IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB12834

This paper appears in the book, *Emerging Trends and Challenges in Information Technology Management, Volume 1 and Volume 2* edited by Mehdi Khosrow-Pour © 2006, Idea Group Inc.

Knowledge Extraction to Improve Information Retrieval in Scientific Documents

Rocío Abascal & Béatrice Rumpler
INSA of Lyon - LIRIS, 7 av. J. Capelle - Bat. Blaise Pascal, F69621, Villeurbanne cedex, France,
{rocio.abascal, beatrice.rumpler}@insa-lyon.fr

ABSTRACT

Annotation is a key way in which documents grow and increase in value. This paper explores the possibility to use concepts extracted from documents by using a Natural Language Processing tool to characterize the content of digital theses. Then, using the results of the study, the paper explores the use of annotated theses in order to access to pertinent information stored in these documents and to extract knowledge by an *"intelligent"* search system.

1 INTRODUCTION

The growth of the World Wide Web and of the corpus of documents it covers has increased the necessity to propose solutions to improve information retrieval. Our proposition is based on a specific semantic annotation process of the documents, made during the writing step and explored during the search session. These semantic annotations also allow knowledge extraction from the documents and lead to an intelligent information processing. Accessing and extracting knowledge from online documents is crucial to develop advanced knowledge services for the Semantic Web.

The scientific library of Doc'INSA set up since 1997 a project named CITHER, which makes possible the diffusion and the access of scientific theses through Internet. Currently, a user can get the contents of only one thesis at the same time without being able to select relevant extracts corresponding to a unit of corpus finer than the chapter. This is the result of the use of an inadequate format, such as PDF (Portable Document Format), the description of the contents by only the keywords added outside the documents, and the use of the tags proposed by the Dublin Core metadata which bring general information of the thesis.

Our research focus is based on the opportunity to use the domain concepts to build the users requests and to organize the document structure. We propose to the users to build a semantic structure for the documents by using a Natural Language Processing (NLP) tool that extracts concepts and by using a base of concepts of the domain field. The selected concepts are stored in the documents as semantic tags. Then, a user's query generates a web access to a page that contains pertinent information. The discovery of relevant contents is done by matching the user's query with the Embedded Semantic Tags (EST's). Once the desired information is found the user only read the pertinent fragments and so, he can select the right documents, in our case scientific theses.

While recent research efforts seek to add relevant markups to the content of the web pages [4], [6], [8], we go a step further by embedding the theses from their creation. We show that this contributes to enhanced automatic semantic-based recovery of information content.

2 Generating Adaptive Annotation to Structure Documents

The Semantic Web aims to create contents that can be manipulated by humans but also by machines [7]. This can be achieved by explicitly adding markups to describe the content of a digital document. The HyperText Markup Language (HTML) was the initial language used to display documents on the web. The main drawback of HTML is its inability to represent semantic contents. This led to the Extensible Markup Language (XML) [5], which allows inserting specific XML tags in the text. These tags permit an automatic exploration of the documents. The Document Type Definitions (DTDs) or a schema like XML Schema can validate the inserted tags.

The annotation of existing digital documents is one of the basic barriers towards the conception of the Semantic Web. Manual annotation is impractical and unscalable, while automatic annotation tools are still in their infancy. Hence advanced knowledge services may require tools able to search and extract the required knowledge from the Web, guided by a domain conceptualization (ontology) that specifies what type of knowledge is needed. In our approach, we propose the author of the thesis to describe his thesis with metadata characterizing the main content.

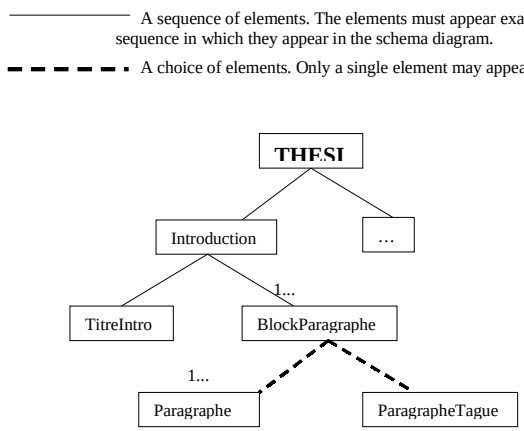
Following a meticulous work of extraction of concepts [3], a study of the corpus and the description of the correlation between the uses of concepts in the corpus, we planned to propose to the author two components to help him during the description step. First, we propose to use a Natural Language Processing (NLP) tool, called Nomino, to automatically extract concepts from a document. We have selected Nomino after a comparative study of four NLP tools [2]. Second, the user can also use Nomino to know and extract the most important concepts included in a fragment of a thesis. So it becomes not necessary to read the entire document.

We have built a knowledge base with the concepts extracted from a corpus of theses. This base must be regularly updated with the new concepts extracted from new theses stored in the digital library. By making some experiments, we have evaluated that the number of the concepts brought by each new thesis of a specified domain does not increase infinitely; it quickly tends towards a constant stabilization. The number of new concepts becomes to be very weak after the evaluation of about 25 theses of the same domain [1].

Our proposal for adaptive semantic annotation can be characterized by the following elements:

- The PhD student is assumed to write his thesis by making semantic annotations. These annotations are generated in XML format. To simplify the student task, a concept extraction tool can be called after the selection of a written fragment, section or chapter. The NLP tool, named Nomino, proposes pertinent concepts and the user can accept or deny then for an insertion in the document
- Another way to add concepts is by selecting them from the base of concepts. In the base, the concepts are ordered by hierarchies according to the computer field.
- While the new tags are currently employed in the document, the annotation system is transparent to the student and it is not necessary to know how XML works.

Fig.1. Example of a model used to validate digital theses



Our proposed annotation system involves the addition of a schema that defines the structure of the document (the thesis). In the next section, we describe the schema used in order to validate well structured documents.

3 VALIDATION OF THE STRUCTURE BY USING XML SCHEMA

In this section we illustrate the use of a model (created in XML Schema) to validate the new format of the theses created by the PhD student.

- A sequence of elements. The elements must appear exactly in the sequence in which they appear in the schema diagram.
- A choice of elements. Only a single element may appear at this position.

A thesis is based on several logical entities, like the introduction, the conclusion, the chapters, the sections, the subsections, the paragraphs and the blocks of text that are the finest textual logical entities. These entities can be *"tagged"* or *"not tagged"*. A set of *"tagged"* paragraphs can constitute a *"BlockParagraphe"*. The *"tagged"* state results from the presence of metadata (concepts) surrounding the *"BlockParagraphes"*. At the beginning of the block we find the heading of the metadata *"EnteteMetadata"*, and at the end, the *"PiedMetadata"* (see Fig. 1). The existence of elements such as the heading of metadata, the lists of paragraphs and the foot of metadata is necessary in a *"tagged"* block, so the minimum cardinality of each element is one. In the same way, the entities named *"EnteteMetadata"* and *"PiedMetadata"*, all the elements such as the opening or closing tags, the list of the concepts and the boolean variables *"Precedent"* (before) and *"Suivant"* (following) must be initialized.

The overlap between different elements is not allowed in XML Schema [9]. However, one or more semantic segments can be defined by interlacing one or more parts of the logic elements. For example, when a prototype is described in a thesis it can be introduced by several chapters about the *"state of the art"*, and it can be described in specific chapters. Thus, in the foot of the metadata of the paragraphs tied to the *"prototype"*, when the paragraphs belongs to the chapter *"state of the art"*, we initialize by *"yes"* the variable named *"Suivant"*. In the same way, in the heading of the metadata of the chapters tied to the prototype, the variable called *"Precedent"* will be initialized by *"yes"*. So, the writer will be able to insert metadata at any part of the body of the thesis to create a well-described document (see Fig. 2). Thanks to these metadata, it becomes possible to extract pertinent information during a search process.

Fig. 2. Example of the metadata tags used in our XML Schema

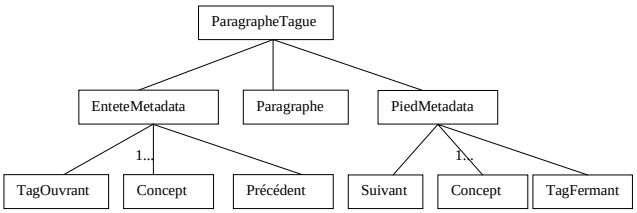
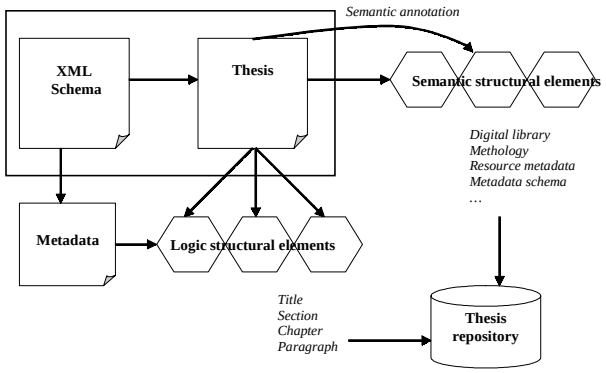


Fig. 3. Combination of logical and semantic structural elements



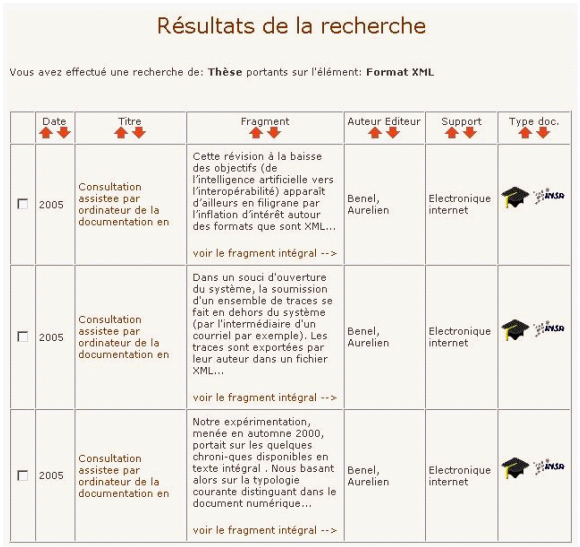
4 EXPERIMENTAL SECTION

In this section we illustrate the use of EST's to identify and extract responses to queries by using concepts names. A digital thesis search tool is used to parse the theses and extract the pertinent fragments. The user request, composed by keywords or concepts-words, is expanded by using narrower and broader concepts found in the base of concepts. These new concepts are proposed to the user in order to clarify his main idea and to reformulate the query by using adequate concepts. The search tool is able to provide the fragments where the concepts of the query physically appear and the fragments surrounded by the pertinent semantic tags even, if the concepts are not explicitly written in the fragments. For example, if we have the following XML paragraph *"<Internet> <Semantic_Web>The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. <Internet/> </Semantic_Web>"* and if the user is searching for all the fragments containing the concept *"Internet"*, by using any research system he will probably not obtain the paragraph presented above. Instead, by using our system, even if the *"Internet"* word is not written in the paragraph, by using our XML tags the user is going to find this paragraph.

The thesis produced by the student is composed by metadata coming from the logical structure (chapter, section, paragraph, image, etc.,) and metadata coming from the semantic structure (paragraph about *"system architecture"*, *"model"* or *"prototype"*, etc.). The metadata used in the semantic structure are very powerful because they give specific information about the fragment or the associated concepts. For example, we can have concepts like: *"digital library"*, *"Fourier model"*, etc. Thanks to these two kinds of metadata the information retrieval tool is able to explore the tags in order to know which chapter is talking about a specific concept (see Fig. 3).

A typical user interaction with the search system consists in inserting a query composed by concepts. If there are more concepts closer to those used in the query then the concepts base will propose to the user other concepts in order to expand the query. The user has the option to select the most adequate concepts to expand his request, as shown in the second screen shot. Finally, once the user has selected the concepts, the system

Fig. 4. Screenshot for the result of a search session



searches the right information in the thesis repository and then shows the pertinent fragments to the user (Fig. 4). The next figure presents different fragments containing the concept *"format XML"*. In this figure we only present the fragments of one these but our system shows all the fragments found in all the theses.

5 FUTUR WORK

In order to improve the research of information within the digital library we propose:

- To use an ontology of the computer field,
- To use the *"user profile"*.

5.1 Using an Ontology of the Computer Field

In our work, the ontology is a complement of research related to the semantic tags that were added into the scientific theses during the writing step. During a research session, by using an ontology it will be possible to seek relevant information based on the semantic tags. The use of the ontology will allow to propose other concepts than those proposed by the concept's base.

The ontology we propose is still very incomplete (it will be completed as soon as new theses are registered into the CITHER system). The ontology can also help the user to build the query during a search session. This way, we study some methods suited to the expansion of queries.

Now, we are testing the opportunity to use an ontology during a search session. The ontology will allow the query expansion by using the concepts related to the ones proposed by the user. The results of this study will allow the evaluation of the possibility to introduce synonyms or others words, to improve our ontology.

5.2 Using the User Profile

Taking into account the needs, the intentions and the cognitive, cultural or different specificities which characterize the *"user profile"* constitutes a determining element to improve the relevance of the answers during a search session in large bases of documents. The modeling of the *"user profile"* and the way to adapt it to different users who do not have a precise idea of the information they seek, enables a personalized access to the contents of scientific documents, based on the exploitation of the *"user profile"*. The *"user profile"* can consist of a whole of characteristics associated to values, containing the user preferences.

The user's profile can be obtained by various ways according to the autonomy of the system and its capacities of observation and adaptation. By using the *"user profile"*, the system is able to select the right information and to adapt the to the user preferences. Thus, we can consider the concept of personalization of information like a process of definition, construction and use of the profiles, in order to answer the request (emitted by users of different profiles) in an effective way.

This way, we plan to define the *"user profile"* in this context. In the same way, thanks to the use of the *"user profile"* we will be able to give relevant answers to the user even when he is in the incapacity to specify his request in a fine way. We study the different typologies of users' knowledge. This study will make possible to build a model of knowledge to characterize some stereotypes. Then this model will be used to us to describe, build and index cases associated with stereotypes. The cases will represent the user's experiments. By accumulating the cases, we will be able to follow the evolution of the user's profiles, but also the tendencies of the behavior of a user group, or various stereotypes of the system. Currently, our team works on the integration of the user's profiles into the system.

6 CONCLUSION

In this paper we present an approach to find pertinent information to extract knowledge by using information retrieval tools in a digital library context. We propose to define a specific structure for the digital document during the creation step. According to this point of view, we have defined a semantic structure of the document by integrating new metadata in significant parts of the corpus. This makes possible to identify semantic segments of the scientific theses stored in our digital library: CITHER. In a search session based on keywords or concepts, the system will compare them with the semantic metadata (delimiting the semantic segments) and with the keywords describing the thesis. Thanks to this approach the user can get pertinent fragments of one or several theses.

REFERENCES

1. Abascal R., Rumpler B., Berisha-Bohé S. Proposition d'une nouvelle structure de document pour améliorer la recherche d'information. Proceedings of the CORIA'05 (COnférence en Recherche d'Infomations et Applications), ISBN: 2-9523810-0-3, IMAG, pp. 389-404, 2005.
2. Abascal R., Rumpler B., Pinon J.M., (2003) An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management. IRMA International Conference, Philadelphia Pennsylvania, USA. Ed. Mehdi Khosrow-Pour, IDEA Group Publishing, ISBN: 1-59140-097-X, pp. 201-204, May 18-21, 2003.
3. Abascal R., Rumpler B., Pinon J.M., (2004) Information Retrieval in Digital Theses Based on Natural Language Processing Tools, J.L. Vicedo et al. (Eds): España for Natural Language Processing (EsTAL'04), LNAI 3230, pp. 172-182, Springer-Verlag Berlin Heidelberg, October 2004, Alicante, Spain.
4. Abasolo J.M., M.Gomez M., An ontology-based agent for information retrieval in medicine ECDL 2000 Workshop on the Semantic Web.
5. Extensible Markup Language (XML) 1.0 - W3C Recommendation.
6. Heflin J., Hendler J., Searching the Web with SHOE. Artificial Intelligence for Web Search. In AAAI Workshop. WS-00-01. AAAI Press, Menlo Park, CA, 2000. pp. 35-40.
7. Heflin J., Hendler J., Semantic Interoperability on the Web. Proceedings of Extreme Markup Languages 2000. Graphic Communications Association, 2000. pp. 111-120.
8. The Semantic Web and its Languages Trends and Controversies November/December 2000.
9. Thomasson J-J. (2002) Schémas XML, Ed. Eyrolles, ISBN: 2-212-11195-9, November 2002, 466 p.