A Prototype Personalization System for the European Library Portal

Marialena Kyriakidi, Lefteris Stamatogiannakis, Mei Li Triantafyllidi, Maria Vayanou, and Yannis Ioannidis

Department of Informatics and Telecommunications, MaDgIK Lab
University of Athens, Panepistimioupolis, Ilissia, 15784 Athens, Greece
{marilou,estama,meili,vayanou,yannis}@di.uoa.gr

Abstract. In this demonstration, we present a flexible system that enables the provision of personalized functionalities to digital libraries. The system has been developed based on the needs of *The European Library* portal and will be demonstrated in that particular context, but could be applied more generally. It implements a broad set of data processing, analysis, and mining techniques over the portal's log files, using an environment called *madIS*. It enables on-line result contextualization and adaptation through the development of REST web services, which are responsible for retrieving and appropriately integrating the extracted information. The demonstration also features a web-based visualization tool for showing the output of the log analysis performed.

Keywords: Log mining, pattern extraction, profiling, personalization.

1 Introduction

The European Library (TEL) portal offers access to the resources of the 48 national libraries of Europe. In the TEL home page, users can initiate simple keyword searches within subsets of library resources, referred to as *collections*. Users may search in a pre-selected set of collections or select particular collections in a customized fashion. In the results page of a query, the relevant collection list is placed on the left and the documents from each individual collection are placed on the main panel.

To personalize this functionality, i.e., customize it to the profile of individual users or groups, we have studied the portal characteristics and have analyzed the TEL usage logs. For example, since query results are grouped per collection, instead of being fused into a unified ranked list, correct use of collection selection features is crucial for users to effectively exploit TEL services and content. Log analysis results, however, show that in almost 65% of sessions, users perform no collection specification but search within the default collection. Likewise, although login functionality is provided for user registration, these are hardly used, imposing a significant obstacle to the extraction of accurate personal profiles.

The above and other important findings of our investigation have formed the basis for the services offered by the Personalization Prototype to be demonstrated. In particular, to address the data sparsity observed, collaborative approaches are employed,

M. Lalmas et al. (Eds.): ECDL 2010, LNCS 6273, pp. 542–545, 2010.

through the specification of group-level user models. Also, usage analysis revealed that users from the same country tend to exhibit several commonalities regarding their preferences [3], leading the way for the definition of a National user group. In addition to the Personal and National levels, a Global profile exploits the "wisdom of the crowds".

Moving in these directions, we have developed a prototype Personalization system that provides personalized and collaborative functionality to TEL portal. It is flexible and easily adaptable to portal changes and new applications requirements, while it can be quickly integrated into other digital library portals as well.

2 System Architecture

Figure 1 depicts the main components of the personalization system, as integrated with TEL. User requests are issued to the portal, which is hosted in TEL server. The results are returned to the user, while all user interaction is recorded in log files.

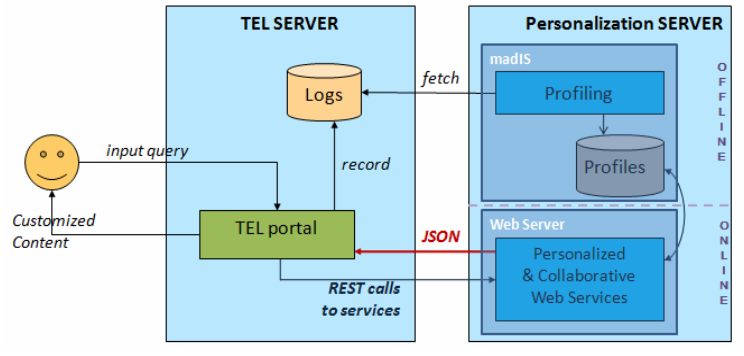

Fig. 1. System Architecture

To adapt the results generated to personal and context-based information, we have followed an implicit feedback approach, using various techniques to process and mine the log data to extract several useful patterns and user preference profiles. All knowledge extraction has been captured by a set of workflows that are executed by the *madIS* prototype processing and analysis environment [2] and are stored in the Personalization server under a relational schema. Due to the high computational complexity of profiling and to avoid any degradation of the system's performance, profile extraction takes place offline, in compliance with the basic principle of online / offline separation [1]. The profiling service is executed periodically for fetching new log data, which is subsequently used for updating existing patterns and profiles.

The information extracted is accessed at run-time by a set of *REST* web services that are responsible for combining related patterns and profiles to provide the corresponding customized information. They use madIS for data processing and retrieval and their results are returned to the invoking entity, encoded in *JSON*. These services are also hosted in the Personalization server and web service functionality and protocols are provided by the *Tornado* web server.

3 Log Mining and Profiling

TEL usage mining has been implemented through madIS workflows, containing a series of queries expressed in an SQL-based declarative language extended with User Defined Functions (UDFs) implemented in Python. As in traditional data mining, workflows include activities for Data Collection, Log Cleaning, Log Transformation, and Pattern Extraction.

In Data Collection, the "Initialization workflow" imports and integrates various types of TEL data (e.g., application log files, collection descriptions, users' registration information, saved queries, and favorites) as well as external data (e.g. GeoIP database, ISO country codes, stop word lists, and statistical language models).

In Log Cleaning and Transformation, several workflows resolve inconsistencies and assemble data into an integrated and comprehensive view. For example, a typical web usage mining activity is *search session reconstruction*, grouping user actions in comprehensive efforts towards one goal. One workflow employs the popular 30 minutes of inactivity timeout, which is shown to be both effective and efficient [4]. Another workflow maps sessions to country codes, which is useful for subsequently extracting national patterns. Finally, another workflow classifies sessions into Expert or Non-Expert (based on ad-hoc session characteristics found critical during log analysis), which is useful for subsequently emphasizing expert behavior more heavily.

In Pattern Extraction, five additional workflows are used to construct specialized patterns or profiles. Depending on the goal of each workflow, it may also include additional processing, e.g., stop word removal, stemming, and language detection. An Apriori-like data mining algorithm has been implemented for extracting correlations among query keywords and is applied at three profile levels: Personal, National, and Global. Expert sessions are emphasized within the computation of group profiles using heuristically defined weights. In addition, two term-indexing table structures are constructed based on query-term frequency for "original query recommendation".

Regarding collection usage, *frecency* metrics are employed, combining frequency and recency, to effectively capture concept drift and temporal trends. Computation is performed again at all profile levels, resulting in three ranked lists of collections. Moreover, correlations between queries and collections are extracted over the group-level profiles, based on frequency measures, while some additional statistics are computed to quantify secondary user actions, such as selection of Advanced Search Fields, Collection Themes, etc. Finally, a user similarity matrix is constructed capturing similarity between each pair of users over a variety of dimensions (user interests, collection usage, queries, favorite object descriptions) that are integrated into a unified similarity score.

4 Demonstration Overview

Addressing the needs and characteristics of the TEL portal, the following five adaptive services are provided, which combine all available profile levels while always emphasizing the finer grained ones:

1. *Personalized Collection Ranking* computes a personalized and context-aware ranking of TEL Collections

2. *Collaborative Query Suggestion* produces "original query recommendations"
3. *Personalized Term Suggestion* generates a related terms cloud
4. *Collaborative Query based Collection Recommendation* generates a list of top recommended collections with regard to the query issued
5. *Personalized User Notification* retrieves top similar users along with their preferred content, thus providing an enhanced, user-aware platform.

A demonstrator tool has been developed, receiving the *JSON* response of each service and presenting it within a web browser window. The graphical interface depicted in Figure 2 has been employed for testing and experimentation during log analysis and it has been extended for demonstrating key statistical results and patterns. During the demonstration, users will be able to set a variety of input parameters and explore the output results of each service using the desired graphical representations. The interface is targeted towards results exploration and it is not meant to be used by TEL end users.

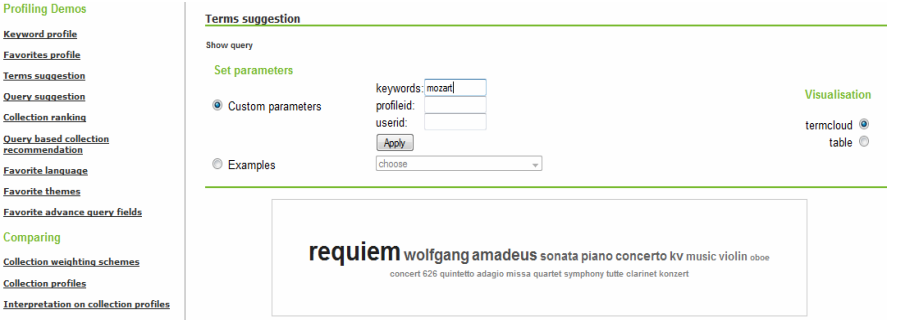

Fig. 2. Graphical Interface of Demonstration Tool

Acknowledgments. This work was done within the TELplus project (www.theeuropeanlibrary.org/ telplus/), funded by the European Commission. We would like to thank the TEL Office for its valuable help during the project, especially Georgia Angelaki and Anna Gos. We are also grateful to our colleagues from the Univ. of Padua, especially Maristella Agosti and Giorgio Maria Di Nunzio for their feedback and valuable discussions during our collaboration within the project.

References

1. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on Web usage mining. Communications of the ACM 43(8), 142–151 (2000)
2. `http://code.google.com/p/madis/`
3. Agosti, M., Crivellari, F., Di Nunzio, G.M., Ioannidis, Y., Stamatogiannakis, E., Triantafyllidi, M.L., Vayanou, M.: Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for The European Library. In: IRCDL, pp. 120–135 (2009)
4. Radlinski, F., Joachims, T.: Query chains: learning to rank from implicit feedback. In: ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (2005)