

CHILE: A Visual Library Catalog Retrieval Prototype

Mari-Carmen Marcos

Information Science Section

Universitat Pompeu Fabra

Ramon Trias Fargas, 25, 08003 Barcelona. Spain

mcarmen.marcos@upf.edu

+34 935422264

Ricardo Beza-Yates

Center for Web research, DCC, Universidad de Chile & ICREA-Dept. of
Technology

Universitat Pompeu Fabra

Passeig de Circumval·lació, 8

08003 Barcelona. Spain

ricardo.baeza@upf.edu

+34 935421452

Carlos Andrés Ardila

Center for Web Research, DCC. Universidad de Chile

Blanco Encalada, 2120

08003 Santiago de Chile. Chile

cardila@dcc.uchile.cl

1. INTRODUCTION

In spite of all the advances in simplifying the access and use of online catalogs (OPACs), in particular with the use of the Web (e.g. metasearch and standard interfaces for programs such as Z39.50), final users still find the same old difficulties: how to find information when they do not know the specific documents that satisfy their needs and how to distinguish which of them are more relevant. This problem comes from the first catalogs and is common to other information retrieval (IR) systems, in particular for subject search.

Research in user interfaces in general has profit from technology advances and the Internet, offering better quality. However, in the world of OPACs, there are few innovative results to improve search interfaces. Ortiz-Repiso and Moscoso [9] show that in spite of the change of media, the organization of the information has been maintained -as Borgman pointed out in 1996 [3]- and there is no special use of the hypertext and multimedia capabilities of the Web.

Looking at standards or guidelines, IFLA (International Federation of Library Associations and Institutions) has elaborated design guidelines for OPAC results [13]. Some authors like Cherry [4] also propose guidelines, but in most cases is work done before the dawn of the Web. A recent result on this problem is due to Babu and O'Brien [1], which also includes a good survey.

Current OPACs, compared to early Web systems, do not show a real evolution regarding access and retrieval techniques, or display of documents. This implies that most users keep having a difficult time finding information needs, particularly in subject searches where they do not know the relevant documents. In this case it is worthwhile to show an overview of the document collection and allow the user to browse.

The main goal of CHILE (Computer-Human Interaction Librarian Experience), is to show a system prototype that includes browsing, clustering and visualization. Our work is an ad-hoc adaptation of the overview-query-preview-answer model for IR [2, 6].

For the development of CHILE, we first did a study of the requirements that a good OPAC interface should have. We defined the phases of the search process (query, display of results and query reformulation), the different interfaces needed and their characteristics [7, 8]. Hence, we propose our own set of guidelines. The main contribution of our system is the integration of several concepts and techniques to OPAC interfaces. We use structured information from a sample of bibliographic records from the University of Chile main library. The architecture of the system includes control access and the overview and result interfaces use clustering as main visualization technique, like other Web systems already do (e.g. KartOO: <http://www.kartoo.com>).

2. MODEL

2.1. Retrieval Stages

The information retrieval process is composed by the following stages and sub-stages:

- A. Query
 1. Collection overview. Global view of the collection (or the subset corresponding to the user profile), that is a generic view with the overall content of the collection and the results the user can expect to find. For this step we propose the use of clustering techniques to display the whole collection as a visual map.
 2. Retrieval by browsing. The user must be able to navigate among related subjects or categories. We also use clustering to create new categories and to visualize them.
 3. Retrieval by querying. Standard interface for users with well defined information needs.
- B. Display of results
 1. The resulting set that corresponds to the user query is presented as a visual map that groups document automatically taking in account the Dewey (DDC) classification of the bibliographic records and the subject terms that they contain.
 2. Each document retrieved shows the basic attributes that identify the document (title, author and year) in addition to the cover image. The user can obtain additional data such as the complete bibliographic record, the summary and keywords, if they are available. We can also offer the comments of users borrowing the document and a list of the most popular documents borrowed by those users.
- C. Query reformulation

If the user wants to improve the retrieved documents, the system must help him/her restricting or enlarging the results.

Current OPAC models usually only allow:

- Retrieval by querying.
- Display of results in a simple list (sometimes not even sorted by relevance, lexicographical or chronological).

- Does not allow to really reformulating the query (just a new query or some fixed set of options).

Our ideal OPAC model would follow our retrieval stages:

- Retrieval by querying or browsing using a hierarchical two dimensional overview map.
- Display of results sorted by relevance and also through a two dimensional map grouped by subject similarity, by using subject terms and a MARC (Machine-readable cataloguing) based classification [5].
- Reformulation of the query by adding/deleting search terms, and finding similar documents. This should exploit search terms and MARC metadata.

We are restricted to current catalog data which is usually in MARC format (either binary or lately in XML). Nevertheless, this restricted structured set of fields can be used to allow new ways to access information.

2.2. OPAC Model for CHILE

Now we will describe in detail some of the characteristics that the OPAC model or our system should have (table 1).

2.2.1. Browsing an overview

As we believe that OPACs should help with a first overview of the overall collection, we use automatic clustering techniques using the hierarchical structure of the underlying bibliographic classification. From here the user can browse the main categories and subcategories, in particular when the search terms are not known for the user.

To start the user obtains a list of main topics of the collection. These will be the first 10 categories in the first level of CDU (Universal Decimal Classification) or DDC (Dewey Decimal Classification). From there we can browse up to three levels or less, if we find a period or another character (e.g. quotes) before in the classification codes.

In any of the screens the user can request a document list belonging to a subcategory, sorted by relevance or by an attribute (e.g. title, author, year, type of document, etc.).

After reaching the last level (which in general would have to be collection dependant, based on collection size and thematic breadth), the user will pass without noticing to the display of results stage. That is, from an overview to a preview. In practice, the only difference of both views is that we go from sets of documents to documents, which at the end depends on the sub-collection size.

2.2.2 Display of results (preview)

After posing a query or browsing in the collection, the user can see a list of the relevant documents (textual or graphically). Here we have two different screens: one grouping documents by CDU codes and another one that gives a list of results sorted by relevance.

For relevance ranking we have considered the following factors: query terms in subject descriptors (for query retrieval), most popular documents in the OPAC (complete record viewed and loan statistics), and recommended documents (in case of university libraries). These data is usually available in the integrated library management system.

2.2.3. *Visual map for results*

After the user obtains a first result, we complement the normal textual list a visual map based in the subject similarity between documents. For this we propose to employ the classification code and the subject terms, to build a two dimensional SOM (self-organizing map).

The visual map would have the following characteristics:

- Documents are represented by small circles grouped by common subject terms, in particular the most frequent of them.
- Just by moving the mouse over each circle opens a window with the title and author.
- By clicking the right button of the mouse we obtain the complete bibliographic record in an overlay window that does not cover the visual map.
- Part of the results can be selected with a zoom-like option, avoiding to loose the context.
- Relevant documents can be marked and extracted in other formats, to save, print or e-mailing them.
- Queries can be stored by the system, such that sessions (or part of them) can be recovered, for example to check for new answers to a query.
- Queries are also stored and can be organized by each user.

2.2.4. *Document View*

In addition to the traditional bibliographic record, we can offer:

- Comments by other users for a given document, including recommendations by experts.
- Table of contents and index of the document.
- Image of the cover, helpful to remind a known document or to remember later a new document.
- Reference to all documents in the same category that have been borrowed by other users that also had been interested in the current document (e.g. as in Amazon, <http://www.amazon.com>).

2.2.5. *Query Reformulation*

As feedback technique a dialog that helps in filtering relevant documents by iterative search. For this, the user may mark relevant documents as well as irrelevant documents, and the system will automatically update the result taking in account the subject and classification code fields.

Each document must allow a query “documents like this one”. In the case of visual maps, we can set a point on the 2D space and use a “build new map” option.

Finally, we have a personalization option, where a registered user can select the subjects of his/her interests. This information can be used later for services associated to user profiles, such as initial overviews or news alerts based on their interests.

Table 1 summarizes the different interfaces from our model of OPAC.

Search method		Display of results		Each retrieved document
querying	browsing	list	map	clustering

Search reformulation

Table 1. Interfaces for our OPAC model.

3. CHILE PROTOTYPE

3.1. Database

To build the database we selected a random sample of the University of Chile library collection, with the restriction that the document should have subject terms, classification code and the word “Chile” appeared somewhere in the bibliographic record. The sample had 5,523 documents with 3,735 different subject terms. All of them were in MARC format, which made things easy for storing the data in XML-MARC, process and storing them in a relational database using MySQL. Our database only kept the fields and subfields that were needed for the prototype, and was designed to facilitate the querying process and the location of subjects on the visual maps. We used two tables, one for the document attributes and their subject terms, and another one to handle user profiles and personalized searches (Figure 1).

The Web front-end of the prototype was implemented using PHP, because was easy to interface to MySQL and is an efficient tool to manage databases, handle the front to back-end communication, and to generate Web pages. The PHP scripts handle all the updates and queries to the database. Each feature is handled by a separate script, to ease code development and maintenance.

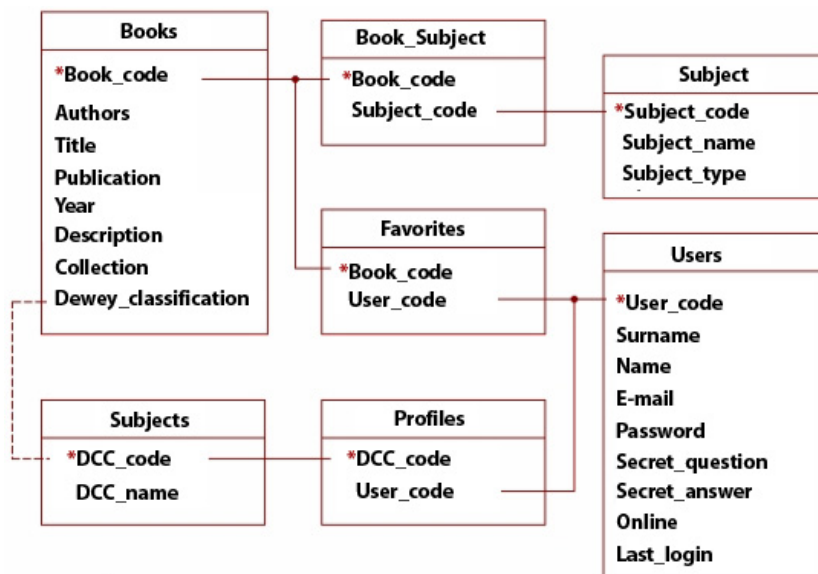


Figure 1. Database structure.

CHILE keeps a log of all subjects and documents queried and viewed by its users, to learn the most frequent subjects and documents requested. An underlying process allows all pages to access the log files, and the news associated to user profiles. This process also does caching to improve the access to previously generated pages (mainly for the frequent visual maps which need more processing time, that is the global overview and the initial overview for frequent users).

Figure 2 shows the structure of the Web front-end.

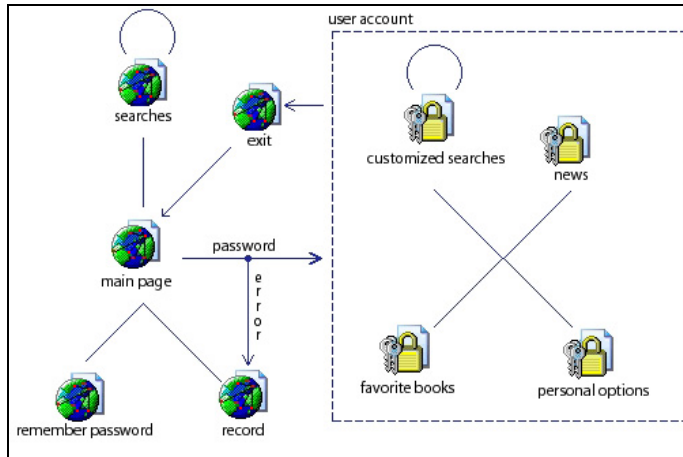


Figure 2. Structure of the Web front-end.

3.2. User Interface

As we mentioned before, we use a visual map for both, the global or user defined overview, and the display of results. A visual map can show many results (much more than 10 as usual Web interfaces) in a reduced space. The proximity of the documents indicates the similarity of subject terms, as we use a variant of self-organizing maps [11], where the location depends on a vector space created from document attributes.

We use a clustering algorithm called PAM (Partitioning Around Medoids) [11], where to find k clusters, a representative object, called medoid, is chosen for each cluster. The medoid must be close to the cluster center and also must contain the subject terms that characterize the cluster. Other objects are classified depending on the proximity to the medoids. Although there are many clustering algorithms, PAM matched well with our retrieval requirements and gave good results. The result of the algorithm over the whole collection (overview) is shown in Figure 3. The display of results for a query is the same as for browsing the categories, and similar to the overview. In figure 4 we show the result of browsing “Native tribes” (“Pueblos indígenas” in Spanish) plus the subject term “Chile”. Notice that new subject terms appear. Queries are handled by standard SQL. For larger catalogs better techniques can be used.

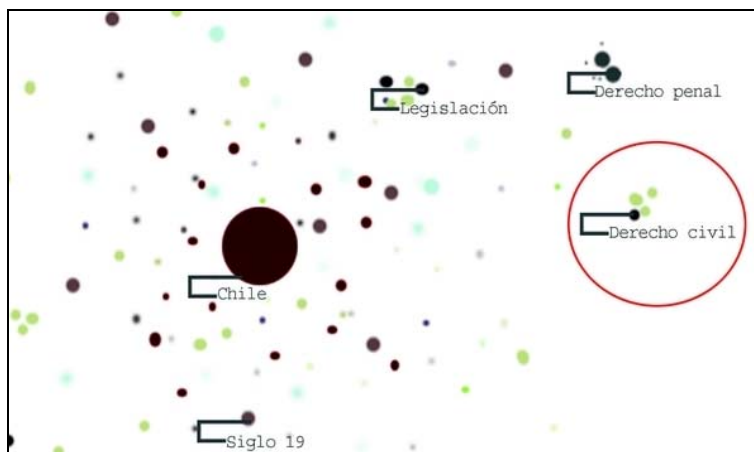


Figure 3. Overview using a visual map.

The algorithm to create a visual map is as follows:

1. Search the subject terms that appear in the collection and their statistics.
2. Select the most important subject terms that will guide the map, using the frequency of occurrences.
3. Find all subject terms that are related to the guiding terms. A circle is drawn with the main subject terms, and each new subject term is positioned in the middle of the largest empty arc in the circle perimeter.
4. Put all other subject terms in a random position in the map borders.
5. Put all documents in the geometrical center of their subject terms.

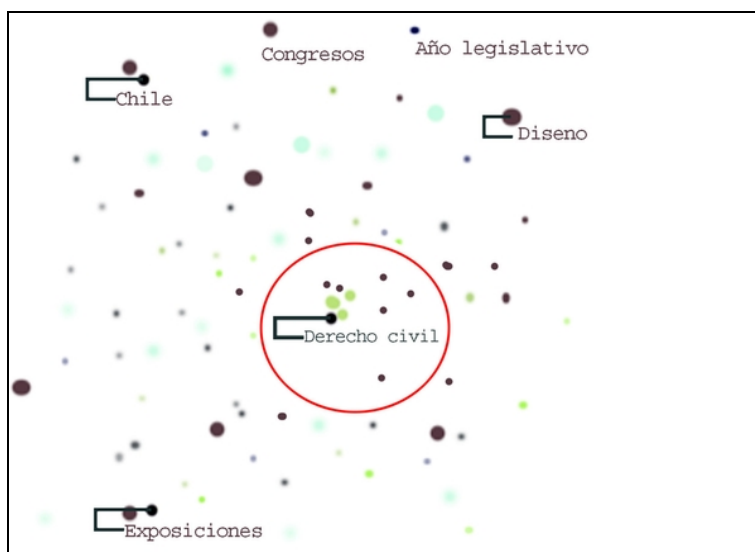


Figure 4. Display of results.

4. CONCLUSIONS AND FUTURE WORK

We have presented a system prototype that allows document visualization based on classification codes and subject terms. The main challenge -and novelty- of CHILE lies in the small amount of information available in OPACs to detect similarities among documents, which is not the case for Web search engines or more complex bibliographic databases with document summaries. We show that in spite of this lack of information, we can avoid useless results thanks to the structured nature of OPAC metadata. Further research is needed to improve interfaces for bibliographic databases.

One problem with our current visual maps is text legibility. We have done no attempts to predict overlaps of subject terms in the screen. As shown in Figure 5, sometimes reading the labels is difficult.

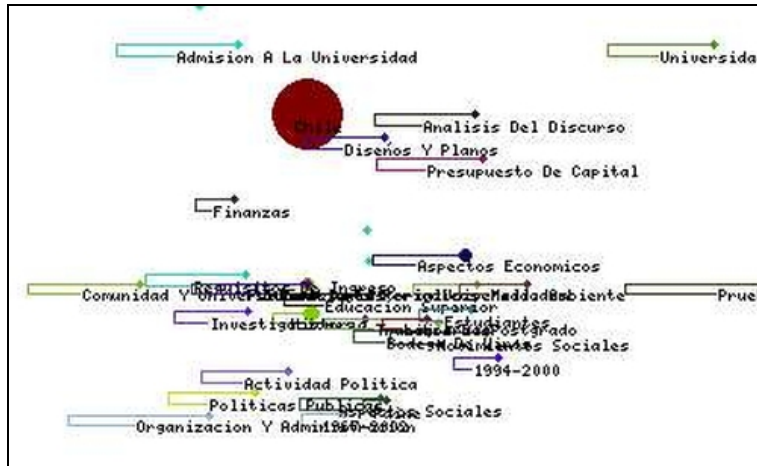


Figure 5. Map legibility problems.

Future work will include:

- Implement all the system requirements, in particular more information about each document in the preview.
- Improve the reformulation stage. Currently, the zoom feature allows redefining the resulting answer but not to pose a reformulated query on that subset or restrict the subject terms.
- Improve the generation and loading time for the main maps. Even caching can be not enough for large maps.
- Change the zoom feature by a fish-eye view such that the context of the visualization is not lost (that is, focus plus context).
- Improve system security. Our current prototype only allows a timeout after which a user must re-enter the user identifier (email address) and password.

REFERENCES

1. Babu, B.; O'Brien, A. Web OPAC interfaces: an overview. *Electronic Library*, 2000, 18:5, 316-327.
2. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. Addison-Wesley, 1999.
3. Borgman, C. Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 1996, 47:7, 493-503.
4. Cherry, J. Bibliographic displays in OPACs and web catalogs: how well do they comply with display guidelines. *Information Technology & Libraries*, 1998, 17:3, 124-137.
5. MARC Standards. Washington: Library of Congress, <http://www.loc.gov/marc/marcspa.html>.
6. Marchionini, G. *Information Seeking in Electronic Environments*. Cambridge University Press, 1992.
7. Marcos Mora, M. C. *Human Computer Interaction for OPAC information retrieval interfaces using subject terms*, Ph.D. thesis (in Spanish), University of Zaragoza, September 2003.
8. Marcos Mora, M. C. *Interacción en interfaces de recuperación de información: conceptos, metáforas y visualización*. Gijón: Trea, 2004.
9. Ortiz-Repiso, V.; Moscoso, P. Web-based OPACs: between tradition and innovation. *Information Technology & Libraries*, 1999, 18:2, 68-77.
10. Pech Palacio, M. A. *Adaptation and use of data mining for spatial and non-spatial information*. Final degree thesis (in Spanish). Universidad de las Américas-Puebla (México), 2002, http://mail.udlap.mx/~tesis/msp/pech_p_ma/
11. Self organizing maps, <http://www.cis.hut.fi/research/som-research/>

12. Staley, E. Graphical interfaces to support information search: an annotated bibliography, 2000, <http://alexia.lis.uiuc.edu/~twidale/irinterfaces/bib-main.html>
13. Yee, M. Guidelines for OPAC displays. IFLANet, Annual Conference (65th Council and General Conference), 1999, <http://www.ifla.org/IV/ifla65/papers/098-131e.htm>