A View on Two Complementary Representations of Documents for Information Retrieval

Béatrice Rumpler, Hassan Nadery

INSA Lyon – LIRIS
7 Avenue J. Capelle Bâtiment Blaise Pascal
F69621 Villeurbanne cedex, France
email: beatrice.rumpler@insa-lyon.fr

Abstract

The indexation of documents is a critical step of the information retrieval process and is often a manual task which highly depends on the indexer's knowledge. We propose to improve the manual indexation of documents by use of a semi-automatic semantic annotation process.

Keywords: indexation; information retrieval; semantic annotation; natural language processing

Introduction

Information retrieval in textual documents is usually based on a process of indexation. Indexation consists in a representation of the document with a list of keywords. These keywords constitute an index by which it is possible to find a document; they are a representation of the concepts connected to the document. The choice of these keywords has a strong impact on the Information Retrieval process and particularly affects the pertinence of the selected documents during a search session. The Indexation Process (IP) takes place once the document is written and just before the documents are stored in the servers for diffusion. Indexation is done manually or partly automatically. Manual indexation is time-consuming and even if it gives a better representation of the document than an automatic indexation, it is not generally performed by the author of the document. It is possible for some important concepts of a document not to be represented in the index. This is why we propose a complementary solution to the indexation process, which is based on semantic annotations of the documents.

Our Solution to Improve Information Retrieval

In our proposal, we allow the author of the document to annotate his document during the writing stage. These annotations, chosen by the author himself, correspond to the most significant concepts of the document. After these semantic annotations have been inserted in the document, it is always possible to index the final document. The semantic annotations can also be used as metadata to describe the document and also to complete a manual index. So an information search session can be done using the index and the semantic annotations. The index permits selecting the most pertinent documents and the semantic annotations allow extracting the most pertinent fragments from the selected documents. In this paper we don't describe the indexation process, this technique being largely detailed in the literature and implemented in all Information Retrieval Systems [1]. We present our solution based on semantic annotations for a corpus of scientific documents (scientific theses and publications).

First, we have defined a base of concepts of the Data Processing field. Our corpus is composed of scientific documents of the Data Processing field. Our approach is based on the use of Natural Language Processing (NLP) tools to extract concepts from documents. After an evaluation of several tools we decided to choose Nomino because of its performances about the recall and precision rates [2]. In the first step, we collected all the concepts extracted by Nomino from our corpus of documents. Then we performed a manual analysis of these concepts to select the most significant ones, and finally we made a classification by topics of the selected concepts. Thanks to a hierarchical organization of the concepts we have built a base of concepts of the Data Processing field [3] [4]. Our corpus is composed of documents written in Microsoft Office Word 2003. Microsoft Office Word 2003 permits using XML reference schemata as "WordprocessingML". Thanks to these schemata it is possible to store the documents in XML format [5]. We have also defined an XML Schema to formalize the documents' structure.

The Annotation Tool

Once the base of concepts and the XML schema were defined, we built our annotation tool. In order to facilitate the author's task, we have decided to include our annotation tool's specific commands in the Microsoft Word software (like Microsoft Word commands). There is no need for the author to learn a new software tool; he/she can easily insert the concepts in his document whilst writing. To insert annotations (also called semantic tags or metadata) in specific parts of the document, the author must select a fragment of document, and our tool offers three possibilities:

1. Insertion of tags chosen by the author himself
2. Insertion of tags proposed by our base of concepts
3. Insertion of tags extracted by Nomino from the fragment of document selected by the author. Nomino is, in this case, used to suggest concepts tied to the current document. The author can select or reject them.

In this part we present an example of a fragment of a document containing semantic annotations (semantic tags) such as:

> <[**user profile**]*Thus, users' information retrieval experiences or instances are saved to be reused in future similar cases. The resulting cooperative memory is utilized for user query expansion. In order to improve the information retrieval experience, we propose in this paper to conceptualize and model both the user profile, and the information retrieval process. This leads us to define some similarity functions between user profiles and information retrieval situations.*[/**user profile**]>

where "[***user profile***]" represents a semantic tag inserted in this fragment of document.

Conclusion

During a search session, the Information Search System can first select the most pertinent documents of the corpus by using the index and in a second step it can extract the most pertinent fragments of the selected documents. The possibility of improving the index with the semantic annotations now exists thanks to our tool.

References

[1] Van RIJSBERGEN, C. J. *Information Retrieval*. 2nd ed., Butterworth, London, UK, 1979.

[2] Nomino Technologies, Document de travail, Copyright Nomino Technologies, 2002. www.nominotechnologies.com

[3] ABASCAL, R.; RUMPLER, B.; PINON, J. M. Information Retrieval in Digital Theses Based on Natural Language Processing Tools. *EsTAL'04, October 2004, Alicante, Spain*. Edit. by J. L. Vicedo et al. *LNAI,* Berlin Heidelberg : Springer-Verlag, 2004, vol. 3230, p. 172-182.

[4] ABASCAL, R.; RUMPLER, B.; PINON, J. M. An Analysis of Tools for an Automatic Extraction of Concept in Documents for a Better Knowledge Management. *IRMA International Conference, May 18-21, 2003, Philadelphia Pennsylvania, USA* Ed. Mehdi Khosrow-Pour. IDEA Group Publishing, 2003, p. 201-204. ISBN 1-59140-097-X.

[5] http://www.xml.org