Computer Science

Boris Lauser

Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment

Diploma Thesis

Diplom.de

Bibliographic information published by the German National Library:

The German National Library lists this publication in the National Bibliography; detailed bibliographic data are available on the Internet at http://dnb.dnb.de .

ISBN: 9783832469054

Boris Lauser

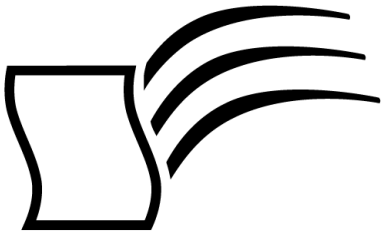
Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment

Diplom.de

Boris Lauser

Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment

Diplomarbeit
an der Universität Fridericiana Karlsruhe (TH)
Fachbereich Wirtschaftsingenieurwesen
Institut für Angewandte Informatik
Januar 2003 Abgabe



Diplomica GmbH
Hermannstal 119k
22119 Hamburg

Fon: 040 / 655 99 20
Fax: 040 / 655 99 222

agentur@diplom.de
www.diplom.de

ID 6905

ID 6905
Lauser, Boris: Semi-automatic ontology engineering and ontology supported document indexing in a multilingual environment
Hamburg: Diplomica GmbH, 2003
Zugl.: Fachhochschule Südwestfalen, Technische Universität, Diplomarbeit, 2003

Diplomica GmbH
http://www.diplom.de, Hamburg 2003
Printed in Germany

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Motivation

The management of large amounts of information and knowledge is of ever increasing importance in today's large organisations. With the ongoing ease of supplying information online, especially in corporate intranets and knowledge bases, finding the right information becomes an increasingly difficult task. Today's search tools perform rather poorly in the sense that information access is mostly based on keyword searching or even mere browsing of topic areas. This unfocused approach often leads to undesired results. The following example illustrates the problem more clearly:

> An agriculture scientist would like to find out which organisation established the Agreement on Agriculture. A simple search for "establish Agreement on Agriculture" might result in a huge list of documents containing these words, but actually none of them containing the desired result: WTO or World Trade Organisation. The problem becomes even worse if the result searched for only appears in a foreign language document.
>
> Figure 1 shows an extract of an ontology, which could solve this problem by following links in a graph. The grey ellipses represent generic concepts, whereas the white ones represent specific instances of these concepts. The two concepts shown here are linked by a relationship. An ontology-enabled search application would first identify "Agreement on Agriculture" as a "standard" and would then detect the relationship "establish" to "international organisation" and its instances, and hence solve the problem by extending the search query. This example shows how ontologies can help to improve the management of information. Furthermore, it could provide added value by detecting other relationships that provide the user with more possibilities: for example, standards of other organisations could be presented.

Semantically annotated documents, i.e. documents that are indexed with ontological terms and concepts instead of simple keywords, provide several advantages. First, the ontological abstraction provides robustness against changes in the document. In the above example, the document representation might change using the term 'Agricultural Agreement' instead of 'Agreement on Agriculture'. However, since the document has been annotated with the ontological semantics, this will not affect the search results. Second, since the ontology used

for annotating the document in this example is domain-specific, the semantic meanings and interpretations of keywords are bound to that domain and therefore the retrieval is likely to be more efficient. A term can have several meanings in different domains. By first mapping the keyword to its semantic representation in a specific ontology and using the ontology's linked knowledge structure, a much more focused search approach can be taken. Third, document specific representations no longer affect the search. This is extremely important in the case of multilingual representations. Keywords of several languages are mapped to the same concept in an ontology and are therefore given the same meaning. Multilingual search portals can be established to produce the same results, no matter which language is used for retrieval.

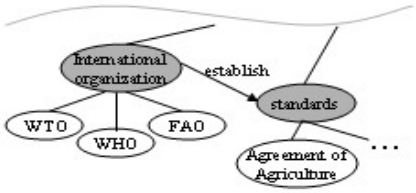


Figure 1: Ontology example, excerpt

An important task in knowledge management facilitating above described search scenario id the classification and indexing of documents. At present, subject specialists are responsible for this time consuming process. However, with today's vast amount of available information on the WWW, automatic support is needed to efficiently manage this task. Ontologies play a critical role in supporting the machine readable semantics needed to facilitate automation. They can be used for providing the categories and keywords needed to describe the content of documents. Automatic text classification tools still lack the necessary precision to replace human indexers and need to be extensively evaluated in different domains.

Before such powerful Semantic Web[1] applications can be built and used within certain domains of knowledge, the basic requirement - a machine readable vocabulary represented by a domain ontology - has to be established. The creation of ontologies is a time consuming task and often carried out in an ad-hoc manner. Only few methodologies exist and existing ones are often extremely complex and need extensive training and expertise. Even less automated tool support is available. Constituting the knowledge base for future Semantic Web applications, domain ontologies have to be created continuously in all possible areas and communities. The need for a reusable methodology is evident.

[1] Refer to [Pal01] for a short introduction to the Semantic Web.

1.2 Approach

The thesis introduces a comprehensive framework for building a domain-specific ontology. The approach combines classical methodologies for human-based ontology engineering with semiautomatic support of a heuristic toolkit. Two methods for ontology acquisition are applied in order to create the domain ontology. The first is to create a small, domain-specific core ontology from scratch. This step is supported by automatically extracting interesting concepts from a corpus of domain texts, which can be used to extend this base ontology. The second acquisition approach takes a well-established thesaurus as a basic vocabulary reference set, and converts it into an ontology representation. Then, a domain specific and a general corpus of texts are used to remove ontology concepts that are not descriptive for the domain from this converted representation. The rational used here is that domain specific concepts are more frequent in the domain-specific text corpus. The results of these steps are assessed to assemble a first version of the domain specific ontology. This ontology is then accessible through a multilingual web portal to be incorporated into other applications, such as document indexing or keyword searching of indexed documents. It could eventually be used to automatically index documents available through this kind of search application.

Carried out in collaboration with the Food and Agriculture Organisation (FAO)[2] of the United Nations (UN), the main focus of this thesis is on the adoption of the proposed framework to the specific environment and needs of this large organisation. The framework has been applied to create a prototype biosecurity ontology for the domain of Food Safety, Animal and Plant Health to be incorporated into an Internet Portal to this domain. Within this context, the conversion of a thesaurus into an ontology and evaluations of two automatic tools especially, constitute the central parts of the academic research work. The first evaluation is on a tree-pruning algorithm used in the ontology creation process to retrieve domain specific concepts from the converted thesaurus. The second evaluation is on a text classification application based on support vector machines, enhanced by a domain specific ontology serving as background knowledge for the classification algorithm.

[2] [http://www.fao.org].

1.3 Outline

The next section gives an introduction and overview about the Food and Agriculture Organisation, and the Agricultural Ontology Service (AOS) Project, which provides the bigger context in which the research work of this thesis is embedded. The current information management structure will be introduced briefly, outlining the overall current status and problems within the organisation.

In section 3, I will give an introduction to the idea of the Semantic Web as well as to ontologies and their various representations and engineering approaches. The comprehensive framework for the creation of a multilingual domain ontology is covered in section 4. The application of the framework will be described in the context of the above-mentioned project to establish an International Portal on Food Safety, Animal and Plant Health. The conversion of an existing thesaurus into an ontology representation as well as the adaptation of a multilingual ontology web browser to be embedded into the system is discussed here in detail.

Sections 5 and 6 describe in detail the adaptation and evaluation of two automatic tools constituting parts of the framework. Section 5 describes the thesaurus pruning algorithm used within the ontology creation framework and discusses the results of an empirical evaluation carried out within the context of the project. Section 6 introduces the reader to the area of automatic text classification and describes the adaptation of an already existent automatic text classifier based on support vector machines to incorporate domain specific ontologies. Several evaluation results are discussed against the question of the applicability of the classifier in the context of the FAO and against results of earlier evaluations. Finally, section 7 summarises the findings and results and provides an overview on future work.