Advanced search and browsing in digital libraries

Sebastian R. Kruk
DERI.Galway, Ireland

Abstract

In the past the main source of reliable information were libraries. Nowadays, the Web is more and more assuming the role of a major information provider. But the move from the well organized libraries to unsupervised and unstructured Internet is causing serious problems. Although it should be possible to find almost anything on the Internet, very often one is unable to find the needle in the haystack. In this paper we present *JeromeDL* - a digital library with semantics. JeromeDL uses Semantic Web technologies together with standard bibliographic description formats to improve efficiency and usability of the searching and browsing process. We show how the DublinCore and FOAF vocabularies work together with well known bibliographic formats like the MARC21 and BibTEX to achieve better search results. Furthermore, *FOAFRealm*, an infrastructure introduced to handle a users profile, provides means to share bookmarks and annotations of the resources within a network of friends. We discuss how this information is used in the search process within JeromeDL.

1 Introduction

In the recent years access to the Internet has become a commodity. The growing number of users with broadband access is causing increased demand for high quality and accurate information. One of the sources of high quality information tend of be libraries. In the Internet, digital libraries are often the islands of high quality and well organized information. Digital libraries not only make electronic information available, but also provide access to legacy information (e.g., old books) which were perviously only accessible to a restricted minority of privileged library users.

Classic libraries are respected for the quality of service they provide. In order to reach this status of trust and reliability, digital libraries require effective information access facilities. But digital libraries are not restricted to conventional means - e.g., information access can also benefit from the development of extensible browsing facilities based on the social connections between readers.

Bibliographic descriptions have been used in conventional libraries for centuries. With semantic web technologies the interpretation of bibliographic description standards as ontologies is a natural choice. The semantic description paradigm that has grown, inter alia, out of the bibliographic formats like DublinCore[2] provide a basis for improving search and access capabilities of standard bibliographic descriptions. To uphold the legacy of the classic library systems, the ontology developed for the digital libraries needs to be compatible with widely used description formats such as the MARC21[3] or BibTEX[6].

2 JeromeDL - e-library with semantics

The Jerome Digital Library[1] is the fully featured digital library system which utilises the semantic description of the resources. JeromeDL utilises the following resources during the searching process consists:

- annotations (provided by the readers) about the resources,
- the MARC21[3] and the BibTEX[6] description of resource,
- the semantic description based on the JeromeDL ontol ogy (compatible with the DublinCore Metadata[2]),
- users' preferences retrieved from statistically extrapolated profiles in the friendship network
- the fulltext index of the content of the resources annotations

The system has been written in the J2EE platform, which enables scalability and extensibility. An administrative application connects to the system with RMI, while other digital libraries and client applications can communicate with JeromeDL using a SOAP extension. The JeromeDL system provides many additional security features, which can be used to protect the access to the resources.

Friendship-based profiles extrapolation

JeromeDL collects information about most frequently read books for individual users. Users are identified using cookies[1]. The mechanism allowed unregistered users to benefit from advanced searching features like result set tailoring by exploiting an statistical analysis of the domains of interests of the particular reader. During the query expansion in the search process the result set has been tailored to the readers' expectations according to their domain of interests.

Additionally users can register to JeromeDL. The explicit identification allows to collect the statistical information more accurately. To improve available profile information users can select their friends that have already register to the JeromeDL system. The friendship graph is described with the FOAF[2] metadata, enriched with trust information trust[4]. Using the FOAF appraoch JeromeDL is able to predict the profile of a new user by extrapolating the profile information of his friends which are known to JeromeDL.

The extrapolated profiles holds the statistical information about a users preferences. This information contains statistical data about frequently viewed resources, bookmarks and annotations about the resources. For creation of extrapolated profiles or to share the bookmarks within the friendship network the FOAFRealm[4] library has been used.

Search algorithm

The first three steps (**A**, **B**, **C**) of the search algorithm (see Figure 1 on the following page) prepare a sketch of the final result set. These three steps are based on the content of the books, the users annotations, MARC21 and BibTEX descriptions. In the next step (**D**) additional information about user preferences (e.g. preferred domains of interests) and detailed semantic descriptions of the resources is exploited. The sketch of the result set is then being tailored to suit the users' requirements. The information about the resources that the user decides to explore is used to update the statistical data in the user profile.

[1] The JeromeDL system (http://www.jeromedl.org/) is an internationalised version of ElvisDL (http://elvis-dl.sf.net/). Both versions are distributed under the GNU Public Licence.

[2] http://www.foaf-project.org/

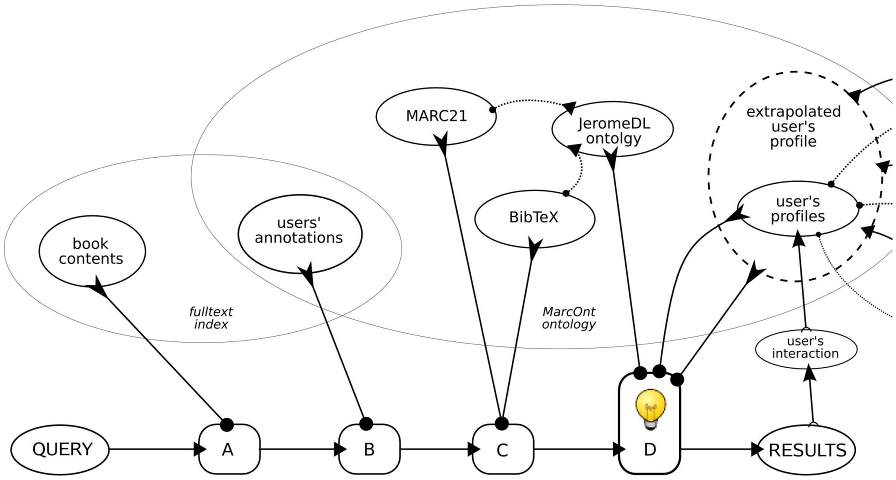
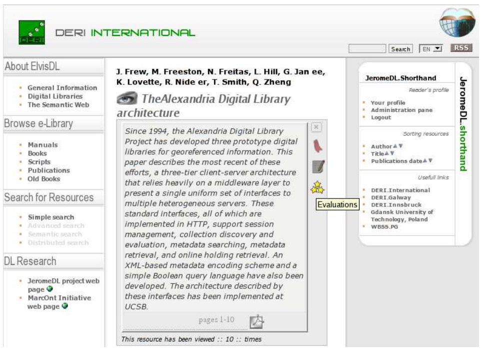


Figure 1: The steps of the searching process in the JeromeDL system

Browsing library content

In addition to the library catalogue (see Figure 2), hte JeromeDL system provides ways to annotate and bookmark interesting resources. The readers can manage their bookmarks in their profiles using a tree GUI. It is also possible to annotate or bookmark the part of the resource (e.g. page set of an old book or a fragment of HTML document presented with Java Applet). Both bookmarks and annotations are specified, together with the extended FOAF metadata, in the FOAFRealm project[4].

The readers can share their bookmarks with their friends. There are two ways how bookmarks are being presented: either by a direct view or by a summary (combination) view of the bookmarks of the reader's friends. Each entry of the bookmarks tree can have a separate access control list. The reader can specify the ACL using information about distance and trust level in the weighted digraph of friendship connections between users, e.g. entry `F[mailto:skruk@jeromedl.org]1.6` denotes that only direct friends of user `sebastian.kruk@deri.org` with trust level above 60% can gain access to the resource.



J. Frew, M. Freeston, N. Freitas, L. Hill, G. Jan ee, K. Lovette, R. Nide er, T. Smith, Q. Zheng

TheAlexandria Digital Library architecture

Since 1994, the Alexandria Digital Library Project has developed three prototype digital libraries for georeferenced information. This paper describes the most recent of these efforts, a three-tier client-server architecture that relies heavily on a middleware layer to present a single uniform set of interfaces to multiple heterogeneous servers. These standard interfaces, all of which are implemented in HTTP, support session management, collection discovery and evaluation, metadata searching, metadata retrieval, and online holding retrieval. An XML-based metadata encoding scheme and a simple Boolean query language have also been developed. The architecture described by these interfaces has been implemented at UCSB.

pages 1-10

This resource has been viewed :: 10 :: times

Figure 2: Browsing JeromeDL's catalogue

3 Future work

To bring the JeromeDL system to the full potential the ontology used to describe the resources has to encompass the bibliographic descriptions like MARC21 and BibTEX. Work on the stated ontology has been undertaken by the MarcOnt Initiative[5]. The developed ontology will be based on the collaboration and negotiation. The JeromeDL system - e-Library with semantics[8] will become a testing platform for the MarcOnt ontology. The new ontology will replace the currently used one and a number of additional metadata standards supported by the JeromeDL.

4 Conclusions

The presented JeromeDL system has several features that increases its usability as a digital library. Apart from an efficient search algorithm[7] that makes use of the semantic description of the resources along with classic bibliographic descriptions (DublinCore, MARC21, BibTEX), the system benefits from the information about relationships between users. This is why a new user is able to benefit from the preciseness of the search process - as long as their likes and dislikes are similar to their friends. The bookmarks and annotations, implemented in the JeromeDL, are also very helpful while browsing the catalogue of the digital library.

References

[1] *Client-side state.* http://www.netscape.com/newsref/std/cookie_spec.html.

[2] DublinCore Initiative, http://dublincore.org/documents/dces/. *Dublin Core Metadata Element Set, Version 1.1: Reference Description.*

[3] Deborah A. Fritz and Richard Fritz. *MARC21 for Everyone: a practical guide.* The American Library Association, 2003.

[4] Sebastian R. Kruk. Foaf-realm - control your friends' access to the resource. In *FOAF Workshop proceedings*, http://www.w3.org/2001/sw/Europe/events/foaf-galway/papers/fp/foaf_realm/, 2004.

[5] Sebastian R. Kruk. Marcont initiative. Technical report, DERI.Galway, Ireland, http://www.marcont.org/, 10 2004. Bibliographic description and related tools utilising Semantic Web technologies.

[6] Leslie Lamport. *LaTeX: A Document Preparation System.* Addison-Wesley, 1986.

[7] Henryk Krawczyk Sebastian R. Kruk. Intelligent resources search in virtual libraries. In Trojanowski Klopotek, Wierzchon, editor, *Intelligent Information Processing and Web Mining*, pages 439–444. Polish Academy of Science, Springer, 2004. Proceedings of the International IIS: IIPWM'04 Conference held in Zakopane, Poland, May 17-20, 2004.

[8] Marcin Synak Sebastian R. Kruk. Jeromedl - e-library with semantics. Technical report, DERI.NUIG - Ireland; Gdansk University of Technology - Poland, http://www.jeromedl.org/, 09 2004.