Semantic cartography: towards helping experts in their indexation task

Eric Kergosien, Marie-Noëlle Bessagnet, Mauro Gaio

Semantic cartography: towards helping experts in their indexation task

Eric KERGOSIEN, Marie-Noelle BESSAGNET, Mauro GAIO
UPPA, Laboratory LIUPPA, 64000 PAU
{eric.kergosien, marie-noelle.bessagnet, mauro.gaio} @univ-pau.fr
http://liuppa.univ-pau.fr

Abstract: Our approach aims at helping experts in their indexation work using the relationship between concepts included in descriptive notices defined by using an external semantic structure (taxonomy, thesaurus, etc).

Keywords: Knowledge Engineering, Thesaurus, Knowledge Representation

1 Introduction

Thanks to the experts' indexation efforts, documents have rich descriptions, notably descriptive notices, described on a well-known semantic structure-base (taxonomy, thesaurus...). We propose to the experts an approach which would make validating their indexation work possible. This approach has two steps: (i) describe information thanks to expert knowledge which is automatically extracted from notices; (ii) give the possibility to navigate within the collection via the identified knowledge representing the indexation work. In a first part (&2), we expose our objectives. Thus, we can develop our approach (&3) to design a specific semantic structure within the PIV[1] project. We describe our experimentations for exploration in a corpus indexed by librarians using RAMEAU[2]. We make propositions (&4) to help the validation of the indexation work within the corpus based on the enriched thesaurus.

2 Objectives

In the aim to propose to experts a validation tool for the use of controlled vocabularies (thesaurus, etc) which they apply to fit their documents analyze, we develop in our approach two preliminary steps. The first step allows extracting and structuring knowledge of our corpus made up of descriptive notices and the controlled vocabulary used to describe these notices. In this step, we connect to research work such as [1] who attaches to mix terms from a thesaurus and terms from other sources in order to better point information retrieval within management system. One of our objectives is the design of a process to transform a classic controlled vocabulary in a knowledge base [2]. The second step proposes a map's representation of this structure, a concepts map [3]. In the end, we use techniques of semantic cartography [4] to tackle in a synthetic way the complete indexation work of a given collection.

---

[1] PIV : Pyrénnées Itinéraires Virtuels

[2] RAMEAU : http://rameau.bnf.fr/ is a french thesaurus defined by the Bibliothèque Nationale de France (BNF) for or the majority of the French libraries

3 The domain-specific thesaurus PIV and a visualization tool

The aim is to represent information thanks to expert knowledge automatically extracted from notices. Indeed, we suppose that we can propose a description of expert knowledge extracting a list of terms (root-words) used in notices by experts to describe documents and the controlled vocabulary used to choose these root-words (RAMEAU for us). According to the AFNOR[3], a thesaurus is a documentary language based on a **hierarchical structuring** for one or more knowledge domains; notions are represented by terms from one or more natural language and relationships between notions by conventional signs. The first step automatically identifies and extracts « root-words » of the controlled vocabulary (named "subject authorities" in RAMEAU) used to describe the content of the document within XML descriptive notices. These root-words are selected by librarians within RAMEAU and used in notices via predefined tags (figure 1).

<**DEE**>Stations climatiques -- Eaux minérales -- Barèges (Hautes-Pyrénées) -- 18e s. </**DEE**>

Fig. 1 – Extract of descriptive notice 1

Each tag DEE contains several root-words separated by element « -- ». For each term found in a notice, we attach within the XML Topics Map structure [5] a link to the document. Authorities represent the conceptual level and document the physical level. This list of root-words is a first step towards the definition of a thesaurus depicting part of librarians' knowledge on the collection; it remains to identify set of relations between these terms in order to develop the sub-thesaurus PIV. By exploiting the thesaurus structure (RAMEAU for the example), we automatically improve the above vocabulary with: « generic », « used for » and « related » (for thesaurus) terms; relations which are linked. The above process gives us an improved organization of knowledge i.e. Domain-specific Thesaurus.

Then, we propose a cartographic representation allowing experts to tackle in a synthetic way the complete indexation work of a given collection, realized by different librarians. Up until now this had been very difficult to represent due to the large number of documents and associated descriptive terms.We propose an interface which proposes a global view (i) and a local view (ii). The global view allows librarians to apprehend the expert knowledge structured in the thesaurus in an integral way; the local view, based-on concepts map representation which highlights a subset of the semantic structure [6]. We explicitly represent the existing relations between the terms of the thesaurus (conceptual level) and the documents of the collection related to these terms (physical level). The combination of both views facilitates navigation and thus re-reading of the indexation work completed on the collection.

4 Suggestions for assistance in the validation of indexation work

The first feedback from experts of the media library is encouraging, showing the possibility to propose a synthetic display of their indexation work.

3 AFNOR is a French institution for standardization. Documentation : règles d'établissement des thésaurus monolingues. NF Z47-100, 1981.

Following the automatic thesaurus creation, an automatic control process begin in order to make sure that the terms used in the descriptive notices indexing the collection, are well labelled as root-words in the given controlled vocabulary. This process identifies 4 types of errors due to the usage of the controlled vocabulary: *(i) Documents without descriptive notices; (ii) Incorrect information on authorities; (iii) Management of non-selected terms; (iv) Management of homonymy.* Concerning the cases linked to name errors and missing words, certain solutions are proposed, by choosing a correct term which is listed. It is possible to index the documents by using terms that are not part of the semantic structure used. If we take the example of a photo of *la Place Royale à Pau*, the expert can choice the term *Place Royale (Pau)*, a specific term which the committee in charge of updating RAMEAU did not choose to include in the thesaurus. Concerning errors linked to homonymy, we limit correction assistance by suggesting, a list of terms containing the term, which reports the error.

5 Conclusion

In this article, we have presented our research work on the structuring and adaptation of a controlled vocabulary as an indexing tool. We worked on a collection which includes 750 files and associated descriptive notices. We validated our first experiments with experts of the domain, i.e. librarians. This feedback encourages us to predict new tools permitting to validate the content of descriptive notices. The defined thesaurus representing this indexation work offers a first step for expert users to navigate in the collection through their own representation. Our structure, including a verification phase of the used terms allows in a second step to offer tools for correcting any errors, thereby facilitating the validation of descriptive notices. Our current work focuses on an approach that aims to define an ontology based on a thesaurus. We hope to integrate new knowledge characterizing the territory (adding "localized named entities" and links between concepts). The aim is to offer an interface which permits all types of user, wanting to discover a territory described by documents, to navigate through the collection thanks to domain ontology.

6 References

[1] Schatz B. R., Johnson E. H., Cochrane P. A. and Chen H. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In Proceedings of the 1st ACM Digital Library Conference, pp 126–133, Bethesda, US, 1996

[2] Elghoul M., Méthodologie de conception d'un SIAD pour la gestion documentaire, aide à l'indexation, aide à la construction du thésaurus, aide à la recherché et aide à l'apprentissage. PhD – Paris : Université de Paris IX, 1990.

[3] Card S.K., Mackinlay J.D. et Shneiderman B., 'Information visualization', Readings in information visualization: using vision to think, Morgan Kaufmann Publishers Inc., pp. 1-34., 1999

[4] Kohonen, T., Self-Organizing Maps. Berlin, Heidelberg, Springer, 2006

[5] Pepper S., Moore G., "XML Topic Maps (XTM) 1.0 Specification", TopicMaps.Org, Aug. 2001. Available at http://www.topicmaps.org/XTM/1.0.

[6] Ziti M., Baudoin-Lafon M., Hypermedia Exploration with Interactive Dynamic Maps, International Journal of Human Computers Studies, 43: pp 441-464, 1995.