Bilingual Subject-based Information Retrieval in NECLIB2 Digital Library

Zoran Constantinescu , Monica Vladoiu

Dept. of Computer Science
PG University of Ploiesti, Romania
zoran@unde.ro
monica@unde.ro

***Abstract*— In this paper we present our work on bilingual subject-based information retrieval in a real world digital library, called NECLIB2, which complements a classic open shelf library that includes around 50,000 materials (many of them being very rare) that are books and e-books, periodicals, audio and video CDs and DVDs, video tapes, scanned microfiche etc., and more than 18,000 items are audio recordings with classical music (in course of digitization).**

***Keywords*— digital library; information retrieval; subject based; thesaurus; natural language processing; inverse index**

I. INTRODUCTION

The main goal of any Digital Library (DL) is to fulfill the needs of its users. Crucial capabilities of digital libraries consist of information management, persistence and reliability, while a general problem is information discovery [1]. Subject–based information retrieval within digital library content relies on well-chosen subject descriptors that are applied to the underlying information resources. Subject analysis and efficient subject-based information retrieval based on these descriptors is essential for users of digital libraries [2, 3]. However, subject-based searching has some shortcomings related to the mismatch between the terminology of the searcher and that of the librarian [4], and to the linguistic problems generated by synonymy and polysemy [2]. One solution for these drawbacks seems to be thesauri-based information retrieval, which can be further sustained by ontologies for modeling their semantic structure, thus leading to development of more successful subject-based information retrieval systems [2, 3]. Moreover, developments in thesauri, social tagging, and guided navigation are expected to contribute to the subject analysis field, and to fulfill suitably user needs, goals, and expectations [3].

Challenges of Information Retrieval (IR) in digital libraries are rooted in the *nature of the content* (amount, variety, multiple sources, multiple languages etc.), the *tasks performed* (information retrieval from this mix of structure and formats, search and retrieval tools that compensate for abbreviated or incomplete cataloging or descriptive information, design tools that facilitate the enrichment of cataloging or descriptive information by including the contributions of users), and *users' profiles, dynamic needs, and goals* (building personalized collections, tracking documents, support for reading of retrieved materials, providing and creating citation trails, automatically creating bibliographies, brief on the fly generated summaries or snapshots of a document's contents, retrieval functionality across a wide spectrum of user types) [5, 6]. Moreover, for a long period of time, research in information retrieval has focused mainly on unstructured text, and has not benefited of available structure or metadata; nor has it acknowledged the broad range of users and their information needs and behaviors, therefore the next challenge for IR research in digital libraries is the design of adaptive, flexible, and interactive IR capabilities [6].

We present here our work on bilingual subject-based information retrieval in a real world digital library, called NECLIB2, which complements a classic open shelf library that includes around 50,000 materials that are books and e-books, periodicals, audio and video CDs and DVDs, video tapes, scanned microfiche etc., and more than 18,000 items are audio recordings with classical music (in course of digitization). Many of the materials are very rare, e. g. editio princeps of books, the same being true for the classical music recordings, which belong to a private collection of a famous literary critic, poet, and radio journalist. The digital library complements the classic open shelf library, which exists at a private foundation that support and finances advanced studies in humanities and social sciences [7].

In this work, we have approached the challenges of information retrieval in digital libraries by putting to practice very simple, yet powerful, ideas that lead to better efficiency, and increased user satisfaction. We present here some of these ideas, revolving around subject descriptors and other search terms, along with the way they work in NECLIB2.

The rest of this paper is structured as follows: the next section presents briefly the socio-semantic model NECLIB2 is based on, while the third section illustrates some particular aspects of the information retrieval process in NECLIB2. The fourth section includes some relevant related works, while the last section is dedicated to conclusions and future work ideas.

II. NECLIB2'S SOCIO-SEMANTIC MODEL

The digital libraries of today, are expected, besides offering classical library services, to be much more *adaptive*, to reflect better their *audiences*, to provide for *collaborative* frameworks that allow users to contribute to the content knowledge that is *socially constructed* - both actively by adding semantic annotations, reviews, ratings, and so on, and passively, by their use patterns, to be *contextual* - namely to be able to express the complex interrelationships and layers of knowledge between the DL's resources, and to provide for a common place of professional and collective wisdom [8].

As the traditional digital library paradigm, which is based on a catalog of metadata records, has failed to capture this rich multi-dimensional information space, mainly because the metadata records cannot provide for the following needs: capturing the complex contextual relationships between the DL resources, discrimination between the multiple entities involved (resources, actors, ontologies, agents etc.), adjusting to the changing information needs, capturing the dynamics of the DL resources, etc. [8, 9, 10]. Moreover, the metadata based model does not allow to deal with both content and metadata seamlessly [8, 11, 12], and does not allow the construction of *collaborative and contextual knowledge environments* that is able to provide for instruction, education, personal growth, and so on [8, 13].

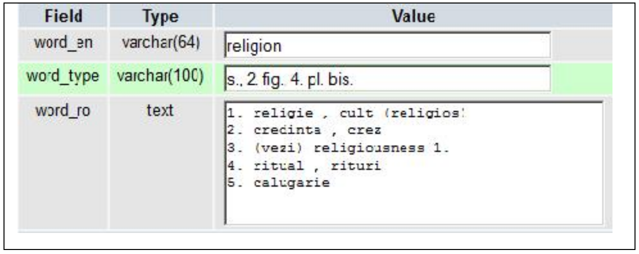
The construction of the NECLIB2 digital library has been based on a socio-semantic contextual model that allows rich bibliographic description of the content, along with semantic annotations, reviewing, rating, knowledge sharing etc. The model is multi-layered and allows both integration of local and distributed information via web services and construction of rich hypermedia documents. Moreover, NECLIB's model can express the complex relationships that exists between various objects (such as information, content, knowledge, and learning objects) and the multi-dimensional spaces, agents, actors, services, communities, scenarios, and meta-information (e. g. ontologies), and, in this way, it represents the information resources in their natural context [14, 15].

The NECLIB2 socio-semantic model subscribes to the Web 2.0 paradigm and provides for the following services: collection management, repository, indexing, semantic purposing, and user interface. The User Interface Service allows users (and communities) to access the digital library's content via typical library services, as searching and browsing, and also to contribute to the content collaboratively by annotating, reviewing, rating etc. the DL resources. Searching provides for metadata-based search, free full-text search, and combinations of these two options, while browsing relies on the rich bibliographical metadata available. Other helpful services are available as well: hypermedia information presentation, alerting services, high level authoring, visualization tools, analytical services, educational discovery etc. The user may be presented with descriptive metadata, e-books, sheet music, audio recordings, lyrics, critical and anecdotic information, various works, author or composer biography and trivia etc. The services available are flexible, manageable, contextual, and proactive. The model is presented in much more detail in [14, 15].

III. INFORMATION RETRIEVAL IN NECLIB2

In this section we present briefly some of the ideas that support the information retrieval process in the NECLIB2 digital library. These ideas have focused on *what is still the main vehicle for conveying information, namely natural language text* [16].

One idea refers to the subject descriptors of the NECLIB2 resources. For the time being, they are available in two languages: our mother tongue and English. Our goal here has been to have the program learning the correlation between the subject in our language and the subject in English. This module uses two dictionaries, one being a two language word by word dictionary, and the other being built continuously, by active learning, as a phrase by phrase dictionary. The former contains around 70,000 words, and the second includes more than 80,000 entries, which is in fact a thesaurus. When the Librarian introduces a new subject in English, she is offered with the options in our language (or other synonyms in English), which are retrieved from the thesaurus, and, if she validates the retrieved word or group of words, the pair constitutes the subject metadata for that particular resource. In Fig. 1, one may see five options available for the English word *religion*, from which the Librarian may choose: (1) *religie, cult (religios)*, (2) *credinta, crez*, (3) *religiousness*, (4) *ritual, rituri*, and (5) *calugarie*. If she is not happy with any of the retrieved options, then she is provided with suggestions taken from the first dictionary, and after validation, the metadata is stored. If she is not pleased with any of the two offerings, the Librarian may introduce her own subject metadata pair.



Field	Type	Value
word_en	varchar(64)	religion
word_type	varchar(100)	s., 2 fig. 4. pl. bis.
word_ro	text	1. religie , cult (religios 2. credinta , crez 3. (vezi) religiousness 1. 4. ritual , rituri 5. calugarie

Figure 1. Synonyms retrieved from the thesauri for the word "religion"

Another idea is based on using natural language grammars to obtain all derived forms of one word in our language, starting with the stem, followed by storing them in a dictionary. Our mother tongue is more complicated than some languages, for example, English, being similar with Latin, French, and Russian from this point of view. This dictionary has been generated using an open source tool, called *ispell*, and a custom built *affix file* that we have developed for our language (at the moment of the development, no such dictionary existed). The resulted dictionary has become a major part of our digital library, and it has contributed significantly to enhancing the efficiency of the information retrieval process.

This affix file contains a comprehensive set of paradigms for the inflectional morphology of the main grammatical categories in our language, namely nouns, adjectives, verbs, numerals, pronouns, and articles. This morphology covers the main declensions/conjugations of each grammatical category. Fig. 2 and Fig 3 illustrate how this looks and works for the word *istorii* (*histories*, in English).

```
### root words list file
istorie/SFGKM

##### affix file rules (partial list)
# plurals of nouns
flag *S:
 E→ -E,I          # istorie → istorii
# articles
flag *F:
 I E→ -E,A        # istorie → istoria
# articles plurals
flag *G:
 E →-E,ILE        # istorie → istoriile
# genitive
flag *K:
 I E→I            # istorie → istoriei
# genitive plurals
flag *M:
 E → -E,ILOR  # istorie → istoriilor
#####
```

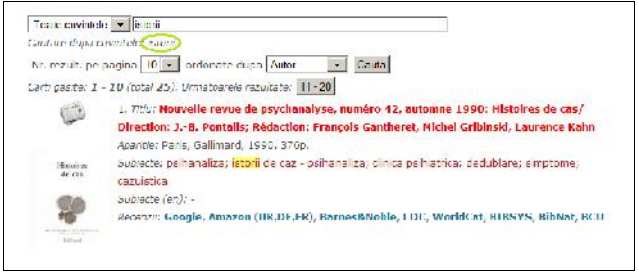
Figure 2. Derivation rules for the word “istorie” (history)

Example: search for "istorii"

step 1 = find the stem from the generated list of all words → "istorie"
step 2 = search all entries containing words derived from this stemm (based on the inverse index of stems for each entry)
step 3 = display results, highlight derived words from the stem word (istorie, istoria, istoriei, istorii, istoriile, istoriilor)

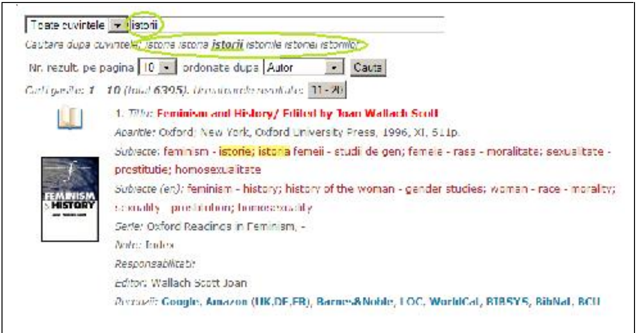
Figure 3. Example of search for the word “istorii”

Fig. 4 and Fig. 5 demonstrate how the search results looks, respectively, without using the affix file for derivation, and when the derivation of various forms of the word is performed, based on the affix file. Particularly, in this example, for the first situation, we get 25 results, while for the second one we get 6395 results, which can be sorted by author, title, publishing house, and language.



Toate cuvintele ▾ istorii
Cautare dupa cuvintele: istorii
Nr. rezult. pe pagina 10 ▾ ordonate dupa Autor ▾ Cauta
Carti gasite: 1 - 10 (total 25). Urmatoarele rezultate: 11 - 20
1. Titlu: Nouvelle revue de psychanalyse, numéro 12, automne 1990: Histoires de cas/ Direction: J.-B. Pontalis; Rédaction: François Gantheret, Michel Gribinski, Laurence Kahn
Aparitie: Paris, Gallimard, 1990, 370p.
Subiecte: psihanaliza; istorii de caz - psihanaliza; clinica psihiatrica; dedublare; simptome; cazuistica
Subiecte (en): -
Recenzii: Google, Amazon (UK,DE,FR), Barnes&Noble, LOC, WorldCat, BIBSYS, BibNat, BCU

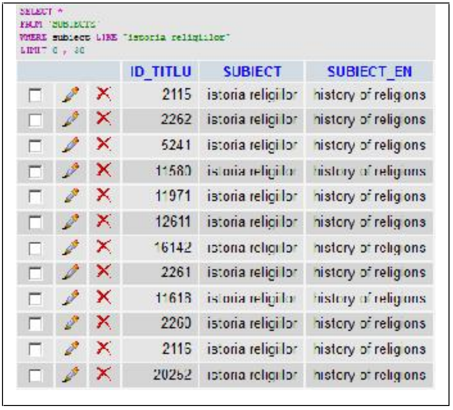
Figure 4. Search with the word “istorii” (histories) *without* derivation



Toate cuvintele ▾ istorii
Cautare dupa cuvintele: istorie istoria istorii istoriile istoriei istoriilor
Nr. rezult. pe pagina 10 ▾ ordonate dupa Autor ▾ Cauta
Carti gasite: 1 - 10 (total 6395). Urmatoarele rezultate: 11 - 20
1. Titlu: Feminism and History/ Edited by Joan Wallach Scott
Aparitie: Oxford; New York, Oxford University Press, 1996, XI, 511p.
Subiecte: feminism - istorie; istoria femeii - studii de gen; femeie - rasa - moralitate; sexualitate - prostitutie; homosexualitate
Subiecte (en): feminism - history; history of the woman - gender studies; woman - race - morality; sexuality - prostitution; homosexuality
Serie: Oxford Readings in Feminism, -
Note: Index
Responsabilitati:
Editor: Wallach Scott Joan
Recenzii: Google, Amazon (UK,DE,FR), Barnes&Noble, LOC, WorldCat, BIBSYS, BibNat, BCU

Figure 5. Search with the word “istorii” (histories) *with* derivation

Finally, using a global inverse index that includes all the possible search terms that are found among the metadata further enhances the efficiency of the information retrieval process in the NECLIB2 digital library. Fig. 5 shows a small portion of this index, where one can identify all the resources that have the subject *istoria religiilor* (namely, *history of religions*, in English).



```
SELECT *
FROM `SUBJECTS`
WHERE subject LIKE "istoria religiilor"
LIMIT 0 , 30
```

			ID_TITLU	SUBIECT	SUBIECT_EN
□	✎	✗	2115	istoria religiilor	history of religions
□	✎	✗	2262	istoria religiilor	history of religions
□	✎	✗	5241	istoria religiilor	history of religions
□	✎	✗	11580	istoria religiilor	history of religions
□	✎	✗	11971	istoria religiilor	history of religions
□	✎	✗	12611	istoria religiilor	history of religions
□	✎	✗	16142	istoria religiilor	history of religions
□	✎	✗	2261	istoria religiilor	history of religions
□	✎	✗	11618	istoria religiilor	history of religions
□	✎	✗	2260	istoria religiilor	history of religions
□	✎	✗	2116	istoria religiilor	history of religions
□	✎	✗	20262	istoria religiilor	history of religions

Figure 6. View in the inverse index

IV. RELATED WORK

Related works on information retrieval in digital libraries present, overview and analyze various solutions offered by particular digital library systems. For example, some emphasize the need of adopting statistical techniques [16, 17], or point out the benefits of using text-based techniques, or technologies like speech processing and natural language processing [16, 18].

Another information retrieval approach in digital libraries uses *tries* structures to index the resources in the Miguel de Cervantes digital library, which had a collection of more than 5000 resources in 2001. However, as the size of this inverse index, the authors consider that this approach works for catalog searches, but not for the whole content of the digital library. So they have proposed a multi-tier *tries* structure, reducing this way the size of the index to one third [19].

Optimized SQL queries and user-defined functions support another project of searching and querying digital libraries based on IR paradigms supported by relational database systems. There are two phases of the process: storing and retrieving. The former includes: document pre-processing, stop-word removal, and term extraction, while the latter is rooted in three IR paradigms: the vector space model, the Okapi Probabilistic Model, and the Dirichlet Prior Language Model. The authors have conducted experiments on the DBLP bibliography and the ACM Digital Library, and have evaluated the performances of various queries when using the three ranking models. They obtained satisfactory performance for medium-size document collections [20].

The authors of [21] identify the fundamental problem of information retrieval in digital libraries as being the fact that large-scale searches can match solely the user-specified terms with the ones that appear in documents of the DL collection. Therefore, intermediate sources (here thesauri and co-occurrence lists) that provide for term suggestions can improve the retrieval process by providing alternative search terms for the user. This way, the recall increases, while the user keeps an eye on not to decrease the precision. The thesauri used in the experiment are human-generated and place the terms in a subject hierarchy, while the co-occurrence lists are computer-generated and lay the terms in the frequency order of occurrence together. The prototype has been built for the University of Illinois Digital Library Initiative (DLI) tested. Multiple views that provide for using and combining of different sorts of term suggesters are supported.

Another interesting approach is taken in [22], where the authors introduce an adaptive tool for giving automated advice by providing suggestions during the searching process in digital libraries. This tool uses case-based reasoning techniques to provide the most useful suggestions for a given situation by comparing them to previous cases stored in the case base, and then adapting the solution. Furthermore, this tool can learn from the interaction with the user. The users involved in the experiment have found the automated, non-intrusive advice to be useful, and have used the suggestions to further improve their searching.

Semantic modelling is also used for IR in digital libraries. Thus a multilingual information retrieval system based on knowledge representation model is presented in [23]. That system provides for information retrieval in a multilingual document repository, in which the materials are written in various languages, however, each individual document may contain text in only one language. To do that, the system relies on a semantic graph-based model, which allows the description of the semantics of a document, in that multilingual context, using two types of knowledge: domain related (concept and relation) and lexical related (that associates terms in a vocabulary to concept or relation type of knowledge).

V. CONCLUSIONS AND FUTURE WORK

Nowadays, every person has the potential to create content, to share it online, ant to mash it up with other content in various ways. While, traditionally, the ability of information searching has made an important part of the reference librarian job description, currently most information seeking is carried out directly by end users themselves, either via library controlled systems or directly on the web. To cope with the challenges risen by these opportunities, digital libraries are expected to fully fulfill the user needs with respect to searching and browsing, which are ought to be performed easily, effectively, and efficiently. Users are likely expected to do nothing more than typing a few words or choosing an appropriate filter, to start the searching process. The information retrieval capability offered by digital libraries is expected to provide them with a proper set of relevant documents, despite the limitations and shortcomings of natural language queries. This desideratum has been approached in this work, in which we have presented some of our solutions for this problem that are at work in a real world library, which includes a very large number of resources in a variety of formats: books, e-books, periodicals, audio and video CDs and DVDs, video tapes, scanned microfiche, (digitized) audio recordings etc..

Further work includes adding capabilities for several activities: recommending, mobile access, collaborative information retrieval, construction of personalized collections, tracking documents, manipulation of citation trails, automatically creating bibliographies, choosing the right algorithm to perform a search (e.g. to produce high precision or high recall; or to show only novelties), more contextualized searches, more options for query extensions and refinements, and collecting more user feedback, to be used for further versions of NECLIB2.

REFERENCES

[1] I. Torvik Sølvberg, "Digital Libraries and Information Retrieval", *ESSIR 2000*, LNCS vol. 1980/2001, pp. 139-156, M. Agosti, F. Crestani, and G. Pasi Eds., Springer-Verlag Berlin Heidelberg, 2001.

[2] I. Papadakis, M. Stefanidakis, K. Kyprianos, and R. Mavropodi, "Subject-based Information Retrieval within Digital Libraries Employing LCSHs", *Dlib Magazine*, vol. 15, no. 9/10, September 2009, http://www.dlib.org/dlib/september09/papadakis/09papadakis.html

[3] C. Schwartz, "Thesauri and Facets and Tags, Oh My! A Look at Three Decades in Subject Analysis", *Library Trends*, vol. 56 (4), pp. 830–842, Winter 2008.

[4] D. E., Bennett, Immaculate Catalogues, Indexes and Monsters Too.., *Ariadne*, 49, Oct. 2006, [Online] Available http://www.ariadne.ac.uk/issue49/cig-2006-rpt

[5] National Digital Library Program, Library of Congress, "Challenges to Buliding an Effective Digital Library", [Online] Available http://memory.loc.gov/ammem/dli2/html/cbedl.html

[6] E. Rasmussen, "Information Retrieval Challenges for Digital Libraries", *ICADL 2004*, LNCS vol. 3334, pp. 95 – 103, Z. Chen et al., Eds., Springer-Verlag Berlin Heidelberg, 2004.

[7] (2012) The NECLIB2 website, [Online] Available http://library2.nec.ro/

[8] C. Lagoze, D. B. Krafft, S. Payette, and S. Jesuroga,"What Is a Digital Library Anymore, Anyway? Beyond Search and Access in the NSDL", *D-Lib Magazine*, Vol. 11, No. 11, November, 2005, [Online] Available http://www.dlib.org/dlib/november05/lagoze/11lagoze.html

[9] P. Parrish, "The Trouble with Learning Objects," *Educational Technology Research and Development*, vol. 52, no. 1, pp. 49-67, 2004.

[10] M. Recker, A. Walker, and D. A. Wiley, "Collaboratively filtering learning objects," *Designing Instruction with Learning Objects*, D. A. Wiley, Ed., Bloomington, IN, AECT, 2000, pp. 243-259.

[11] R. Daniel Jr. and C. Lagoze, "Extending the Warwick Framework: From Metadata Containers to Active Digital Objects," *D-Lib Magazine* November, 1997.

[12] M. Recker, J. Dorward, and L. M. Nelson, "Discovery and Use of Online Learning Resources: Case Study Findings," *Educational Technology and Society*, vol. 7, no. 2, pp. 93-104, 2004.

[13] D. Greenstein, "Lessons in Deep Resource Sharing from the University of California Libraries", Council on Libraries and Information Resources, [Online] Available http://www.clir.org/pubs/reports/pub119/greenstein.html

[14] Z. Constantinescu, and M. Vladoiu, "MNECLIB2 – A Classical Music Digital Library", in *Proc. International Conference on Electrical, Computer, Electronics and Communication Engineering (ICECECE 2011)*, pp. 682-689, 2011.

[15] M. Vladoiu, and Z. Constantinescu, "Open Digital Library on Digital Libraries", in *Proc. 10th Int'l Conference Romanian Educational Network - RoEduNet*, pp. 209-214, 2011.

[16] K. S. Jones, "Information retrieval and digital libraries: lessons of research", in *Proc. International Workshop on Research Issues in Digital Libraries (IWRIDL 2006)*, Kolkata 2006, ACM, 2007.

[17] B. Schatz, "Information retrieval in digital libraries: bringing search to the net", *Science*, January 17, vol. 275(5298), pp. 327-334, 1997.

[18] E. M. Voorhees, “Natural Language Processing and Information Retrieval”, *Information Extraction Towards Scalable, Adaptable Systems*, pp. 32-48, M. T. Pazienza, Ed., Springer-Verlag, London, UK, 1999.

[19] A. Bia, A. Nieto, (2012) “Information Retrieval in Digital Libraries: efficient catalog searches using tries”, [Online] Available http://bib.cervantesvirtual.com/research/articles/tries.pdf

[20] C. Garcia-Alvarado and C. Ordonez, “Information retrieval from digital libraries in SQL”, in *Proc. 10th ACM workshop on Web information and data management (WIDM '08)*, ACM, New York, USA, pp. 55-62, 2008.

[21] B. R. Schatz, E. H. Johnson, P. A. Cochrane, and H. Chen, “Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval”, in *Proc. 1st ACM international conference on Digital libraries (DL '96)*, pp. 126-133, E. A. Fox and G. Marchionini, Eds., ACM, New York, USA, 1996.

[22] S. Kriewel and N. Fuhr, “Adaptive search suggestions for digital libraries”, in *Proc. 10th International Conference on Asian Digital Libraries: looking back 10 years and forging new frontiers (ICADL'07)*, pp. 220-229, D. H.-L. Goh, T. H. Cao, Ingeborg Torvik Sølvberg, and E. Rasmussen, Eds., Springer-Verlag, Berlin, Heidelberg, 2007.

[23] C. Roussey, S. Calabretto, and J.-M. Pinon, “SyDoM: A multilingual Information Retrieval System for Digital Libraries”, in Proc. *5th International ICCC/IFIP Conference on Electronic Publishing, ELPUB’2001*, p. 150-160, 2001.