Documents Content Indexing for Supporting Knowledge Acquisition Tools

Marek Ciglan

Institute of Informatics, Slovak Academy of Sciences
Dúbravská cesta 9
845 07 Bratislava, Slovakia
marek.ciglan@savba.sk

Abstract. In this article we describe several aspects of the documents content indexing system developed in the context of project NAZOU to support knowledge acquisition tools. We compare two methods of the word lemmatization, process of identifying the base form of the words; one based on the Porter algorithm, the other using dictionary created by human experts and discuss the advantages and disadvantages of both approaches. The tool for document indexing keeps detail information about the document content, including the positions of the words in the documents. On a case study, we show how this aspect helps to support tools for automatic semantic annotations.

1 Introduction

Information retrieval is a process of identifying the text resources of interest from a large collection of documents, that would satisfy the user needs. Full text search is a wildly used concept in todays information systems for information retrieval. Full text search engines usually operates over the index structure which keeps information about documents content. Indexes of the documents content are exploited because of the time efficiency of retrieving information from those structures. In this article, we briefly describe an implementation of a full text document indexing and search tool developed under the NAZOU project [8]. We than discuss the problem of obtaining the base for of words in documents and compare two approaches – Porter stemming algorithm and WordNet thesaurus lemmatization functionality. Concerning the stemming and lemmatization we focus on the practical issues from the viewpoint of the document indexing mechanism, rather than on other issues of the methods such as the performance of stemming. The tool for document indexing keeps detail information about the document content, including the positions of the words in the documents. This allow us to extend the described tool from simple document retrieval to a system for information mining. We describe the use of the tool in the application for automatic semantic annotation.

2 daiRFTS tool

Our tool for document indexing and document search is named daiRFTS. The name stands for Data Access and Integration: Rich full-text search. The prefix dai refers to the ogsa-dai middleware [1] which we have integrated our tool to. The ogsa-dai framework allows us to expose the tool via Web-Service interface in a standardized way. The functionality of the tool can than be accessed remotely, providing better accessibility and interoperability.

Detailed information about documents content are stored in the index structure, including the positions of the word in the documents. The words in tool's dictionary are kept in the basic, lemmatized form. The tool exploit relational database to store all the information about documents and its contents.

3 Words base forms

Different morphological variants of the natural languages words have in most cases the same or very similar semantic interpretations and can be considered as equivalent for the purpose of information retrieval systems. This means that different morphological forms can be represented by a single representative term. The words in the index structures describing documents content and also key terms of the queries can be represented by their representative terms. Queries can than produce more relevant responses and the dictionary size needed for representing a set of documents decrease. A smaller dictionary size results in a saving of storage space and processing time. A number of stemming algorithms have been developed. Those algorithms are designed to reduce a word to its stem or root form.

The Porter algorithm [4] [3] is most widely used stemming algorithm and de facto standard for English language. It is a process for removing the commoner morphological and inflexional endings from words in English. It is used for term normalization done usually for Information Retrieval systems.

Another approach to finding base forms of the words is to use dictionary produced by human experts. WordNet [5] [6] is an lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets. WordNet is publicly available. Even it can do a lot more, it can be exploited to obtain lemmas of the word forms.

In our work on document indexing tool we have compared those two approaches: the stems produced by Porter algorithm and lemmas obtained from WordNet dictionary. The goal of this composition is to determine if the Porter algorithm is sufficient enough to replace lemma dictionary created by human expers (as Porter algorithm is faster compared to the lemma retrieval from WordNet dictionary). We have used three sets of documents containing jobs offers, each containing about 20 MB of plain text files. In the following text, we will refer to the Porter's stems and WordNet lemmas as the representative terms(RP).

According to performed experiments WordNet lemma retrieval produced 7% more representative terms than Porter's algorithm. For 6.8% of the WordNet RPs there were more than than one related Porter RPs (e.g. for words WRITTEN and WRITE WordNet produces RP: WRITE, while Porter alg. produces RPs: WRITTEN, WRITE). In the majority the of the cases different variants of non regular verbs caused generation of additional RPs.

For the 11% of the Porter RPs there were more than than one related WordNet RPs (e.g. for words BUSY and BUSINESSES Porter's alg. produces one RP: BUSI, while WordNet produces RPs: BUSY, BUSINESS). Porter algorithm joined into one class of equivalence different word classes in the majority of the cases and in some cases words with different semantics.

Use of the WordNet provide us with sematicaly more precise information and provide us with better categorization of words to representative terms. However, if we consider nouns as the most important word class for the full-text search information retrieval, than multiple RPs for irregular verbs and joining words with simmilar base structure from different words classes are not the key factors; while the performance benefit of using Porter's algorithm can be significant.

4 Integration with semantic annotation

On the example of the integration of DaiRFTS tool with the semantic annotation application we show the potential of extracting useful information and statistical values from detailed data in the index structure. We first describe the semantic annotation application and than explain how the functionality provided by DaiRFTS improves it.

Ontea [7] is a tool developed under project NAZOU. It aim is to automatically identify the instances of the ontological concepts in the input plain text files and fill the new instances to the ontology. Identification of the ontological instances is based on the rule-based system defined by the system user. The rules are defined in the form of regular expressions.

To illustrate the mechanism, we provide following example. The pilot application of NAZOU project is a system for retrieval and representation of information about job offers in the knowledge oriented structures, ontologies. Ontea is one of the tools for automatic filling of the ontologies with the information from acquired data. When Ontea finds a string matching the regular expression of its rule, it creates a new instance of a concept in ontology (if this instance does not exist yet). For example, the rule for ontological concept "Location" is filling the ontology with the words following the word "location" from the documents identified as job offers.

This method might have some serious disadvantages, notably if the rules for Ontea are not tuned properly. A part of automatically generated instances suffer from being inaccurate or misleading. However, Ontea alone does not have any mechanism for evaluating the correctness of the new instances. Let's say Ontea finds word A followed by B where A is the key word bound to a rule and B is identified as a new instance for a concept in the ontology. This is where DaiRFTS

tool can help by providing useful statistical information. To verify correctness of the new instance, using DaiRFTS interface we can gather information about the number of occurrences of words A and B within certain distance from each other or within the same phrases in the document collection. We can also retrieve number of occurrences of the word B in the documents collection. From those values, we can derive the probability of occurrence of word B after word A and can use the value to evaluate the correctness of the new instance in the ontology.

5 Conclusions

In this paper, we have briefly presented the tool for documents content indexing and full text search and we have discussed the problem of finding the base word forms for index structures by comparing the words grouping obtained by Porter algorithm and from WordNet dictionary. We have also described the usability of our full-text search for statistical information gathering useful for other tools for knowledge acquisition, especially for semantic annotation.

Acknowledgements

This work was partially supported by the Slovak State Programme of Research and Development "Establishing of Information Society" under the contract No. 1025/04.

References

1. OGSA-DAI: `http://www.ogsadai.org.uk`
2. Hovy, E. H., Hermjakob, U., Lin, C. Y: The Use of External Knowledge of Factoid QA, Text REtrieval Conference, 2001
3. Jones, K. S., Willet, P: Readings in Information Retrieval, San Francisco: Morgan Kaufmann,1997,ISBN 1-55860-454-4.
4. Van Rijsbergen, C. J., Robertson, S. E., Porter, M. F. New models in probabilistic information retrieval. British Library Research and Development Report, no. 5587, British Library, London, 1980.
5. Miller, G.: WordNet: An On-line Lexical Database, Special Issue, International Journal of Lexicography, Vol. 3, Num. 4, 1990
6. WordNet: http://wordnet.princeton.edu/
7. Laclavik, M., Seleng, M., Gatial, E., Balogh, Z., Hluchy, L.: Ontology based Text Annotation OnTeA; In: Proc. of 16th European-Japanese Conf. on Information Modelling and Knowledge Bases, EJC'2006, Y. Kiyoki et. al. eds., 2006, Dept. of Computer Science, VSB - Technical University of Ostrava, pp. 280-284, ISBN 80-248-1023-9. Trojanovice, Czech Republic.
8. Návrat, P., Bieliková, M., Rozinajová, V.: Methods and Tools for Acquiring and Presenting Information and Knowledge in the Web. In: CompSysTech 2005, B. Rachev, A. Smrikarov (Eds.), Varna, Bulgaria, June 2005. pp. IIIB.7.1-IIIB.7.6.