Defining User Profile to Improve Knowledge Extraction in a Digital Library of Scientific Documents

Rocío Abascal-Mena,[†] Béatrice Rumpler[††] and Suela Berisha-Bohé[††],

[†] Universidad Autónoma Metropolitana – Cuajimalpa, México
[††] LIRIS – Institute National des Sciences Apliquées – Lyon, France

Summary
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 defining different user's profiles.

Key words:
Annotation, Digital Library, User profile, Case-Based Reasoning.

Introduction

As the amount of online documents increases by leaps and bounds, the design of solutions to improve information retrieval has become of great interest. Two important aspects motivate our work. First, the documents need semantic annotations for a better selection. This way, the documents, in our case PhD theses, are annotated with pertinent concepts. The user annotates his thesis during the writing step by using a tool that helps to select these concepts [8]. These semantic annotations allow knowledge extraction from the theses and lead to an intelligent information processing. Second, to provide a personalized access to the knowledge, it becomes necessary to take into account the user's profiles during the search sessions.

The scientific library of Doc'INSA sets up since 1997 a project named CITHER [1], 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: (1) the use of an inadequate format, such as PDF (Portable Document Format), (2) the description of the contents by using only some keywords (title, name of the author, date, university, etc.) added outside the documents, and (3) the use of the tags proposed by the Dublin Core metadata which bring general information of the thesis.

In order to achieve pertinent information retrieval we have used a Natural Language Processing tool to extract pertinent concepts of our corpus of documents. The concepts are used as *"semantic tags"* for annotation. In Section 2, we present the process to generate adaptive annotation. We have also defined a new model of document based on Schema XML to generate enriched documents, suited to the logic and semantic structure of the thesis.

In Section 3, we show the importance of using a model to extract pertinent information. To satisfy the second requirement, we personalize the access and the search of information by defining user's profiles. In this way, we have created a model for the user. The way to adapt the user model to different users (who need help to build their request during a research session) allows the use of a personalized access to the information based on the user profile. Section 4 discusses and shows the importance of the definition of a system based on Case-Based Reasoning (CBR) to capture user's knowledge and preferences. This way, we are able to structure and to manage the user profile evolution. Finally, in Section 5, we decline conclusions and draw future work.

2. Semantic Annotations on Digital Theses

The Semantic Web aims to create contents that can be manipulated by humans but also by machines [5]. This can be achieved by explicitly adding markups to describe the content of a digital document. The annotation of existing digital documents is one of the basic barriers towards the conception of the Semantic Web. Manual annotation is impractical, while automatic annotation tools are still in their childhood. Hence advanced knowledge services may require tools able to search and extract the required

[1] http://docinsa.insa-lyon.fr/these/

Manuscript received July 5, 2006.
Manuscript revised. July 20, 2006

knowledge from the Web, guided by a domain conceptualization that specifies what type of knowledge is needed. In our approach, we propose to the author to describe his thesis with metadata characterizing the main content. Thus, we propose to build digital theses including *"semantic tags"*. These tags are going to allow the extraction of the most pertinent fragment(s) of the thesis (or the theses) related to the user's needs.

Our research is based on the use of a base of concepts (of the computer field) to build the user's requests and to organize the logic and semantic structure of the document. Once the base of concepts is defined, we propose the users to build a semantic structure for the documents by using a Natural Language Processing (NLP) tool. The semantic concepts are inserted in the document as *"semantic tags"*. Then, a user's query, based on these concepts or their synonyms, will generate a web access to a page that contains the pertinent information. The discovery of relevant contents is done by matching the user's query with the paragraphs surrounded by the pertinent concepts. Once the desired information is found, the user has only to read the pertinent fragments and so, he can select the best documents, in our case the PhD scientific theses.

While recent research efforts seek to add relevant markups to the content of the web pages [4], [6], [13], we go a step further by embedding the theses from their creation. This approach contributes to the recovery of pertinent information based on the use of semantic concepts.

We propose to use a NLP tool, called *"Nomino"*, to automatically extract concepts from a document [11]. We have selected Nomino after a comparative study of four NLP tools [10]. The user can also use Nomino to know and extract the most important concepts from a fragment of a thesis. So, it becomes not necessary to read the entire document to know if it is pertinent or not.

We have built a knowledge base with the concepts extracted from a corpus of theses (corpus composed by 25 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 extracted 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 very weak after the evaluation of about 25 theses of the same domain [9].

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

- The PhD student is assumed to write his thesis including *"semantic annotations"*. These annotations are generated in XML (eXtensible Markup Language) [3] format. To simplify the author task, a concept extraction tool can be called after the selection of a written fragment, section or chapter. The NLP tool, Nomino, proposes pertinent concepts and the user can accept or deny them for an insertion in the document.
- Another way to add concepts consists in selecting them from the base of concepts. In the base, the concepts are ordered by hierarchies according to the computer field.
- The system allows to the user the employment of new tags as *"semantic annotations"* without needing to know how to use XML.

We propose, also, the use of an ontology of the domain in order to help the author of the thesis to select concepts related to those already used [8], [9]. An ontology is a formal description of the concepts and relationships that can exist between these concepts. Our ontology has been constructed by using the concepts extracted from the theses that compose our corpus.

Also, our proposed annotation system involves the addition of a schema that defines the structure of the document (the thesis) [8]. In the next section, we show the importance of using this schema in order to validate well-structured documents.

3. Exploring Well-Structured Documents

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. The block of text is a fragment of text that appears in any part of the paragraph. In our proposition, a block of text, of undefined size, will become the finest logical entity considered. All these entities can be *"tagged"* or *"not tagged"*. The *"tagged"* state results from the presence of metadata (concepts) surrounding one or several paragraphs.

We use XML Schema [7] to create well-structured documents. By using the schema we validate the correct use of the metadata. Also, we validate the parts that are required to compose the logic structure of the thesis (title, name of the author, date, introduction, chapter, etc.). Thanks to the use of metadata, it becomes possible to extract pertinent information during a search session.

The thesis produced by the author (PhD student) is composed by metadata coming from the logic 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 is very powerful because it gives specific information about the fragment that contains associated concepts. Thanks to the metadata, the information retrieval tool is based in the exploration of the logic and semantic tags of each chapter (or part, or fragment or section of the document) is talking about a specific concept.

A digital thesis search tool is used to parse the theses and to 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 more 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 because the word *"Internet"* does not appears in the paragraph (in bold we have the concept *"Internet"* but by using a simple system we are not be able to search in the *"semantic tags"*). Instead, by using our system, even if the word *"Internet"* is not written in the paragraph, by using our XML tags the user is going to find this paragraph. This means, that by using the annotation system, the author of the thesis is able to use related concepts to clarify the different fragments of his thesis.

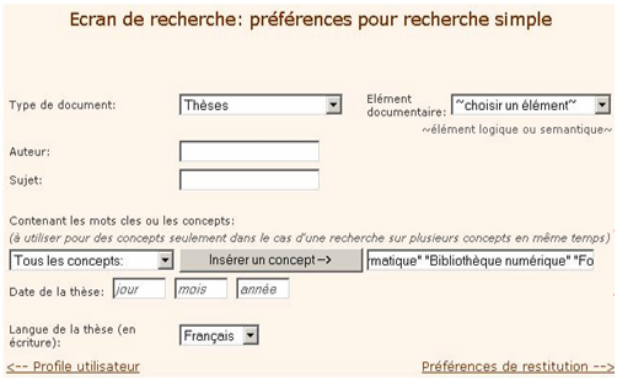
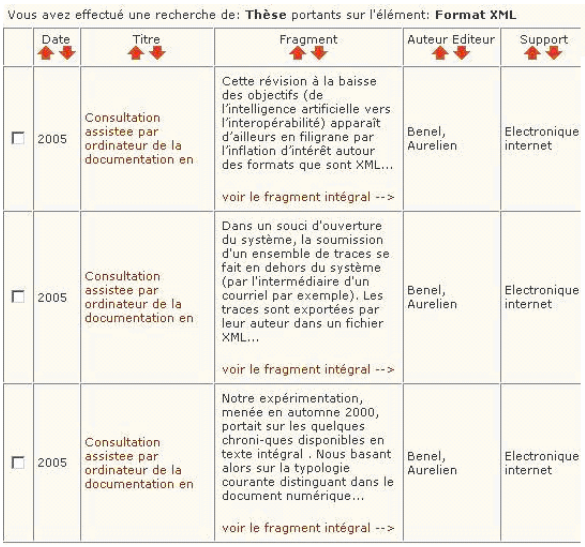


Fig. 1 Screenshot for a request using several concepts.

A typical user interaction with the search system consists in inserting a query composed by concepts (Figure 1). If there are more concepts (in the base of 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. Finally, once the user has selected the concepts, the system searches the right information in the thesis repository and then it shows the pertinent fragments to the user (Figure 2).


Vous avez effectué une recherche de: **Thèse** portants sur l'élément: **Format XML**

	Date	Titre	Fragment	Auteur Editeur	Support
☐	2005	Consultation assistee par ordinateur de la documentation en	Cette révision à la baisse des objectifs (de l'intelligence artificielle vers l'interopérabilité) apparaît d'ailleurs en filigrane par l'inflation d'intérêt autour des formats que sont XML... voir le fragment intégral -->	Benel, Aurelien	Electronique internet
☐	2005	Consultation assistee par ordinateur de la documentation en	Dans un souci d'ouverture du système, la soumission d'un ensemble de traces se fait en dehors du système (par l'intermédiaire d'un courriel par exemple). Les traces sont exportées par leur auteur dans un fichier XML... voir le fragment intégral -->	Benel, Aurelien	Electronique internet
☐	2005	Consultation assistee par ordinateur de la documentation en	Notre expérimentation, menée en automne 2000, portait sur les quelques chroni-ques disponibles en texte intégral . Nous basant alors sur la typologie courante distinguant dans le document numérique... voir le fragment intégral -->	Benel, Aurelien	Electronique internet

Fig. 2 Screenshot for the result of a search session.

The Figure 2 presents different fragments containing the concept *"format XML"*. In this figure we only present the fragments of one thesis but our system shows all the pertinent fragments found in all the theses. In this example, we have three fragments (from the same thesis), these are:

(1) *"Cette révision à la baisse des objectifs (d'intelligence artificielle vers l'interopérabilité) apparaît d'ailleurs en filigrane par l'inflation d'intérêt autour des formats que sont XML…"*,

(2) *"Dans un souci d'ouverture du système, la soumission d'un ensemble de traces se fait en dehors du système (par l'intermédiaire d'un courriel par exemple). Les traces sont exportées par leur auteur dans un fichier XML …"* and

(3) *"Notre expérimentation mené en automne 2000 portait sur les quelques chroniques disponibles en texte intégral. Nous basant alors sur la typologie courant distinguant dans les document numérique ..."*.

We can notice that in the first fragment, for the request *"format XML"*, we have underlined the concept

"formats" and the concept XML. For the second fragment we only have *"XML"*. Instead, for the third fragment we have neither *"format"* nor *"XML"*, but in this case the author of the thesis has considered that this fragment is related to *"format XML"* maybe because it talks about *"digital documents"* (in French: *"document numérique"*). This is why our system is so interesting. We can find fragments either if the concept doesn't physically appear. We are producing a semantic search, a more intelligent search based on *"semantic tags"*.

3. Defining the User Profile

The *"user profile"* takes into account the needs, the intentions and the cognitive, cultural or different specificities, that characterize the user. This way, 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. The *"user profile"* is composed of attributes, containing the user preferences associated to values.

The *"user's profile"* can be explicitly defined by the user or implicitly by the system [1], [2]. By using the *"user profile"*, the system is able to select the right information and to adapt it to the user preferences. Thus, we can consider the personalization of information like a process including the definition, construction and use of the profiles, in order to answer the request (sent by users of different profiles) in an effective way.

This way, we define the *"user profile"* in our context. In the same way, thanks to the use of the *"user profile"* we are able to give relevant answers to the user even when he doesn't precisely formulate his request.

We have defined the different typologies of user's knowledge [12]. This study has made possible to build a model of user's knowledge. Our model is based on the CBR (Case Based Reasoning) approach where the cases represent the user's experiments. By updating the cases, we are able to follow the user's profiles evolution, but also a user group behavior. When a user sends a request, first the system tries to find the closest case. If one case is found, the system uses this case to perform the search session; else a new case is created. We have defined *general* cases named *"stereotypes"* which are used at the starting of the application.

The integration of the user's profiles into the system allows a personalized access to the theses at the same time that it takes into account: the user's expertise of the system and the previous user's requests. For example, during a search session, the user expresses his needs for the specific search. These needs are related to his interests and his knowledge which are represented by keywords or documents selected by the user. These needs are deduced from his behavior within the system. However, it is difficult to anticipate all these characteristics in order to help the user and to bring him the necessary assistance in all the possible cases and contexts. Instead, for our study we have defined some categories of knowledge which appears like the most important. We have organized them in five groups: *"General Knowledge"* (GK), *"Knowledge of the Field"* (KF), *"Knowledge of the System"* (KS), *"Knowledge of the research"* (KR) and *"Knowledge of the restitution"* (KE) (Figure 3).

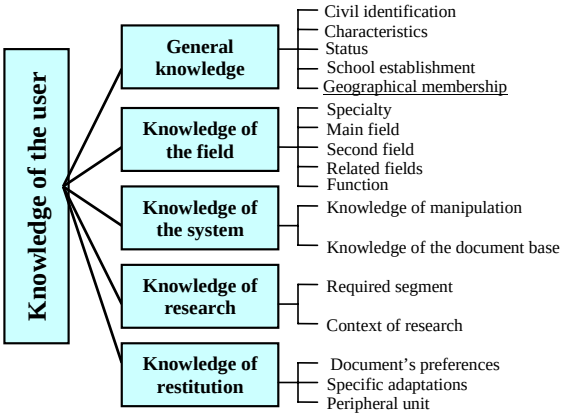


Fig. 3 The user model based on the knowledge.

In the next section, we are going to explain each of the five groups that compose the *"Knowledge of the user"*.

4.1 General Knowledge

The *"General knowledge"* is related to the next attributes: *"Civil identification"* of the user (civility, name, first name), with the *"Geographical membership"* (addresses, city, country), with *"Characteristics"* of the user like the handicap (visual, deafness, etc.) and with the socio-cultural membership (*"Status"*) of the user. Each attribute is defined by a value. For example, *"Status"* can have the next values:

- The *"maker"*, who is a user identified and recognized by the system. In this user group we find the students and the professors. Among the students, we find the *"writers of the theses"*. Then,

we find, the readers and correctors of the thesis, as well as the members of the jury. Some of these members form part of the group of professors.

- The *"librarian"*, who is also an identified user to the system and which has certain number of rights during the consultation of the digital theses.
- The *"administrator"*, also an identified user to the system, is charged of the management of the document's base.
- The *"invited user"*, who is not identified by the system and which, casually, makes a research.

The main characteristic of this knowledge lies in the fact that it remains practically immutable by the time.

4.2 Knowledge of the Field

The *"Knowledge of the field"* is composed by five attributes, which are: *"Specialty"*, *"Main field"*, *"Second field"*, *"Related fields"* and *"Function"*.

- The *"Function"*, means the paper or role that the user makes in the system. The mains roles are: *"professor"*, *"student"* and *"user of the system"*. We are interested in the different knowledge of the users. Specially, in the knowledge that has an important influence in the way that the user acts. This type of knowledge allows us the definition of stereotypes and the construction of actions plans [1], [2].
- The *"Main field"* concerns the user, such as for example *"computer science"*. We are especially concentrated on this field for the practical needs of our study because of the expertise that we have and that helps us to compare the behaviors of the system.
- The *"Second field"* corresponds to the specialties of teaching exempted to the future researchers.
- The *"Specialty"* corresponds to the group(s) of research in which the user is evolved.
- The *"Related fields"* correspond to the fields of application in which the user is interested.

These sublevels remain faithful to the real practices of the memberships and the groups of the users of the INSA of Lyon. Also they satisfy our present needs for the first experimental phase.

4.3 Knowledge of the System

A user, who knows well the system, defines, in theory, his needs in a better way. We estimate that a user will be more powerful at a session of research, if he knows the way in which the theses are stored in the base of documents by knowing the functionalities of the system.

By defining this knowledge, the idea is to determine in advance the needs of the user. This way, the system will have an idea of the session of research that it will offer to the user.

Once that the user knows the possibilities offered by the system he will be able to describe in a better way his needs of research.

4.4. Knowledge of the Research

The user must be able to make precise researches and the system must be able to give relevant answers. Two types of data will be combined to define the request of the user, the *"Context of research"* and the *"Required segment"*.

- The *"Context of research*" relates to the type of document treated and the way in which it will be treated.
- The *"Required segment"* will indicate the element to be treated. The part of the document required by the user during a precise research.

The type of research is a major element of the model of *"Knowledge of the research"*. For example, if the user proceeds to a research by topic, the system will start a research by fields and under-fields.

The precise research will be carried out by using concepts. This research is based on the logic and semantic structure of the documents [8].

Once the request has been sent, the user waits for the answers whose methods of restitution will depend on his preferences. This is why we have defined another group of knowledge, the *"Knowledge of restitution"*.

4.5. Knowledge of Restitution

This knowledge will contain the documentary preferences, the necessary adaptations for the handicapped people or the specific needs for the user, as well as the peripherals of restitution.

A certain number of elements were represented in this typology; however we are interested only in the

documentary preferences of restitution on a fixed computer of office.

5. Construction of the User Model

For the modeling of the user and the taking into account of the evolution of his profile, we chose to use the Case-Based Reasoning (CBR), which is one of the methods of resolution of problem by the machines within the framework of the Artificial Intelligence [12]. The base of this reasoning is the idea that an experiment is represented by a case. By making a comparison of the experiment (or case) in progress with the reports of the preceding experiments of the same user or others of users, the system must deduce the present needs for the user. This way, the system will be able to give the pertinent information to the user by taking into account his needs. Finally, we are not going to make a simple comparison between the attributes of each case, but a process of comparison, improvement and memorizing of experiments like it is done in the human training.

This way, a case corresponds to an experiment. Concretely, in our study, an experiment corresponds to the request emitted during a session of research.

The model of the user is formalized by a case represented by a list of attributes - values. We don't have listed all the attributes - values associated with all the knowledge quoted in the preceding paragraphs. Some of them (attributes - values) will be necessarily indicated by the user. Instead, others could be deduced according to the characteristics of the stereotypes. The goal of our work was to draw up a list, as far as possible, which represents in a significant way the characteristics and behaviors of almost all the population of users [12]. Based in the cases, the query of the user is ameliorated by taking into account his needs and his experience.

6. Integration of the User Profile in the Research System

From this typology of knowledge of the user (defined above), we will be interested in the aspects of formalization, construction and evolution of the user profile. This way, we can generate stereotypes that are used to retrieve the pertinent information by taking into account the user's needs.

By defining this knowledge, the idea is to determine in advance the user's needs. For example, a new user will not have the same attributes in *"Knowledge of the system"* as a user who already has written and sent a thesis to the digital library: CITHER. The PhD student knows that in this system he cannot consult all the theses supported during the current year, but only those that were diffused, whereas an occasional user ignores this detail.

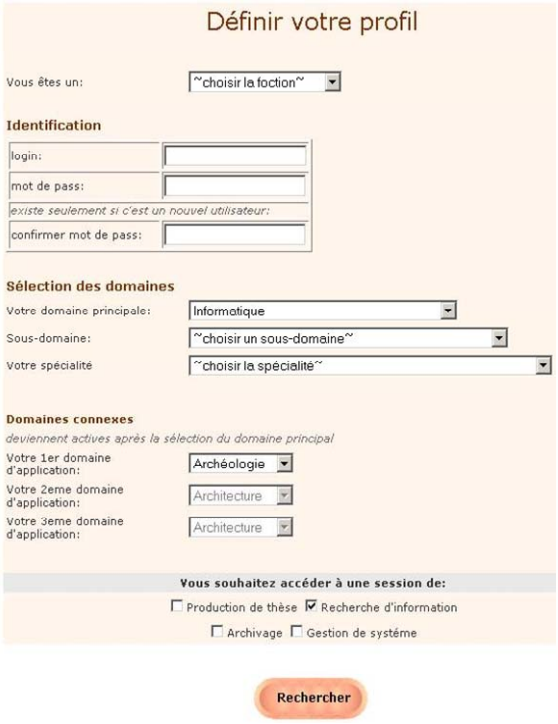


Fig. 4 Example of the definition of the user profile.

Currently, we are working on the integration of the *"user profile"* into the search system (Figure 4).

In the screen of search for this system we added a check box besides the heading *"According to the information of the preceding page"* (in French: *"En fonction des renseignements de la page précédente"*). Indeed, *"the preceding page"* relates to the screen of the Figure 4. This screen can be useful either during a drafting session of a thesis by selecting the box *"Production of the thesis"* (in French: *"Production de these"*) at the end of the page, or during a session of search by selecting the box *"Search for information"* (in French: *"Recherche d'information"*).

When the user specifies a field (domain) or a related field, and he wants to write his thesis, the system will be able to present the parts of an ontology of the domain (constructed for the computer domain) and our base of concepts which precisely are related to these specifications. Thus, the user is assisted in the task of choice of relevant metadata in order to insert them in the system.

If the user only asks for one research according to the data of the field, the system will bring closer the experiences of search of other users who have the same profile (they worked in the same fields) as the user. So these experiences will give more precise details on the requests concerning this profile.

If the user has used some keywords to carry out his research, the system will improve the request while carrying out the use of ontology.

Then, the request rewritten by the system will act on the base of theses available and marked out with *"semantic tags"* as indicated previously throughout this paper.

In the same way, the system will take into account, also, the preferences of the users such as: the language, the format of restitution, the number of results posted by page, etc. The Figure 5 shows an example of the screen that manages these preferences.

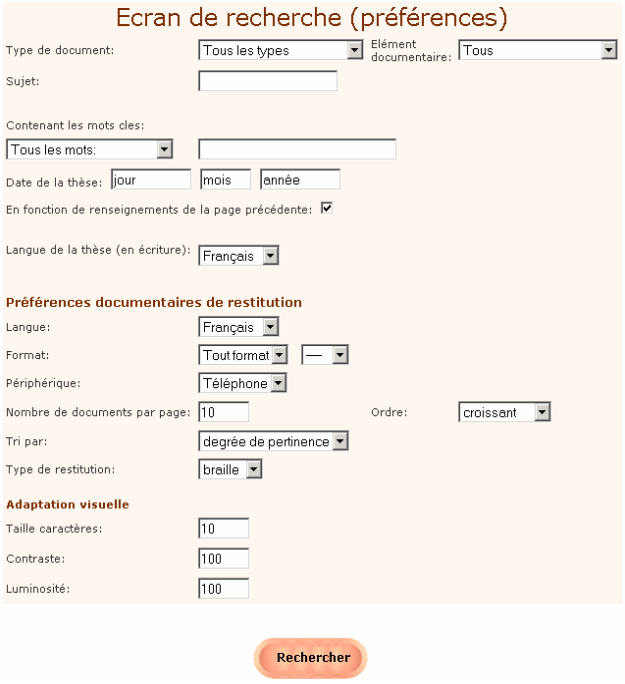


Figure 5. Example of the preferences that the user can select in a search session.

7. 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.

During our study we tried to deduce a certain number of stereotypes being based on our personal experiment from the use of the system CITHER. It is certain that at the present time it is impossible for us to connect all this knowledge in order to establish action plans for each stereotype. However, our system will be able to gain all this expertise lasting on the tests on a more important corpus and with a great number of users. From this expertise, it will be certainly possible to deduce other characteristics and other action plans to highlight new and more precise stereotypes.

In our work, we use also an ontology to complement the research step. During a research session, by using an ontology we are able to seek relevant information based on the semantic tags. The use of the ontology allows the definition of 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 use of 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.

Acknowledgments

This research was partially supported by the French Ministry of Research and New Technologies under the ACI program devoted to Data Masses (ACI-MD), project #MD-33.

References

[1] A. Kobsa and J. Fink. *"Performance Evaluation of User Modeling Servers Under Real World Workload Conditions"*. In 9th International Conference on User Modeling, 2003.

[2] E. Rich. *"Stereotypes and User Modeling"*. In: A. Kobsa, and W. Wahlster (eds.), User Models in Dialog Systems. Springer, Berlin, Heidelberg, 1989, pp. 35-51.

[3] Extensible Markup Language (XML) 1.0 - W3C Recommendation.

[4] J. Heflin and J. Hendler. "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.

[5] J. Heflin and J. Hendler. "Semantic Interoperability on the Web", in *Proceedings of Extreme Markup Languages 2000*. Graphic Communications Association, 2000. pp. 111-120.

[6] J.M. Abasolo and M. Gomez. "An ontology-based agent for information retrieval in medicine", in *ECDL 2000 Workshop on the Semantic Web*.

[7] J-J. Thomasson. *"Schémas XML"*, Ed. Eyrolles, ISBN: 2-212-11195-9, November 2002, 466 p.

[8] R. Abascal and B. Rumpler. "Using Embedded Semantic Tags to Explore Digital Libraries", in *International Journal Transaction on Computer Science and Engineering*. Vol. 9, No. 1, ISSN: 1738-6436, 2005, pp. 27-38.

[9] R. Abascal, B. Rumpler and J-M. Pinon. "An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management", in *IRMA International Conference*, Philadelphia Pennsylvania, USA. Ed. Mehdi Khosrow-Pour, IDEA Group Publishing, ISBN: 1-59140-097-X, 2003, pp. 201-204.

[10] R. Abascal, B. Rumpler and S. Berisha-Bohé S. "Proposition d'une nouvelle structure de document pour améliorer la recherche d'information", in *Proceedings of the CORIA'05 (COnférence en Recherche d'Infomations et Applications)*, ISBN: 2-9523810-0-3, IMAG, 2005, pp. 389-404.

[11] R. Abascal, B. Rumpler B and J-M. Pinon. "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, Springer-Verlag Berlin Heidelberg, 2004, Alicante, Spain, pp. 172-182.

[12] S. Berisha-Bohé, B. Rumpler and R. Abascal. "A Semantic Structure to Improve Information Retrieval Using XML", *9th ICCC International Conference on Electronic Publishing – Elpub'05*, 2005, pp. 319-321.

[13] The Semantic Web and its Languages Trends and Controversies November/December 2000.