Using LDAP in a Filtering Service for a Digital Library

João Ferreira[(**)]
IST – Instituto Superior Técnico
(Universidade Técnica de Lisboa)
Erreur! Source du renvoi introuvable.

José Luis Borbinha[(*)]
IST – Instituto Superior Técnico
(Universidade Técnica de Lisboa)
INESC – Instituto de Enghenharia de Sistemas e Computatores
Erreur! Source du renvoi introuvable.

José Delgado[(*)]
IST – Instituto Superior Técnico
(Universidade Técnica de Lisboa)
INESC – Instituto de Enghenharia de Sistemas e Computatores
Erreur! Source du renvoi introuvable.

Abstract

This paper describes how an LDAP directory service can be used to support a filtering service for a digital library. The directory stores and manages profiles of registered users and authors, which are used to implement a filtering service concerned with the submission and change of documents and document annotations, the registration of new users and changes in registered users profiles. The same user profiles are also used to rank results from search tasks as also for user authentication.

1. Introduction

The volume of electronically available information has been increasing in a way impossible to follow by an individual. Filtering services are one possible answer to this problem, and some of those services have been announced, based on user profiles [1]. In ArquiTec, a networked digital library, user profiles are managed in a directory service using LDAP (Lightweight Directory Access Protocol), a standard directory service for Internet. Based on that directory an information filtering service has been built.

The next section briefly resumes the information filtering perspective, presenting a few paradigmatic projects. Section 3 introduces the ArquiTec system, and section 4 introduces the main concepts of the X.500 model, from which LDAP derives. Section 5 describes how an LDAP based solution was used in ArquiTec, and section 6 explains how that directory is being used to implement a filtering system. Finally, the most important open issues are presented in section 7.

2. Information Filtering

Information filtering is an actual subject, with numerous systems appearing and raising important questions. In December 1992 ACM recognized the importance of this new field and published a Communications issue on filtering information. The subject returned again in the March 1997 issue, now focused on a new perspective called «recommender systems». The first filtering systems were targeted for electronic mail and USENET news filtering, but soon those systems were applied to other sources of information, such as the World Wide Web.

Basically, filtering systems use information retrieval techniques in which user queries are replaced by user long term interests, or profiles. These profiles can be created using explicit or implicit methods. Currently, user profiles is one of the richest areas of exploration, specially in the implicit approach (there are experiences, for example, using the time spent reading, analysis of users bookmarks and server log files, etc).

Due to human subjectivity and to achieve better results, several systems involve also humans in the filtering process. For example, in some cases user reactions to the documents are recorded (such as ranking, notes, etc.) and later used to help other users. Those kinds of systems are known as recommender, collaboration or social filter systems.

One of the first historical systems was the TAPESTRY project [2], which coined the term «collaborative system» and raised a new perspective to the problem. TAPESTRY gave two approaches for filtering: *automatic*, where the system evaluates what is interesting to the user, and *social*, where users help each other.

Table 1 summarizes a few paradigmatic systems developed until now or under development, emphasizing the users profile and matching techniques.

Sift and Newsweeder represent two examples of *automatic* filtering systems. The basic difference between them is the way profiles are defined, where Newsweeder uses also an implicit method based on past user experience. *Automatic* filter has had success only in very simple systems. The main problem is that it has to deal with the issue of automatic creation of representatives of documents (or surrogates), a complex task even for well-defined areas.

[(**)] Work supported by the JNICT fellowship PRAXIS XXI/BD/5968/95

[(*)] Work partially supported by the JNICT contract PRAXIS/2/2.2/TIT/1667/95

Table 1: Some paradigmatic filtering systems

As examples of *social* filtering systems we have Grouplens and ReferralWeb. Those systems are in general more successful than the automatic ones, but unable to provide information in documents that have never been read. Another weakness is the problem of finding the correct tools to keep out (or to minimize the effect) of disruptive users (such as, for example, users who are not really collaborative but only interested in giving high rates to themselves or related friends).

System	Information Source (IS)	Profile		Matching		Remarks
		Explicit	Implicit	Tecniques	Arguments	
Grouplens (1992)	Usenet	Numeric Vector	Numeric Vector (reading time)	Cosine measure	(user profile) versus (users profiles)	Collaborative filter system
Sift (1994)	Usenet	Keywords lis	-	Boolean	(IS) versus (user profile)	Filter system
Newsweeder (1994)	Usenet	Numeric Vector	Numeric Vector (user history)	Cosine measure	(IS) versus (user profile)	Content based-filter system
Fab (1994)	Web	Numeric Vector	Numeric Vector (user history)	Cosine measure	(IS) and (users profiles) versus (user profile)	Collaborative and content-based filter system
ReferralWeb (1994)	Web	***mention of a person or a document***			***(user profile) versus (community profile)***	Collaborative and social filtering

Concerning the matching techniques, two main approaches have been tested: (i) to match the profile against other profiles and to choose the information in the nearest one (collaborative) or (ii) to match against community standard profile and to use the nearest standard to get information (social). Fab is a system that tries to combine both approaches.

3. ArquiTec

The ArquiTec project aims to develop a digital library for the Portuguese scientific and research community [3]. It started in the beginning of 1997, and a first phase will end with a working prototype, scheduled for public release in the first quarter of 1998.

ArquiTec is accessible over the Internet, through a WWW interface. It provides access to different kinds of technical documents (such as papers, reports, theses, dissertations, etc.), in different fields of knowledge, while special services will also be provided to the community.

The system was conceived around three main entities, as shown in Figure 1: *documents, users* and *concepts*. Informal and formal documents exist in local repositories, managed by a structure of distributed servers based in the NCSTRL technology [4]. To address the problem of long term preservation, the Portuguese National Library will maintain a PURL service [5] and a central official archive with a copy of selected formal documents (such as thesis and dissertations).

ArquiTec users can be authors, readers, or both. Users are managed in a global X.500 like directory [6], where their identity, contacts, affiliations and a special profile are registered. Anonymous access is possible for search, browse or even retrieval, but users are always suggested to identify themselves for profile management.

The concept space, or ontology, is based in the integration of possible multiple statistical and formal thesauri, as well as user contributions. Two important components of this space are user and collection statistical thesauri, created from the document collections and also from the user directory (profiles). In that sense our thesauri perform functions well beyond their usual roles as auxiliary tools for classification and search. Matching these thesauri with the collection makes it possible to identify document clusters, for example, but it makes also possible to identify virtual user communities (defined as groups of users sharing common interests).
Documents, users and concepts are interactive and dynamic entities, which means that they can change over time. For example, documents can have new releases or attachments (submitted as annotations), users can become interested in new subjects, new subjects can be included in concept space, new relationships can be established between existing subjects, etc. Indexes, user profiles and the relations between documents and users (authors or just readers) associate these entities among themselves.
Users are identified in ArquiTec by their interests and contributions, which relate to subjects in the concept space (likewise for documents). In ArquiTec users are viewed not only as authors and patrons but also as important sources of information, with their profiles becoming part of the contents. Profiles serve also to provide special services to the users, such as filtering (automated notifications) and ranking of search results.
Conceptually, a *catalog* makes it possible to explore, in an integrated perspective, the above six concepts (comprising the three main entities and the relationships between them). In that sense it becomes possible and has an equivalent meaning, for example, to search for documents or for users related to a specific subject (in an integrated perspective, it is also possible to search for both users and documents related to specific subjects, and so conceptually «sharing common interests»).

4. X.500 and LDAP

ArquiTec uses an X.500 directory in an LDAP implementation.
X.500 is an OSI directory service, which defines an information model, a namespace, a functional model and also an authentication framework. An X.500 directory is based on entries, which are collections of attributes as defined in RFC 1779 [7]. Each entry has a type (or class), typically defined by one or more mnemonic strings, and can have one or more values.
The attributes required and allowed in an entry are controlled by a special object class attribute in every entry. The information is supposed to be structured in a tree, accessible by servers possibly distributed over a network.
As shown in Figure 2, at a top level there are entries representing countries, below that there are entries representing national organisations, and so on. At the lowest level it is supposed to find entries representing any desired class of objects, such as people, computers, printers, etc.
X.500 defines the Directory Access Protocol (DAP) to access the service, a full, complex and heavy OSI protocol supporting operations in three areas: search/read, modify and authenticate. The search is possible at any level, based in a filter query involving attributes and returning requested attributes from each matching query.
The problem of the excessive complexity of the DAP protocol has been addressed by the Network Working Group of IETF, which has been proposing the Lightweight Directory Access Protocol (LDAP) as an alternative for the Internet.
LDAP is a client-server protocol that runs directly over TCP/IP, and it was conceived to remove some of the burden of X.500 access from directory clients, such as taking out some of the less-often-used service controls and security features.
LDAP is being positioned as the directory standard for the Internet, with leading industry players like Microsoft, Netscape, IBM, Lotus, Novell and Banyan supporting it or intending to support it in the near future [8]. There are also plans to develop LDAP access for several database and index machines, such as Glimpse, for example).

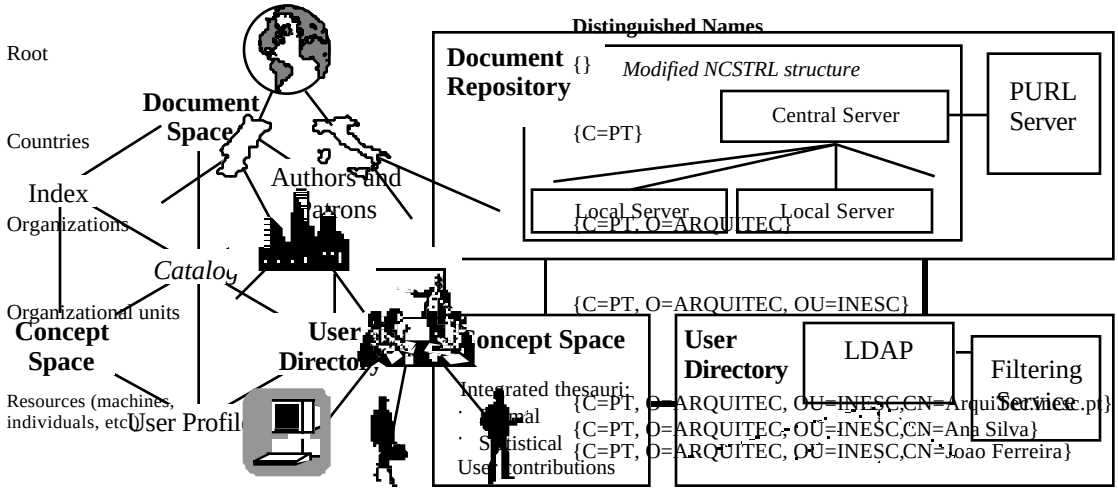

Figure 1: Main Components and X.500 structure of ArquiTec

Generic attributes	Profile attributes
Name	**Explicit Fields:**
Institution	List of interesting subjects
User identifier (ArquiTec)	List of non interesting subjects
Password	**Implicit Fields:**
Password tip	List of identifiers of archived documents which the user has authored
Email address	List of subjects of documents which the user has authored
Telephone	List of identifiers of submitted annotations
Fax	List of identifiers of retrieved documents
WWW home page	List of subjects of retrieved documents

Table 2: ArquiTec user entry in the user directory.

5. LDAP in ArquiTec

ArquiTec uses the Directory Server package, an LDAP implementation from Netscape and based in an original work from the University of Michigan [9]. This LDAP implementation has three main components:

Server: our server runs on a Unix machine as a stand-alone daemon.

Client library: a powerful C language API for accessing and using LDAP, with LDAP clients and a backend to handle database operations [10].

Gateway: a special WWW interface is available for directory and server administration.

Users access ArquiTec in one of two modes: anonymous or identified. Identified users have profiles composed of explicitly provided data (their explicit interests) and data implicitly extracted from the history of their interactions with the system (such as submitted and retrieved documents).

ArquiTec users are managed in a structure such as presented in Figure 2, where each user entry has a list of fields as presented in Table 2.

At the moment, the user directory is implemented in only one server. However, to provide flexibility and fault tolerance it will be distributed and replicated it in the near future by other servers within the national academic network (a feature supported by LDAP).

1. Filtering in ArquiTec

In ArquiTec the filtering service follows both the *automatic* and *social* approaches. It is a social system because document classification gets richer with annotations submitted by users. It is also an automatic system because it automatically matches new documents and annotations with the existing user profiles and new profiles with the existing documents.

More generically, user profiles serve three main purposes in ArquiTec, as shown in Table 3:

Filtering: profiles are used to provide an information filtering service, supported by electronic mail, through which users can receive automatic notification of new events.

Searching: profiles can be used to rank search results, for example to highlight documents that best match user's interests (but ranking will never hide or restrict the access to other documents that also match the queries).

Retrieval: the access to different kinds of documents or to special user information can depend of the user profile. This is a scenario not yet implemented in ArquiTec, where privacy protection concerns have to be taken in account, since it requires defining profiles fields not controlled by the user but by an administrative authority (in the current scenario user profiles are public and fully controlled by the users).

The filtering service tracks five kinds of events:

Notification of new documents: any user whose profile matches the classification of a new document is informed about it (to submit a document, a metadata form has to be filled).

Notification of changes in stored documents: if a new version of a document is submitted, users that, for example, had retrieved that document, will receive a notification.

Notification of new annotations: any user whose profile matches a new annotation will be notified about it (in

Event sources	User profiles usage in ArquiTec		
	Filtering Service	**Information Search**	**Information Retrieval**
Documents	- New documents - Document changes	Ranking of query results	Control Access
Annotations	- New annotations		
User Profiles	- New users - Changes in profiles		

Table 3: Usage of user profiles in ArquiTec

fact, an annotation in ArquiTec is just a document metadata form, similar to the form filled in the submission of the document).

Notification of new users: when a new user is registered, users with similar profiles will be notified.

Notification of changes in user profiles: when a user profile changes, users matching the new profile will be notified.

User profiles can be used also to rank search results, giving more relevance to results that best match the profile of the user. For this task, it is also possible for the user to choose to identify him/herself with a virtual profile created by the system, instead of its own.

From the user directory it is possible to identify groups of users with similar profiles, and so to create virtual profiles of possible communities. In the future, this feature will be exploited for collaborative services, such as mailing lists (automatically created).

1. Future Work

ArquiTec is work in progress. The structure of the user profiles still need to be tested and tuned (it was defined until now in a mixture of implicit and explicit methods). Access restrictions to information (documents and user profiles) will be also implemented based in different criteria, namely in administrative fields in the user profile.

Work has to be done yet in the conceptual space based on the collection statistical thesauri and user directory. An important open issue here is the creation and maintenance of authority lists, vital to control the integration of thesauri. Finally, an exciting issue is the development of strategies for the (semi-)automatic identification of user communities and the conception of new services based on that perspective.

References

[1] Resnik, P.; Varian, H.R. (1997). **Recommended systems**. Communication of ACM, March 1997, Vol. 40, N. 3

[2] Goldberg, D.; Nichols, D.; Oki, B.M.; Terry D. (1992). **Using collaborative filtering to weave an information TAPESTRY** Communication of ACM, December 1992, Vol. 35, N. 12.

[3] Borbinha, J.L.; Ferreira, J.; Jorge, J; Delgado, J. (1997). **A Digital Library for a Virtual Organization.** Proceedings of the **Erreur! Source du renvoi introuvable.**.

[4] Davis, J.R. (1995). **Crating a Networked Computer Science Technical Report Library**. D-Lib Maganize, September 1995. *Available on-line in 27 September 1997 at http://www.dlib.org/september95/ 09davis.html*

[5] Weibel, S.; Jul, E. (1995). **PURLs to improve access to Internet**. OCLC Newsletter, November/December 1995, 19. *Updated version available on-line in 27 September 1997 at http://purl.oclc.org/ OCLC/PURL/SUMMARY*

[6] CCITT (1988). **X.500 The Directory: Overview of Concepts, Models and Service**. CCITT Recommendation X.500, 1988.

[7] Yeong, W.; Howes, T.; Kille, S. (1995). **RFC 1777: Lightweight Directory Access Protocol**. IETF Network Working Group, March 1995. *Available on-line in 27 September 1997 at http://ds.internic.net/rfc/rfc1777.txt*

[8] Cooper, J.; Ratcliffe, N (1997). **The role of LDAP and X.500**. Data Connection, August 1996. *Available on-line in 27 September 1997 at* **Erreur! Source du renvoi introuvable.**

[9] Howes,T.; Smith, M. **LDAP Programming Directory-enabled Applications with Lightweight Directory Access Protocol.** Macmillan Technology Series (1997)

[10] Howes, T; Smith, M. (1995). **RFC1823: The LDAP Application Program Interface**. IETF Network Working Group, August 1995. *Available on-line in 27 September 1997 at http://ds.internic.net/rfc/rfc1823.txt*