An Automated Approach to Ontology Building in Text Analysis Problems

I.V. Zakharova
Department of Mathematics
Chelyabinsk State University
Chelyabinsk, Russia
e-mail: iren@csu.ru

A.V. Melnikov
Department of Mathematics
Chelyabinsk State University
Chelyabinsk, Russia
e-mail: mav@csu.ru

J.A. Vokhmitsev
Department of Mathematics
Chelyabinsk State University
Chelyabinsk, Russia
e-mail: jakow@csu.ru

Abstract[1]

An approach to how to automatically build an ontology for complex tasks of full-text document classification using UDC is discussed in the article.

1. Introduction

Developing industry of electronic document exchange with growing number of full-text documents requires new methods of data access organization, and many of them could be classified as artificial intelligence systems – knowledge-based systems.

One of the major tasks that full-text databases should handle is search of documents by their contents. Traditional methods of context search (implemented in web search engines, for example) often don't provide adequate response. The main problem of these methods is difficulty to formulate query precisely – to select key words that are to be looked up in documents. This problem arises from a number of reasons, among those: user's insufficient knowledge of subject area terminology, presence of synonyms and polysemic words in natural language, and even misspellings of key words in the query and source documents.

Another fundamental reason for this problem is that sometimes the user doesn't know exactly what information she wants to get, having only a general idea about what she's interested in. So, for example, trying to extend one's knowledge in computer linguistics, one could get from AltaVista search engine hundreds and thousands of references to documents containing words "computer" and "linguistics". One could wish to group these documents by topic, select documents about milestones in this subject area, major events, get names of main researchers and links to their publications, etc. This example shows that the problem of finding information is closely related to a problem of text classification. The use of ontologies to explain implicit and hidden knowledge is one of approaches to this task [3].

2. Approach to Automatically Build Basic Ontologies

The term "ontology" is commonly used in Gruber's definition: explicit, i.e. express specification of conceptualization, where conceptualization is defined as description of a set of objects and links between them [1]. Formally, ontology consists of notions (concepts) organized in a taxonomy, their descriptions and inference rules. For our problem we will use a definition introduced by Hotho: ontology O is a set,

$$O = (L; F; C; H; Root),$$

where: L – dictionary (set of terms), C – set of notions (concepts), F – interpretation function,

$$F(L) \rightarrow C,$$

describing relationship of the set of terms to the set of notions, H – taxonomy. Concepts are connected by directed, non-cyclic, reflexive relationship H ($H \subset C \times C$). Root – the main concept [2].

Proceedings of the 8th International Workshop on Computer Science and Information Technologies CSIT'2006
Karlsruhe, Germany, 2006

Subject area ontologies are being created by subject area experts to formalize knowledge, definitions, and rules to extract new knowledge. There are specifications and tools to support the process of ontology creation and maintenance [4,5]. One of the important points is the ability to merge ontologies created by different experts into one meta-ontology, because the data format is specified.

We tried to build our own meta-ontology tailored to text document classification tasks. The method we used is based on a simple principle – an expert in some subject area writes books, articles and reviews on that subject. Those publications are collected in libraries and get bibliographic descriptions in databases. Published books also get mandatory classification codes under classification system standardized in the country of publication. It's obvious that the author of the publication is capable of determining the subject classification of the said publication very precisely.

Let's look more closely at the bibliographical description of a book. Apart from author's name and title it consists of classification code and key words, which are usually taken from annotation and table of contents. The basis of any ontology is concept taxonomy. Our ontology is based on Universal Decimal Classification (UDC), which contains about 133 000 concepts, organized in a taxonomy. UDC was chosen because in Russia UDC code is mandatory for every published book. The next step in ontology building process is creating a description for every concept. In our ontology descriptions were gathered automatically from bibliographic databases by retrieving all the key words corresponding to every UDC code, i.e. those key words that were assigned by a cataloguer as a part of cataloguing process. Bibliographic data were gathered from distributed bibliographic databases maintained by more than 200 libraries around our country [6]. At present time the ontology contains 133 151 concepts and includes from 50 to 1000 terms for each concept.

3. Example of Produced Ontology

An example concept from the ontology is shown on Figure 1.

This ontology could be used for text classification tasks and building queries on large datasets. For example, in return to query "philosophical systems and concepts" addressed to "usual" search engine we could get loads of unsystematized references. In case the query was first passed through ontology, we could have gotten a classification of philosophical systems and names of their originators, i.e. ability to substantially narrow the search. The broadening of a search is also possible, for example, the query "Teilhard de Chardin" would result in knowledge that he is a philosopher of XX century, originator of a philosophical system, and references to other philosophical systems.

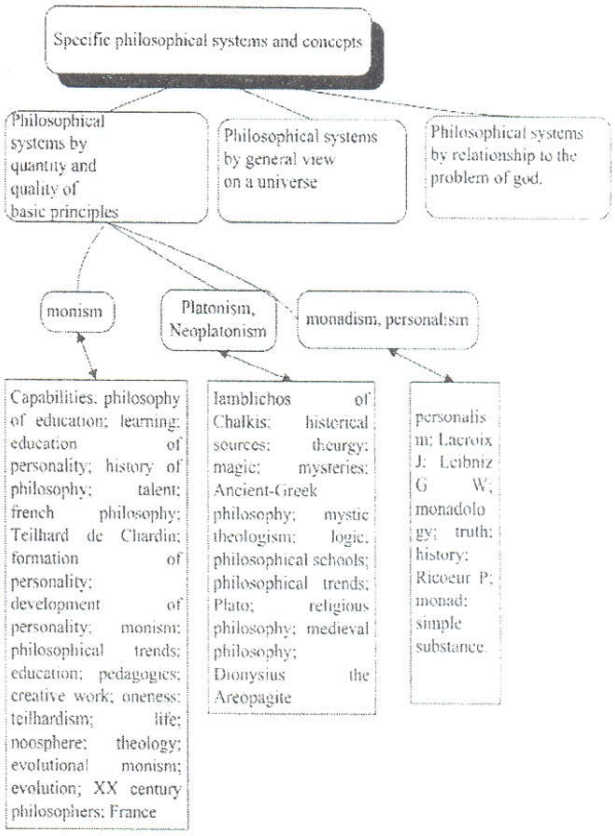

Figure 1

Ontologies are more than just complex approach to description and classification of information. They could be used to support development of a new kind of digital libraries – implemented as distributed intellectual systems

References

1. Gruber T.R. "A translation approach to portable ontology specifications". *Knowledge Acquisition*, 1993; 5(2).
2. Hotho A., Maedche A., Staab S. "Ontology-based Text Clustering". In: *Proc. of IJCAI*, 2001.
3. Nguyen M.H., Adzhiev A.S. "Description and use of thesauri in information systems. Approaches and implementation". *Electronic libraries*, 2004; 7(1).
4. Ovdej O.M., Proskudina G.Yu. "Review of ontology toolkits". *Electronic libraries*, 2004; 7(4).
5. http://www.w3.org/2004/OWL/
6. http://www.arbicon.ru/services/index.asp