Mining Online Users' Access Records for Web Business Intelligence

Simon Fong
Faculty of Science and Technology
Universidade de Macau
Macao
simon@enetique.com.sg

Serena Chan
Omron Asia Pacific
e-Business Development Team
Singapore
serenachan@omron.com.sg

Abstract

This paper discusses about how business intelligence on a website could be obtained from users' access records instead of web logs of "hits". Users' access records are captured by implementing an Access-Control (AC) architectural model on the website. This model requires users to register their profiles in an exchange of a password; and thereafter they have to login before gaining access to certain resources on the website. The links to the resources on the website have been modified such that a record of information about the access would be recorded in the database when clicked. This way, data-mining can be performed on a relatively clean set of access records about the users. Hence, a good deal of business intelligence about the users' behaviors, preferences and about the popularities of the resources (products) on the website can be gained. In this paper, we also discussed how the business intelligence acquired, in turn, can be used to provide e-CRM for the users.

1. Introduction

Nowadays most companies have moved into e-business where they use Internet as a platform to interact with their customers. This has created more touch points than the traditional face-to-face visits, and therefore we have a new realm called Web Business Intelligence to deal with.

In Web Business Intelligence, for targeting the business areas of decision support and personalized service, we have identified three main technologies namely Data-warehousing (that includes data-mining modules and decision support modules), Web log analyzer [1] and Web spy [2]. Each one of them provides certain insight or understanding of one of the many aspects of a business respectively. In particular, we state that having a web log analyzer alone is not sufficient to provide a complete picture of business intelligence. Nevertheless, web-log analysis gives very good detailed reports on web site traffic statistics. Other information such as competitors, market positions, and most importantly the customers and their behaviors on the website needed also to be known. In our case, e-Personalization that means personalizing web-content according to the customer's interest/needs will be guided by the knowledge that we have on him.

To realize the objectives of Web Business Intelligence, we propose an Access-control (AC) architecture that can progressively built upon a basic website. The objective is to distribute the functions of e-CRM and business intelligence using an agent-architecture and to monitor users' accesses on the website.

2. Hybrid AC Architecture

We present in this section our hybrid model for e-CRM and data-mining for business intelligence, that is tailored to meet the requirements for easy integration and upgrade for some primitive web-server to operate in e-business environments. The distinguishing features of this architecture are the integration of the distributed software agents each takes a specific role for producing a particular result, and the users' access monitoring system on the website. It supports the ability for mining to be performed at remote sites using mobile agents provides insightful business intelligence reports, and it also feeds back the results to some rule-based systems residing at the server for doing e-Personalization. This helps it to be easily adopted by most existing web-sites who want to have complete functions of e-CRM and BI, but yet reluctant to replace the whole e-business platform.

The hybrid model operates on the principle of letting the website with only some slight modification to act as the portal for capturing the user's particular, their trails, and also for providing them a better online service via e-Personalization. Thus, it has the option of using the client-server model or the mobile-agent model or an integrated approach involving both. We suggest that a web business should evolve progressively by taking a phrasal approach. The difference in performance between the three models are how fast the BI can be obtained and how much real-time the e-CRM can take place. The components of the hybrid AC architecture illustrated in Figure 1 are as follows:

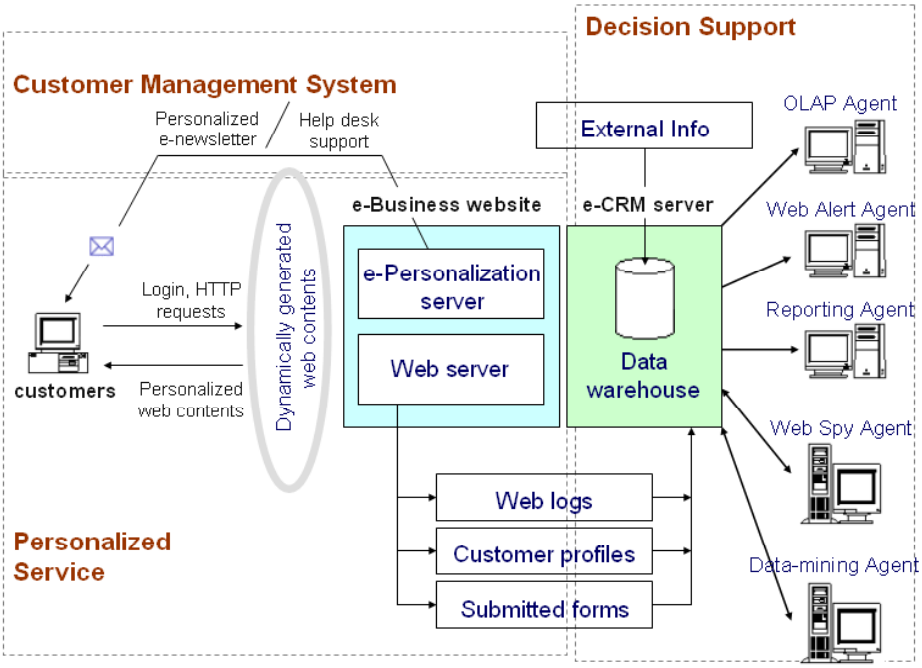

Figure 1. Agent-based Framework for BI and e-CRM

It can be seen that with all the components working together, the three business areas are covered, as shown in the figure. The highlight of this paper is on about how a login and access control mechanism is implemented on the web site; thereafter Web business intelligence can be discovered.

3. Access-Control Mechanism

AC mechanism is a prerequisite for implementing an e-CRM enabled website with customer-centric business intelligence. It works, basically by requiring a user to login, in order to identify himself before he is granted access to certain protected resources. It is assumed that a user would think it is fair that they would have to be bothered about logging in for getting some invaluable resources in return. At the first time, he would have to sign up a registration form with a user-name and a password of his choice.

On the website, the hypertext links to the protected resources such as a free-trail of software, a product's manual, etc., would have to be modified. Upon clicking on any one of these protected hypertext links, not only a user would be challenged with a password for checking access rights, the event of that "click" will be recorded in a database. Specifically, we record down who, at what time, on which website (in the case of a multi-regional website), have accessed to what resources from which web page. This recording event happens from when the user has logon (both success failure cases) to every time he clicks on a protected link. The idea of doing so is that we want to be able to keep track of which resources have been accessed by whom, when, and how often, on our website. Hence we can derive insights on what is happening on the website rather in terms of accesses than merely traffic hits per visit. The information obtained are more insightful than those from web-logs, since we beforehand have kept a full demographic details about the users when they registered, compare to the information obtained from web-logs which are only IP addresses. We will later discuss about how the records of user access will be analyzed for web business intelligence. The technical implementation flows for login procedure and thereafter access procedure are shown in Figures 2 and 3 respectively. For both cases, a record is generated in two different database tables namely login and access. The AC architecture has three distinguishing features:

1. Instead of using web log data "GET" that is relative only to the name of a web page, we can more accurately pinpoint which "resources" on the page that we want to monitor. It can be a link, a JPG file, a PDF file, or a movie clip. Access to them requires the user to login. We have the flexibility to design which links require the users to login and therefore we can monitor the access patterns on them.
2. Users can be uniquely identified by their login ID
3. Before the analysis, we can have relatively clean user profiles because the data fields of an access record are carefully chosen.

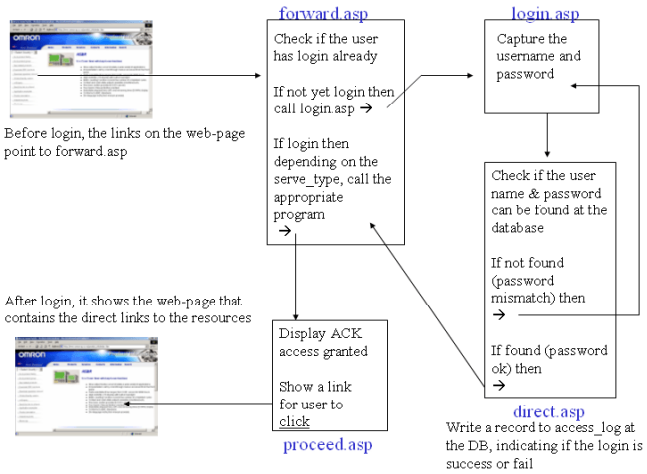

Figure 2. Logical flow for login procedure

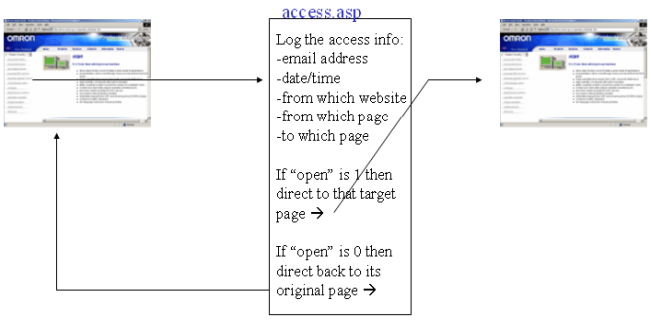

Figure 3. Logical flow for access procedure

4. Web Business Intelligence

Web business intelligence refers to the ability to make better business decisions in running a website through intelligent use of collected data, mostly from the same website. It is about gathering, managing, and analyzing data. Then the data are transformed into useful, actionable information to run an information-driven e-Business.

With the AC mechanisms implemented, at the appropriate links at which the company wants to monitor about their popularity, a wealth of good business intelligence (BI) can be obtained. In essence, we would be able to obtain a new realm of BI with the information about when and what the users have access and login to the website. We suppose that the web business intelligence can be obtained at two levels; one with and the other one without the aid of data-mining. At the first level, where simple statistics are used, we can show the following information as business intelligence about the products and the website:

(1) Customer profiling (demographics). When integrated with web traffic, one can tell the online customers behaviors. However, this requires some advanced techniques like web log mining. At the first level, for example, we can tell about the top 10 (or top 5, 100, etc) of the countries, the industries where the users come from, or their designations;

(2) Customers' interests. The top customers' interests on the products can be listed out. In a simple way, this information is obtained from the registration form when the users signed up. There is other more sophisticated technique to estimate the user's interest of a web page without directly asking the user. This method [3] is proposed for locating multi-word phrases in order to estimate interests on certain pages.

(3) Product popularity. The web pages of products that have been accessed most could be ranked and displayed as vital business information. This is a relatively fair approach when compare to measuring web page hits as each page may have different hit count per access

(4) Technical knowledge base. Over a certain period of time, the online inquiry forms posted by the customers have been collected, organized that form a searchable technical knowledge base. The base can made to be searchable that stores the past problem cases and the appropriate solutions. In a simpler form, it could be provided to the users as a set of frequently asked questions.

So far the task of obtaining the BI at the first level is limited to a single or only a few sources of data. Data-mining plays an important role in analyzing complex sometimes unorganized data from multiple sources. For instance, we can merge the data from different dimensions such as the attributes about the customers, the aspects about the products, different places or the schemes of the website, traffic patterns and time. From this huge pool of complex merged data with a diversity of dimensions, data-mining could help discovering some hidden patterns that may be of business values. Finding correlation among data and doing prediction are beyond the capability of simple statistical calculation.

5. Pattern Discovery under AC Architecture

Web usage mining, which is also known as pattern discovery is a process of mining for user browsing and access patterns. With our AC architecture, mining is an important backend process for deriving business intelligence at the second level. In AC environment, we record only those user accesses taken from the links that we chose. It is assured that the user must have logged in, given us his unique user-id and then granted the access.

So each access record stored in the database contains vital information of the user's identification, time of access, the website's identity, from which page he is making the access, and which "item" he is accessing. The item or the place of tracking can be a graphics, a movie clip, a PDF file or a hypertext link. Each user-id can be easily linked to the user registration table in the database; that implies a full set of information about that user is available. The problems of user identification as with web log processing no longer exist. We can take for granted that the user is readily identified as he has passed the password challenged. The recording of his trail is made almost instantly by our program code embedded at the link as soon as he mouse clicked on it. Besides the flexibility of choosing whichever resources we want to monitor, the biggest advantage is that the effort for data preprocessing is minimal. This AC method is believed to be more reliable because there is no missing or ambiguous value, and no need to remove any reluctant information. However, this requires a good design in setting up the links that we want to trace in advance. A non-technical drawback however, would be the difficulty in luring the users to sign up, and then get them perform the login action at every first time he wants to access a resource during a session. The value of the resources that the user seeks must be able to justify the trouble of login.

6. A Generalized Data-mining Model for AC

With our AC model, we attempt to generalize the discovery attributed to the following format: *who -> when/how visited which parts of the website -> for what -> why*. Obviously, the format contains pivots of natural clauses such as who, when, what, etc. These pivots would be easily derived from the attributes or fields of the AC records, except the last two: for what and why.

Who: attribute of the user-id in AC record. Knowing the user-id, the rest of the demographic information about that user can be obtained from the user registration records.

When: time stamp information in AC record. A hierarchical time dimension model will need to be created later on, e.g. year, quarter, month, week, etc. prior to mining.

How: this information can either be a case in the form of patterns obtained through the analysis discussed in the previous section, or a simple path formed by the "From" and "To" fields in the AC records. Other abstract forms of taxonomical representation could be "through a promotion", "from a link in a newsletter" or "from a search engine".

Which parts: this is usually the destination information in the AC records plus the preceding trails if any. The destination information could be a single resource file or a page that is obtained from the "To" field of the access table. The preceding trails would have to be obtained from Path Analysis. Each one of the preceding trails is a paths a user during one visit has traversed within the physical web site layout.

For what: this is tacit information inferred from the results of Association and Clustering analysis. Alternatively, an analyst may have a conjecture for the groups and explicitly set a value for this variable with his subjective judgment. In an information-only website, the ultimate purposes for the users may be to obtain certain information in the form of a report, a product brochure, some free software download or to post a technical question. Furthermore, the users may want to purchase something in the case of an online shopping website.

Why: this is the concluding remark that an analyst draws for the patterns. This is a continuous review process with largely human intervention since the data-mining system would not be able to explain why such has happened. It would only reveal the hidden patterns from a large pool of data. The explanation of the outcomes would have to be first answered by a human expert. Subsequently, prediction using a decision tree will be possible as an extension to this model.

7. Conclusion

Access Control (AC) technique is based on this key innovation: obtaining data from the resource links at the web server instead of web-server log files. It is a password-based forcibly method for capturing customers' trails. It is based on the assumption that the users are willing to login in exchange of some valuable resource from the website. By virtue of this innovation, AC architecture provides better information than log-file analysis, gives site owner more reliable business intelligence. Web log tells mostly about traffic statistics and treating each user anonymous (because it is difficult to identify a user by its IP address). Access records with unique user identification and exact navigated locations, would be more suitable for data-mining for customer-centric Web Business intelligence.

8. Reference

[1] WebTrend Weblog Analyser, www.webtrend.com

[2] S. Fong, A. Sun, K. Wong, "Price Watcher Agent for E-Commerce", Int. Conference on Intelligent Agent Technology, World Scientific, pp.294-300, 2001.

[3] P. Chan, "A Non-invasive Learning Approach to Building Web User Profiles", KDD-99 Workshop on Web Usage Analysis and User Profiling, pp.7-12, 1999.