A Framework for Web Usage Mining in Electronic Government

Ping Zhou, Zhongjian Le
School of Information Management,
JiangXi University of Finance and Economic, NanChang ,China 330013
Zp_jx@126.com

Abstract. Web usage mining has been a major component of management strategy to enhance organizational analysis and decision. The literature on Web usage mining that deals with strategies and technologies for effectively employing Web usage mining is quite vast. In recent years, E-government has received much attention from researchers and practitioners. Huge amounts of user access data are produced in Electronic government Web site everyday. The role of these data in the success of government management cannot be overstated because they affect government analysis, prediction, strategies, tactical, operational planning and control. Web usage miming in E-government has an important role to play in setting government objectives, discovering citizen behavior, and determining future courses of actions. Web usage mining in E-government has not received adequate attention from researchers or practitioners. We developed a framework to promote a better understanding of the importance of Web usage mining in E-government. Using the current literature, we developed the framework presented herein, in hopes that it would stimulate more interest in this important area.

1 Introduction

The recent years have seen the flourishing of research in the area of Web usage mining from both the research and practice communities. With the rapid growth and development of electronic government as well as the ease and speed with which government affairs can be carried out over the Web, one of the important application fields of Web mining is electronic government systems. Electronic government is one of the most appropriate applications of data mining, that is because electronic government domain is very easy to suit the conditions of data mining: richest and the most common source of data, automatically generated data. The result of data mining can be quickly converted into the government behavior, at the same time the policies of government derived from the mining can be evaluated in time.

Web Mining is that area of Data Mining, which deals with the extraction of hidden and interesting knowledge from the large volume of Web documents and records[1] . It is a comprehensively integrated technique, involving Internet, Artificial intelligence, Computer language, informatics, statistics etc. Web Mining can be broadly divided into three classes[2]: content mining, structure mining and usage mining. Web content mining is that part of Web Mining, which focuses on the raw information available on Web pages or the searched results(e.g. words); Web Structure Mining is that part of Web Mining, which focuses on the structure of Web site including intra-page structural information and inter-page structural information presented on Web pages(e.g., links to other pages).

Web Usage Mining is that part of Web mining, which deals with the extraction of knowledge on users' access patterns and user behavior from data collected from the main sources: Web servers, proxy servers, Web clients (including registration data and user profile information) using some kind of data mining techniques.

Web content mining and Web structure mining focus on the raw information on web pages. In Web usage mining, the focus is on data describing the usage pattern of Web pages, including: Web server side access log files, proxy side log files, client side log records, user registration information, user suggestions and user request information etc, which can be used to track the behavior, the goal and the motivation of users producing these data. Exploiting these usage data can largely help government to identify the citizen's or the business' needs, requests, requirements and behaviors etc and make corresponding policies. Hence in E-government, Web usage mining is the main Web mining. The remaining of this paper is organized as follows: Initially, in section 2 we propose a framework for Web usage mining in e-government. Then in section 3 some concrete applications of Web usage mining in e-government platform are presented. Finally, in section 4 the conclusions are drawn.

2 Web Usage Mining in Electronic Government

As shown in figure 1, a framework for Web Usage Mining in E-government is presented. There are four main tasks for performing Web Usage Mining in E-government. This section presents an overview of the tasks for each step.

2.1 Data Collection

Server (including Web server and proxy) side, client side and user registration information are the present main three sources of usage data on E-government Web site.

2.1.1 Server side

Web server side usage data mainly consist of: Web server log files, Cookies, submission data and the statistic information from the external third side, which all implicitly record the browsing behavior of site visitors.

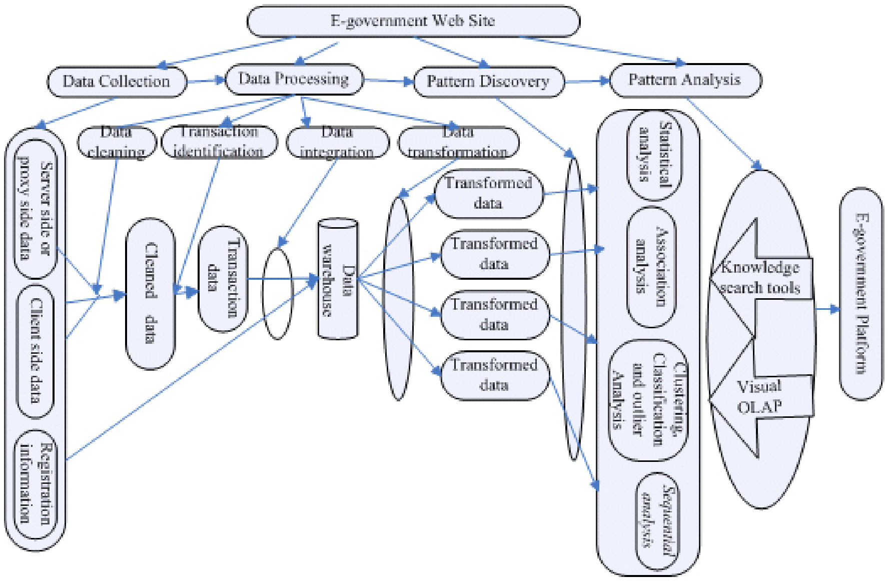

Fig. 1. Web usage mining in E-government framework

(1)Web server side log files: important sources of data for performing Web usage mining. The data recorded in server logs reflect the access of single server by multiple users. These log files can be stored in various formats such as common log or extended log formats.

(2) Cookies: Cookies are tokens generated by the Web server for individual client browser in order to automatically track the site visitor. When the user visits the same Web site again, she/he can be identified immediately.

(3)Submission data: various openly entered and submitted data by users, which basically reflect the user's interest and preferences, merged with the government Web site's structure, content, key words semantics and domain knowledge can perfectly discover the visitor's behavior and motivation.

(4) The statistic information: information on some users can be bought from external channels, for example, the third database.

Proxy side usage data: collecting navigation data at the proxy level in many respects is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers, which help to reveal the visitors' behavior and action behind the proxy.

2.1.2 Client side

Usage data on client side, which records data of single user accessing groups of web servers, provide detailed information on actual user behavior [3]. The client side data are more credible than that of the server side, because it avoids the high caching and IP errors.

2.1.3 Registration information

Registration data refers to the relevant information submitted to the Web server through Web pages, including logon message, suggestion on Web etc., which must be integrated with all accessing log files in order to increase the accuracy of Web usage mining in E-government.

2.2 Data Processing

In practice, data in E-government systems provided by the data sources described above are usually inconsistent, incomplete, redundancy and obscure. Data processing means converting the data contained in the various available data sources in E-government Web site into the data necessary for useful pattern discovery. Only the clean, accuracy and simple abstract data can be used for mining analysis. Data processing usually comprises data cleaning, transaction identification, data integration and data transformation.

2.2.1 Data cleaning

Data cleaning refers to removing all the data tracked in Web logs that are invalid for web usage mining purposes [4,11,12,14]. We only want to keep the entries that carry relevant information. Therefore, data cleaning is used to eliminate the irrelevant entries from the log file, e.g. requests for graphical page content; requests for any other file which might be included into a web page. The data collected from Web servers or Proxy servers, which record the interactions between groups of users and multiple servers, usually need to be cleaned before use for mining. In contrast, data collected from client side are relevantly cleaner because of less user interference. Moreover, the data input by users should be confirmed, restructured and formatted for pattern discovery.

2.2.2 Transaction identification

Before performing Web usage mining in E-government, transaction should be predefined according to the characteristics of pattern mining. Usually, different user sessions analysis can produce different transaction to extract different useful information.

2.2.3 Data integration

Web usage data in E-government, which distributed over various sources of data as described above, are regularly localized or even personal and difficult to share. Only merging these data can it be applied to Web usage mining to extract the truly useful information for government.

2.2.4 Data transformation

Data transformation refers to mobilization and conversion of the existing integrated data for different analysis and decision making tools.

2.3 Pattern Discovery

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition etc.

Methods developed from other fields must take into consideration the different kinds of data abstractions and prior knowledge available for Web Mining. For example, in association rule discovery, the notion of a transaction for market-basket analysis does not take into consideration the order in which items are selected. The following are common methods to extract knowledge about visitors to government Web sites.

2.3.1 Statistical Analysis

Statistical methods are the most common method to extract knowledge about visitors to a government Web site. By analyzing the session file, one can perform different kinds of descriptive statistical analysis (frequency, mean, median, etc.) on variables such as page views, viewing time and length of a navigational path. Despite lacking in the depth of its analysis, this type of knowledge can be potentially useful for improving the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for government's decision making.

2.3.2 Association analysis

Association analysis can be used to relate pages that are most often referenced together in a single server session[4]. In the context of Web Usage Mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. These pages may not be directly connected to one another via hyperlinks. In actual government Web site design, applying these related pages can help Web designers to restructure their Web site so that it is easy for citizens to access their wanted pages. The association rules may also serve as a heuristic for pre-fetching documents in order to reduce user-perceived latency when loading a page from a remote site.

2.3.3 Clustering, Classification and outlier Analysis

Clustering is a technique to group together a set of items having similar characteristics[5]. Clustering of users tends to establish groups of users exhibiting similar browsing patterns. Such knowledge is especially useful for inferring user demographics in order to perform service segmentation in E-government applications or providing personalized Web content to the citizens. On the other hand, clustering of pages will discover groups of related Web pages. This information is useful for Internet search engines and Web service providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs[4].

Classification is the task of mapping a data item into one of several predefined classes [6]. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category.

Outlier refers to the data that do not comply with the general behavior or model of the data. Data mining in many fields often try to minimize the influence of outliers or eliminate them all together. This, however, would result in the loss of important information because the outliers may be of particular signal reflecting abnormal events, irregularities, such as the case of criminal activity, fraud etc. Outlier detection can reveal points that behave “anomalously” with respect to other observations. Examining such points can reveal clues to solve the problem at hand. In other cases, the sudden appearance of a large number of outliers can point to a change in the underlying process that is generating the data.

Thus, in E-government, outlier detection and analysis are a very important and interesting data miming task. Analyzing these outliers, government departments can take action accordingly in time, predict the trend of society development, consequently enhance the government service ability and the capability of reining the complicated events etc.

2.3.4 Sequential Patterns analysis

The technique of sequential pattern analysis attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes[1]. By using this approach, government Web observers can predict future visit patterns which will be helpful in placing especial messages aimed at certain user groups. Other types of temporal analysis that can be performed on sequential patterns include trend analysis, change point detection or similarity analysis.

2.4 Pattern Analysis

Pattern analysis is the last step in the overall Web Usage mining process in E-government as described in Figure 1. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

2.5 Applying to E-government platform

The previous sections described how the knowledge on the public we have been discussing in this paper can be obtained.

Examples are:

- cleaning, tracking, browsing and discovering usage data,
- being alerted to abnormal outliers,
- knowledge visualization,
- querying knowledge in similar natural language.

To be really effective in every use, the end-user functionality needs to be integrated into, for example, desktop applications. We can also see these discovered knowledge as part of an organization, being integrated into more specific government applications and solutions. A few examples of the possibilities are given in the next section. This is by no means an exhaustive list.

3 Applications of Web usage mining in E-government

In this section we present the main applications of Web usage mining in E-government, as shown in Figure 2. The application of Web usage mining to E-government is a procedure which translates citizen or business' usage data on government Web site into valuable knowledge which can provide various decision supports in government affairs, such as: finding out the preferences/interests/desires of citizen and improving the citizen or business satisfaction; Restructuring the Government Web Site and Increasing the System Performance; enhancing the government planning and promoting government innovation; improving the analysis and decision making of government etc.

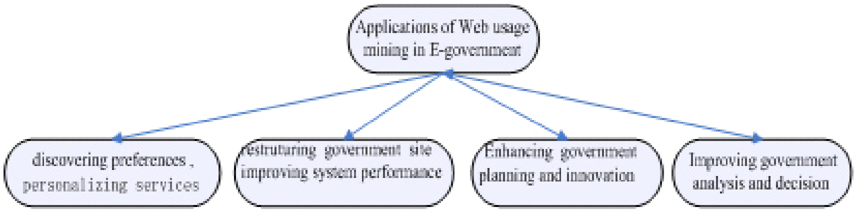
Applications of Web usage mining in E-government
discovering preferences , personalizing services
restrutaring government site improving system performance
Enhancing government planning and innovation
Improving government analysis and decision

Fig.2. Application areas for Web Usage Mining in E-government

3.1 Discovering the Preferences/Interests of Citizens and Providing Personalization Services

User's actions can observe his/her behavior and derive his/her preferences/interests [7] [8]. For example, Where does he click, how long does she remain on certain pages, what words does he search for, from which websites did she come, interactions done with this website, and so on. A list of keywords from pages that a user has spent a significant amount of time viewing is compiled and presented to the user. Through feedback analysis of the keyword list and his profile, recommendations for other pages within the site are made. Personalizing the Web experience for a user is the holy grail of Web-based applications based on her/his registration data and usage behavior which were often used to discover clusters of users having similar access patterns and their respective preferences. Every user will be assigned to a single cluster based on their current traversal pattern. The links that are presented to a given user are dynamically selected based on what pages other users assigned to the same cluster have visited. Tracking a user as he or she browses the government Web site and identifying the links that are potentially interesting to the user are necessary for understanding citizen behavior, their preferences and desires and providing each citizen excellent and personalized services and corresponding management responding to their needs[14]. The whole process can be described in figure 3.

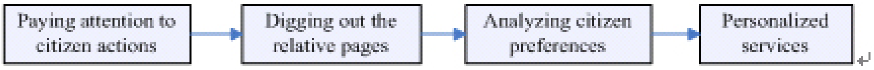
Paying attention to citizen actions
Digging out the relative pages
Analyzing citizen preferences
Personalized services

Fig.3. The process of providing personalization services

3.2 Restructuring the Government Web Site and Increasing the System Performance

The content and structure of a government Web site are not constant. The designers of a government Web site can not only completely rely on specialists in the field to design the structure of the Web site, they should dynamically restructure the structure and content of government Web site according to the result derived from Web usage mining in E-government in order to convenience the citizen and business access[9]. For example, you should restructure the Web site with respect to the frequent access paths of visitors, on the one hand, user access time is saved. On the other hand, expenditure of Web site is significantly economized.

The following three suggestions can be used to optimize the government Web site structure:

(1)Mining the Web log files can extract the related access pages, then new links are added between them to facilitate the visitors.

(2)Using the path analysis techniques to find out the most frequent access path and putting the important message over there will enhance the appetency of government to citizens and improve the service qualities.

(3)Mining the Web log files can discovery expected information position by citizens. If the expected access position frequency is higher than that of the actual, a navigation link can be established between them to optimize the structure of the Web site.

Web usage mining in E-government can effectively observe and analyze the users of government Web site and their behaviors and actions, which inscribe the knowledge in human or social domain essentially related to human action and benefit the system improvement, e.g., security is an important issue for government site, by tracking the users' access pattern and access paths, the hidden intrusion can be detected fleetly and easily.

3.3 Enhancing the Government Planning and Fostering Government Innovation

Through employing data mining technology, government can manage with reason the human resources, material resources and information resources to harmonize the relation between resources inside government and those outside government, such as, the whole process from program planning to program implementing can exchange and share the same data. OLAP(Online analysis and process) can optimize the project flow to best fit together the society resources, which will largely reduce the cost that social resources circulate and information transfers, promote the government planning to be more scientific, informational and intelligent. A scientific and effective government planning can be achieved by intensifying government's real time control and management with intelligent technology and visualization technology.

Mining Web usage data on government site can quickly obtain the information about government affairs to make the government grasp the society development trends in time, which at the same time make management and redeployment of society resources more systematic, macroscopical and dynamic to improve the government innovation capability.

Innovation, which does not only limit to administrative method and government affair process flow, and that including constituting government development strategies and public policies, is the basic requirement that change government from functional organization to service organization. Beside that, mining usage data on government site can significantly improve governmental responsiveness to all the issues, from day-to-day citizen request to paroxysmal events, also foster the government work innovation and personnel making increasing.

3.4 Improving the Government Analysis and Decisions

Mining and analyzing the vast quantities of usage data on government site are of great benefit to the government analysis and decisions. Government decisions making through analyzing information about citizens(e.g. suggestion, request, desire, etc) are more likely to be acceptable to the majority of people. Citizen usage data on government site provide a source of special insight, information, knowledge, and experience, which contributes to the soundness of government solutions to public problems[10]. Citizen participation in public affairs also serves to check and balance political activities. A cross section of citizen participation in the decision-making process reduces the likelihood of government leaders making self-serving decisions. Mining these citizen usage data on government Web site can also legitimize a program, its plans, actions, and leadership. Unsupported leaders often become discouraged and drop activities that are potentially beneficial to citizens. On the other hand, citizen's voluntary suggestion can reduce the cost of personnel needed to carry out many of the duties associated with community action. In a word, mining citizens' usage data can identify and extract the hidden important information to provide all levels government department with effective decision support[13]. For example, mining the client side log files can obtain the opinion of citizen and effectively assist government department to make scientific and rational adjustment to their desires.

4 Conclusions

In this paper, a framework for Web usage data in E-government was presented. This framework consists of five parts. We believe that this framework can provide a number of benefits to the different stakeholders within public authorities who need to capture the hidden and tacit knowledge on the citizens, businesses or other organizations. At the same time, important applications of Web usage mining in E-government platform were described, which provide analytical help for decision making, for monitoring or for revisions. Decision makers in government agencies should fully make use of these mined valuable information to modulate strategy and tactics accordingly, dynamically design government site to satisfy citizen and business and improve the government affair service efficiency.

References:

1. Jiawei Han and Micheline kamber, *Data Mining Concepts and Techniques*(second edition)(China Machine Press, Bei Jing, 2006).

2. Federico Michele Facca, Pier Luca Lanzi, Mining Interesting Knowledge from Weblogs: a survey, *Data & Knowledge Engineering (*53), 225－241(2005).

3. K.D. Fenstermacher, M. Ginsburg, Mining Client-side Activity for Personalization, *in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems* (WECWIS_02), 205－212(2002).

4. Jaideep Srivastava , Robert Cooley , Mukund Deshpande et al ,Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data[J] . *SIGKDD Explorations* , 1 (2), 12 – 23(2000).

5. MOBASHER B, COOLEY R, SR IVASTAVA J, Creating Adaptive Web Sites Through Usage-based Clustering of URLs[C], *Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop* (KDEXp99) , 19 – 25(1999).

6.LIU JG, WU W P. Web Usage Mining for Electronic Business Applications [J], *Machine Learning and Cybernetics*(2) ,1314-1318(2004).

7. Mobasher B ,Cooly R. Srivastara J . Automatic Personalization Based on Web Usage Mining[J], *Communications of the ACM*, ,43 (8) ,142- 151(2000).

8. Przemysław Kazienko , Michał Adamski., AdROSA—Adaptive Personalization of Web Advertising. *Information Sciences* (177) ,2269－2295*(*2007).

9.Elisabeth N. Bui, Brent L. Henderson and Karin Viergever, Knowledge Discovery from Models of Soil Properties Developed Through Data Mining, *Ecological Modeling* (5), 431-446(2006).

10. Milakovich, M. and Gordon, G. ,Public Administration in America (7th Ed.),Bedford/St. Martin's (New York, 2001).

11.C.R. Anderson, A Machine Learning Approach to Web Personalization, Ph.D. thesis, *University of Washington*(2002).

12. B. Diebold, M. Kaufmann, Usage-based Visualization of Web Localities, in: *Australian symposium on information visualization*, 159－164(2001).

13. Paul Beynon-Davies, Constructing Electronic Government: the case of the UK inland revenue, *International Journal of Information Management (*25),3－20(2005).

14.Haibin Liu, Vlado Keselj, Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users' Future Requests, *Data & Knowledge Engineering* (61), 304－330(2007).