KNOWLEDGE TECHNOLOGIES FOR INFORMATION ACQUISITION AND RETRIEVAL

doc. Ing. Marián Mach, CSc., Ing. Kristína Machová, CSc.
Department of Cybernetics and Artificial Intelligence,
Technical University of Košice, Letná 9, 042 00 Košice,
Slovak Republic.
Tel: +421-95-6022571; Fax: +421-95-6253574
E-mail: machm@tuke.sk, machova@tuke.sk

ABSTRACT

The aim of this paper is to present an overview and achievements of the national VEGA project 1/8131/01 of Scientific Grant Agency of Ministry of Education of the Slovak Republic. It is based on technologies and tools developed within the ESPRIT KnowWeb project (1998-2001) and ESPRIT Enrich project (1998-2000) and extends some of the results achieved within these two projects.

1 INTRODUCTION

The efficient utilization of information gathered from various sources inside and outside a commercial organization has become an important competitive factor. Information in the organizational environment of a company passes several phases during its life cycle. The central role within this cycle is played by information preservation and the ultimate goal is the presentation of relevant information in the right time. In order to achieve such a goal it is necessary to organize available information in a way that supports effective storage and retrieval of information.

There is a limited number of technologies and methodologies that are currently used for the organization and sharing of information. In addition, the functionality of those methods that are in wide use is heavily limited. These methods lead toward the utilization of traditional algorithms for information retrieval, for example search based on attached metadata (e.g. author, creation date), key words or full-text search. However, most traditional methods are not very effective, they typically provide a lot of information of low relevancy.

A balanced solution contains technologies built on the knowledge modeling. This approach assumes the existence of a conceptual model describing selected application domain so called domain model. A domain model can be roughly understood as a shared vocabulary consisting of organization-specific terms that are typically used to represent, organize and transfer knowledge within an organization.

2 PROJECT AIMS

The main aim of the project is development of methods for searching and retrieval of information from different sources using a unified interface. From the point of view of using knowledge modeling techniques it is very important to support automatic building of information repositories in connection with searching for relevant documents with subsequent annotating. Using methods for knowledge discovery in databases enables to define relationships between existing and newly discovered concepts and to code them in a knowledge model. That may lead toward sustainable extensions of a domain model within a company. Project aims were defined in more details in [1].

3 PROJECT DATA

The presented project is still currently running as a three year VEGA project. It has been launched on January 1, 2001. Planed scope of the project is over 7,75 thousand hours of research capacity in 2003. In this year the project team is composed of eleven people – majority of them from Dept. of Cybernetics and Artificial Intelligence FEI TU Košice.

4 ACHIEVEMENTS

Results of the project achieved so far cover creation of domain models, information annotation, and intelligent information retrieval:

a) Various techniques for knowledge discovery in databases were analyzed and used for clustering and classification of textual documents. Knowledge discovery in textual documents consists of two major phases. Within the first phase a document collection is transformed into an internal or intermediate representation form, which presents already structured data suitable for text data mining. Within the second phase text data mining itself is performed. Three main approaches to text mining from several possibilities have been used in the project: clustering/visualization based on self–organized neural networks, association rules and predictive classification models (rules induction, instance based learning and Naive Bayes classifier) [2].

A supervised approaches for building classification model was used. Knowledge discovery techniques for indexing and annotation of information contained in databases was used. When using a domain model it is possible to link the discovered rules (or other knowledge structures) to the elements of a domain model [3].

b) Methods for semi-automatic construction of a domain model based on the acquisition and modification of relevant knowledge concepts within existing databases were analyzed and used in the project. The system MEBL (Modification of Explanation Based Learning) was proposed. It modifies the domain model, in order to adapt it to a user environment. The system is based on the modification of domain models using reduction. Two kinds of reduction are presented: reduction in the horizontal and in the vertical directions [4].

c) A method for retrieving relevant textual documents from a set of documents using the WordNet lexical database was explored. The method adds semantics information of the words to the classic word-based indexing process. Every document can be represented not only by any lexical representation, but can also include sense-based information about itself [5].

d) Web-based communication forum for exchanging and delivering information was proposed. One-way and two-way communication modes were used, including information publication, contributing to discussions, and opinion polling. Proposed system plays the role of an advanced communication forum enabling to create, store and share information among different kinds of users. In order to organize information available within the system, it uses knowledge modeling and document annotation techniques [6].

e) In the project there was proposed a method for semantic searching in textual documents based on ontology. This approach requires documents to be linked onto ontology concepts and the definition of a representation of semantic distance between texts. Existing methods of automatic linking of documents were used [7].

f) An original approach to supporting knowledge management within an organization was proposed. Special attention was paid to organizations with distributed environment. For this purpose an experimental system for support of mobile agents that combines the power of high-level distributed programming with the mobile agent paradigm has been proposed and tested [8].

5 EXPECTED ACHIEVEMENTS

Since the project is not finished yet, some more achievements are expected. In the last third of the project duration the project team is being focused on the following:

- Creation of domain models represented as ontological models based on the analysis of textual documents
- Identification/extraction of concepts from textual documents using statistic and semantic-based methods
- Classification of textual documents against elements of an ontological model based on machine learning and statistic methods.

6 CONCLUSIONS

The paper discusses the VEGA project 1/8131/01 “Knowledge Technologies for Information Acquisition and Retrieval” of Scientific Grant Agency of Ministry of Education of the Slovak Republic and its achievements.

REFERENCES

[1] Mach, M.: Knowledge Technologies for Information Acquisition and Retrieval. In: Proc. of the II. Internal Scientific Conference of the Faculty of Electrical Engineering and Informatics ISC’2001, Košice, 2001, ISBN 80-88964-84-9, 29-30.

[2] Paralič, J. – Bednár, P.: Knowledge Discovery in Texts. Technical Report, Dept. of Cybernetics and AI, Technical University, Košice, 2002, 26 pages.

[3] Paralič, J. – Bednár, P.: Knowledge Discovery in Texts Supporting e-Democracy. In: Proc of the Int. Conference on Intelligent Engineering Systems INES2002, University of Zagreb, Opatija, Croatia, 2002, ISBN 953-6071-17-7, 327-332.

[4] Machová, K. – Janovský, J.: Knowledge Acquisition and Modification in the Context of Prenatal medicine. Lekař a technika, Vol. 32, No. 6, Česká lékařská společnost J.E. Purkyňe, 2002, ISSN 0301-5491, 147-151.

[5] Hudák, S.: Text Retrieval Using WorldNet Lexical Database. In: Proc. Of the II. Doktorantskej konferencie, TU Košice, 2002, ISBN 80-968666-2-1, 35-36.

[6] Mach, M. – Macej, P. – Hreňo, J.: Ontology-based Communication Forum. In: Knowledge-based Intelligent Information Engineering Systems & Allied Technologies, IOS Press, Amsterdam, 2002, ISBN 1-58603-280-1, 1544-1548.

[7] Bednár, P. – Hudák, S.: Automatic Linking of Textual Documents. 5 ved. konf s medzin. účasťou “Informatika a algoritmy 2002”, Herľany, 2002, 215-219.

[8] Paralič, J. – Paralič, M. – Mach, M.: Support of Knowledge Management in Distributed Environment. Informatica, Vol.25, No.3, 2001, ISSN 0350-5596, 319-328.