ICIT2007

The 3nd International Conference on Information Technology
May 9-11, 2007
AL-Zaytoonah University, Amman, Jordan

Paper ID: 332

Using Data Mining Techniques for the Improvement of the 'Adaptive Website Agents' process

www.**alzaytoonah.edu.jo**/ICIT2007/Conf Program.doc

Prof. Dr. Alaa Al-Hamami
Amman Arab University for Graduate Studies, Zip Code: 11953, P.O.B. 2234, Amman, Jordan.
alaa_hamami@yahoo.com

Mazin S. Al-Hakeem
Computer Science Department, University of Technology, Baghdad, Iraq.
mazin_ictc@yahoo.com

Using Data Mining Techniques for the Improvement of the 'Adaptive Website Agents' process

Prof. Dr. Alaa Al-Hamami
Amman Arab University for Graduate Studies, Zip Code: 11953, P.O.B. 2234, Amman, Jordan.
alaa_hamami@yahoo.com

Mazin S. Al-Hakeem
Computer Science Department, University of Technology, Baghdad, Iraq.
mazin_uot@yahoo.com

ABSTRACT

This research introduces the notion of using Data Mining – web usage techniques as a tool to influence how 'Related Web-page Recommendations' (RWR) are made. The Data Mining – web usage techniques will be implemented in the 'Adaptive Website Agents' – which is introduced in 2002 by [9]- to improve the work of these agents to recommend relevant web pages.

The 'Adaptive Website Agents' are agents which are used to help visitors in finding information at a particular web site, this goal is achieved by discovering the relationships between accessed web pages to learn about web site visitors and from their patterns.

However, The 'Improved Adaptive Website Agent' is implemented using Microsoft ASP (Active Server Page) Technology as a server side programming language, to run at MS IIS (Internet Information Services), and use the ADO (ActiveX Data Object) to access the MS Access database from the web pages.

Key Words: Data Mining, Boolean Associated Rules, Adaptive Website Agents, Web Server Logs File, Related Web-page Recommendations (RWR).

1. Introduction

The impact of the World Wide Web as a main source of information acquisition is increasing dramatically, the existence of such abundance of information, in combination with the dynamic and heterogeneous nature of the web, makes web site exploration a difficult process for the average user [8]. To address the requirement of effective web navigation, web sites provide personalized recommendations to the active users [8]. In pervious paper presented in 2002 by [9], it is described to help the visitors explore the web site(s) by recommends personal recommendations. However, the 'Adaptive Website Agents' has been used to help visitors in locating interest information on the web site(s) [9]. There are two important viewpoints must be considered when using agents, to help the visitor to surfing web sites and to locate the interest information (to recommend the personal recommendations) [11]:

a. The visitor's viewpoint: The agent should help the visitors to make sure

that interest information is not overlooked.

b. The web site developer's viewpoint: The designer wishes to increase the amount of interest information that accessed by the visitor(s).

However, a variety of systems have been proposed to help adaptation of a web site to visitors.

- **Perkowitz and Etzioni's System** [12], have discussed visitor access patterns that can be taken into account to re-construct new "index pages" by grouping pages that commonly accessed together in a single page.
- **The Footprints system** [15] allows a visitor to a web site to visualize the paths through a web site that are commonly traversed.
- **AVANTI's System** [2] uses a set of adaptation rules that customize the appearance of a web site for groups of users.
- **Michael J. Pazzani and Daniel Billsus's System** [9], have presented several strategies to represent the relationships between documents (web page), these strategies, which described in section 2, are: Similarity, Referenced, Referenced-by, Downloaded-with, (i.e. the Downloaded-with strategy is depend on the idea of the document that frequently accessed in combination with another during the same session, and then followed Perkowitz and Etzioni by estimating the probability of relationship).

Each of these systems uses a different strategy to influence how related web pages (interest web pages) are recommended for visiting.

2. Adaptive Website Agents

The goal of the 'Adaptive Website Agents' [9] is to assist the visitor with navigating the web site. When the visitor views a web page, then it is possible to receive recommendations for related web pages (receive 'related web-page recommendations' (RWR)). Michael J. Pazzani and Daniel Billsus [9] have deployed two agents with different topics and audiences, and monitored how visitors interact with these agents.

- The first agent at a publications web page at the web site of the university, recommends scholarly publications to visitors.
- The other agent at a web page presents information on raising goats and other livestock.

The agents share the same engines for analyzing web server logs file, determining similarity between documents (web pages), making personal recommendations, and learning about visitors. The agents used many ways to identify how web pages may be related, how a personalized profile is created for each, or group, of visitor(s), and how this profile is used to make recommendations. However, these ways are:

1. ***Similarity***: The document is similar to another as determined by comparing the TF-IDF representation [13] of the documents using the cosine similarity metric.
2. ***Referenced***: The document contains a hypertext link to another.
3. ***Referenced-by***: There is a hypertext link from another document to this one. Since the agent has knowledge of the inverse of every link, it can recommends documents that link the current document.

4. ***Downloaded-with***: The document is frequently accessed in combination with another during the same session. This information is obtained from web server - logs file.

However, the visitor has viewed a document, and when the agent detects the return of the visitor to the browser, it recommends that the visitor visit or download another documents. For example, the agent can recommend a document that is on a topic similar to that of other items seen by the visitor. Alternatively, a document could be recommended that has been accessed frequently by other visitors in combination with documents seen by the visitor.

3. Web Usage Mining

Data mining - web usage mining technique is an automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of active user interactions with web resources on one or more web sites [2, 14, 10]. The goal of web usage mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a web site [2]. The discovered patterns are usually represented as collections of pages, objects, or resources that are frequently accessed by groups of users with common needs or interests [2].

The overall process of recommend the personal recommendations based on web usage mining consists of three phases: data preparation and transformation, pattern discovery, and recommendation. Of these, only the latter phase is performed in real-time [2].

Association rule discovery techniques, such as the A Priori algorithm [6], were initially developed as techniques for mining supermarket basket data but have since been used in various domains including web mining [1].

4. The Proposed System

This research describes a data mining – web usage mining framework to improve the process 'Adaptive Website Agents' model to recommend the related web-pages for the visitor(s).

The key idea is to mine web server logs file (web log records data) using the 'Boolean Association Rules' based on the popular A Priori algorithm.

Through the mining, the agent has access to the relationship knowledge of various documents in terms of the downloaded-with, and to knowledge of the popularity of various documents in terms of the number of downloads. The agent uses this information to decide upon the strength of a recommendation.

The associated rules will be apply to estimate P(D2/D1) from the log data (i.e., the probability that document D2 is downloaded given that D1 has been downloaded). The strength of the relationship is calculated by using A Priori algorithm for all documents. The general algorithm and all the details would be explained in the following sections briefly.

4.1. Overview of the General Data Mining Framework

The general web usage mining framework overview, as shown in figure (1), it presents the main parts and the main proposed operations. The overall process of proposed framework can be divided

into two components. The offline components are data mining – web usage mining tasks and part of 'Adaptive Website Agent', online and offline, tasks. The online component is another part of 'Adaptive Website Agent' tasks.

The offline-Data Mining tasks divided into two separate stages. The first stage is that preprocessing and data preparation, including data cleaning, filtering, and transaction identification. The second is the mining stage in which patterns are discovered via strong Boolean association rule mining. The offline-'Adaptive Website Agent' tasks will be applied on discovered association rule up on accessed web page to make a 'related web-page recommendations' (RWR). Online-'Adaptive Website Agent' tasks, is to compute a RWR for the current session, consisting of links to pages that the visitor may want to visit based on similar patterns. The RWR will be send using HTTP protocol to client-side for display front of the current visitor (current session) as a recommended web pages.

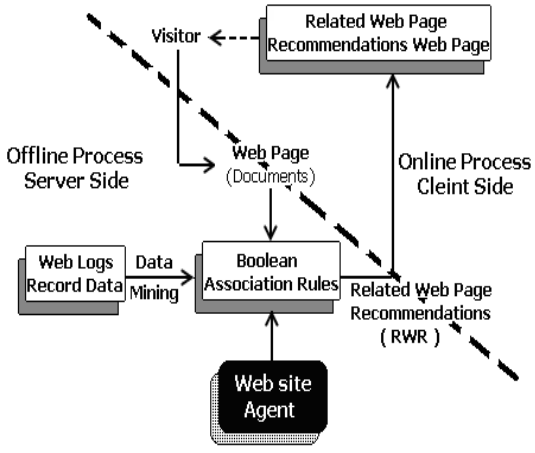

Figure (1): The general Data Mining – Web Usage Mining Framework Overview.

4.2. The Preprocessing and Data Preparation Stage

The first stage in the offline-Data Mining tasks is that preprocessing and data preparation; including process the web server logs file data by conversion this web logs data to a relational database. Also, the preprocessing stage includes log file data cleaning (from redundant entry-records) and filtering (from non-existent pages).

4.3. Applying the Basic Algorithm

The second stage in the offline-Data Mining tasks is the mining stage in which patterns discovered via strong Boolean association rule mining. After conversion of web logs in the web data file to relational database is completed, this relational database must be submitted to Data Mining –web usage mining tool. However, the basic (A Priori) algorithm that applied to discover the new strong associated rules is shown here. Let D be a set of documents (D1, D2, D3, … , Dn), and I be a set of values on D, called items. Any subset of I is called an item-set. The number of items in an item-set of I is called length.

Let L be a series of accessed (downloaded) documents in the same session, and NL is the number of web logs, and let A be a database with attributes (NL, L). The number of items in L is number of downloaded document in the same session, which called width.

Define support (X) as the percentage of transactions (records) in A that contain item-set X. An association rule is the expression X → Y, c, s. Here,

- X and Y are item-sets and
- X Φ Y = ∩.
- s = support (X U Y) … is the support of the rule (is the all

transaction under the analysis rule), and

- c = support (X U Y)/ support (X) … is the confidence.

Boolean association rules are considered interesting if they satisfy both minimum support threshold and minimum confidence threshold. Threshold can be set by web site administer.

To generate a strong association rules, must be find all frequent item-set. Here, an influential algorithm for mining frequent item-set for Boolean association rules called A Priori algorithm will be used. The A Priori algorithm declared completely [6].

For example, if there are a very-simple web site that continents of five web pages (D1, D2, D3, D4, D5), as shown in figure (2), and the visitor can be access this web site through Home-page (D1) or another main page (D2).

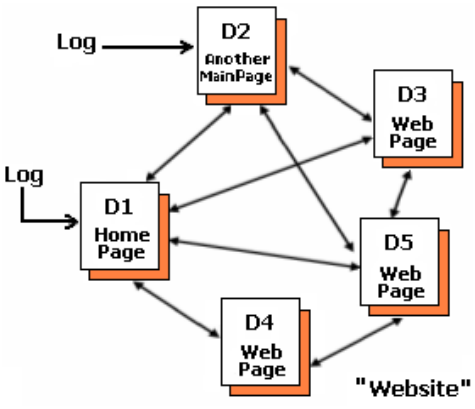

Figure (2): A Simple Web Site with Five Web pages.

And if there are nine transaction in database A (A is a database with attributes (NL, L)), and assume the minimum support threshold is equal to 2. The algorithm will be operating to find the frequent item-set in A, which shown in figure (3), as following tables in as shown in figure (4):

Database "A"

NL	L
Log100	D1, D2, D5
Log101	D2, D4
Log102	D2, D3
Log103	D1, D2, D4
Log104	D1, D3
Log105	D2, D3
Log106	D1, D3
Log107	D1, D2, D3, D5
Log108	D1, D2, D3

Figure (3): The Frequent Item-set Table.

Sample Web transactions involving Web page view D1, D2, D3, D4, and D5.

Finally, these patterns ({D1,D2,D3}, {D1,D2,D5}) considered as a basic scheme for making related web-pages recommendations as a records entered to result inside recommendations web page. If the visitor visits D1 and D2, the 'Adaptive Website Agent' will be recommend D3 and D5 as a possible recommendation.

However, figure (5) shows the frequent item-sets graph constructed based on the frequent item-sets in figure (5).

Now, given visitor active session window {D1,D2}, the recommendation that generate by A Priori algorithm recommend D3 and D5 as candidate recommendations. The recommendation scores of D1 and D2 are 1 and 4/5, corresponding to the confidences of the rules {D1,D2}→{D3} and {D1,D2}→{D5}, respectively.

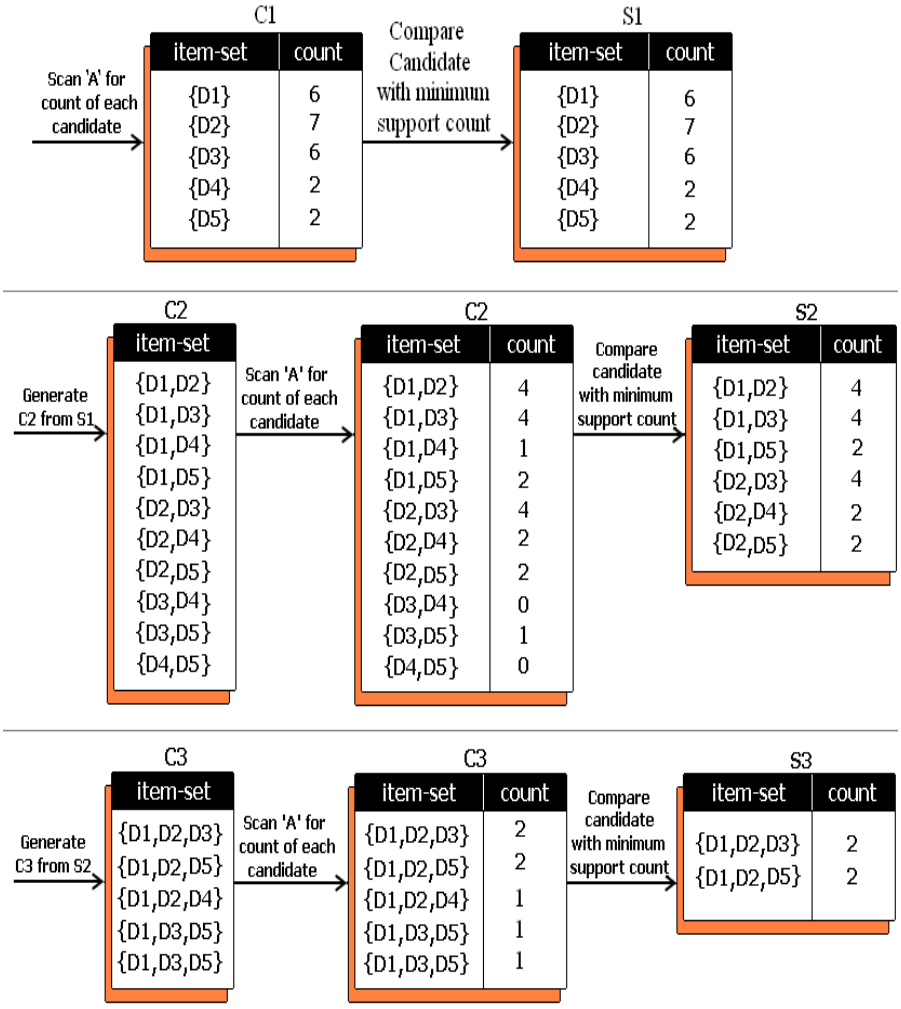

Figure (4): Using A Priori Algorithm to find the frequent item-set in Database (A).

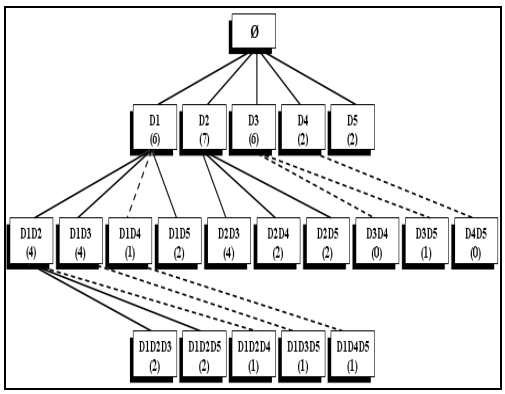

Figure (5): The frequent item-sets graph for the example.

4.4. Making 'Related Web-page Recommendations' (RWR)

The agent makes a related web-pages recommendation (RWR) to the visitor when explicitly requested by the visitor or when focus is returned to the web browser after the visitor has activated another program or window and then returns to the web browser.

When making a recommendation using A Priori algorithm, the agent can be viewed as a personal representative of the web site author.

The agent infers from the visitor's actions whether the visitor followed the recommendation and increases or decreases the probability that future recommendations are made for that same reason for that visitor.

5. Evaluations

Like using Download-with method (which describe in [9]), the 'Adaptive Website Agents' based on the A Priori algorithm increases the amount of information accessed by users.
Table (1) shows the average numbers of web pages (documents) viewed (downloaded) by users for 3-days and 3-days after adding engine to 'web site's library of computer sciences department - university of technology in Baghdad).

Table (1): Average numbers of web pages downloaded by users before agent and with agent installed (during 3-days).

Web Pages	Before Agent	With Agent
PhD Thesis	1.2 Pages	3.0 Pages
MSc Thesis	3.8 Pages	5.4 Pages

However, mining the cleaned-and-filtered web server logs file using A Priori algorithm would perform the process of making related recommendations in early time, rather than using the Download-with method (which describe in [9]) in audit data. Figure (6) shows the results of required time vs. amount of data (viewed web pages).

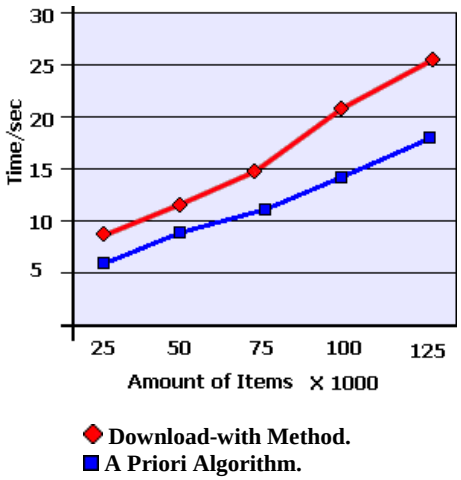

Figure (6): Required Time vs. Amount of Items (Viewed Web Pages).

6. Conclusion

From the improving of the process of 'Adaptive Website Agents', many conclusions which are of certain signification are researched. They are:

1. The 'Adaptive Website Agent' serves as a personal representative of the web site visitors and web site developers. The goal of the agent is to help the visitors to find interest information at the web site by recommending additional related documents to the visitors. The another goal of agent is to help the developer find the popular sections in its site by mining the web server logs file and downloaded pages and for reconfigure the server organization according to it.
2. Mining the web server logs file data (using A Priori algorithm) would perform the process of making related recommendations in early time, rather than using the Download-with method (which describe in [9]) in audit data, as shown in figure (6).

3. It is very useful to cleaning the web server log file from redundant entry-records and filtering it from non-existing pages, which consider as a noisy data, especially in order to make strong recommendation sets.

References:

[1] Baumgarten, M., Buchner, A.G., Anand, S.S., Mulvenna, M.D., Hughes, J., "User-Driven Navigation Pattern Discovery from Internet Data", Proceedings of the WEBKDD'99 Workshop. Lecture Notes in Computer Science 1836. Springer-Verlag, 2000, 74–91.

[2] Cooley, R., Mobasher, B., Srivastava, J., "Web Mining: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), Newport Beach, CA, 1997, 558–567.

[3] Fayyad,U.M., G. Piatesky-Shapiro, P.Smyth and R. Uthurusamy, "Advances in Knowledge Discovery and Data Mining". AAAI/MIT Press, 1996.

[4] Fink, J., Kobsa, A., Nill, "User-oriented Adaptively and Adaptability in the AVANTI Project", Designing for the Web: Empirical Studies, Microsoft. Usability Group, Redmond (WA), 1996.

[5]Gediminas Adomavicius and Alexander Tuzhilin, "Personalization Technologies: A Process-Oriented Perspective", 2004.

[6] Han,J., M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2000.

[7] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2002.

[8]Magdalini P. Eirinaki, "New Approaches to Web Personalization", Ph.D. Thesis, Athens University of Economics and Business, Dept. of Informatics, 2006.

[9] Michael J. Pazzani, Daniel Billsus, "Adaptive Website Agents", Autonomous Agents and Multi-Agent Systems, Kluwer Academic Publishers, 2002.

[10] Mobasher, B., "Web Usage Mining", InWong, J., ed.: Encyclopedia of Data Warehousing and Data Mining. Idea Group Publishing, 2005, 1216–1220.

[11] Olfa Nasraoui, "World Wide Web Personalization", Department of Computer Engineering and Computer Science, University of Louisville, USA, 2005.

[12] Pazzani, M. & Billsus, "Learning and Revising User Profiles: The identification of interesting web sites", Machine Learning, 27, 1997, pp313-331.

[13]Salton, G., "Automatic Text Processing. Addison-Wesley", 1989.

[14]Srivastava, J., Cooley, R., Deshpande, M., Tan, P., "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKDD Explorations 1(2), 2000, 12–23.

[15] Wexelblat, A. & Maes, P., "Using History to Assist Information Browsing". RIAO'97: Computer-Assisted Information Retrieval on the Internet, Montreal, 1997.

Published at:
www.uotechnology.edu.iq/dep-cs