will not be "visible" from the server side, and the only verifiable method of tracking ached page views is to monitor usage from the client side [5].

The *session identification* is also a part of the usage preprocessing. The goal of it is to divide the page accesses of each user, who is likely to visit the Web site more than once, into individual sessions. The simplest way to do is to use a timeout to break a user's click-stream into session. The thirty minutes is used as a default timeout by many commercial products. Another problem is named as *path completion*, which indicates the determining if there are any important accesses missed in the access log. The methods used for the user identification can be used for path completion. The final procedure of the preprocessing is *formatting*, which is a preparation module to properly format the sessions or transactions. For the details of the data preparation for the Web mining, please refer to [4].

3.2   Pattern Discovery

This is the key component of the Web mining. Pattern discovery converges the algorithms and techniques from several research areas, such as data mining, machine learning, statistics,

and pattern recognition. According to the techniques adopted in this area, I will introduce this process in the separate subsections as follows.

3.2.1 Statistical Analysis

Statistical techniques are the most powerful tools in extracting knowledge about visitors to a Web site. The analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic Web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions [5].

3.2.2 Association Rules

In the Web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions [4]. The authors of [5] pointed that in the term of the Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for prefetching documents to reduce user perceived latency.

3.2.3 Clustering

Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development

and execution of future marketing strategies [4]. Clustering of users will help to discover the group of users, who have similar navigation pattern. It's very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The clustering of pages is useful for Internet search engines and Web service providers, since it can be used to discover the groups of pages having related content.

3.2.4 Classification

Classification is the technique to map a data item into one of several predefined classes. In the Web domain, Web master or marketer will have to use this technique if he/she want to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe te properties of a given class or category. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc [5].

3.2.5 Sequential Pattern

This technique intends to find the inter-session pattern, such that a set of the items follows the presence of anothers in a time-ordered set of sessions or episodes. It's very meaningful for the Web marketer to predict the future trend, which help to place advertisements aimed at certain user groups. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis [5].

3.2.6 Dependency Modeling

The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the Web domain. The modeling technique

provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future Web resource consumption.

3.3   Pattern Analysis

Pattern Analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of Web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. This can be done with the help of some analysis methodologies and tools. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations [15]. All these methods assume the output of the previous phase has been structured. There are more techniques coming out in recent years, such as visualization etc.

This is also a fertilized research area. Although there are quite a few commercial analysis applications available and many more are free on the Web, most of them are dislike by users, considered too slow, inflexible, difficult to maintain and limited in the functionality. To develop the efficient, flexible, and powerful tools, lots of work need to be done for both researcher and developer.

4. WebSIFT: The Web Site Information Filter System

The Web Site Information Filter System is a Web usage mining framework, that uses the content and structure information from a Web site, and finally identify the interesting results from mining usage data [6]. The WebSIFT system is designed to perform usage mining from the serverlogs in the extended NSCA format.   The preprocessing algorithms include identifying users, server sessions, and inferring cached page references through the use of the referrer field. Besides creating the server session, WebSIFT system performs content and
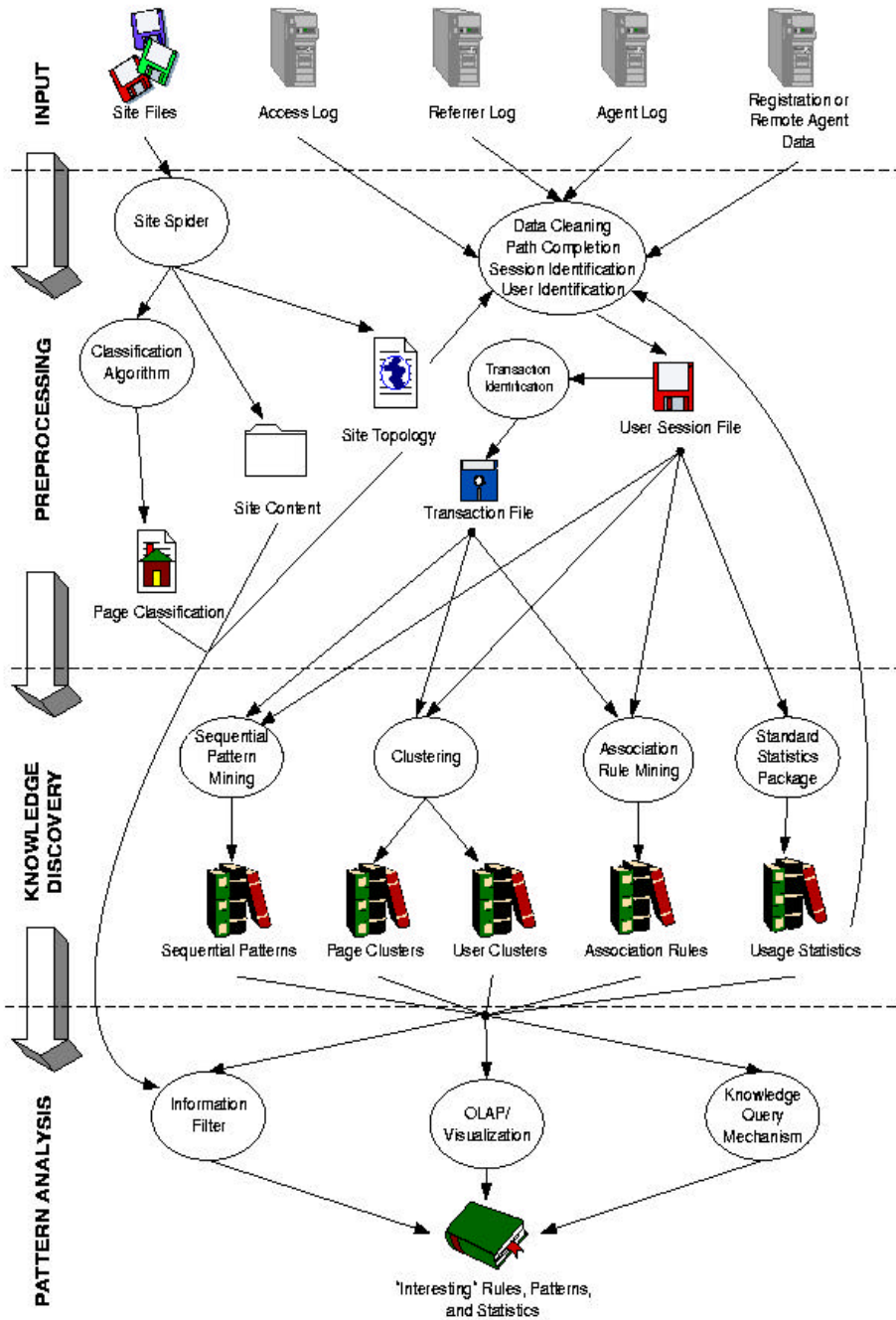
14

Figure 2: WebSIFT Architecture

– 3 –

15

structure preprocessing, and provides the option to convert server sessions into episodes. The server session or episode files can be run through sequential pattern analysis, association rule discovery, clustering or general statistics algorithms.

The WebSIFT system is based on the WEBMINER prototype [4], and divides the Web usage mining process into three principal parts that are corresponding to the three phases of usage mining I described in Section 3. Figure 1 is also the high level architecture of the WebSIFT. Figure 2 [5], [6] provides a more details to show how to do usage mining in a particular Web site.

In input of the mining process includes three server logs – access, referrer, and agent; the HTML files that make up the site; and the optional data such as registration files, remote agent logs. In the preprocessing process, the input data is used to construct a user session file, to derive a site topology and to classify the pages of a site. The user session file will be converted to the transaction file and output to next phase – Pattern Discovery. Both the site topology and page classifications are fed into the information filter, which belongs to the Pattern Analysis process and makes use of the preprocessed content and structure information to automatically filter the results of the knowledge discovery algorithms for patterns that are potentially interesting [5]. The pattern discovery phase uses the existing data mining techniques as mentioned in Section 3 (statistics, association rules, clustering, sequential) to generate rules and patterns. The discovered information is then fed into various pattern analysis tools, which includes the information filtering, OLAP, and knowledge query mechanism like SQL, to generate the final mining results.

The WebSIFT system has been implemented using a relation database, procedural SQL, and the Java programming language. Java Database Connectivity (JDBC) drivers are used to interface with the database. To the reader who is interested to know the experimental evaluation, please refer to [6].

5. 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 [8]. 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.

5.1 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 e-commerce development. 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 tailored products, 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 [11]. B. Mabasher [9] 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.

Mabasher et al. also provided a system model for mining Web log files to discover profile for the provision of recommendations to current users based on their browsing similarities with previous users. There are several principal elements consisting of Web personalization in their framework. They are the modeling of Web objects (products, service, pages etc) and subjects (users), categorization of objects and subjects, mapping between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. The overall process of usage-based personalization is divided into two components: offline component vs. online component. The offline component is consisted of the data preparation and specific usage mining tasks that have been introduced in the previous sections. Online component uses the discovered patterns to provide personalized content to users, based on their current navigational activity. The authors introduced a personalization system based on the architecture they propose in the same paper – WebPersonalizer System. Currently, the system relies on only anonymous usage data provided by Web server logs and hypertext structure of a site, and provides a list of recommended hypertext links to a user while browsing through a Web site. Please refer to [9] for the further details.

Some current open issues in this area are mentioned in [9], such as the problems of the profile data being subjective, as well getting out of date as user preferences change over time.

5.2 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 [2]. Good knowledge on the way of visitors navigate in a web site could prevent disorientation and help the provider to place the information properly.

Analysis of user behavior has two aspects, one concerning the interests of the users and the accessed information, the other concerning the way of accessing the information. The first aspect is solved by techniques for the construct of user profiles and is not specific to the Web usage, while the second one is address by analyzing Web server logs, which falls in the field of the Web usage mining [12]. In the paper, M. Spiliopoulou et al. proposed the exploitation of mining technology to discover access patterns with "interesting" statistical properties and presented Web Utilization Miner (WUM) – a tool designed for the purpose. The mining model of WUM is in two aspects. First, it predicts that the "importance" indicators in user behavior go far beyond than frequent access to some pages, such that the pattern discovery can be done in the statistical domain, but also supports the subjective specification. Second, by processing aggregated sequences and applying optimization steps during the mining process, the high performance can be achieved.

B. Berendt [1] commented the WUM in term of that WUM not only discovers frequent sequences, but it also allows the inspection of the different paths through the site. WUM performs sequence mining, which is the branch of data mining investigating the temporal characteristics of Web usage. Compared with most sequence miners, WUM not only discovers sequences that are frequent and exceed a specified confidence threshold, but also allows the analyst to investigate the details of the visit path of users. However, WUM is limited in its ability to support statistical tests for significance, and it is site semantics independent by design.

6. 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.

6.1 Legislation efforts

In 1995, the European Union passed its Directive on Data Protection that introduces privacy protection applying to the private sector. The Directive required member countries to adopt national data protection laws that meet the standards of the Directive within three years (by 1998). The European Union's European Data Protection Directives limits access to Internet-based customer information. European companies can use data about customers to profile, but those profiles are encrypted to block out customers' names. Meanwhile, the Directive prohibits member countries from transferring personal information to a non-member country or to a business located in a non-member country, if the non-member country's laws do not provide adequate protection for personal information. (European commission. The directive on the protection of individuals with regard of the processing of personal data and on the free movement of these data. http://www2.echo.lu/)

Unfortunately in U.S. there is no unifying framework in place, although a Congress develop legislation has been recommended by U.S. Federal Trade Commission to regulate the personal information being collected at Web sited [5].

6.2 Technology development

While there are great efforts to address privacy issues by the legislative and regulation bodies, many researchers are working on new technologies to better protect consumers' privacy. Researchers at Xerox Corp.'s Palo Alto Research Centre have created an algorithm that designed to keep the behavior of online shoppers hidden from Web site operator (www.znet.com). Encirp, a vendor of marketing software designed to work in the electronic billing environment, uses an engine that function on the *consumers' desktop* to sidestep the privacy trap. It protects consumers' privacy by avoiding centralized data storage by the service provider and provides personalized interface through the engine and data stored at the consumers' desktops [16].

As J. Srivastava pointed in [5], the main challenge is to come up with guidelines and rules. With the rules and guidelines, site administrator may perform various analyses on the usage data without compromising the identity of an individual user. W3C has initiated a project called Platform for Privacy Preferences (P3P), which provides a protocol try to solve the conflict between Web users and the site administrators. P3P is also in proceeding to provide guidelines for independent organization which can ensure that sites comply with the policy statement they are publishing (http://www.truste.org/). Please go to http://www.w3.org/P3P/ for the details.

It is expected a complete solution for the privacy issues around Web mining will not be easily found for many years to come. However, the process is sure accelerating with the public attention, the efforts of the companies, the breakthrough technologies and the regulation of the government agencies. The key issue for all sides is to maintain a balance in privacy concern and the use of data mining including both the results implementation and the data collection. Only by maintaining a careful balance can the beauty of Web mining be fully explored.

7. Related Works

As many researchers believe, it was Etzioni who first came up with the term of Web mining in his paper [7]. He brought out a question: is it practical to mine Web data? He also suggested dividing the Web mining to three processes. The paper opened up a new active research field. There are increasing number of researcher working on this field and do some surveys around the data mining on the Web. The Web mining was clearly categorized as Web content mining, Web structure mining and Web usage mining in [2, 17] till 1999. The research works have been well classified since then. There have been some works around content mining, and structure mining, based on the research of Data mining and Information Retrieval, Information Extraction, and Artificial Intelligence. In the usage mining research area, several groups did distinguished work. R. Cooley et al. in University of Minnesota [3, 4, 5, 6, 8] did in-depth research to all the procedure of usage mining. They proposed a mining prototype WebMiner and derived a system WebSIFT to perform the usage mining, which is relatively practical. O. Zaiane et al. [15] proposed the idea of how to implement the OLAP technique on the Web mining. Their works on the multimedia data also provided a valuable solution for content mining. M. Spiliopoulou et al. [12, 13, 14] focused on the applications of the usage mining. His works on the navigation pattern discovery and web site personalization has special meaning for the e-commerce society and the Web marketplace allocation, and will be very helpful for both Web user and administrator. The Web Utilization Miner system is an innovative sequential mining system. J. Borges et al. has explored some algorithms to mine the user navigation pattern in [2] and his other papers. He proposed a data mining model to achieve an efficient mining, which captures the user navigation behavior pattern by using N-grammar approach.

8. Conclusion

In this paper, we survey the researches in the area of Web mining with the focus on the Web Usage Mining. Three recognized types of web data mining are introduced generally. Around the key topic of this paper - usage mining, we provide detailed description of the three phases of the process. An example of usage mining system is given to illustrate the overall usage

mining process. Moreover, the research of major applications of usage mining – personalization and navigation pattern discovery are discussed. Finally, we wrap up this paper with the most controversial topic - the user privacy.

Besides the generalization of the current research work, we also try to clarify some confusion and reveal the up-to-data research issues.

Acknowledgement

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

[8] R. Kosala, H. Blockeel. Web mining Research: A Survey

[9] B. Mobasher, R. Cooley, J. Srivastava. Automatic Personalization Based on Web Usage Mining. Communications of the ACM, Volume 43, Number 8 (2000)

[10] S.K.Madria, S.S.Bhowmick, W.K.Ng, and E.P.Lim. Research issues in Web data mining. In Proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK '99, pages 303-312, 1999

[11] M.D.Mulvenna, S.S.Anand, A.G.Buchner. Personlization on the Net using Web Mining Introduction. Communicaitons of the ACM, Volume 43, Number 8 (2000)

[12] M. Spiliopoulou, L.C.Faulstich, K. Winkler. A Data Miner analyzing the Navigational Behaviour of Web Users

[13] M. Spiliopoulou. Web Usage Mining for Web site evaluation

[14] M. Spiliopoulou. Data mining for the Web. In Proceedings of Principles of Data Mining and Knowledge Discovery, Third European conference, PKDD'99, P588-589

[15] O. Zaiane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by applying OLAP and Data Mining Technology on Web Logs. In Advances in Digital Libraries, pages 19-29, Santa Barbara, CA, 1998

[16] Report, "Personalized marketing without privacy invasion", ABA Banking Journal, August 2000

[17] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E. P. Lim. Research issues in web data mining. In proceedings of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK'99, pages 301-312, 1999

[18] O. Zaiane, J. Han, Z. Li, S.H. Chee, J.Y. Chiang. MultiMediaMiner: A system prototype for MultiMedia Data Mining

[19] S. Chakrabarti. Data Mining for hypertext: A tutorial survey