Journal of Engineering and Applied Sciences 2 (3): 585-590, 2007


# Streamlined User Navigation of Web Logs and its Privacy Issues and Tools-A Study Report

[1]Meera Gandhi and [2]S.K Srivatsa
[1]Department of Computer Science and Engineering, Sathyabama University,
Jeppiaar Nagar, Old Mahabalipuram Road, Sholinganallur, Chennai-119, India
[2]Department of ECE, M.I.T Campus, Anna University Chennai-44, India

**Abstract:** Recently computer systems have become a critical part of network-connected system, possessing essential economic and human values to individuals and organizations. This key role of the systems has increased the requirements for their protection. They have to be more resistant against malicious activities .Web mining refers to the whole of data mining and related techniques that are used to automatically discover and extract information from web documents and services. Web mining does, however, pose a threat to some important ethical values like privacy and individuality. Web mining makes it difficult for an individual to autonomously control the unveiling and dissemination of data about his/her private life. To study these threats, we distinguish between `content and structure mining" and `usage mining." Web content and structure mining is a cause for concern when data published on the web in a certain context is mined and combined with other data for use in a totally different context. Web usage mining raises privacy concerns when web users are traced, and their actions are analyzed without their knowledge. Sophisticated analytic and data mining software tools enable firms to use the data contained in these logs to develop and implement a complex relationship management strategy Although there are a variety of solutions to privacy-problems, none of these solutions offers sufficient protection. The values of privacy and individuality should be respected and protected to make sure that people are judged and treated fairly. People should be aware of the seriousness of the dangers and continuously discuss these ethical issues. This should be a joint responsibility shared by web miners (both adopters and developers), web users, and governments. This study gives a detailed study report on privacy issues on the streamlined navigation of users, which helps them in privacy and individuality.

**Key words:** Ethics-individuality, threat, privacy, user navigation, web data mining

## INTRODUCTION

Because of increased online activities of companies' data mining by means of web server access logs is becoming more powerful tool of collecting personal information from customers. Data gathered by this method can support implementation of complex relationship management strategies. Based on retrieved customer information consumers can be grouped in high lifetime valuable and low lifetime valuable, which would influence the services provided to them. In our study we explore major social issues of web server logs data mining.

**Benefits for the companies:** Today, the enormous content of the Internet has made it difficult to find relevant information on a subject. Methods helping user navigation and retrieving information have become particularly important. Online shops need to offer personalized products to clients but before being able to do that they have to personalize the web sites to the clients. This is where the data mining techniques in web server logs are coming in. Companies can use the basic data retrieved from the data logs (Anthony, 2002) to analyze customer behaviors, evaluate the current usage, if the customers liked or disliked it and so on. To create adaptable web sites to each user, first, the user navigation patterns in the web have to be found and analyzed (Smith and Ng, 2003). Data mining is a method extracting valuable information from the data for statistical purpose.

### Literature survey

**How does it work? Discussion on data mining techniques:** One of the most known data mining techniques is WEBSOM (Smith and Ng, 2003) (Web Based Self-Organizing Map) , which organizes documents into two-dimensional map according to their content

**Corresponding Author:** Meera Gandhi, Department of Computer Science and Engineering, Sathyabama University, Jeppiaar Nagar, Old Mahabalipuram Road, Sholinganallur, Chennai-119, India

rather than by keyword. There are several benefits using SOM (Abdi, 2003; Smith and Ng, 2003) as it clusters the documents not people. The whole process is automatic, it can be done in a larger scale saving labor costs and it employs search by context instead by key word. However, in order to get the full benefits of the needs of the users and to organize the web pages in a web-user-oriented way instead in a content-driven way the feedback from the users is essential. Nevertheless, as a rule of thumb, the web pages are designed in a way that individuals can search information in effective and efficient manner or at least, this is the goal in the most cases.

Unlocking the usage patterns of the web users hidden in the log files has therefore been a challenge for several researchers. Fu, Perkowitz and Etzioni (Smith and Ng, 2003) have demonstrated that web users c an be grouped into meaningful clusters (Abdi, 2003), which help the web designers to provide high-value customized services as the data mining of the web server logs provides them with the information needed to understand users better. These systems can also be used to improve current Web Pages.

Mobasher (Smith and Ng, 2003) created a Web Personalization system that organizes web usage data not the content of the data mentioned above into clusters. The system analyzes the web server logs, it identifies to which user group the current user belongs to and makes suggestions to links that would interest the user. These suggestions are based on the past experience of a particular user group. However, this approach has its limitations. It cannot provide sufficient information for the user to search by concept; it does not help the user to move from one concept to another and restricts the user to see a web from a broad perspective. Instead of providing efficient means for information retrieval, it provides suggestions what page to visit next.

LOGSOM (Log Based Self-Organizing Map) combines the benefits of the both of these systems. It keeps a track and organizes the web pages according to the user navigation behaviors and interest not to the web content. In this way, self-organizing maps (Smith and Ng, 2003) can be updated regularly according to user's interest. In addition, it does not only group web pages into clusters but it also illustrates the graphical relationships among these. It is flexible enough to present a large amount of data. The input data is gathered not about the content of the web pages but which users have visited which page. The limitation of the system is that it has been tested only on a sample of data, the School of Business Systems web server at Monash University (Anthony, 2002) However, it has proven to be effective and the next challenge is to use it in a larger context.

**More detailed view how they actually do it - data caching algorithms:** The general idea is to make the full use of the web log data using data mining applications. Data mining is aimed to discover the standards, structure, content, usage patterns and so on of users and web pages. Intelligent web caching algorithms are the tools to predict the web requests. These web-caching algorithms (Bonchi *et al*., 2001) are able to adapt their behavior on the basis of the user access patterns, which in step are extracted from the historical access data recorded in the log files by means of data mining techniques (Abdi, 2003).

**The goal of web catching algorithms:** The aim is to increase the number of web pages that are retrieved directly from the cache instead requesting them from the server. There are several web caching algorithms, we list three of them:

**Frequent patterns:** In the case of frequent patterns, we extract from the web logs the patterns that follow the form A and B (if A, then B). If A has been requested then B is likely to be requested next.

**Decision trees:** In the case of decision trees, we develop a decision tree in a basis of the historical data in the web logs but on this case concentrating on the time needed until the next request.

**Page gather:** This algorithm uses the data mining clustering to group a collection of web sites that are regularly visited close to each other and distant from other groups. Data mining clustering differs from the traditional clustering in a way that it can place one document in a multiple overlapping clusters.

Whatever technique is used, first, the data gathered has to be put into a data mart, where it is consolidated, cleaned, selected and prepared for the data mining analysis. Data mart is populated by the set of the basic fields available from web servers through SQL commands. URL text messages are encoded to a numerical form. Advantages of the numeric form are that it optimizes the disc space, enables efficient comparison, makes it easy to sort the data and also employs privacy concerns. For instance, for all the reasons mentioned above, the fields of site, path, file name etc. of the URL are encoded and decoded using hash functions.

**Which type of data can be collected from where?:** Everything you do while surfing in the net can provide useful information, for instance, for web site designers. Click stream data is the path that the user creates when steering through the sites and following links. It can be used to evaluate the traffic and popularity of the page.

Shopping chart can provide information in e-business where the purchases were made and where the customer left the order unfinished.

Psychographic data would include data on user's attitudes towards topics, products etc., buying behavior and beliefs. Access data counts the time between the last and next access to the same URL. Time data gives information on amount of time a user spends exploring the site, the product or topic he or she is interested in.

All that provides us with the useful information about the users but how far can we go until it becomes a privacy concern. Is it appropriate to record all the user activities in order to find out how users perceive the site?

## WHAT DO USERS THINK?

As data mining tools and algorithms become more sophisticated and widely available, customers. Privacy concerns are constantly increasing. These concerns are especially high due to opportunity of World Wide Web to easily automatically collect consumer data and add it to databases. With organizations increasingly building comprehensive consumer databases and applying sophisticated data-mining techniques privacy and ethics issues become more pressing.

**Does internet data mining violate users? privacy?:** According to the study conducted by the Federal Trade Commission, among 1400 web sites, only 14% of them post notices or disclosures on web pages. Current research reports that disclosure rate has improved. But still there are improvements to be made, because, according to same study only 10% of the 361 organization Web sites practice all 4 substantive fair information practices of: notice, choice, access and security (Milne, 2000).

Web server log data mining implies gathering user information without immediate knowledge of consumer, which includes using cookies and tracking software, click-and-viewing patterns. That is the reason for raised consumer concerns about profiled web sites (Spiliopoulou, 2000) and junk e-mail. Because consumer information is stored on a database, it is potentially accessible for the whole Internet. So, personal data can be accessed later and used for purposes different from initial. Many companies, such as double-click can merge cookies to develop complex views of consumer's online behavior (Yen, 2003).

Many consider customer information acquired by firms by means of data mining server logs to be privacy violation, because information is being collected without notification of the person. Of course, there is an option to block cookies by selecting a certain option in a browser. Recently, the use of click stream data has come under court investigation. The most recent juridical case in this area is concerned with the Real Networks Company that produces a popular software program for listening to music on computers. The company was recently accused in using software to transmit information about users listening patterns in Internet to its headquarters. The software contained details about music preferences of each customer, the number of songs they copied, and a serial number that was matched with a consumers e-mail address. The lawsuit claimed that the transmission of user music habits was not disclosed in a Real Networks privacy practices.

**What is the user persistence of internet data mining?:** In USA there was conducted a survey by Hanrick Associates, which aim was to gain the understanding of consumer attitudes toward online privacy. It explored the impact of privacy policies on consumers purchasing behavior (Lee *et al*., 2004), users opinion about acceptable levels of data mining and, as a result of those, 4 ways were suggested for companies to improve user trust. Three hundred and fifty responses were received by May 2001 with some follow up after September 11th, an event, which played an important role in users security concerns.

As was found out, control over personal data remains central for consumers. Thus, people like to be sure, that their names will not be added without notification to spam lists or given to other organizations without their permission. Benefits of data mining, such as personalized content and streamlined navigation override privacy (Milne, 2000) concerns only for some users. Though, survey has found out that mostly people are satisfied with companies online privacy policies and don't mind data collection and mining practices. Fifty four% of survey respondents think that Internet companies are moving in the right direction in terms of personal information and privacy issues. For the question about how privacy concerns influenced the willingness to visit or purchase from Web site, 62 and 49% of users, respectively, answered, that it depends on the site. So, nearly two-thirds of users surf in the web selectively and almost half shop online selectively.

According to Forrester Research and Active Medias Real Numbers behind Net Profits, three-quarters of respondents agree or strongly agree that Web sites should be allowed to analyze Web site traffic on an anonymous, aggregate level and 71% agree or strongly agree that Web sites should be allowed to analyze anonymous user demographic information on the site. From information above we can conclude that majority of

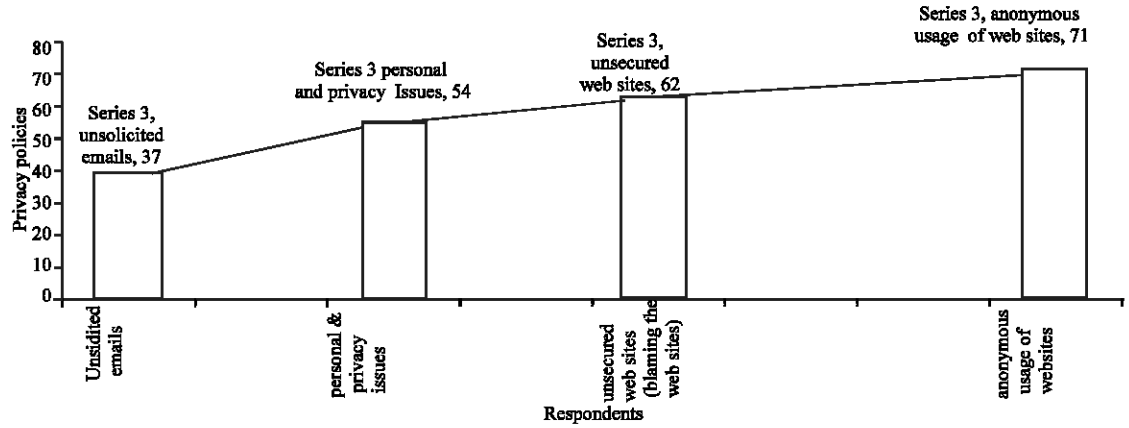


Fig. 1: Survey on internet usage and privacy policies

consumers don't mind companies data mining practices once they are of general, aggregate nature, probably concerning traffic level, sources of referral, or popularity and sequencing of page views.

Even though customers seem to have generally positive opinion on Web site data mining, control over their personal information still is very important. It was found out, that 39% of consumers noted lack of control over who gets the information, and 37% pointed to unsolicited e-mails as strong privacy concerns.

Based on results received, it is possible to draw a general picture of user's persistence on the issue of Web server log data mining. They allow companies to gather aggregate, anonymous, group-level data on Web site and believe most companies are moving in the right direction in terms of user concerns over privacy. Fig 1.

**What can marketers do about internet data mining privacy?:** From the results above 4 tips for online marketers can be suggested to improve user trust and data mining techniques privacy.

**Explain the purpose of data mining:** Data mining always has a certain goal. The survey found that at least some segments of online users lose their negative approach to Internet data mining when they better understand the purpose of it.

**Control of information distribution:** Users are in majority strongly against the sharing of mined data among different companies, for example.

**Provide key trust points to improve e-commerce practice:** That implies clear statements of privacy policy, also privacy software and description or communication about that.

**Adjust privacy methods in respect of different customer groups:** Customers have different opinions and concerns. For example, increased age generally correlates with increased concerns about online privacy. While 51% of respondents age 45 and older were extremely concerned about spam, only 23% of age 16 to 24 year olds felt the same way. Also technology-aware users are less concerned about privacy than technologically unaware people. Among respondents who are more or less familiar with web and non-web technologies only 21% said that they are worried about lack of control and only 13% about companies having access to browsing habits.

These tips will help the companies to improve their privacy performance. But are there any tools for users to protect themselves from access without notification to their personal information by online companies?

**What can users do about internet data mining privacy?:** Several tools have been developed to help Internet users surf the Web anonymously. Anonymity agents and pseudonym agents are useful for Web surfing in which users have no need or willingness to be identified. Negotiation agents and trust engines can assist users in reviewing a services request and determining whether or not to provide the requested data or access.

**Anonymity agents:** The anonymity agents ensure that users. IP address cannot be identified through web server logs.

**The anonymizer:** One of the most well known Web anonymity tools is the Anonymizer, a service that submits

HTTP requests to Web sites on behalf of its users. As a result, the only IP address revealed to the Web site is Anonymizers. But the users IP address is open to the very Anonymizer and to its own ISP.

**The crowds:** There are also anonymity tools that do not require user to trust a single third party to maintain its anonymity. Crowds are an anonymity agent based on the idea that people can be anonymous when they are in a crowd. Rather than submitting HTTP requests through a single third party, Crowds user submits its requests through a crowd, a group of Web surfers running the Crowds software. Crowd's user forwards HTTP requests to a randomly selected member of its crowd. The origin of request can be identified neither by the end server, nor by any crowd member.

**The onion routing:** Another anonymity agent called Onion Routing uses the following algorithm user submits encrypted HTTP request using an onion a layered data structure that specifies symmetric cryptographic algorithms and keys to be used as data is transported to the intended recipient. As the data passes through each onion router along the way, one layer of encryption is removed according to the recipe contained in the onion. The request arrives at the recipient in plain text, with only the IP address of the last onion-router on the path.

**Pseudonym agents**

**The LPWA:** Sometimes users want to establish anonymous but persistent relationships with a Web site to take advantage of customized services, for example. The Lucent Personalized Web Assistant (LPWA), a pseudonym agent helps user do that. LPWA can be used to insert pseudonyms into Web forms that request a user's name or email address. It uses the same pseudonyms every time same user returns to the same site, but uses a different pseudonym at each site.

**Negotiation agents**

**The P3P:** When user wants goods delivered home or something like that he or she just has to provide some personal information. So anonymity or pseudonym agents won't help in this case. Negotiation agents and trust engines can assist users in making decision about providing personal data. For example, the Platform for Privacy Preferences Project (P3P) can provide a rich vocabulary for services to express their information practices and for user to express its privacy preferences. Thus, P3P helps user make informed decisions about when to release personal data. But P3P doesn't protect data itself. User must be sure that services will use data only as was stated initially.

**Trust engines**

**The TRUSTe:** Another privacy tool, a trust engine, is the TRUSTe, a self-regulatory privacy initiative dedicated to building consumers trust and confidence on the Internet through a program in which Web sites can be licensed to display a privacy seal or trust mark on their sites. These marks are currently visual, but can also take shape of digital certificates that can be dealt by negotiation agents and trust engine.

## CONCLUSION

As information exchange as well as mining the information from the web increases, interest in Internet and amount of services and goods online increases, data mining activities can expand rapidly allowing firms to retrieve highly personalized data about customers, which as well implies high privacy violations and concerns. Both marketers and users should follow privacy policy rules. Marketers should pay more attention to level of user trust and couple their data mining efficiency with respect to user privacy.

As for as consumers are concerned, they desperately need purely sensitive information transmission, they should select secure channels and store data securely by tools provided as a result of technological achievements. Based on the survey report, awareness has been created for users, and also aimed to construct a prototype model for securing the computer system protects its data and resources from unauthorized access (intruders, anomalies), tampering and denial of service attacks.

## REFERENCES

Anthony Danna, 2 002. All that Glitters is not gold: Digging beneath the surface of data mining, Journal of Business Ethics, C Kluver Academic Publishers printed in Netherlands, 40: 373-386.

Abdi, H., 2003. Neural Networks. In M. Lewis-Beck, A. Bryman, T. Futing (Eds). Encyclopedia for research methods for the social sciences. Thousand Oaks (CA): Sage, pp: 792-795.

Bonchi, F., F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C.Renso and S. Roggieri, 2001. Web log data warehousing and mining for intelligent web caching, Data and Knowledge Engineering, Vol. 39.

George, R. Milne, 2000. Privacy and ethical issues in database/interactive marketing and public policy: A research framework and overview of the special issue, J. Public Policy and Marketing, 19: 1-6. Electronic ESSN: pp: 1547-7207.

Kate, A. Smith, Alan Ng, 2003. Web page clustering using a self-organizing map for user navigation patterns, Decision Support Systems, 35: 245-256.

Mike Perkowitz, Oren Etzioni, 1999. Towards adaptive web sites: Conceptual framework and case study, Computer Networks, Issues, pp: 11-16.

Myra Spiliopoulou, August, 2000. Web usage mining for Web site evaluation, Association for Computing Machinery, Communications of the ACM, New York, pp: 127-134.

Show-Jane Yen, 2003. an efficient approach for analyzing behaviors in a web based training environment an Int. J. Distance Edu. Tech., pp: 55-71.

Yue-shi lee, Show-Jane Yen, Ghi- Hua Tu, 2004. Mining traveling and purchasing behavior of customers in electronic commerce environmen, Proceedings of IEEE International Conference on E-technology, E-commerce and E-services, pp: 227- 230,