A Framework for Personal Web Usage Mining

Yongjian Fu
Department of Computer Science
University of Missouri-Rolla
Rolla, MO 65409-0350
yongjian@umr.edu

Ming-Yi Shih
Department of Computer Science
University of Missouri-Rolla
Rolla, MO 65409-0350
mingyi@umr.edu

Abstract

In this paper, we propose to mine Web usage data on client side, or personal Web usage mining, as a complement to the server side Web usage mining. By mining client side Web usage data, more complete knowledge about Web usage can be obtained. A framework for personal Web usage mining is proposed. Some related issues and applications of personal Web usage mining are also discussed.

Keywords: Personal Web usage mining, Web usage mining, Web mining, Web log.

1. Introduction

With the increasing popularity of the Web, it is no surprise that Web mining has attracted lots of attentions. An important area in Web mining is Web usage mining, the discovery of patterns in the browsing and navigation data of Web users. Web usage mining has been an important technology for understanding user's behaviors on the Web.

Currently, most Web usage mining research has been focusing on the Web server side. The main purpose of research is to improve a Web site's service and the server's performance. Data sources for Web usage mining are primarily Web server logs. Although it is very important and interesting to investigate server side issues, we argue that an equally important and potentially fruitful aspect of Web usage mining is the mining of client side usage data. Since this studies individual's Web usage, rather than group behaviors, we term this as *personal Web usage mining*.

In this paper, we advocate personal Web usage mining as a complement to the server side Web usage mining. Like server side Web usage mining, personal Web usage mining will apply data mining algorithms to Web usage data. However, in personal Web usage mining the data source for mining and the goal of mining are different from these in server side Web usage mining.

We first explain what personal Web usage mining is and then give a framework for personal Web usage mining. Some related issues and applications are discussed. Considerations for implementation will also be presented.

The paper is organized as follows. In Section 2, related work and background are discussed. An introduction to and a framework for personal Web usage mining are proposed in Section 3. In Section 4, applications of personal Web usage mining and implementation issues are discussed. Section 5 concludes the paper.

2. Related work and background

Currently, Web usage mining finds patterns in Web server logs. The logs are preprocessed to group requests from the same user into sessions. A session contains the requests from a single visit of a user to the Web site. During the preprocessing, irrelevant information for Web usage mining such as background images and unsuccessful requests is ignored. The users are identified by the IP addresses in the log and all requests from the same IP address within a certain time-window are put into a session [6]. Different heuristics have been developed to deal with the inaccuracy due to caching, IP sharing or blocking, and network congestion [6,17].

Previous studies in Web usage mining include association rules and sequential patterns mining [1,5,16], user clustering [9,15], personalization [5,15], adaptive Web sites [8,13,14], and Web OLAP and warehouse [2,20].

Some common characteristics of previous studies in Web usage mining are described as follows.

- Their goal is to improve Web services and performance through the improvement of Web sites, including their contents, structure, presentation, and delivery.
- They focus on the mining of server side data. Their data sources are almost exclusively server logs, sometimes with site structure and/or page contents.
- They target groups of users instead of individual users. It is overwhelming for a Web site to deal with users on an individual basis.

Based on these observations, we propose to mine Web usage data on the client side. By looking into a user's Web usage data, we hope to understand the user's interests, behaviors, and preferences. In other words, we are building the user's Web profile. We call this personal Web usage mining since it focuses on personal Web usage, in contrast to previous Web usage mining which focuses on group Web usage.

Some of the reasons we advocate personal Web usage mining are as follows.

- The goal of personal Web usage mining is to help and enhance individual users Web use. It intends to make the Web easier to use from a single user's point of view.
- Client side data provide a more accurate and complete picture of a user's Web activities. The client side data are clean of the uncertainties that plaque server data such as caching and network congestion. Besides, we will be able to collect user's footprints across the Web (on tens or even hundreds sites), rather than a single Web site. Moreover, we may include other actions besides "point and click", such as "save as", "print", "email", and so on.
- We can achieve true individualism and personalization. Although we can find large groups with similar interests, it is safe to say that no two persons' needs are the same.
- Since it is done at the client's side, users have full control of what, when, and how their data can be used for mining. Unlike previous Web usage mining, the privacy of users will be protected.
- Personal Web usage has increased significantly recently, to a level that it generates sufficient data to perform meaning mining tasks. For example, a recent survey showed that Americans on average spent 13 hours browsing 720 pages per month [19].

Some researchers are building intelligent agents or Internet agents that will help individuals use the Web. For example, many agents were built for information filtering and gathering on the Web. WARREN [18] is a multi-agent system for compiling financial information. WEBMATE [4] edits a personal newpaper. WebSifter [10] is a meta-search agent which uses taxonomy to improve search on the Web. Other examples include home page finder [7], user interface learning agent [3], and Web browsing assistant [11].

While our proposal shares similar goals with many those agents, our approach is automatic that it does not require user's explicit input. Moreover, we take a systematic approach to collect and comprehend user activities. We provide a general framework for collecting, mining, and search/query personal usage data, which may be employed by various agents. In addition, the applications of our framework are not limited to agents, but may include many others such as Web personalization, learning, and security as we discussed in Section 4.

An applet based user profiler is proposed in [15] which can find the accurate time a user spent on a page. In [12], a Web warehouse is proposed and some data mining issues are discussed. The warehouse stores Web pages and links visited by a user, which may be used for data mining such as association rules mining.

Although some aspects and pieces of personal Web usage mining may be around in various areas such as intelligent agent, Web warehousing, and Web usage mining, the important of this research direction deserves a more dedicated focus and study.

In this paper, we try to define the scope and relevance of the research. We then give a framework for personal Web usage mining. It is our hope that this position paper will attract more interests in personal Web usage mining.

3. Personal Web usage mining

In personal Web usage mining, two kinds of user Web activities are recorded for analysis: remote activities and local activities. The remote activities include requests sent by a user to a Web server. Such kind of click stream data includes the URLs of pages as well as any keywords, queries, forms, and cookies sent with the URL. The local activities include actions the user can take at his or her desktop without the knowledge of Web servers. They include, but are not limited to, the following.

- Save a page
- Print a page
- Click **Back** on browser
- Click **Forward** on browser
- Click **Reload** on browser
- Click **Stop** on browser
- Email a link/page
- Add a bookmark
- Minimize/maximize/close window
- Change visual settings such as font size

The remote activities can be captured by almost all Web browsers. Besides, the browsers also cache the Web pages in most cases. The local activities can be recorded by an activity recorder, which is a client side program running on top of the browser. These two kinds of activities are put together into an activity log. Each entry in the activity log will contain a timestamp and an activity. Some will contain extra information such as URL, cache address, keyword, cookie, email address, and font size. The schema of the log looks like this:

(timestamp, activity, [URL], [cache address], [keyword], [cookie], [email address], [other optional fields])

The framework for personal Web usage mining is given in Figure 1. There are four major modules in the framework: logging, data warehousing, data mining, and tool/application. In the logging module, user Web activities are stored into the activity, as well as the cached pages. In the data warehousing module, the logs and cached pages are cleansed, extracted, transformed, aggregated, and stored in a data warehouse. The data warehouse will facilitate search, query, and OLAP operations, in the mean time providing data sources for mining. In the data mining module, various data mining algorithms are applied to the data in the data warehouse, whose findings will be used by the tools and applications in the tools/application module. Some example applications are discussed in Section 4.

A user's actions are recorded by the Web browser and the activity recorder. The browser will also cache the Web pages requested by the user. As shown in Figure 1, the user's activities will be the source for data warehousing and data mining, whose results will be employed by the tools and applications, which, in turn, aim to help the user with his or her Web activities. In such a user-centric way, personal Web usage mining has great potentials in bringing Web to people.

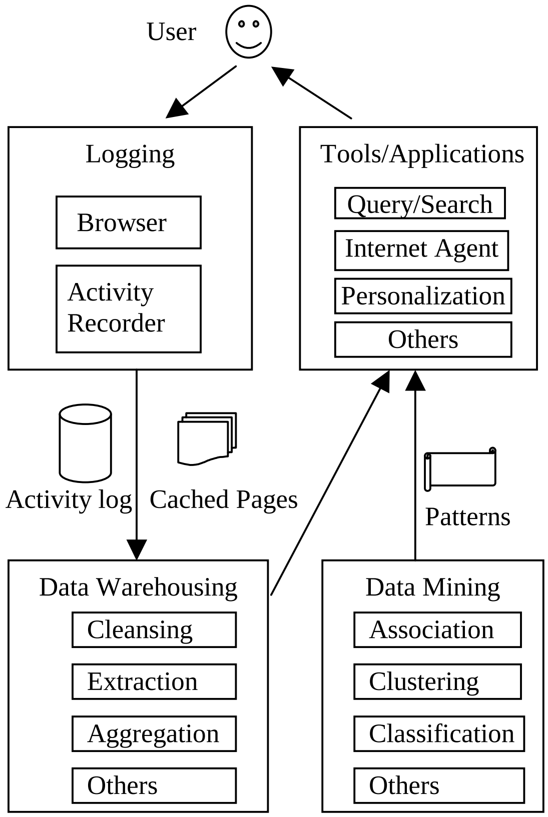

Figure 1. A framework for personal Web usage mining.

Several kinds of patterns can be discovered from the data using various data mining techniques. The following is a list of patterns that may be mined. Each of them can be a component of the data mining module in the framework.

- Summarization
  The data can be summarized and abstracted to find general patterns. For example, if the user prints ten pages about music from the Yahoo site on February 22, 2002, the individual activities can be summarized as “10 pages on Yahoo related to music printed on February 22, 2002”. This can be done using the OLAP operations and data summarization techniques such as attribute-oriented induction.
- Classification
  Various classification techniques can be used to classify the pages. For example, a page can be defined as interesting if it is saved, printed, emailed, or viewed for more than five minutes. A classifier can be built to classify pages as interesting or uninteresting based on its keywords, URL, access time, etc.
- Association
  Associations among the activities can be found using the association rule mining algorithms. For example, an association rule may be found from the activity log “when a page on computer games is browsed, 80% of times a page on computer games is printed also”.
- Clustering
  Various clustering techniques may be employed to cluster pages, links, and activities. For example, the pages with similar topics can be clustered to find groups among pages.
- Sequence
  A sequence is a list of events that happen sequentially, e.g., “if a page on a concert is browsed, 80% of times another page on tickets is printed within 10 days”. Techniques for sequence mining are derived from techniques for association rule mining.

Other kinds of patterns may also be discovered. It should also be pointed out that our framework does not restrict the techniques for mining these patterns.

4. Discussions

Personal Web usage mining provides a basis for developing further applications. Many applications are described in Section 4.1. In Section 4.2, some considerations for implementing a prototype system are outlined.

4.1 Applications

Based on the patterns found and the original cache and log data, many applications can be developed. Some examples are given as follows.

- Personalization of Web
  A user’s view of Web can be personalized dynamically and incrementally based on the patterns. Many Web sites let users personalize their view of the sites. A user usually selects or de-selects a set of pre-defined sections and topics. It cannot really achieve true individualism. Besides, it requires explicit user intervention whenever a user preference changes.
- Internet agents
  Internet agents are intelligent agents that work on the Internet. Such agents can be built to perform various tasks assisting Web usage, including recommendation, automation, scheduling, reminding, and monitoring. The agents will build a user profile from the patterns and logs, to be used in assisting the user.
- Study of learning process
  Since the activity log records every step a user takes to achieve his or her goal. It essentially records the user’s learning process. It may provide insights as how a user learns to navigate, search, and explore Web sites.
- Intrusion detection and security
  A template of a user’s regular usage can be established by analyzing the patterns. A security system can be developed which will cause an alert if the current user’s behavior is widely different from the template. The security system can

also restrict the user’s access to certain online contents.

Other applications can also be developed in the framework for personal Web usage mining. Our framework is flexible to easily accommodate new applications.

4.2 Implementation

Currently, we are investigating implementation issues for a prototype system in personal Web usage mining. Besides, an intelligent agent for Web page recommendation based on personal Web usage mining is being developed.

The activity recorder in the logging module can be implemented as a Java wrapper of a browser. The wrapper will capture the local activities taken by the user such as **save** and **print**. The data warehouse will be implemented using the IBM DB2 system. The data mining tool Intelligent Miner from IBM will be used to mine the data. Besides, a document processing component will be developed to process the cached pages, such as keyword extraction.

A Web page recommendation agent will also be built as a sample application of personal Web usage mining. Given a list of previously browsed pages, the agent will extract browsing patterns such as frequency, time, and page size. The agent will then recommend a set of Web pages that may be interested to the user. The recommendation procedure of the agent is outlined as follows.

```
1. For each candidate page
   1.1. Extract keywords of the page
   1.2. Compute the interestingness of the
        page by comparing it with every
        previous page. The interestingness
        measure is based on the keywords
        and browsing patterns.
2. Order the pages based on their
   interestingness
3. Recommend the pages in decreasing
   order of interestingness or only those
   pages which meet a minimum
   interestingness threshold.
```

Various features, interesting measures, ranking schemes will be investigated. In the next phase, an incremental algorithm will be developed which updates the user profile incrementally.

5. Conclusion

We propose the mining of client side Web usage data, which is termed personal Web usage mining. Based on our analysis, it is an interesting and important research area. A framework for personal Web usage mining is developed. Some applications and implementation issues are discussed.

Currently, we are working to implement the prototype system and the recommendation agent. Some part of the prototype system has been designed. In the future, more applications of personal Web usage mining will be developed.

References

1. J. Borges and M. Levene, *Mining Association Rules in Hypertext Databases*, Proc. 1998 Int'l Conf. on Data Mining and Knowledge Discovery (KDD'98), 149-153, 1998.
2. A. Buchner and M. Mulvenna, *Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining*, SIGMOD Record, 27(4), 1998.
3. A. Caglayan, M. Snorrason, J. Jacoby, J. Mazzu, R. Jones, K. Kumar, Learn Sesame -- a Learning Agent Engine, Applied Artificial Intelligence, 11:393--412, 1997.
4. L. Chen and K. Sycara, Webmate: A Personal Agent for Browsing and Searching, Proc. of Second International Conf. On Autonomous Agents (Agents 98), Minneapolis, MN, May, 1998.
5. R. Cooley, B. Mobasher, and J. Srivastava, *Web Mining: Information and Pattern Discovery on the World Wide Web,* Proc. Int. Conf. on Tools with Artificial Intelligence, Newport Beach, CA, 558-567, 1997.
6. R. Cooley, B. Mobasher, and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, Journal of Knowledge and Information Systems, 1(1), 1999.

7. O. Etzioni. Moving up the information food chain: Deploying softbots on the world-wide web. In Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI96) , Portland, OR, 1996.

8. Y. Fu, M. Creado, and M. Shih, Adaptive Web Site by Web Usage Mining, International Conference on Internet Computing (IC'2001), Las Vegas, NV, 28-34, June, 2001.

9. Y. Fu, K. Sandhu, and M. Shih, *Clustering of Web Users Based on Access Patterns*, International Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), San Diego, CA, 1999.

10. L. Kerschberg, W. Kim, and A. Scime, WebSifter II: A Personalized Meta-Search Agent Based on Weighted Semantic Taxonomy Tree, International Conference on Internet Computing (IC'2001), Las Vegas, NV, 14-20, June, 2001.

11. H. Lieberman. Letizia: An Agent That Assists Web Browsing. In 1995 International Joint Conference on Aritifical Intelligence, Montreal, CA, 1995.

12. S. Madria, S. Bhowmick, W. K. Ng, and E. P. Lim, *Research Issues in Web Data Mining*, Proc. DAWAK'99, Florance, Italy, Sept. 99.

13. B. Mobasher, R. Cooley, and J.Srivastava, *Creating Adaptive Web Sites Through Usage-Based Clustering of URLs*, Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999.

14. M. Perkowitz, and O. Etzioni, *Adaptive Web Sites: Automatically Synthesizing Web Pages*, Proceedings of Fifteenth National Conference on Artificial Intelligence, 727-732, 1998,

15. C. Shahabi, A. Zarkesh, J. Adibi, and V. Shah, *Knowledge Discovery from Users Web-Page Navigation*, In Proceedings of the IEEE RIDE97 Workshop, April 1997

16. M. Spiliopoulou, L. Faulstich, C., and K. Winkler, *A Data Miner analyzing the Navigational Behaviour of Web Users*. Workshop on Machine Learning in User Modelling of the ACAI'99 Int. Conf., Creta, Greece, July 1999

17. M. Spiliopoulou, The laborious way from data mining to Web log mining, International Journal of Computer Systems Science and Engineering, Vol. 14, No. 2, 113-125, 1999.

18. K. Sycara, K. Decker, A. Pannu, M. Williamson, and D. Zeng, Distributed Intelligent Agents, IEEE Expert, 11(6), 1996.

19. http://www.computeruser.com/news/00/11/18/news6.html.

20. O. R. Zaiane, X. Xin, and J. Han, *Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs*, Proc. Advances in Digital Libraries, 19-29, 1998.