Towards personalization in digital libraries through ontologies

Núria Ferran[1], Enric Mor[2] and Julià Minguillón[2]

[1]Digital Library, Universitat Oberta de Catalunya (UOC)
[2]Computer Science and Multimedia Studies, Universitat Oberta de Catalunya (UOC)
{nferranf,emor,jminguillona}@uoc.edu

Keywords

Digital libraries, ontologies, semantic web, navigational patterns, personalization, community involvement

Abstract

In this paper we describe a browsing and searching personalization system for digital libraries based on the use of ontologies for describing the relationships between all the elements which take part in a digital library scenario of use. The main goal of this project is to help the users of a digital library to improve their experience of use by means of two complementary strategies: first, by maintaining a complete history record of his or her browsing and searching activities, which is part of a navigational user profile which includes preferences and all the aspects related to community involvement; and second, by reusing all the knowledge which has been extracted from previous usage from other users with similar profiles. This can be accomplished in terms of narrowing and focusing the search results and browsing options through the use of a recommendation system which organizes such results in the most appropriate manner, using ontologies and concepts drawn from the semantic web field. The complete integration of the experience of use of a digital library in the learning process is also pursued. Both the usage and information organization can be also exploited to extract useful knowledge from the way users interact with a digital library, knowledge that can be used to improve several design aspects of the library, ranging from internal organization aspects to human factors and user interfaces. Although this project is still on an early development stage, it is possible to identify all the desired functionalities and requirements that are necessary to fully integrate the use of a digital library in an e-learning environment.

INTRODUCTION

Distance education is becoming one of the most attractive methods for incorporating all kinds of people into higher and university degree education levels, moving towards a "blended" technology approach deploying multiple technologies. The introduction of new technologies of information and communications with the intensive use of e-learning environments, such a virtual campus, for example, allows students to break through the barriers of space and time, and to design their own lifelong curricula, adapting it to their particular necessities and preferences, according to their possibilities as students, changing the usual way of both teaching and learning (Jonassen, 1995).

The students of an e-learning environment have access to a predetermined repository of learning resources, which are part of the learning process designed by the team of instructional designers and teachers for each course. But usually, these students might need more additional learning resources and documents to successfully follow the

recommended learning itineraries, so there is the necessity of providing mechanisms for accessing such resources in a structured manner. On the other hand, researchers and teachers in a collaborative e-learning environment are also usual customers of the services offered by an academic library, although under a completely different approach. These kinds of users have different experiences of use of the digital library than students, as their goals are clearly different, in both content and context aspects. All this richness needs to be captured somehow in order to better understand the way users of a digital library perform their navigation. Digital libraries must change in order to be proactive, more responsive to possible changes and to include new services such as personalization in order to increase user satisfaction and fidelity.

Therefore, in an ideal scenario, the digital library should adapt to the specific characteristics of each user profile, but also to the particular necessities and preferences of each user, combining both user and profile level personalization capabilities. Personalization is one of the key factors which are directly related to user satisfaction (Riecken, 2000) and, therefore, linked to the failure or success of the performed activity, although must be carefully introduced (Nielsen, 1998). Personalization has been shown useful in several areas such as e-commerce (Kasanoff, 2001), business to business companies (Colkin, 2001), and obviously reproduced in other environments such e-learning (Mor, 2004), for example. Regarding the library management field, there exist remarkable approaches such as the recommendation system for electronic journals of MyLibrary from Los Alamos National Laboratory or MyOpenLibrary from the Open University; the personalization librarian initiatives are showing successful results as appears at the research being done about the user satisfaction (Sinha & Swearingen, 2001).

In order to build such personalization system, several multidisciplinary aspects must be addressed: first, there are cognitive and behavioral aspects (Ford & Ford, 1993) that determine the way users perform searches and examine the obtained results. The “I’m Feeling Lucky” button in the Google search engine home page is a good example of such fact. Second, personalization issues must be addressed from a user-centered point of view, under the approach of human computer interaction, as it is well known (Schonberg, 2000) that most personalization systems fail, not because of the personalization system in itself, but in the interaction with the user and the way recommendations are presented. Third, there are technological and knowledge engineering aspects related to the way all this information is structured for both updating and querying purposes. In this paper we describe the set of desired functionalities and requirements of an ideal scenario for a digital library which includes personalization capabilities by means of ontologies. The use of ontologies for describing the possible scenarios of use in a digital library brings the possibility of predicting user requirements in advance and to offer personalized services ahead of expressed need. This suggests that there is a need for further focus on the interoperability of objects which in turn requires well developed ontologies to describe the properties of both objects and individuals and the relationships between them (Brophy, 2004).

DIGITAL LIBRARIES

As stated previously, one of the most important resources for supporting users in a distance e-learning environment is the possibility of accessing to a digital library, which allows the users to collect and organize the necessary information for achieving their particular goals. Furthermore, the search of information can be a learning but also an assessment activity by itself, so it is important to ensure and facilitate a proper use of the library.

There are several terms being used interchangeably when we approach the concept of a library with digitized data and accessible remotely. Among these we can find hybrid library, digital library and virtual library. An informal definition of a digital library is "a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network" (Arms, 2001). The hybrid library provides electronic information sources too but also paper-based information. The hybrid should be considered as a model by itself not as a transitional phase from a conventional library to a digital one (Brophy, 2001). And finally the virtual library has been defined as the concept of a "remote access to the contents and services of libraries and other information resources, combining an on-site collection of current and heavily used materials in both print and electronic form, with an electronic network which provides access to, and delivery from, external worldwide library and commercial information and knowledge sources" (Gapen, 1993).

Nowadays, we are finding new types of libraries coming up from long-term personal digital libraries, as well as digital libraries that serve specific organizations, educational needs, and cultural heritage and that vary in their reliability, authority and quality. Besides, the collections are becoming more heterogeneous in terms of their creators, content, media, and communities served. In addition, the user communities are becoming heterogeneous in terms of their interests, backgrounds, and skill levels, ranging from novices to experts in a specific subject area (Callan & Smeaton, 2004). This growing diversity has changed the initial focus of providing access to digital content and transforming the traditional services into digital ones to a new handicap where the next generation of libraries should be more proactive offering personalized information to their users taking in consideration each person individually (his or her goals, interests, level of education, etc.).

While data and information are captured and represented in various digital formats, and rapidly proliferating, the techniques for accessing data and information are rudimentary and imprecise, mostly based on simple keyword indexes, relational queries, and/or low-level image or audio features (i.e., research results of the 70s and 80s). In the current context of explosive availability of data, there is a need for a knowledge discovery approach, based on both top-down knowledge creation (e.g., ontologies, subject headings, user modeling) and bottom-up automated knowledge extraction (e.g., data mining, text mining, web mining), promises to help transfer digital library from an institution of data and information to an institution of knowledge (Chen, 2003).

The UOC virtual library

The UOC Library was born in 1995 as a virtual academic library to support an e-learning university model and, since then, provides online services and information resources both print and digital, owned by the library or by other libraries. Users can access the library from any computer and do not need to do any travel in order to get any information resource. All authorized users are able to use remotely the exclusive content of the library as subscribed databases like the Electronic Management Research Library Database (Emerald) or the Web of knowledge from ISI and benefit from the services of the library such as the Selective Diffusion of Information, Bibliographic searches or Loans all of them performed through the web.

The UOC virtual library can be accessed in different ways. One of them is from the virtual campus where the user finds the whole content and services of the library. But the main entrance to the library can be found in the campus's virtual classrooms where the teachers and the librarians bring a selection of the most interesting resources for

every subject, for instance the learning material, recommended bibliography where each book is linked with the loan form, electronic articles, self-assessment exercises, a selection of Internet resources, databases and electronic journals, e-books, exams from previous years, etc. This has been the first step for tailoring information for a very well defined community.

The main objective of the library is to provide the students, but also lecturers, researchers and management staff, access to the information relevant to the fulfillment of their basic functions: learning, teaching, research and management. There exist several profiles: undergraduate student, Ph.D. student, teacher, learning process manager, among others. Each profile can be partially identified by the tasks related to the digital library that it performs. For example, students usually browse the digital library looking for documents related to the activities such as exercises, exams and recommended articles and in very specific periods of time, when the delivery date of the exercise is due. On the other hand, teachers can navigate among the content of the library in order to mentor a student in doing his or her homework or providing content to the digital library associated to the virtual classroom. A final example could be the researchers who usually perform more focused searches during a larger period of time. Each of these users may have common goals in certain times but their knowledge, tasks, social activities and preferences are totally different.

BROWSING, SEARCHING AND PERSONALIZATION

The web has become a very common tool for information browsing and searching, and the success of search engines such as Google or A9, for example, has facilitated the diffusion and access to repositories of digital documents. Nevertheless, one of the main problems of such search engines is that the generated results are not always of interest for the users performing the search, as these engines use a generalist approach based on several criteria which might not match the criteria of a specific user. On the other hand, several e-commerce web-based services, such as Amazon or e-Bay, for example, also provide browsing and searching services, but focused on categories. Both approaches can be combined to facilitate the way users browse the contents in search of information.

Several techniques are used for guidance and for providing recommendations to users; among others, collaborative filtering (Herlocker, 2004) is one of the most successful ones. Briefly, collaborative filtering is selecting content based on the preferences of people with similar interests, basically by pooling and ranking informed opinions (or experiences of use) on any particular topic. That is to say, an automatic system collects information about user actions (explicit, such as voting or answering a question; or implicit, such as noticing which offered links are visited and which are not, and how much time) and determines the relative importance of each content by weighting all the collected information among the large amount of users.

Both navigational techniques are also valid in a digital library scenario of use: simple searches starting from a single search term or advanced searches using multiple criteria, but also a recommendation system based on guided navigation through an ordered set of categories. The basic idea of this paper is that the efforts for finding a useful piece of information in a digital library carried out by an individual can be stored in a structured way and then shared for future users with similar necessities. Furthermore, if such information searching and browsing combines several web-based resources with different approaches (access mechanisms, query languages and interfaces, and so on), it is important to describe a common strategy (Sadeh, 2003) for

minimizing the necessary efforts to fight against duplication (thus inconsistency) and source diversity.

Identification of personalization system functionalities

Two elements determine the functionalities of the desired personalization system: first, the user's profile, including navigational history and user preferences; and second, the information collected from the navigational behavior of the digital library users. User profile should include all the information relevant to user: personal information, which is publicly made available by each user in order to facilitate the discovery of similar interests, and navigational history and behavior records, which will be used altogether with the personal information by the personalization system to build the set of recommendations that will help each user in browsing and searching the digital library. This information should help the user to improve his or her searches, by obtaining additional information when searching or browsing. It is remarkable to say that this information has been validated by the ontology, and that is not biased by any non-academic purpose of use (such as commercial-supported recommendations in Google or Amazon, for example).

Two different behavior types can be identified, depending on the users' navigation, exploratory navigation and goal-oriented navigation. The exploratory navigation can be mainly oriented to obtain a general vision of the available resources in the library. Depending on user profile, the exploratory navigation would have different implicit intentions. In the case of goal-oriented navigation, it is usually considered that the user is looking for a resource. These searches can be classified in different use cases. For example, in the situation of searching for an author, if the user is a student, the recommendations associated to search results should be oriented to the area of the course subjects, taking into account the navigation of other students and also the recommendations of the teachers. If the user is a researcher, recommendations should be oriented by different criteria depending on the searches that have been carried out by other investigators, or to the magazines, books and conferences where the searched author had published, understanding that the same magazine, conference or book might contain other interesting resources. Recommendations are generated using the knowledge extracted from the searching and browsing profiles of users with similar interests, knowledge integrated in the ontology such as course bibliography, or by following citations of similar documents, for example. Regarding the sources of information, using the library for accessing selected free Internet resources will be of particular interest, because the recommendation system ensures the users receive the opinions of a large set of experts (that is, the collaborative filtering system), therefore giving authenticity to such electronic information sources. On the other hand, regarding the library exclusive content, the user will get into the external databases (commercial) or internal databases (such as digital repositories or catalogues) in a transparent manner.

For library managers, the creation of an ontology will help them to construct tailored libraries for each subject. Every library is built on the explicit recommendations from a teacher, but in an unstructured manner. With an ontology, these specialized libraries could be built from the use that previous students gave to that resources and new information could be added from the use of the library by experts.

Privacy issues

A very important aspect that cannot be ignored is the fact that users are always under control, in the sense that all taken actions are monitored and registered. This might seem a very invasive setup which harms user privacy and, therefore, undesirable.

Nevertheless, there are several remarkable facts that need to be clarified: a) users know in advance that, in a virtual e-learning environment (or any other web based environment), all actions are logged; b) the recommendation system must be designed in a non-intrusive manner and be user-friendly, including the possibility of disconnecting it or minimizing its participation in the browsing or searching activities; and c) the participation of each individual user in the final recommendation system is completely anonymous. Finally, it is also important to remark that the collected information is only used with personalization purposes, and it is not meant for commercial reasons, and that the library (a non-profit organization) will use the data rationally and in a transparent way.

On the other hand, the browsing and searching history of each user is part of his or her private profile, and only the user can modify (delete) his or her history records in order to update his or her preferences and navigational profile, according to the directions given by the personalization system. Some parts of this profile can be made publicly available in order to create strong links with other users sharing the same navigational interests. There is evidence that people are sometimes eager to be identified and become part of a community (Lynch, 2003), and the use of ontologies could be one of the strengths of digital libraries for pursuing such purpose.

ONTOLOGIES AND THE SEMANTIC WEB

Therefore, it is necessary to build a complete and complex structure for describing all the richness of the possible scenarios of use of the digital library and the relationships which can be established among all the participants. This can be achieved by means of ontologies and the use of the semantic web services.

An ontology is, taking the meaning adopted by the semantic web community, a formal description of a possible scenario or context; that is, what "exists" is what can be represented by an ontology. Formally, an ontology is the statement of a logical theory, but by “formal description” we also mean that it can be automatically queried and updated, as the main users of ontologies are (or should be) computers, not humans, in order to explicitly represent the objects, concepts and other entities that are assumed to exist in some context, altogether with the relationships that hold among them, although ontologies must be also human-readable. “Ontologies and taxonomies are, in functional terms, often used as synonyms. Computer scientists call hierarchies of structured vocabularies *ontologies* and librarians deploy the term *taxonomy*” (Adams, 2002). Nevertheless, ontologies also include a set of semantic rules which are used to infer knowledge from a structured hierarchy of information, giving to the complete structure a semantic meaning, not only syntactic (Gruber,1993).

Building the digital library personalization ontology

First, it is important to clarify that we are not building an ontology for describing the contents of a digital library, but an ontology for describing the way users browse and search such contents, with the aim of building a personalization system based on accurate recommendations. Therefore, more than building a low-level ontology for describing a particular concept, we are trying to describe a complex scenario of use.

Creating an ontology (Denny, 2002) for describing the richness of such complex scenario as is a digital library should follow these recommendations: 1) acquire domain knowledge: assemble appropriate information resources and expertise that will define, with consensus and consistency, the terms used formally to describe things in the domain of interest. This is being carried out by several teams in the UOC, in particular

digital library managers, computer scientists and usability experts take part of the project described in this paper; 2) organize the ontology: design the overall conceptual structure of the domain, identifying the domain's principal concrete concepts and their properties, the relationships among the concepts, creating abstract concepts as organizing features, referencing or including supporting ontologies, and distinguishing which concepts have instances, and applying other guidelines of the specific scenario of use. In a digital library, this means specifying formal definitions for the following concepts: user profile, electronic (digital) information resource, navigational profile and learning process activities. These are the main high-level concepts, and it would be interesting to reuse previous ontologies as much as possible, in order to minimize overlaps and inconsistencies. Regarding the relationships between these concepts, it is necessary to define the actions that users perform in the digital library, but also the possible actions that the personalization system takes in order to make recommendations; 3) elaborate the ontology: add concepts, relations, and individuals to the level of detail necessary to satisfy the purposes of the ontology. This consists in creating the initial user profiles and navigational records, and tagging all the information resources according to the definitions which have been determined in the previous step. This should be performed trying to reuse the maximum information from the current environment; for example, all the documents in the digital library are already partially or totally catalogued, thus this information should be incorporated into the ontology automatically; 4) consistency checking: reconcile syntactic, logical, and semantic inconsistencies among the ontology elements, involve automatic classification that defines new concepts based on individual properties and class relationships. In the particular case of an ontology associated to a scenario of use, this means collecting data during a period of time, and validating the recommendation system using a reduced set of users which have been previously trained; and finally, 5) validate the ontology: incumbent on any ontology development effort is a final verification of the ontology by domain experts and the subsequent commitment of the ontology by publishing it within its intended deployment environment. Once again, it is necessary to compare the obtained results (i.e., the recommendations given by the system) with the desired goals, in order to detect any possible mistake or misuse of the personalization system, using the results collected in the previous step. As usual, these steps overlap and must be taken in a recursive way: depending on the results of the evaluation undertaken in the fifth step, several definitions in the second and next steps might be modified. In fact, the ontology in itself will evolve with the new apparition of desired functionalities and requirements.

It is also remarkable that the use of ontologies can be extended to implement and transfer the concept of user profile and user navigational behavior to other digital libraries and databases, so when a digital library user leaves one service to connect into another one, the user profile (including preferences and navigational behavior) can be transferred from one database to another through the appropriate semantic web services, because all databases share a common domain of discourse that can be interpreted further by rules of inference and application logic.

Regarding implementation issues, ontologies are usually described by means of one or more descriptive languages based on XML (W3C, 2004a). Basically, RDF (W3C, 2004b) is used for describing resources, while DAML+OIL (W3C, 2004b) which is currently being evolved into the Web Ontology Language (OWL) standard (W3C, 2004c), is becoming the standard for describing ontologies and accessing resources through the web. The use of XML and description language standards ensures the interconnectivity with other existing ontologies and the possibility of upgrading the ontology for new requirements and functionalities. The widespread of XML for describing data (but also information and knowledge, with the help of ontologies) has made possible the apparition of new applications into classical areas of knowledge.

CONCLUSIONS

In this paper we have described a personalization system which uses all information relevant to the process of searching and browsing a digital library to build a complete navigational profile for each user. Then, all these profiles are then combined with the help of an ontology that establishes the possible relationships between all the elements present in a typical scenario of use in a digital library integrated in an e-learning environment. We have described the basic functionalities of the personalization system by means of use cases, and a methodology for building and ontology which describes the complete scenario of use.

Ontologies are a powerful tool for describing complex scenarios of use such as a digital library, where several concepts and relationships between these concepts can be identified and formally represented. The use of ontologies promotes the integration of new services into existing ones, and the interoperability with other systems through the appropriate semantic web services.

Current and further research in this subject include the integration of the digital library personalization services with other personalization mechanisms provided by the virtual campus, towards a unique and complete user model. The inclusion of new concepts related to the temporal validity of the ontology instances (resources, users and so) and their relationships should be also addressed. Finally, the definition of a validation rating algorithm combining both automatic but also user explicit rating systems is also under consideration.

REFERENCES

Adams, K. (2002), “The Semantic Web: differentiating between taxonomies and ontologies”, Online, Vol. 26, No. 4, pp. 20-23

Arms, W.Y. (2002), Digital Libraries, The MIT Press, Cambridge, MA

Brophy, P. (2001), The library in the twenty-first century, Library Association Publishing, London, UK

Brophy, P. (2004), “Networked Learning and Networked Information: Towards a Theoretical Basis for the Development of Integrated Information Environment”, Proceedings of the 4th Networked learning conference, Lancaster, UK. Available at http://www.shef.ac.uk/nlc2004/Proceedings/Symposia/Symposium9/Brophy.htm

Callan, J. and Smeaton, A. (2003), ”Personalization and Recommender Systems in Digital Libraries”, Joint NSF-EU DELOS Working Group Report. Available at http://www.dli2.nsf.gov/internationalprojects/working_group_reports/personalisation.html

Chen, H. (2003) “Towards Building Digital Library as an Institution of Knowledge”, NSF Post Digital Library Futures Workshop, Chatham, MA. Available at http://www.sis.pitt.edu/%7Edlwkshop/paper_chen.html

Colkin, E. (2001), “Personalization tools dig deeper”, Information Week, No. 852, pp. 49-51

Denny, M. (2002), “Ontology Building: A Survey of Editing Tools”. Available at http://www.xml.com/pub/a/2002/11/06/ontologies.html

Di Giacomo, M. and Mahoney, D. and Bollen, J. and Monroy-Hernandez, A. and Ruiz-Meraz, C.M. (2001), “MyLibrary, A Personalization Service for Digital Library Environments”, Proceedings of the Second DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries. Available at http://www.ercim.org/publication/ws-proceedings/DelNoe02/Giacomo.pdf

Ford, N. and Ford, R. (1993), “Towards a Cognitive Theory of Information Accessing: An Empirical Study”, Information Processing & Management, Vol. 29, No. 5, pp. 569-585

Gapen, D.K. (1993) “The Virtual Library: Knowledge, Society and the Librarian”, in Saunders, L. (Ed.), The Virtual Library Visions and Realities, Meckler Corporation Westport, CT, pp. 1-14

Gruber, T.R. (1995), “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, International Journal of Human-Computer Studies, Vol. 43, pp. 907-928

Herlocker, J.L. and Konstan, J.A. and Terveen, L.G. and Riedl, J.T. (2004), “Evaluating collaborative filtering recommender systems”, ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 5-53

Jonassen, D. and Davidson, M. and Collins, M. and Campbell, J. and Haag, B. (1995), “Constructivism and computer-mediated communication in distance education”, The American Journal of Distance Education, Vol. 9, pp. 17-23

Kasanoff, B. (2001), Making it Personal, Perseus Publishing, Cambridge, MA

Lynch, C. (2003), ”Digital Library Opportunities”, Journal of Academic Librarianship, Vol. 29, No. 5, pp. 286-289

Mor, E. and Minguillón, J. (2004), “E-learning Personalization based on Itineraries and Long-term Navigational Behavior”, Proceedings of the 13th International World Wide Web Conference, Alternate Track Papers & Posters, pp. 264-265

Nielsen, J. (1998), “Personalization is over-rated”, Alertbox. Available at http://www.useit.com/alertbox/981004.html

Riecken, D. (2000), “Personalized views of personalization”, Communications of the ACM, Vol. 43, No. 8, pp. 27–28

Sadeh, T. and Walker, J. (2003), “Library portals: toward the semantic web”, New Library World, Vol. 104, No. 1184/1185, pp. 11-19

Schonberg, E. and Cofino T. and Hoch R. and Podlaseck M. and Spraragen S.L. (2000), “Measuring success”, Communications of the ACM, Vol. 43, No. 8, pp. 53-57

Sinha, R. and Swearingen, K. (2001), “Comparing Recommendations Made by Online Systems and Friends”, Proceedings of the Second DELOS Network of Excellence Workshop on Personalization and Recommender Systems in Digital Libraries. Available at http://www.ercim.org/publication/ws-proceedings/DelNoe02/RashmiSinha.pdf

W3C (2004a), eXtensible Markup Language (XML) 1.0 (Third Edition). Available at http://www.w3.org/TR/REC-xml/

W3C (2004b), Resource Description Framework (RDF) Primer. Available at http://www.w3.org/TR/rdf-primer/

W3C (2004c), Web Ontology Language (OWL) Guide. Available at http://www.w3.org/TR/owl-guide/