elpub2002 Proceedings — J. A. Carvalho et. al. (Eds.) — VWF Berlin, 2002

Improving Information Retrieval in Digital Theses Using Metadata

Rocio Abascal, Béatrice Rumpler, and Jean-Marie Pinon

INSA de Lyon – LISI
7 Avenue J. Capelle Bât 502 – Blaise Pascal
F-69621 VILLEURBANNE CEDEX – FRANCE
{rabascal@lisi., beatrice.rumpler@}insa-lyon.fr

Abstract. In this paper, we present an approach to improve information retrieval in digital scientific theses. This approach consists in defining and using "metadata" to help the users to find relevant information during search sessions. This research is one part of the CITHER project (Consultation en Texte Intégral des THÈses En Réseau) developed by the INSA of Lyon (France). CITHER concerns the online publishing of the INSA's scientific theses. In a first step, CITHER has permitted the distribution of the theses in PDF format (Portable Document Format), via a server of documents. However, this system does not permit to select only the pertinent parts of the theses during a search session. It is necessary to read the entire document to find them.

In the first part of this paper, we present the initial structure of a thesis stored in the CITHER's server. Then, we describe our method to define a new structure of document based on "*semantic metadata*". We propose to introduce the concept of ontology to define the semantic "metadata" and to use XML (eXtended Markup Language) to structure the document. To define the semantic "metadata", we extract the main concepts such as "model", "theorem", "method", "tool", etc. found in almost all the theses. We formalize these concepts according to an ontology. Therefore, the new model of document includes "*classical metadata*" (Dublin Core's, etc) and "*semantic metadata*", which are both included in the document by the way of XML tags. This model of document is based on a logical and semantic structure. The information retrieval, in our new format of theses, takes place by the use of XML tags.

In the second part of this paper, we will present our prototype and some examples to illustrate our proposition and the results. We will finish this paper with our conclusion and the future propositions to improve the prototype.

1 Introduction

The digital library (DL) has the potential to make fundamental changes in the routine of the research institutions have assembled, organized, and provided access to research materials for generations. The collections in a DL represent a large part of the accumulated documentary knowledge wealth of the research world. To make this knowledge accessible to all the users it is important to provide *intellectual access* to the information. Research in this initiative is concerned with developing concepts, technologies and tools to access to the knowledge and meaning inherent to digital collections. For example:

- For the users, this means intelligent information retrieval, adapted organization of information and intelligent tools and interfaces to access the information.
- For content and collections providers, this means new information types, new structures of documents, document encoding and metadata for enhancing context.

Now, one of the main difficulties for the users in using DL, is to locate relevant information. The tools used to access DL do not use contextual information to make the difference between the terms and the meanings [11]. Nowadays, many projects are working to allow the distribution of theses online. To find relevant information, generally these projects only propose a full text search engine that works in combination with keywords such as the author name, the title and the abstract of thesis.

Our project CITHER[1] (Consultation en texte Intégral des THÈses En Réseau) takes place in the INSA of Lyon's[2] scientific library. CITHER makes a distribution of the theses in PDF format (Portable Document Format) via Internet. In the CITHER system, we encountered the same difficulty as in other DL's: during a research session, it is impossible to extract the pertinent contents of the theses. We only get one thesis at a time during a search session. To evaluate the pertinence of the thesis, we have to read the entire thesis or at least, several chapters. To solve these problems, we propose to study:

- The "*semantic*" structure of the document in order to offer a more detailed and pertinent retrieval.
- Software for information extraction able to identify the key concepts of a document.
- The use of XML (Extensible Mark-up Language) to translate the contents into a structured form.
- The use of new "metadata", information contained in the data and the information about this data [2], to describe the contents of the thesis.

[1] http://csidoc.insa-lyon.fr/these/
[2] http://www.insa-lyon.fr/

Following this presentation, in Section 1, we present the initial CITHER project and the methodology to define new metadata for the theses. In Section 2, we present the extraction of concepts and some tools that extract the terms and concepts located in a document. In Section 3, we present the notion of ontology to represent the concepts in the scientific theses. Section 4 presents our prototype and the results of our approach. Finally, in the Section 5 we discuss the related works to finish with the conclusion and the future work.

2 New Metadata for the Scientific Document

Describing data by using "*metadata*" has proven to be useful for integrating data from different sources and to have better information retrieval methods. *Metadata* is rarely used. As practice shows, only 3% of data available in the Web uses Dublin Core[3] as a content description model [8]. The correct employment of metadata in DL's is crucial to provide effective access to the pertinent information.

The initial CITHER project proposes the online access to the scientific doctoral theses of the INSA of Lyon, since January 1997. It allows the consultation, the conservation of the theses and the promotion of the research of the laboratories of the INSA of Lyon. The distribution of the theses, in PDF format, is done by the way of a server. CITHER proposes various classifications for information access: by alphabetical lists, by computerized catalog (Doris Web) and by full text. CITHER makes the digitalization of the theses; the information to be stored in electronic shape is coded with a specific CEN (Chain of Digital Edition). The theses deposited in Word or Latex format, are then translated into PDF format by using the CEN. Nowadays, CITHER permits to get the complete thesis corresponding to a demand, but the main need is to have a tool that allows making the consultation of the thesis by the way of pertinent extracts. To structure the theses, CITHER uses the specification proposed by Dublin Core. These metadata allows extracting the title of the thesis, the author and some key concepts but it does not permit to select the most pertinent theses from the collection.

In our current research, we are developing methods and concepts to try to improve information retrieval in the DL. We are constructing ontology providing an effective way to conceptualize the scientific theses.

Therefore, our work consisted, in a first step, to include new *metadata* in a thesis, by using a document editor [1]. This editor allows the writer to define styles and add tags according to the physical structure of the thesis during the writing process. In a second phase, we decided to analyze each thesis to find the recurrent semantic elements that are present in almost every thesis. This permits to define new tags or "*indicators*" corresponding to the semantic structure of the document such as: "*model*", "*tool*", "*author*", "*modeling*", "*application*", "*method*",

[3] http://www.dublincore.org/

etc. An example of the differences between one document without metadata in the CITHER project and the one with the new metadata is presented in the next figure (Fig. 1). We can notice that in the first document there are few tags, coming from the transformation of RTF to XML.

The new metadata are included in our prototype by the way of XML tags in combination with the proposed ontology (see: Section 3). This new model of document permits the user to select pertinent information.

In the following section, we describe the process for extracting the recurrent concepts found in the scientific theses. In addition, we present a study of some tools that can help us in such process.

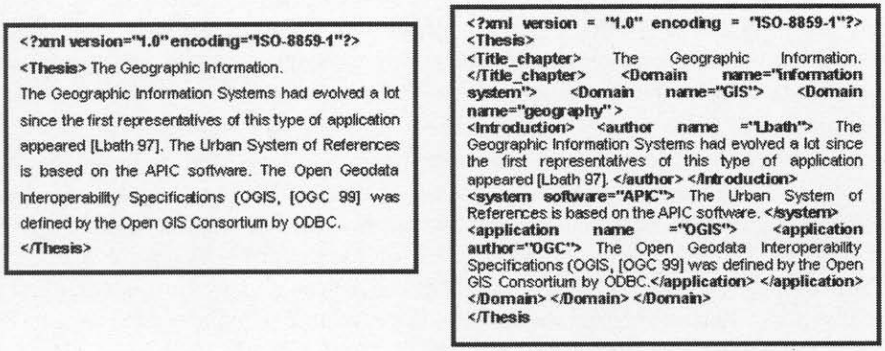

Fig. 1. *Example of a document including new metadata.*

3 Concepts Extraction

One of the causes of the problem of the document retrieval deals with the fact that information retrieval systems rarely focus on exploiting the semantic content of the documents [6]. Our work aims at providing a model of structured scientific documents that does not focus on how to represent the data physically, but on how to represent the elements of such documents that can be implicated during an information retrieval session. To make this possible it is necessary to model the knowledge represented in the scientific document.

The modeling of knowledge has been the objective of many works around the world. The main problems about the modeling of knowledge are the selection of concepts, the choice of their properties and their relations, their grouping and the influence of the application in making these choices. Nowadays, several tools make the extraction of concepts from a document possible.

To model the scientific document, the theses, our approach is based on the extraction of all the concepts found in almost all the theses. For this, we have made a study of the different tools able to extract "*keywords*" or "*key phrases*" from documents. By the term *key phrase*, we denote a list of phrases composed of two or more words. The task we consider here is to take a document as an input and automatically generate a list of key phrases as outputs. We evaluate the tools by comparing their output phrases or words with an initial list. The principal objective is to select one tool to work with it along all our research.

The main tools for extracting terms from documents work by making a statistic analysis of the number of times that a word appears in the text. The problem is that some of these tools make the extraction of words that are not concepts and it is therefore necessary to use a dictionary, a glossary or a thesaurus to find the concepts that are associated to these terms.

We are testing some tools such as: XTS (Xerox Terminology Suite) of Xerox, TerminologyExtractor of Chamblon Systems Inc., Nomino of Nomino Technologies and Copernic Summarizer which use the algorithm of the NRC (National Research Council). These tools are appropriate to our project because they use a semantic approach to make the analysis. In our study, we are testing several articles and chapters of theses. We compare the keywords given by the author with the concept extracted by the tools. The results of our study will be published soon.

4 The Notion of Ontology

Nowadays, many researchers propose to improve information retrieval by using the semantic representation of a document and to introduce concepts associated to significant words of the documents. This approach is based on ontology. Ontology is the description of the concepts and the existing relations [9]. An ontology is defined as an explicit specification of a conceptualization [10]. Ontology offers possibilities to navigate by concepts between different documents. The ontology is tied to a model, which represents the document structure, through a tree of objects connected with relations [5].

4.1 The Methodology

The ontology we proposed is based on the analysis of the CITHER's scientific theses using the concepts found by the tools listed in the previous section (Section 2). The ontology is built on two structures of the document: the logical structure and the semantic structure.

To model the document, in a first step, we use the concepts as new metadata, this correspond to the semantic structure of the thesis. During our experience, we also noticed that in the first chapter of a thesis, we often find elements such as:

"*state of the art*", "*modeling*", "*method*", "*tool*", "*application*", etc. We have decided to use these elements (called "*metadata*"), according to the notion of ontology, to highlight parts of the thesis. This modification affects the logical and the semantic structure of the thesis. These "*metadata*" will be implemented as new "tags" in the XML document. One of the main ideas of our approach is to permit the user to include semantic metadata during the process writing of the thesis.

The relations existing between different concepts and the knowledge stored in the theses express the semantic representation of the theses. For example, the element "*State of art*" is composed by the element named "*Modeling*". This element is tied to several indicators such as "*tool*", "*method*", and "*application*". Each element possesses several attributes like: "*name*", "*date*", and "*author*". The logical representation of the thesis is expressed by a set of rules, which can be compared with the grammar of a programming language. The rules associated to this structure are expressed by the XML language, which allows the representation of the document structure. We are going to present a new DTD (Document Type Definition) in the next figure (Fig. 2). This new DTD, uses new metadata to model a thesis.

```
<!ELEMENT State_of_art (Modeling)*>
<!ELEMENT Modeling (Tool | Method | Application)*>
<!ELEMENT Tool name CDATA #IMPLIED>
<!ELEMENT Tool author CDATA #IMPLIED>
<!ELEMENT Tool date CDATA #IMPLIED>
<!ELEMENT Method name CDATA #IMPLIED>
<!ELEMENT Method author CDATA #IMPLIED>
<!ELEMENT Method date CDATA #IMPLIED>
<!ELEMENT Application name CDATA #IMPLIED>
<!ELEMENT Application author CDATA #IMPLIED>
<!ELEMENT Application date CDATA #IMPLIED>
```

Fig. 2. *Example of rules in XML.*

The ontology has a taxonomy and a set of "inference rules". The taxonomy defines classes of objects and relations between them. By using the ontology, we want to conceptualize specific concepts of a domain, the knowledge, and the relationships between concepts. To make this possible we have to normalize the list of concepts. This means, to separate the concepts by their differences and to establish the relationships between every concept. A large number of these relationships can be expressed by assigning properties to classes. These classes have subclasses, which can allow inheriting properties [4].

In the next section, we present the prototype developed to validate our approach.

5 Prototype

The prototype has been made by using the documentary features of ORACLE such as the facility to find the information stored between two XML tags or the information included in the attributes of the XML tags.

To evaluate our approach, we compare the functioning of the initial version of CITHER to our new proposition. The initial interface of CITHER only allows a request containing the author name, the date and the theme of the thesis. This is very restrictive because it only allows getting one complete thesis at a time, related to our query.

Our prototype proposes three different types of requests: "*simple*", "*combined*" and "*personalized*".

The "*simple*" research looks for the existence of only one keyword in the document. The result appears in a window with the title, the name of the thesis, different pertinent parts of the thesis related to this keyword and the summary.

The "*combined*" research allows the user to make a request composed of several keywords. This is possible by using logical operators like "AND", "OR", and "NOT". As with the "simple" research we obtain the title and the name of the thesis, the paragraphs containing the keywords and the summary of the thesis.

The "*personalized*" research is the most performing research and corresponds to the objective of our work. This research uses all the new metadata we had introduced in the thesis. The interface proposes the list of the metadata included in the thesis to the user. The user only has to select the metadata to build his request. For example, if the user wants to find a thesis about the *domain* "GIS" (Geographic Information System) written by "Robertson" in the year "2000" and containing references to a *system application* named "APIC", he has to select the metadata as shown in the next figure (Fig. 3).

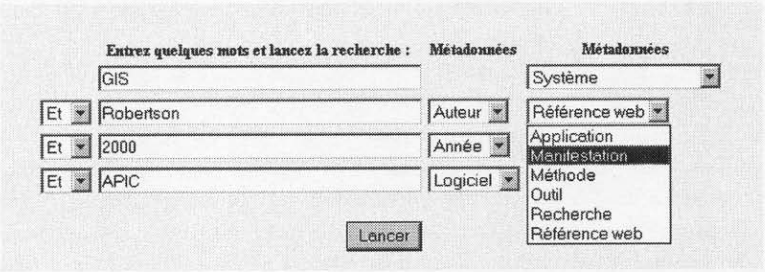

Fig. 3. *Example of the "metadata research".*

Unlike the results obtained with the initial version of CITHER, which sends back the complete thesis, our proposition only returns the title, the author name, the most pertinent paragraphs and the summary of the thesis for every type of requests. The interface offers a global view of the metadata added to the thesis, so that the user can create the requests more easily. One of the objectives of this interface is to mask the complexity of the system to the user. Our prototype also allows, in a single request, to get several paragraphs from various theses, the most relevant ones for the user request. The ORACLE system, which makes an important part of our prototype, allows to archive the theses in a database and to select the most relevant document for the user. Thanks to the features of ORACLE 8i, it is also possible to store the URL number of a thesis. ORACLE 8i manages connection with Internet and permits to develop software with Java language.

In the following section, we are going to show the works related to our approach. These works are addressed to the diffusion of doctoral theses in DL's. We are going to present the way they structure the information to solve the document retrieval problem.

6 Related Work

There are different DL's projects whose aim is to collect, store, organize and diffuse doctoral theses in digital forms via Internet. For example, we can talk of projects such as: Cyberthèses of the University of Lyon II (France) and the University of Montreal, the NDLTD project of the University of Virginia Tech (United States) [7], Webthèses of the University of Lille II (France), the University of Waterloo (Canada) and the CITHER's project of the INSA of Lyon.

All these projects propose a documentary research form the title, the author name, the scientific director name, the keywords and the date of the thesis. This research only sends back one complete thesis so that the user has to read every chapter to find the relevant information. Some of these DLs use metadata to structure the information. The choice of metadata is a very important step in the process because it allows a better classification of the information and the document [3]. In a library, it is possible to make an enormous number of descriptions available from Internet by converting the information or by using specific XML's DTD [12]. For example, the University of Waterloo uses an XML DTD to describe the scientific thesis. This approach seems to be close to our needs but we do not find in the DTD the special elements used in the CITHER format of document. The CITHER project uses 15 tags proposed by Dublin Core specification. The DL generally use metadata to define the physical structure of the thesis but we never find a DTD describing the semantic structure, the metadata are not used yet to permit more pertinent retrieval information from a doctoral thesis. For this reason, in our project, we propose an original approach to find relevant information. The

new metadata or "indicators" added by the writer of the thesis represent the base for a semantic model of the thesis. This model gives many perspectives for a better access to DL's.

7 Conclusions

The project CITHER of DOC'INSA has the objective to diffuse the theses of the INSA of Lyon, in complete text and in PDF format. Therefore, during a research it is possible to reach the contents of only one thesis by selecting every chapter.

We have described in this paper a way to model scientific documents by adding new metadata to achieve their retrieval.

Our approach has permitted to build a prototype able to select pertinent thesis from the database and to extract the most relevant paragraphs using the semantic structure of the thesis. The researcher can create this semantic structure while writing the thesis, simply by adding metadata in the text. This project also shows the interest of a research using metadata and is a first step to automate the building of semantic representation of scientific documents.

Many future works in this direction are possible to improve our approach. Now, the user can only insert in the thesis the metadata included in the system. It would be interesting to use a thesaurus combined to an expert system, to automatically propose to the user the suited metadata to include in his document. To do it, it is necessary to integrate a thesaurus into the system to improve the precision of the research.

References

1. Abascal, R. (2001) Modélisation du contenu sémantique de documents scientifiques à partir des métadonnées. Application au serveur de thèses CITHER. Mémoire de DEA, Septembre 2001. LISI-INSA de Lyon, France, 32 pages.
2. Arms W. (1995) Key Concepts in the Architecture of the Digital Library. Corporation for National Research Initiatives. Reston, Virginia .D-Lib Magazine, July 1995.
3. Beaudry G., Gauvin J. (2000) Rapport de la phase pilote du projet de publication et de diffusion électroniques des thèses de doctorat. Université de Montréal.
4. Berners-Lee T., Hendler J., and Lassila O. (2001) The Semantic Web. Scientific American Feature Article. May 2001.
5. Desmoulins C., Grandbastien M. (2000) Des ontologies pour indexer des documents techniques pour la formation professionnelle (Loria Nancy).
6. Fourel F., Mulhem P. (1996) Modelling Multimedia Structured Documents: A Retrieval Oriented Approach. Database and Expert System Applications (DEXA) Workshop. Zurich, Switzerland.
7. Fox E., Eaton J., McMillan, et al. (1997) Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources, in D-lib Magazine (September 1997).

8. Gertz, M. (2000) Achieving Semantic Interoperability Through Controlled Annotations. US-Korea Joint Workshop on Digital Libraries. Aug 10-11, 2000, San Diego Supercomputer Center, San Diego, C.A.
9. Gruber T. (1993) A translation Approach to Portable Ontology Specification. In Knowledge Acquisition, 5(2), pp. 199-220.
10. Guarino, N., Giaretta, P. (1995) Ontologies and knowledge Bases: towards a terminological Clarification. Towards Very Large knowledge Bases: Knowledge Building and Knowledge Sharing, Eds. N. Mars, IOS Press, pp. 25-32.
11. Heflin J., Hendler J., Luke S. (1999) Applying Ontology to the Web: A Case Study. In J. Mira, J. Sanchez-Andres (Eds.), International Work-Conference on Artificial and Natural Neural Networks. IWANN'99. Proceedings, Volume II, Springer, Berlin, pp. 715-724.
12. Hernandez F., Linde P., Mulrenin B., Yeates R. (2001) Converting heterogenous cultural catalogues and documents to XML-strategies and solutions of the COVAX project, in Electronic Publishing'01 – 2001 in the Digital Publishing Odyssey. A. Hubler et al. (Eds). pp. 65-82.