Integration of Multilingual Classification Systems with the Dienst digital library system

Nuno Freire

Instituto de Engenharia de Sistemas e Computadores (INESC)
Rua Alves Redol, 9, 1000 Lisboa, Portugal
Email: Nuno.Freire@inesc.pt

Abstract

This document aims to provide a presentation of the current stage of the digital theses and dissertations project. This project started in the beginning of August 1998 and comprises the processing of thesis and dissertations. We are currently working in the integration of Multilingual Classification Systems with the Dienst digital library system. One of the key functions of the classified space will be the information space normalization and to allow a cross-language information retrieval service.

INTRODUCTION

We intend to promote a digital circuit in co-operation with the university libraries to cover the digital processing of theses and dissertations. A national trial for an optional process for digital versions, parallel to the actual paper-based one, is under development with the collaboration of a group of universities.

This project has the following objectives:

- To improve graduate education by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- To increase the availability of student research for scholars and to preserve it electronically
- To lower the cost of submitting and handling theses and dissertations
- To empower students to convey a richer message through the use of multimedia and hypermedia technologies
- To empower universities to unlock their information resources
- To advance digital library technology

A customised network of servers based in the DIENST technology and running in Unix machines will support the system. The system is known to work on Linux and Solaris platforms, although any system running Apache web server and a PERL interpreter should support it.

Client-side requirements are a typical "web browser".

Local DIENST servers will be installed at the local university libraries, from which the documents are captured to the central server using HTTP.

We are currently trying to solve the problem of correct document classification. Usually, the librarian classifies the document but, if we want to obtain a correct document classification, the classification should be performed by the author. This is the reason why we are integrating the classification systems with Dienst and creating user interfaces to allow both the author and the user, to submit, browse and search the collection, using the classification schema.

THE CLASSIFICATION SYSTEMS

We use the existing classified systems stored in LDAP due to its hierarchical and distributed proprieties. This space will be used to normalise users and information spaces and to solve multi-language information retrieval.

1.1 LDAP interface (Distributed directory)

All information about the classified systems is stored in a directory consisting of an X.500 directory, using an LDAP implementation. This decision allows a fast and hierarchical access to this information.

X.500 is a standard directory service that 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 [1]. 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.

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, that 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 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 [2]. There are also plans to develop LDAP access for several database and index machines, such as Glimpse.

LDAP stores this information and we built interfaces to browse easily classification. We propose a LDAP implementation on a Linux machine with a freeware version offered by this University of Michigan [3]. This LDAP implementation has three main components:

- *Server*: We will run our server on a Linux machine as a stand-alone daemon. However, to provide more flexibility and fault tolerance, it will be distributed and replicated by other servers within the National academic networks (a feature supported by LDAP).
- *Client library*: a powerful C language API for accessing and using LDAP, with LDAP clients and a backend to handle database operations. With these tools we will build a user interface to browse easily the classified space and give also an interface for administration of this space.
- *Gateway*: a special web interface is available for directory and server administration.

The document-classified space is stored as shown in figure 1. At the top level there are entries representing the available classification systems, in the next levels there are entries representing general terms, and so on. At the lowest level there are entries representing subject descriptions. Each entry contains a unique identifier and a term in English, Portuguese and, in the future, translations to other languages.

Access to perform writing inputs is only given to Human authority. Users can only browse this space.

This classified space is important for normalisation purposes.

1.2 Cross-language information retrieval service

The existence of classified systems in different languages will allow a cross-language information retrieval service. When the document is classified, the system indexes the terms and the corresponding identifiers. So, if the user searches for relevant documents using the classification systems to formulate his query, the system will retrieve all documents with the selected classifications. The documents are retrieved independently from the language in witch they were classified. The system searches for the selected terms and respective identifiers.

In this project we use English and Portuguese for the document classification systems, but this framework is designed for an easy extension to other languages as well. With this tool we can provide a information retrieval service for differed languages only with the requirement of having classified systems translated to different languages.

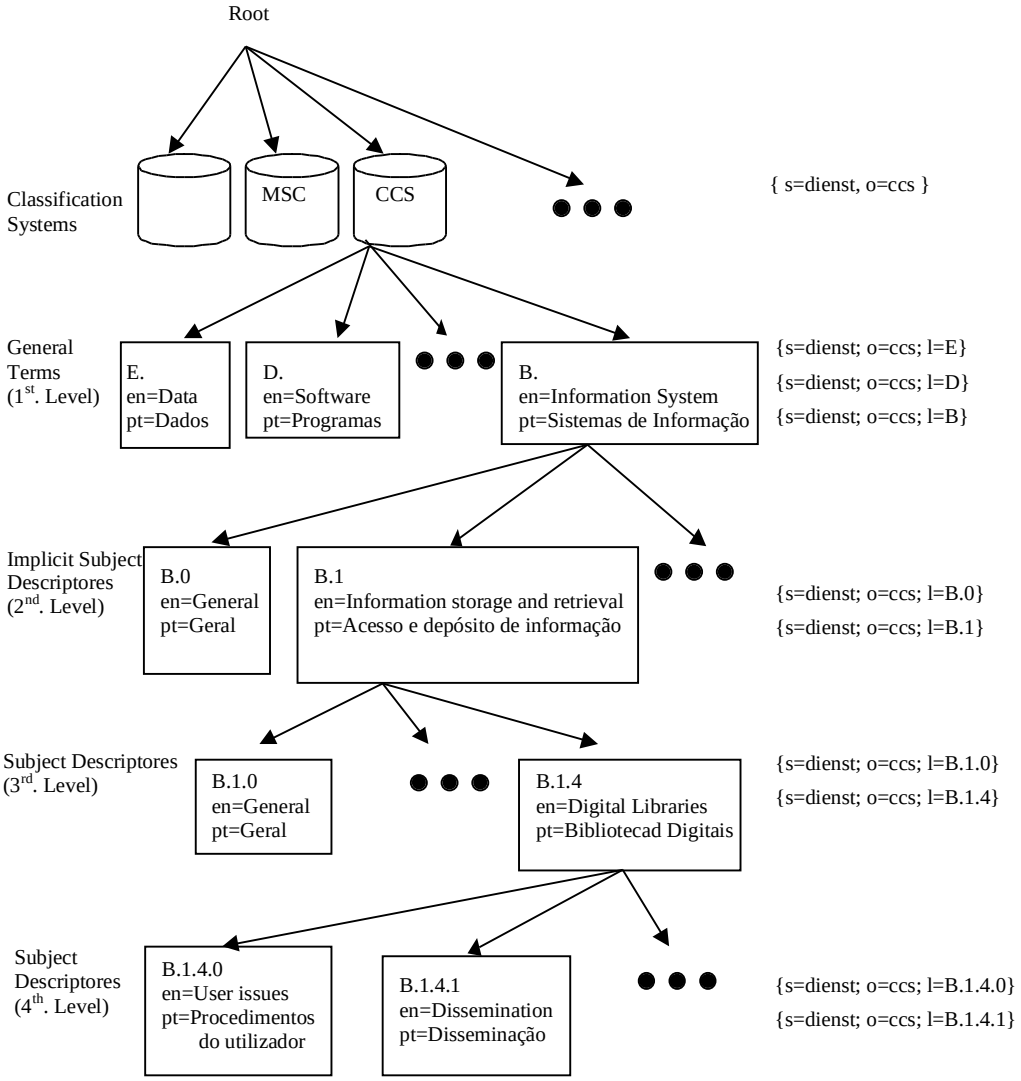

Figure 1: Classified spaces are stored and maintained in a LDAP space.

SYSTEM ARCHITECTURE

A user can access the system with one of two purposes: to submit a new document to the collection, or to search/browse in the collection. In both the situations it is possible to take advantage of the Classification Server.

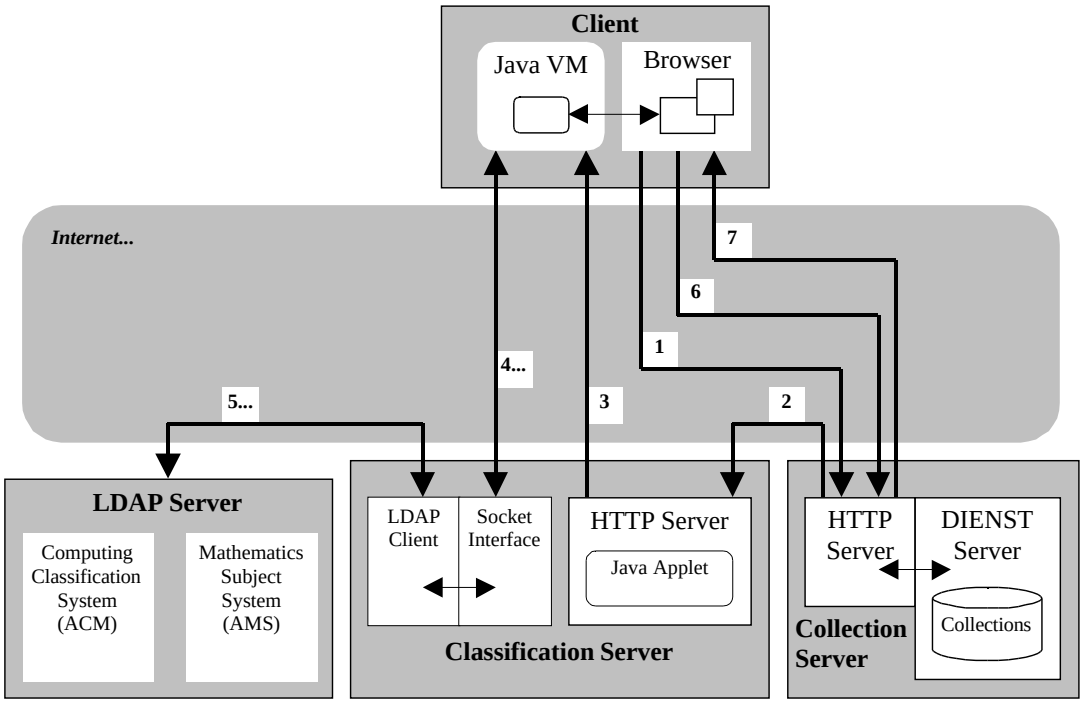

Figure 2 : Integration of Multilingual Classification Systems with Dienst

An interaction starts with a contact to the Collection Server, by HTTP, where the desired operation is selected (1).

If the Classification Server is requested, that request is transferred to it (2), which replies to the user sending him/her an applet (3).

With this applet the user can browse the classification systems available from the Classification Server (4). Those classification systems are stored in a X.500 directory, accessible by LDAP (5). The LDAP server used in this system is provided by the University of Michigan.

Actually we have available from the Classification Server the Computing Classification System from ACM, and the Mathematics Subject System from AMS, both in English and partially translated to Portuguese.

The directory was conceived to host also other structured classification systems, as also their translations in multiple languages. On the other side, the Java applet is completely independent of the contents of the directory, being configured according the information received from it.

In the interaction with the Classification Server the user can select the desired terms, in one or more the languages, and return to the Collection Server with those selections (6) to be used in the task in course (7).

Depending of the actual task, the selected terms can be used in the Collection Server in a pre-coordination purpose, to classify a new document, or in a post-coordination purpose to search in the indexes. The actual Classification Server can maintain several indexes, depending of the metadata structure of the collection. In the actual system, we support one index for each classification schema, but the terms included in those indexes are also used as generic keywords and indexed in the keywords' index (in this sense, a user can perform a free searching in that index using terms from the classification systems and find the right documents in the same way).

CONCLUSION

Using the classification systems to improve the document's metadata allows better results in information retrieval. But the classification should be made by the author, to assure a correct document classification. In this paper we have described a user interface for document classification and retrieval using the available classification systems. The classification systems are stored in LDAP directories. These directories are accessed by HTTP, through a java applet, where the user can select the terms from the classification systems, either for browsing/searching, or document classification.

The cross-language information retrieval service is achieved by including the terms from the classification systems and their respective identifiers in the document's metadata. Using the identifiers, it is possible to find documents classified in any language because the identifiers are common to all languages.

REFERENCES

[1] Cooper,J.;Ratcliffe,N. The role of LDAP and X.500. Data Connection, August 1996. *Available on-line in* 3 July 1998 *at* *http://www.datcon.co.uk/docs/press/mdwhite1.htm*

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

[3] Howes, T;Smith,M. (1995). **RFC1823: The LDAP Application Program Interface**. IETF Network Working Group, August 1995. *Available on-line in* 3 July 1998 *at* http://ds.internic.net/rfc/rfc1823.txt