Association-Rules-Based Recommender System for Personalization in Adaptive Web-Based Applications

Daniel Mican and Nicolae Tomai

Babes-Bolyai University, Dept. of Business Information Systems,
Str. Theodor Mihali 58-60, 400599, Cluj-Napoca, Romania
{Daniel.Mican,Nicolae.Tomai}@econ.ubbcluj.ro

Abstract. Personalization systems based upon the analysis of users' surfing behavior imply three phases: data collection, pattern discovery and recommendation. Due to the dimension of log files and high processing time, the first two phases are achieved offline, in a batch process. In this article, we propose Wise Recommender System (WRS), an architecture for adaptive web applications. Within this framework, usage data is implicitly obtained by the data collection submodule. This allows for the extraction of usage data, online and in real time, by using a proactive approach. For the pattern discovery, we efficiently used association rule mining among both frequent and infrequent items. This is due to the fact that the pattern discovery module transactionally processes users' sessions and uses incremental storage of rules. Finally, we will show that WRS can be easily implemented within any web application, thanks to the efficient integration of the three phases into an online transactional process.

Keywords: Adaptive web-based applications, Web usage mining, Recommendation systems, Web personalization, Association rules.

1 Introduction

The ability of a web application to offer personalised content and to adapt is determined by its ability to anticipate users' needs and to provide them with the information and content they need. Adaptive web applications [10] can do this only after having analysed data resulted from the users' current and former interaction with the system. Based upon the similarities discovered between different types of content and different user groups, one can make a series of recommendations enhancing the web applications' capacity of adaptation and personalisation. Personalisation systems based upon the analysis of the user's surfing behaviour imply three phases [9]: data collection and preparation, pattern discovery and content recommendation. Thus, a new research branch, called Web Usage Mining came into being, its goal being to discover useful information and knowledge as a result of the analysis of these interactions. The WUM techniques use data extracted from log-files and provide information about activities undertaken by users while surfing. In order to discover new useful information, WUM applies a series of techniques, like classification, clustering, discovery of association rules or sequential patterns [9].

F. Daniel and F.M. Facca (Eds.): ICWE 2010 Workshops, LNCS 6385, pp. 85–90, 2010.

In our research, we have used the technique of association rules, in order to discover correlations between the pages of a web application, based upon the analysis of the user's surfing sessions. Our efforts were channelled towards finding an efficient solution for implementing a recommendation system within a web application capable to synthesize and to store only those data that are relevant for the recommendation process and within which all three phases could be realized online. Thus, we have proposed a new framework to fulfill the high personalisation needs of actual web applications. The Wise Recommender System (WRS) uses a proactive approach, allowing the extraction of data about the users' interaction with the web application. The collecting of usage data is being implicitly achieved, without the need of an explicit request from the part of the users' opinion. Our approach allows the incremental finding and storing, both of the existing connections between frequently visited pages and those less frequently visited. As a result, the WRS is able to offer a list of personalised pages to each user, depending on the pages he is currently surfing on, without the need for a surfing history or a minimal number of visited pages.

2 Adaptive Websites and Web Personalisation Systems

The concept of Adaptive Websites was proposed by Perkowitz and Etzioni in [10]. Adaptive websites are defined as those sites using information about the way in which users access them, in order to improve their organisation and presentation. The Web Personalization System is defined in [9] as any action that adapts information and services provided by a web application to the needs of one user or of a group of users. A personalisation system must be able to provide users with the information they need, without them having to explicitly request it.

During the last years, more and more researchers paid a special attention to WUM domains and to the personalisation of web applications. Among the first researchers who channelled their efforts towards Web Usage Analysis and WUM were Cooley et al. [7], who proposed WebMiner, one of the first systems offering an overview of WUM. The PageGather [10], is a synthesising algorithm that uses clustering in order to find collections of similar pages within a website. The WebPersonalizer [9] has the goal to make recommendations to the users, based upon the similarity of the surfing behaviour with that of users in the past and contains an offline module, whose role is to filter log files data and to extract the most interesting surfing models. We can also mention SUGGEST [2] and Smart-Miner [3] which can efficiently process terabytes of web log files.

In order to be able to use data residing in log files, it is absolutely necessary that these be cleaned and filtered. The analysis of the log files raises a series of problems, namely: the existence of a high number of irrelevant records for the process of web usage mining, the difficulty in identifying users and sessions, the lack of information about the content of accessed pages, and the fact that data processing is a batch processing, which takes up time and resources. Literature mentions several methods helping to identify and delimit a user's sessions. The most popular approache is the 30 minutes time limit threshold [4], followed by reference length [7], and maximal forward reference [6]. The multiple drawbacks connected to the users and sessions identification have led to the development of reactive and proactive strategies. Reactive

strategies want to associate requests with users, based upon web server logs, following their interaction with the website. On the other hand, proactive strategies want to associate requests with users, during their interaction with the website [11].

The goal of the association rules mining [1], [5] is to discover correlations or relations of association between existing records in a dataset. In [5] fundamental association rules have been mentioned, from their emergence and up to the present moment. These works present both classic algorithms like: Apriori, Eclat, Clique, FP-Growth, a.s.o., as well as the generic optimisations they were provided with. In [8], practical and efficient methods are presented, whose aim is to find association rules in the case of less frequent items.

3 Wise Recommender System (WRS), the Proposed Architecture for Web Personalisation

In order to increase the capacity of adaptation and personalisation of web applications, we have integrated several submodules in an innovative manner. Thus, the collecting submodule allows the extraction of usage data, online and in real time, by using the proactive approach. The extraction of data about the users' surfing behaviour, preferences and activities is implicitly accomplished, without the necessity of explicitly involving these into the collection process. The data extracted in this manner are quality data, complete, noiseless and error-free. Moreover, WRS also takes into consideration the content very rarely or occasionally accessed. In figure 1, one can see the architecture of the WRS system we propose.

The crawler identification submodule allows the identification of search engines from human users. This submodule has access to a table which contains the names and the IPs, respectively the IP intervals that are allotted to the main web crawlers. Its role is to filter web crawlers and to send to the users' identification submodule only the traffic generated by human users.

The goal of the user identification submodule is to identify, within a web application, human users exclusively. In order to achieve this goal, the submodule implemented by us successfully uses the newest web technologies allowing session work and uniquely identifies each user, by the means of the IP address, while also taking into consideration the fact that it could be behind a proxy. Thus, it will associate to each user a unique session ID, valid from the moment the user accesses the application and up to the moment the user will close the web browser.

The content identification submodule must uniquely identify the content accessed by the user. The identified content will be stored in a table, together with the number of hits it had over time. Should a page be accessed several times by a user within a surfing session, its accessing value will be incremented by a unit.

The goal of the session identification submodule is to identify surfing sessions for each user. This is quite simple, due to the fact that these ones are already uniquely identified by the users' identification submodule. Furthermore, all pages the web application generates for a user will be accompanied by their session ID. In order to identify and delimit users' sessions, we based our implementation on the W3C approach, one session being made of the totality of pages accessed by a user, from the moment the user opens the web browser and up to the moment the user closes it.

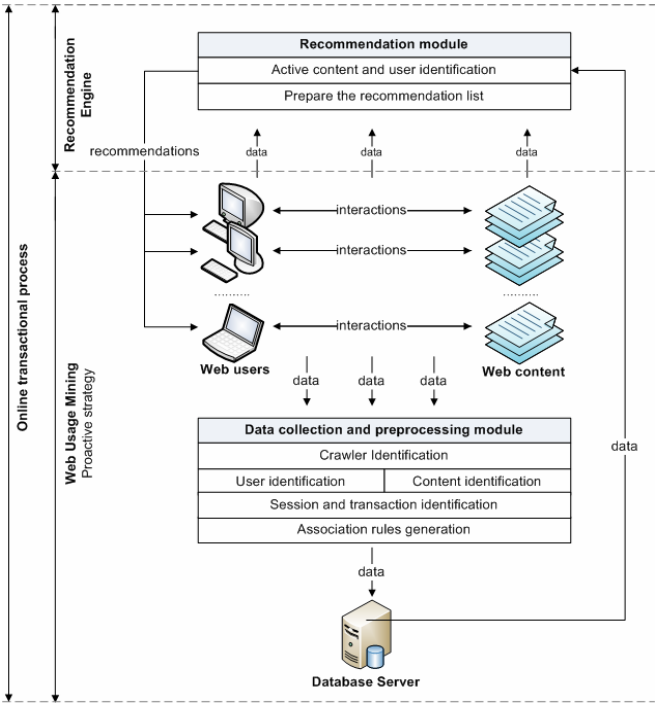

Fig. 1. WRS. The proposed architecture for web recommendation and personalization

The submodule generating association rules can access in real time data taken over by the other submodules. Thus, it will connect to the sessions identification submodule and will take over from this one data related to the users' sessions, in order to generate association rules. The generating of association rules is being done online, in real time, in a transactional process. Once the session has been processed and rules extracted and inserted, it will be deleted from the sessions table. Due to the innovating method of processing sessions and storing of association rules, we succeeded in achieving a scalable model, able to work in real time with a large volume of data.

The active content and user identification is a submodule whose role is to identify the active user and the page this one is visiting. The moment a user accesses a web page, an identification ID is sent, in order to identify the active page of the recommendation list generation submodule. The recommendation list generation receives an identification ID of the current page and has the role of selecting from the database the recommendation list that will contain the pages in a descending order, according to the degree of confidence.

4 Description of Experiments We Have Carried Out

In order to prove the scalability of the proposed model, we have undertaken a series of experiments on two popular Romanian websites: Intelepciune.ro and BizCar.ro.

The two websites total over 380.000 unique online visitors, respectively over 1,6 million page views per month. We have implemented the proposed model on the two above-mentioned sites and in figure 2 one can notice the dynamics of content, sessions and rules over a period of 32 days in classifieds section.

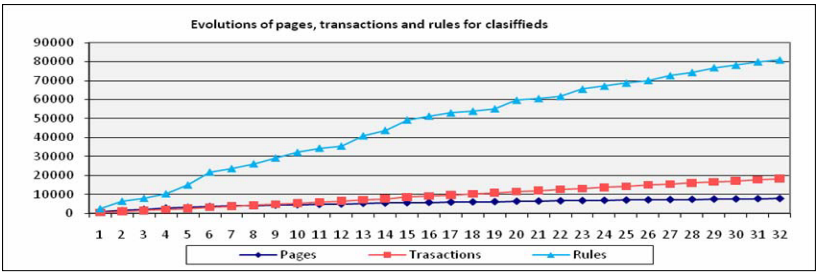

Fig. 2. The evolution of pages, sessions and rules in classifieds section

Unlike other approaches, in the model we are proposing, the recommendation period is not influenced by the number of recorded sessions. In table 1, we can notice the time necessary in order to generate a recommendation list depending on the evolution in time of the number of pages, sessions and rules generated for Intelepciune.ro.

Table 1. The time, number of pages, sessions and generated rules

Page No.	1072	3589	5234	7525	9669	17230	20453	22564
Sessions	667	2687	4922	9723	10799	19952	25637	33628
Rules	3497	18774	33337	53558	63711	99135	116110	142131
Time	0.002	0.008	0.013	0.039	0.052	0.085	0.139	0.159

By looking at the data from the table 1, one can notice that the time necessary to generate a list of recommendations is influenced by two factors: the number of generated pages and rules. As a result, we undertook to analyse the dependence between the recommendation time and the two factors of influence with the help of a regression model. As a result of a statistical analysis, we obtained Multiple R = 0.988572, which shows that there is a strong connection between the two variables, the number of generated pages and rules. We obtained R Square = 0.977274, which shows that 97% from the variation of the recommendation time is explained by the two variables. The average square variation (Standard Error) = 0.009048, the result being that the points on the regression are approaching a straight. Due to the fact that in the case of the number of pages, the P-value is 0.00399 < 0.05, the result is that this coefficient is of significance. For the generated rules of the P-value = 0.884357 > 0.05, the result translates in the fact that the coefficient is insignificant. From here, we can conclude that this variable can be eliminated from the model, thus resulting a simple linear regression model.

5 Conclusions

One of the most important goal of an adaptive web application is the content recommendation in a period of time as short as possible. In this article, we proposed WRS for content recommendation and we used association rules in order to model existing connections between the pages of a web application. The proposed system brings an additional benefit, because it allows the finding and maintaining in the system of the rules existing between those pages that are not frequently accessed by users, too. In this article, we proposed a different approach for a recommendation system, by integrating the pattern discovery phase and that of data collection and filtering into a single module. Following the undertaken experiments, it resulted that the proposed model is very efficient in recommending the content and can be easily implemented within any web application. In the future, we wish to continue the optimisation of the recommendation system by incorporating different particularities resulting from the type of content existing in different web applications. Likewise, we would like to exploit knowledge extracted over a longer period of time, in order to see the evolution of the intensity of connections discovered over time.

Acknowledgement. This work is supported by the Romanian Authority for Scientific Research under project IDEI_2596.

References

[1] Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C, pp. 207–216 (1993)
[2] Baraglia, R., Silvestri, F.: Dynamic personalization of web sites without user intervention. ACM Commun. 50(2), 63–67 (2007)
[3] Bayir, M.A., Toroslu, I.H., Cosar, A., Fidan, G.: Smart Miner: a new framework for mining large scale web usage data. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 161–170. ACM, New York (2009)
[4] Catledge, L.D., Pitkow, J.E.: Characterizing browsing strategies in the World-Wide Web. Comput. Netw. ISDN Syst. 27(6), 1065–1073 (1995)
[5] Ceglar, A., Roddick, J.F.: Association mining. ACM Comput. Surv. 38(2) (2006)
[6] Chen, M.S., Park, J.S., Yu, P.S.: Efficient data mining for path traversal patterns. IEEE Transactions on Knowledge and Data Engineering, 209–221 (1998)
[7] Cooley, R., Mobasher, B., Srivastava, J.: Data preparation for mining World Wide Web browsing patterns. Knowledge Information Systems 1(1), 5–32 (1999)
[8] Ding, J., Yau, S.S.: TCOM, an innovative data structure for mining association rules among infrequent items. Comput. Math. Appl. 57(2), 290–301 (2009)
[9] Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. ACM Commun. 43(8), 142–151 (2000)
[10] Perkowitz, M., Etzioni, O.: Adaptive sites: Automatically learning from user access patterns. In: Proc. of the Sixth International WWW Conference, Santa Clara, CA (1997)
[11] Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M.: A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. INFORMS J. on Computing 15(2), 171–190 (2003)