Hypatia - Towards a Support Expert System for Documentation Departments

Angel L. Garrido, Sergio Ilarri, Eduardo Mena
IIS Department
University of Zaragoza
Zaragoza, Spain
Email: {***garrido, silarri, emena***}***@unizar.es***

Today, the vast amount of information stored in public and private organizations require different approaches and new tools to be able to manage adequately. While the popularity of indexers, search engines and document management software is widespread today, these tools still suffer several deficiencies that force documentation departments to make significant manual work to classify, store, and retrieve information. Moreover, if we analyze the activities performed by documentation departments of these organizations, we find out that the production of reports and dossiers is one of theirs most important missions. Actually, for this task there is no useful computer systems able to exploit the latest advances in natural language processing, semantics and ontologies [1]. Therefore, our proposal is to study the different modules that could form an expert system called ”Hypatia” to do these tasks.

The processes of classification of documents are the most time-consuming tasks to people within documentation departments. Any professional of these departments used to classify documents by using a list of terms, a thesaurus [2] or sometimes even ontologies. Despite the existence of many standard thesauri and ontologies in different disciplines, it remains usual to find that each department documentation performed an adaptation of any of them, or even their own developed from scratch. The existence of these custom thesauri prevents the application of standard software for automatically classifying documents. Moreover, the contextual information of each organization is one of the most important guidelines when categorizing the texts, which makes it even more difficult the application of standard tagging techniques. To deal with this first challenge the idea consists of applying lemmatization of the documents, obtaining keywords by statistical techniques, and finally using a combination of Support Vector Machines [3], ontologies and a set of rules to deduce appropriate tags for the annotation [4].

Also, from a documentary point of view, an important aspect when we are conducting a rigorous labeling is to consider the geographic locations related to each document. Although there exist tools and geographic databases, it is not easy to find an automated labeling system for multilingual texts specialized in this type of recognition and further adapted to a particular context. So, we propose a method that combines geographic location methods using a Gazetteer like Geonames[1] with Natural Language Processing (NLP), statistical techniques, and semantic disambiguation tools to perform an appropriate labeling. The method can be fine-tuned for a given context in order to optimize the results [5].

Sometimes the problem itself is the development of a thesaurus or ontology that serves to classify documents, since its definition can be very time-consuming. Here our proposal is to conduct a statistical, linguistic and semantic analysis over the documents, so that we obtain entities and relationships from which to build a topic map [6] and finally an ontology [7].

The more complex process, which is where we are currently working in, is to obtain custom dossiers on a particular topic. For this, the idea is to combine the techniques discussed above with the use of external resources located on the Web such as DBpedia[2], Freebase[3] or OpenCyc[4] and also take advantage of bringing the techniques of creating automatic summaries.

Furthermore, the enrichment of databases queries incorporating linguistic and semantic tools is another important line of work within our project. The idea is to take adventage from traditional relational data stores to obtain additional information, by enriching the standard queries that users launch agaisnt these kind of databases. This enriching will be done by using lexical databases and disambiguation technincs.

We have carried out a detailed evaluation of our methods by comparing our results with the work of a real documentation department in several media, obtaining really promising results [8].

ACKNOWLEDGMENT

This research work has been supported by the CICYT project TIN2010-21387-C02-02 and DGA-FSE.

[1] http://www.geonames.org/
[2] http://dbpedia.org/
[3] http://www.freebase.com/
[4] http://www.cyc.com/platform/opencyc

REFERENCES

[1] T. R. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[2] A. Gilchrist, "Thesauri, taxonomies and ontologies - an etymological note," *Journal of Documentation*, vol. 59, no. 1, pp. 7–18, 2003.

[3] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Tenth European Conference on Machine Learning (ECML'98)*, pp. 137–142, Springer, 1998.

[4] A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "Nass: News Annotation Semantic System," in *23rd International Conference on Tools with Artificial Intelligence (ICTAI 2011)*, pp. 904–905, IEEE, 2011.

[5] A. L. Garrido, M. G. Buey, S. Ilarri, and E. Mena, "GEONASS: A semantic tagging experience from geographical data on the media," in *17th East-European Conference on Advances in Databases and Information Systems (ADBIS 2013), Genoa (Italy)*, vol. 8133, pp. 56–69, Springer, 2013.

[6] S. Pepper and G. Moore, "XML Topic Maps (XTM) 1.0 - TopicMaps. org specification," *TopicMaps. Org Authoring Group, http://www.topicmaps.org/xtm*, 2001.

[7] A. L. Garrido, M. G. Buey, S. Escudero, S. Ilarri, and E. M. anda Sara B. Silveira, "TM-gen - a topic map generator from text documents," in *25th International Conference on Tools with Artificial Intelligence (ICTAI 2013)*, IEEE, 2013.

[8] A. L. Garrido, O. Gomez, S. Ilarri, and E. Mena, "An experience developing a semantic annotation system in a media group," *Natural Language Processing and Information Systems*, pp. 333–338, 2012.