Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval

Mohamed Koutheaïr KHRIBI [1], Mohamed JEMNI [1], Olfa NASRAOUI [2]
[1] Technologies of Information and Communication Lab,
Higher School of Sciences and Technologies of Tunis,
University of Tunis
5, Av. Taha Hussein, B.P. 56, Bab Mnara 1080, Tunis, TUNISIA
mk.khribi@uvt.rnu.tn, mohamed.jemni@fst.rn.tn

[2] Knowledge Discovery & Web Mining Lab,
Dept. of Computer Engineering & Computer Science,
Speed School of Engineering,
University of Louisville, Louisville KY 40292, USA
olfa.nasraoui@louisville.edu

Abstract

The World Wide Web (WWW) is becoming one of the most preferred and widespread mediums of learning. Unfortunately, most of the current Web-based learning systems are still delivering the same educational resources in the same way to learners with different profiles. A number of past efforts have dealt with e-learning personalization, generally, relying on explicit information. In this paper, we aim to compute on-line automatic recommendations to an active learner based on his/her recent navigation history, as well as exploiting similarities and dissimilarities among user preferences and among the contents of the learning resources. First we start by mining learner profiles using Web usage mining techniques and content-based profiles using information retrieval techniques. Then, we use these profiles to compute relevant links to recommend for an active learner by applying a number of different recommendation strategies.

1. Introduction

Up to the very recent years, most e-learning systems have not been personalized. Several works have addressed the need for personalization in the e-learning domain. However, even today, personalization systems are still mostly confined to research labs, and most of the current e-learning platforms are still delivering the same educational resources in the same way to learners with different profiles. In general, to enable personalization, existing systems used one or more types of knowledge (learners' knowledge, learning material knowledge, learning process knowledge, etc). A number of these systems have relied on explicit information given by a learner (demographic, questionnaire, etc) and have applied known methods and techniques of adapting the presentation and navigation [4]. In fact, as explained in [2], two different classes of adaptation can be considered: *adaptive presentation* and *adaptive navigation support*. Later, in [3], the taxonomy of adaptive hypermedia technologies was updated to add some extensions in relation with new technologies. Then, the distinction between two modes of adaptive navigation support became a necessity, especially with the growth of recommender systems. In fact, adapting links that were already prepared and presented on a certain page is quite different from generating *new* ones. Automatic recommendation implies that the user profiles are created and eventually maintained dynamically by the system without explicit user information. Examples include amazon.com's personalized recommendations and music recommenders like Mystrand.com in commercial systems [7], and smart recommenders in e-learning [12], etc. In general, such systems differ in the input data, in user profiling strategies, and in prediction techniques. Several approaches for automatic personalization have been reported in the literature,

978-0-7695-3167-0/08 $25.00 © 2008 IEEE
DOI 10.1109/ICALT.2008.198

such as content-based or item-based filtering, collaborative filtering, rule-based filtering, and techniques relying on Web usage mining, etc [8]. In the e-learning area, one of the new forms of personalization is to give recommendations to learners in order to support and help them through the e-learning process. In this paper, we present our proposed personalization approach taking into account both the Web access history of learners as well as the content of the learning material. Our approach is based on applying Web usage mining techniques in combination with an open source Web information retrieval system to enable an implementation that is not only open and scalable, but also fast to deploy. The following section describes the proposed approach and the corresponding phases of profiling and recommendation. In Section 3, we present some implementations of the proposed methodologies. In Section 4, we make our conclusions.

2. A framework for building automatic recommendations in e-learning platforms

Our proposed framework is composed of two modules: an off-line module which pre-processes data to build user and content profiles, and an on-line module which uses these models on-the-fly to recognize user goals and predict a recommendation list. Recommended URLs are obtained by using a range of recommendation strategies based mainly on content based filtering and collaborative filtering, each applied separately or in combination. The recommendation procedure is performed using the following tasks:

- **Preliminary offline mining of usage profiles** based on Web usage mining techniques. First, we apply a clustering approach to directly cluster user sessions. Each cluster contains similar sessions, showing similar interests of different learners. Each cluster can also be viewed as one user *profile* ;
- **Preliminary offline mining of association rules** (e.g. “Resource A → Resource B”) from clustered sessions;
- **Preliminary offline crawling and indexing of learning resources**: this step consists of crawling the entire learning resources available in a course repository and forming an *inverted* index mapping each keyword to a set of pages in which it is contained;
- **Extracting user preferences from the learner’s active session** (set of URLs or list of terms extracted from these URLs);
- **Computing relevant links to recommend for the active learner** by applying a number of recommendation strategies.

The proposed approach, with main features depicted in Figure. 1, is essentially based on two components: *the Modeling phase* and the *Recommendation phase* [5].

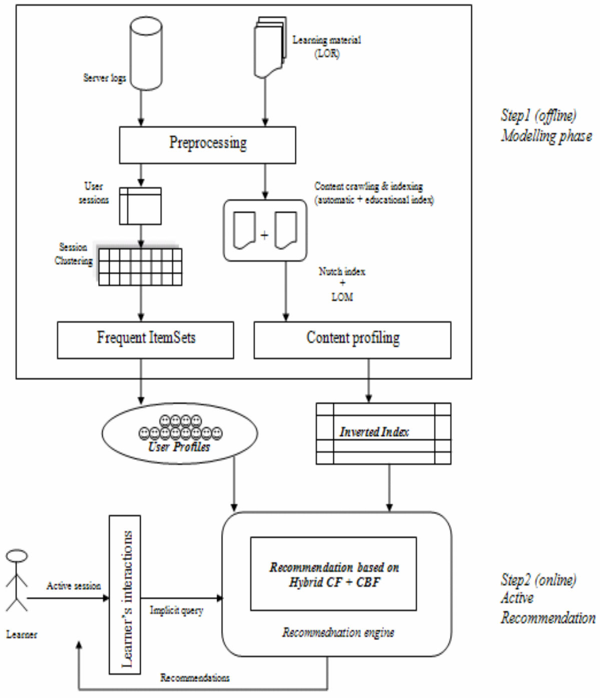


Figure 1: Proposed personalization approach

2.1. Modeling phase

2.1.1. User profiling. We apply data mining techniques to build user profiles, where the prediction of the user model is accomplished not using explicit user interaction, but rather implicit information collected from all past usage sessions. The input data for this first step consists mainly of Web server access log files. In order to extract useful information from log files and build user profiles, we apply Web Usage Mining techniques. First, the server log data is pre-processed, using tasks that include: data cleaning, session identification, etc. The obtained sessions are considered as sets of URLs visited by a learner. Let U be a set of n unique URLs appearing in pre-processed log files: $U = \{u_1, u_2, u_3, \ldots, u_n\}$, and S a set of m users’ sessions extracted from pre-processed log files: $S = \{S_1, S_2, S_3, \ldots, S_m\}$, where each session $S_i \in S$ is a subset of U. Once sessions are delimited properly, we can apply data mining methods to build user profiles.

2.1.2. Content profiling. *G*enerally, content profiling involves applying indexing and text mining *techniques* (which are part of Web content mining). The originalities of our approach are twofold: (1) the use of the Nutch[1] open source search engine in the content modeling phase, followed by content based filtering as a recommendation strategy. Using a search engine's powerful capabilities can *automate* and *scale* the *crawling and indexing* of the learning material; (2) we automate the indexing of *educational* content based on norms and standards used in *e-learning*. Thus, our addition of an index used in *LOM (Learning Object Metadata)* for learning content (if available) to the preliminary crawling and indexing phase done by Nutch is expected to improve the accuracy of the final index, and likewise, improve content search and recommendations. Finally, we note that a byproduct of our crawling and indexing with Nutch is an available interface to use for explicit searches over the material if needed.

2.2. Recommendation phase

2.2.1. Formulating a learner's implicit query. This initial step in the recommendation process consists of transforming a new user session into a set of URLs expressing user preferences and interests in visited learning resources. This task is accomplished in two phases (1) delimiting the current active learner session, and (2) extracting URLs of interest from this active session. Since the active user session will be extracted from the Web log file, we identify only the log records representing the last *W* visited pages in the active user session, which is called the sliding window. The recommendation process will then depend on these pages or on the relevant terms that they contain. On the other hand, to make selected pages more closely express learner interests, a weight can be associated with each URL contained in the learner's session. This weight can be binary (existence or non-existence of a URL referenced in a session), or it can be computed as a function of a number of features based essentially on the frequency of occurrence of the URL within a session and/or the time that a learner spends on a particular page, which could express implicitly the fact that a learner liked or disliked the URL [10], [11]. However, since such measures have been found to be inaccurate as indicators of the user interest [6], we did not consider the URL weight *in the present work*.

2.2.2. Recommendation process. This task is principally based on a content-based filtering approach (CBF) and a collaborative filtering approach (CF). Several recommendation strategies based on these approaches have been investigated in our work. First, we applied the (CBF) approach alone using the search functionalities of the Nutch search engine. We first extract the top *K* relevant terms from a *sliding window consisting of* the last *W* pages in the current session, and then submit them to the search engine in order to compute recommendation links. We also applied the (CF) approach alone using Association Rules (*ARs*) (mined in the offline phase), and we compared the sliding window pages to these *ARs* in order to find relevant recommendations. The Apriori algorithm [1] was used as the AR mining technique. Finally, following our initial goals, we included the possibility to combine *both* of the recommendation approaches (CBF and CF) in order to improve the recommendation quality and generate the most relevant learning objects to learners. Hence, two approaches were considered: *Hybrid content via profile based collaborative filtering* with *cascaded/feature augmentation combination*, which performs collaborative recommendation followed by content recommendation (the reverse order could also be considered); and *Hybrid content and profile based collaborative filtering* with *weighted combination*, where the collaborative filtering and content based filtering recommendations are performed *simultaneously*, then the results of both techniques are combined together to produce a single recommendation set [9].

3. Experimentation and results

To implement and evaluate the proposed personalization approach, we used the course repository of the Virtual University of Tunis, which is accessible via the RPL platform[2]. We conceived a system composed by a set of components, where each component is performing a number of services. The main features of the proposed recommender system are shown in Figure. 2. *Component 1* and *Component 2* represent the input for the recommender system in two different forms. Component 1 extracts the sliding window pages (the *W* last visited pages from the active learner session). These pages are extracted from the Web log file, starting from the time that the learner connected to the platform until he/she asks for recommendations. We consider, here, a fixed window size *W* =3, so that only the last three visited pages may affect the recommendation. The second input form is performed by Component 2 which extracts from Component 1's results (sliding window) the top *K* relevant terms. Each URL of the sliding window is transformed into a set of content terms which are the most relevant terms characterizing this URL. This task is performed using Nutch's built-in functionalities for parsing HTML pages, and a plug-in for stop word elimination. Component 1 and Component 2 are

[1] http://lucene.apache.org/nutch/

[2] http://cours.uvt.rnu.tn

performed in an on-line mode. Component 3 concerns the phase of user profiling. This phase requires that the usage sessions (from which the profiles are to be extracted) be extracted first from the access log data. Therefore, we collected Web usage logs from the RPL log files (Apache web access log files[3]) in the period between February and July 2007.

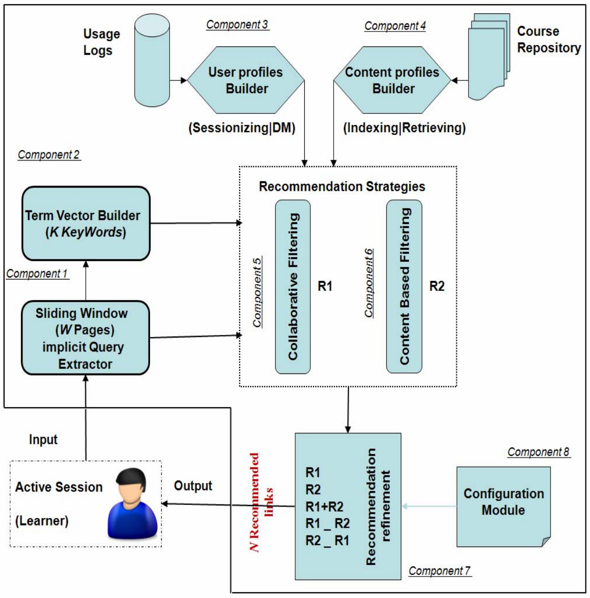


Figure 2: Architecture of the hybrid e-learning recommender system

RPL Log files were rotated based on daily directives, thus generating 180 large log files with a total of 3,049,986 requests. The log format was ECLF (Extended Common Log Format), further enriched with added user authentication information to make the session extraction error free (using an embedded session Id mechanism, added via the Apache configuration and our RPL code). Since the collected log data contains many uninteresting elements (graphics, icons, requests generated by crawlers/bots, etc), it must first be pre-processed. Starting with 3,049,986 requests, the cleaning operations resulted in only 594,325 nontrivial requests. Figure. 3 shows the variation of request numbers before and after data cleansing per month. After cleaning the original logs, we sessionized the remaining requests. Figure. 4 shows the variation of the number of resulting unique usage sessions per month.

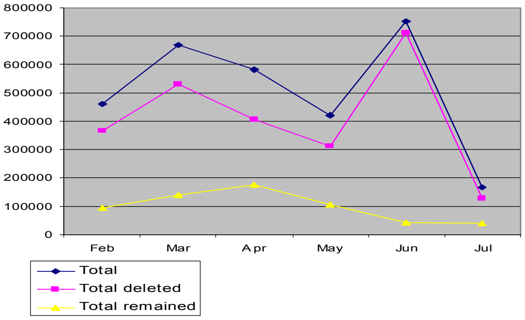


Figure 3: Pre-processed RPL logs per month

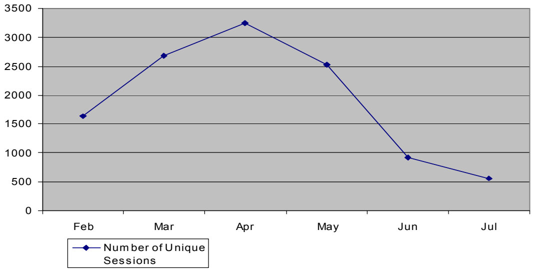
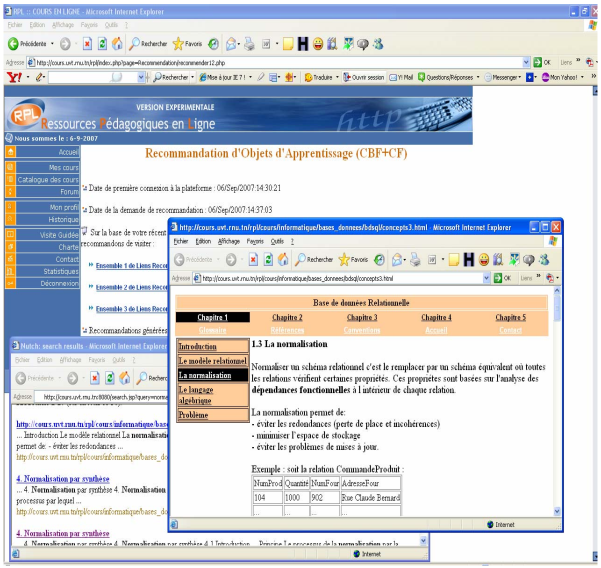


Figure 4: Variation of unique sessions per month

Component 4 concerns the phase of content profiling. Our initial crawling process results in a URL file containing 73,469 URLs representing all learning objects available in the RPL repository. The Nutch crawler was invoked using the following command line: `nutch crawl urls -dir crawldir -depth 6`. The crawler uses the URL file for fetching, parsing and indexing the URLs, thus creating an inverted index which will be used to represent the model of educational content. Moreover, in order to make content models more adapted to the pedagogical area, we added other attributes to the inverted index. These additional attributes are given by the educational metadata providing descriptions and additional information (author, title, technical requirements, rights management, etc) about learning resources. This information is added automatically to the inverted index thanks to *imsmanifest* files -available in the *SCORM* Learning Objects- and Nutch capability to crawl and parse XML files. Component 5 and Component 6 represent the two main recommendation strategies used to compute what to recommend to the learners. Each recommendation strategy uses as input results returned by Components 1 and 3 and/or Components 2 and 4. Component 7 finally performs the task of delivering recommended links by combining the use of various recommendation approaches or by using them separately based on guidelines given by Component 8 which represents the configuration module specifying a set of entry details to the recommender system, such as recommendation strategy and the variation of related parameters (*K*, *W*,

[3] http://httpd.apache.org/docs/2.0/logs.html

N, etc). Figure 5 shows a screenshot explaining the obtained recommended links within the RPL platform based on a cascaded hybrid recommendation strategy.



http://cours.uvt.rnu.tn/rpl/index.php?page=Recommendation/recommender12.php

Nous sommes le : 6-9-2007

Recommandation d'Objets d'Apprentissage (CBF+CF)

Date de première connexion à la plateforme : 06/Sep/2007:14:30:21

Date de la demande de recommandation : 06/Sep/2007:14:37:03

Sur la base de votre récent
recommandons de visiter :

Ensemble 1 de Liens Recon

Ensemble 2 de Liens Recon

Ensemble 3 de Liens Recon

Recommandations générées

http://cours.uvt.rnu.tn/rpl/cours/informatique/bases_donnees/bdsql/concepts3.html - Microsoft Internet Explorer

http://cours.uvt.rnu.tn/rpl/cours/informatique/bases_donnees/bdsql/concepts3.html

Base de données Relationnelle

Chapitre 1 Chapitre 2 Chapitre 3 Chapitre 4 Chapitre 5
Glossaire Références Conventions Accueil Contact

Introduction
Le modèle relationnel
La normalisation
Le langage algébrique
Problème

1.3 La normalisation

Normaliser un schéma relationnel c'est le remplacer par un schéma équivalent où toutes les relations vérifient certaines propriétés. Ces propriétes sont basées sur l'analyse des **dépendances fonctionnelles** à l intérieur de chaque relation.

La normalisation permet de:
- éviter les redondances (perte de place et incohérences)
- minimiser l'espace de stockage
- éviter les problèmes de mises à jour.

Exemple : soit la relation CommandeProduit :

NumProd	Quantité	NumFour	AdresseFour
104	1000	902	Rue Claude Bernard
...	...	...	...

Nutch: search results - Microsoft Internet Explorer

http://cours.uvt.rnu.tn:8080/search.jsp?query=norma

http://cours.uvt.rnu.tn/rpl/cours/informatique/bas
... Introduction Le modèle relationnel La **normalisati**
permet de: - éviter les redondances ...
http://cours.uvt.rnu.tn/rpl/cours/informatique/bases_do

4. Normalisation par synthèse
... 4. **Normalisation** par synthèse 4. **Normalisation**
processus par lequel ...
http://cours.uvt.rnu.tn/rpl/cours/informatique/bases_do

4. Normalisation par synthèse
4. **Normalisation** par synthèse 4. **Normalisation** par synthèse 4.1 Introduction ... Principe Le processus de la **normalisation** par la

Figure 5: Recommendation using a cascaded hybrid approach (CF followed by CBF)

4. Conclusions

In this paper, we have outlined the general principles of a new approach to perform personalization in e-learning platforms by resorting to a recommender system relying on web mining techniques and scalable search engine technology to take care of one of the crucial steps in personalization, which occurs in the "online" phase to compute the recommendations against a possibly massive repository of educational resources in "real time". In the modeling phase, we used Nutch's automated crawling and indexing techniques as well as standardized educational content metadata to build content profiles, and Web usage mining techniques (clustering and association rule mining) to build user profiles. Hybrid recommendations (based on CBF and CF) were used in the recommendation phase. We are currently exploring several techniques and strategies in the modeling and recommendation phase in more detail, and performing more evaluations.

Acknowledgments:

This research was partially supported by National Science Foundation CAREER Award IIS-0133948 to O. Nasraoui, and the Virtual University of Tunis.

5. References

[1] R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules", In Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, September 1994.

[2] P. Brusilovsky, "Methods and techniques of adaptive hypermedia", *User Modeling and User-Adapted Interaction*, 1996, 6 (2-3), pp. 87-129.

[3] P. Brusilovsky, "Adaptive hypermedia", *User Modeling and User Adapted Interaction*, Ten Year Anniversary Issue (Alfred Kobsa, ed.), 2001, 11 (1/2), 87-110.

[4] H. Chorfi, and M. Jemni, "PERSO : Towards an adaptative e-learning system", *Journal of Interactive Learning Research, 2004,* 15 (4), pp 433-447.

[5] M.K. Khribi, M. Jemni, and O. Nasraoui, "Toward a Hybrid Recommender System for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval", In Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, 2007, (pp. 6136-6145). Chesapeake, VA: AACE.

[6] J. Konstan, B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens : applying collaborative filtering to usenet news", *Communications of the ACM* (40) 3, 1997.

[7] B. Mobasher, "Data Mining for Web Personalization", *The Adaptive Web: Methods and Strategies of Web Personalization,* Lecture Notes in Computer Science, New York, 2006, Vol. Springer-Verlag, Berlin-Heidelberg.

[8] O. Nasraoui, "World Wide Web Personalization", Invited chapter in *"Encyclopedia of Data Mining and Data Warehousing"*, 2005, J. Wang, Ed, Idea Group.

[9] O. Nasraoui, Z. Zhang, E. Saka, "Web Recommender System Implementations in Multiple Flavors: Fast and (Care) Free for All", In Proceedings of the ACM-SIGIR Open Source Information Retrieval, 2006.

[10] C. Shahabi, A.M. Zarkesh, J. Adibi and V. Shah, "Knowledge Discovery from User's Web-page Navigation", in *Proc. 7th IEEE Intl. Conf. On Research Issues in Data Engineering,* 1997, 20-29.

[11] T. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking", In Proceedings of the 5th International Worl Wide Web Conference, Paris, France, 1996.

[12] O.R. Zaiane, "Building a Recommender Agent for e-Learning Systems", in Proc. of the 7th International Conference on Computers in Education, Auckland, New Zealand, December, 2002, 3 – 6, pp 55-59.