TOWARDS AN ONTOLOGY BASED SEARCH MECHANISM FOR THE EDMS AT CERN*

Antonio Jose Jimeno Yepes†, Bertrand Rousseau‡,
CERN, Geneva, Switzerland

Abstract

CERN is building its new accelerator, the LHC. Its lifecycle dataflow is stored in the EDMS system. Due to the size of the collection and the diversity of people, organizations and divisions it is difficult to find documents without prior domain knowledge. To overcome this problem, an approach based on a hand-crafted domain specific ontology has been tested in order to improve information retrieval for the LHC Equipment Catalog. Experiments have shown that the use of an ontology improves average precision.

INTRODUCTION

CERN[1] (European Organization for Nuclear Research) is building its new accelerator, the LHC[2] (Large Hadron Collider). The LHC lifecycle dataflow is stored in the EDMS[3] (Engineering Data Management System). This dataflow involves the design, manufacturing, testing, installation, integration, maintenance and dismantling steps.

The EDMS is based on proprietary applications and databases that store: the documents, the documents' metadata and the information needed in each step. A search mechanism is provided for the documentation system but the current results are not yet satisfactory and some strategies are being tested. Due to the size of the system and conventions coming from the diversity of people, organizations and divisions it becomes very difficult for a normal user to get the right document without prior knowledge of the domain. The solution proposed in this article is to use a hand-made ontology that collects the domain knowledge to refine the user query.

DEVELOPMENT

Information Retrieval

An EDMS document object deals with a specific project or item, it contains meta-data (like title, description, keywords, authors, approval list, version, document status). There are many types of document files (drawings, PDFs, MS Word) only PDF and MS Word files are considered. There are two official languages, English and French, only English documents are considered. We have been using the very well known SMART[5] system that works on the bag-of-words model. In this model each document is represented by a vector that lives in a high-dimensional space where each dimension represents a word and the value in each dimension indicates the relevance of the word for that document. Computing the relevance of the query against a document is done usually by calculating the angle between the query and the document in the high-dimensional space, being more relevant the document with smaller angle. Query Expansion is a technique that has been used for improving an inadequate or incomplete user query[7]; this can be supported by an ontology.

Subsection 2.2 gives a brief description of the ontology we built concerning the LHC and subsection 2.3 discusses the query expansion mechanism used.

Ontology

An ontology gives a formal description of the concepts and their relations for a given domain. There are generic ontologies like Wordnet[9] or more specific ones like UMLS[6]. Such ontologies are of no utility to us because LHC concepts are very specific. As the LHC lifecycle is very complex, we have limited the scope of the ontology to the equipment types in the main tunnel, their possible locations, and different operations like testing and installation. The concepts in the ontology are in a taxonomy of concepts, contain all the possible acronyms and synonyms and the relations between concepts (like made_of, connected_to, located_at). The ontology has been prepared with the support of experts on the domain of the LHC project, mainly engineers.

Ontology Based Query Expansion

Previous work on ontology-based query expansion has been based mostly on generic ontologies like Wordnet. Voorhees[10] used it in combination with a TREC[4] collection; she found some improvement with short queries but some degradation in longer ones. Mandala[8] obtained better results refining Wordnet with some learned relations form the corpus.

In this experiment we tried an approach based on Voorhees' manual query expansion, replacing Wordnet with our ontology, that looks more like Mandala's one. Based on a query, a set of concepts are chosen from the ontology. For each of these concepts, the synonyms, hypernyms, hyponyms and related concepts are extracted. At the moment, only one level of the concept taxonomy is searched. Five bags-of-words are built, corresponding respectively to the query terms, synonyms, hypernyms, hyponyms and related concepts. These are fed into the

*I would like to thank you Christian Pellegrini and Melanie Hilario from the University of Geneva and Roberto Saban and Samy Chemli from CERN, for the opportunity and the means to perform this study

† antonio.jimeno@cern.ch

‡ bertrand.rousseau@cern.ch

SMART system which returns a ranked list of documents for each. These five lists are combined using document scores computed by SMART and weights assigned to each bag-of-words by the system designer. The weights are used to compare the contribution of each one of the five bag-of-words.

EVALUATION

Introduction

The Cranfield paradigm is used to compare the performances of the configurations tested. This paradigm needs a benchmark made of: a fixed set of documents, a fixed set of queries and a set of relevant judgements for the documents and the queries. The measures used are standard ones in information retrieval: precision (number of relevant documents retrieved against the number of retrieved documents) and recall (number of relevant documents retrieved against the number of relevant documents in the collection). In addition, 11th point average precision, that measures precision at different levels of recall, is used because the information retrieval systems use to retrieve long lists of documents. It can measure the position of relevant documents against the non-relevant ones; a higher value in 11th point average precision means that the relevant documents are closer to the first positions of the list of retrieved documents.

Benchmark Preparation

We needed to build our own benchmark because we do not use a standard collection. As said before we need three components:

1. Set of documents. We considered only English documents and MS Word and PDF documents. Due to some problems with Oracle Intermedia Text we had to work with a subcollection of documents so, based on the EDMS structure, we extracted the documents talking about the domain of the ontology. As afterwards we used SMART, we could work as well with the whole collection so we could have a collection of around 20.000 documents and a subcollection of 3.000 documents.

2. Set of queries. We extracted the queries from the log table of the current system. The queries that contained references to meta-data about the documents (interested on a given document id, ...) were discarded. Their length is in a range between 2 words and 6 words, being small. It was not possible to collect the person that posed the query because it used to be the guest user.

3. Set of relevant judgements. It was not possible to obtain the judgements from the log table and we could not find the user that posed the query. The experts on the domain of the query were interviewed. To make it easier, lists of candidate documents were prepared so the expert only made binary judgements (relevant or not-relevant) on the documents in the lists and added the documents that were not there and the expert considered as relevant. The configuration used with SMART: we have used stemming, several weighting schemes (statistical calculation of the relevance of a word for a given document and in the collection and on a normalization factor to compensate the differences among the documents).

Table 1: Results with the subcollection

W. Scheme	Baseline	QE	Change
ntc	0.25	0.37	%47
lnx	0.32	0.47	%49

Table 2: Results with all the documents

W. Scheme	Baseline	QE	Change
ntc	0.19	0.30	%60
lnx	0.23	0.40	%73

Results

The results were based on the two collections, the SMART configuration and the expansion being applied. We observed that the recall was almost perfect without query expansion and with query expansion we retrieved all the relevant documents. Concerning precision, as soon as we added terms to the query we retrieved more documents than before. As soon as we could not retrieve much more relevant documents from the collections, the results for the precision are worst with query expansion, we have to highlight that no threshold has been used.

On table 1 and table 2 the results for 11th point average precision for the two collections of documents is shown. On figure 1 and figure 2 the precision at different levels of recall are shown. For each collection we compare the results using ontology query expansion against not using it. We observed (not shown in the article) that the weights used in the ontology query expansion indicate that the most relevant bag-of-words added to the query are coming from the the hyponyms and the related concepts, as shown already in the literature[8].

For each collection two weighting schemes are compared: cosine normalization (ntc) and pivoted cosine normalization (lnx). It can be observed that the results are improved independently of the weighting scheme. Comparing the two collections, the collection containing all the documents obtained higher improvement on 11th average precision because the documents from the subcollection are closer semantically.

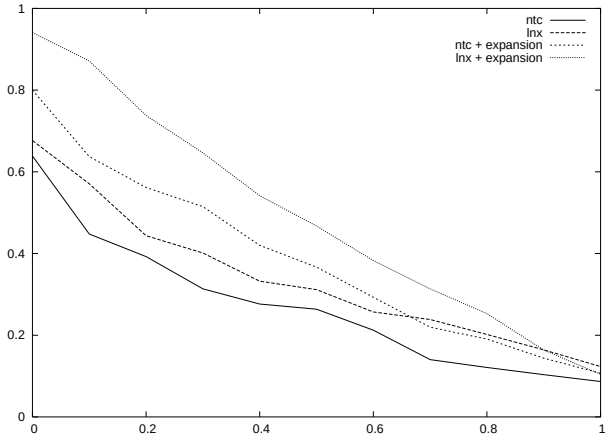

Figure 1: 11th point average precision subcollection

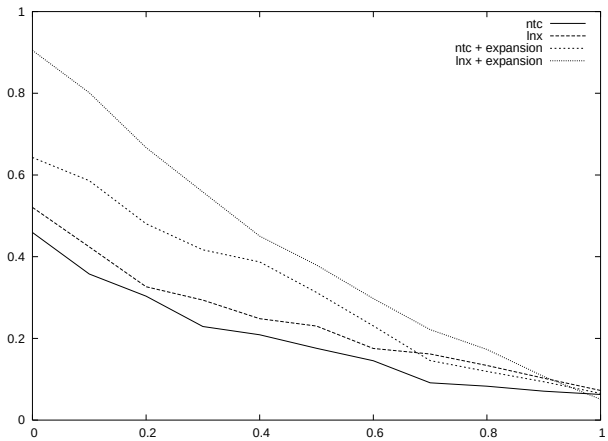

Figure 2: 11th point average precision all the documents

CONCLUSIONS

Current results suggest that ontology based query expansion improves document ranking, at least for this configuration. The improvement can be justified because the expanded query carries a more expressive content about the user need, mainly because the original user query is short. On the figures 1 and 2 it can be seen that higher precision is obtained at low recall but as we retrieve more relevant documents the curves for expanded and non-expanded queries get closer and even cross when recall is next to 1. It means that the terms that are contained in the ontology and used for the expansion do not perform the same on all the relevant documents, so further research on the relation between the expansion and the documents is suggested to improve the ranking for all the relevant documents.

The problem of the method is that, since recall was already near perfect before query expansion, adding more terms cannot retrieve more relevant documents but on the contrary only harms precision. The goal is then to try to improve precision without hurting recall.

FUTURE WORK

Based on the results from the experiments there are several possibilities for solving the current precision problem.

One solution could be to limit the number of documents retrieved by setting a threshold either on document scores or on the number of documents to be retrieved, since no training data is available a good approach is to look at the distribution of the retrieved documents.

A second solution would consist in refining the ontology based on an analysis of its impact on document retrieval, the main problem is that the bag-of-words model do not consider relations between the words being difficult to make the distinction between relevant and non-relevant documents due to the compound words, that could be a reason for the low precision (asking for the *separation dipoles* the system retrieves documents about *main bending dipoles*) Methods for named entity recognition may be used to overcome this situation. In addition, this is a time-consuming and labor-intensive task;so a more interesting approach would be to look for automated or semi-automated ontology refinement.

A third solution would be to use a different model from the bag-of-words that could gather the particularities of the domain terminology, like the compound words, and/or more elaborated expressions like syntactic relations between the different entities.

Once the document search cannot be improved will be the time to expand the search over other EDMS object types (projects, items, ...), so the ontology can be like a central point where the spars ed databases are gathered together with a common access point. Depending on the relevance of the French language, the ontology could be translated in order to perform cross-lingual retrieval.

REFERENCES

[1] http://www.cern.ch

[2] http://cern.ch/lhc-new-homepage

[3] http://edms.cern.ch

[4] http://trec.nist.gov

[5] ftp://ftp.cs.cornell.edu/pub/smart/SMART

[6] http://www.nlm.nih.gov/research/umls/umlsmain.html

[7] Efthimiadis, Efthimis, Query Expansion. In Williams, Martha E. (Ed.), Annual Review of Information Systems and Technologies (ARIST), v31, pp 121-187.

[8] Mandala, Rila, Takenobu, Tokanuga, Hozumi, Tanaka, The Use of WordNet in Information Retrieval,Use of WordNet in Natural Language Processing Systems: Proceedings of the Conference,31–37,1998

[9] G. Miller. WordNet: A Lexical Database for English. Communications of the ACM, vol. 38, no. 11, Nov, 1995.

[10] E.M. Voorhees. Query expansion using lexical-semantic relations. In proceedings of the 17th ACM-SIGIR Conference, pp. 61-69, 1994.