SYSTEM OF KNOWLEDGE EXTRACTION FROM ONTOLOGY OF ELECTRONIC EDUCATIONAL RESOURCE

Tatyana Balova[1], **Natalya Rokhas Kriulko**[1], **Andrzej Kotyra**[2]
[1]East Kazakhstan State Technical University n. a. D. Serykbayev, [2]Lublin University of Technology

***Abstract**: This article describes the system having the main function of extraction of knowledge from the document corresponding to the knowledge stored in ontology of electronic educational resource.*

Keywords: ontology, knowledge extraction, semantic analysis, electronic resource

SYSTEM POZYSKIWANIA WIEDZY Z ONTOLOGII CYFROWYCH ZASOBÓW EDUKACYJNYCH

***Streszczenie**: W artykule przedstawiono opis systemu, którego podstawową funkcją jest wydobywanie z tekstu dokumentu nazw odpowiadających nazwom przechowywanym w ontologicznym elektronicznym zbiorze edukacyjnym.*

Słowa kluczowe: ontologia, ekstrakcja nazw, analiza semantyczna, zasoby elektroniczne

Introduction

Knowledge, intellectual capital, intellectual property gain growing recognition as a new source of development [2]. This is why innovative higher education institutions tend to control and manage knowledge that they possess in more effective manner. One of the current trends in development of modern university is a conversion of information space into space of knowledge and competences. The most current issue is elaboration of formalized models for knowledge representation that would allow processing of scientific, educational and methodic information at semantic level in university knowledge management system. The solution of this problem is related to development of structured information storage with multiple inference rules using Semantic Web technology.

The core of Semantic Web technology is ontology that is used for formal specification of concepts and relations, which characterize a certain area of knowledge. The advantage of ontology as a mean of knowledge representation is its formal structure, which is easier to implement [3].

1. Ontology creation of "Electronic University" domain

Creation of unified ontology for detailed description of knowledge base model of electronic university is comparatively difficult and time-consuming task. In order to build electronic university ontology, the following structure of ontology model is suggested:

$$O = \{O_u, O_r, O_k\}, \tag{1}$$

where O_u – electronic university ontology, O_r – information resources ontology, and $O_k = \{O_1, \ldots, O_m\}$ – hierarchical, consistently expansible system of main knowledge areas ontologies O_i, valuable for work of the electronic university. Emphasis of knowledge areas hierarchy provides for possibility to create separate ontologies of different knowledge subareas, which might have different detalization depending on modeling demands. Electronic university ontology O_u includes main concepts that describe structure, composition of elements and work of university (department, professors, students, curricula, etc.). Information resources ontology O_r includes description of all types of data and information of the organization (documents, files, information bases, programs, etc.).

Since „Electronic University" domain is relatively massive, ontology of this work will include areas related to educational and methodic field complexes (EMFC), including educational and methodic discipline complexes (EMDC) and working curriculum (WC).

A part of hierarchical taxonomy of main classes in ontology of electronic educational resource (EER) developed is presented in Figure 1.

After the main concepts (classes) of „Electronic educational resources" domain are determined, formed into classes taxonomy and corresponding properties are set for each class, we can implement ontology practically using Protégé 4.1. Web Ontology Language OWL was chosen as a language for ontologies description. It is based on DARPA Agent Markup Language/ Ontology Interchange Language (DAML/OIL), that has strong both theoretical and practical support and is a standard in Semantic Webs. What is more, the language dialect OWL DL was applied for it supports descriptive logic and provides maximum possibilities for logical inference of a new knowledge and the knowledge put into ontology.

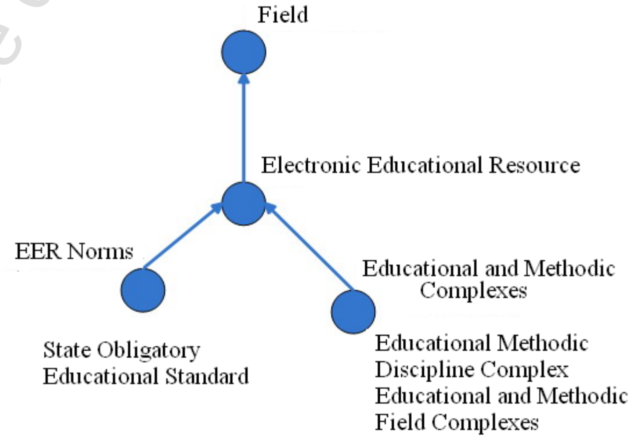
Field
Electronic Educational Resource
EER Norms
Educational and Methodic Complexes
State Obligatory Educational Standard
Educational Methodic Discipline Complex
Educational and Methodic Field Complexes

Fig. 1. Fragment of taxonomy of electronic educational resource

Formal model of ontology is a variation of network knowledge model and is represented as three finite sets – classes, relations and properties. To develop taxonomy of classes of electronic university ontology combined method was chosen. Microsoft Visual Studio 2008 was used to develop the dedicated software. Ontological knowledge base is a reference base, that is why full-text search method can be used to extract knowledge from electronic educational resources.

2. Algorithm of knowledge extraction from electronic educational resources ontology

The basic system function is to extract knowledge document text [1], corresponding to knowledge stored in ontology. Text downloaded by a user is represented in a pdf file. Two types of chains of triplets are extracted from the fragment of text – facts describing properties – values and properties – objects.

Artykuł recenzowany otrzymano: 01.06.2012r., przyjęto do druku: 31.07.2012r.

The algorithm proposed below consists of the following steps.

At the **first** step, occurrences in ontologies are identified on the base of its properties (values). It is described as <Class_Id,PropertyId,Value>.

Step 2. It is necessary to find class occurrence having name Class, which property – value with name PropertyId has value Value. Note that for some occurrences there might be few properties – values that define it.

For example,
<AKS,hasName,"Computer system architecture">
<AKS,hasKredit,"9">
<AKS,hasPrak," Arithmetic and logical basics of computer.">

In the case being discussed, a new occurrence is searched with properties having appropriate values.

Step 3. To make text fragment analysis, we choose property-value corresponding with user's request. If this property is found, then the analysis is continued (**Step 4**). Otherwise, the adequate message is generated and the program terminates.

Step 4. A search request is started.

Search conditions are values of property value found at the first step. Full text search method described earlier is used at this stage. The result of the request is a group of numbers $result=\{k_1,k_2,..k_i\}$, where k_i – quantity of compliances founded in the text and corresponding to i-th search condition. This group is used to determine final result. In order to describe the final result as a linguistic variable, a production model of data representation was used, which is based on rules allowing to represent knowledge in form of proposals „If (condition), then (conclusion)". Result and conclusion will be given as the analysis protocol.

The algorithm described above is implement as a software module of logical inference in semantic analysis system. Its structure is presented in Figure 2.

The central part of knowledge extraction and evaluation system is Text Analysis Server. Reference to this server comes from the program used for text processing, which provides user's working text to the server. As a result, the server returns text, in which compliances were founded.

Function of text work comes to PDF-document conversion into the text that is transferred to text analysis server.

Text analysis server management model has a full row of functions, which are interactions with program module for working with texts, as well as with models of management and administration.

Knowledge model of the system is based on ontology described in OWL language. Basic structural block – confirmation represented by the triple: resource (class occurrence for OWL), named property and its value.

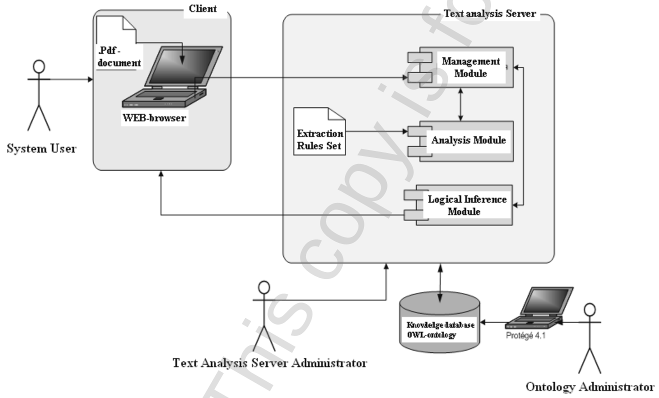

Fig. 2. Semantic analysis system structure

The main task of logical inference module is finding compliances of facts in text and facts described in ontology. As it was mentioned above, a few extraction variations come for inference (set of facts chains). It is necessary to consider all the extraction variations and make a conclusion regarding facts reliability based on the most „acceptable".

Library PDFBox.dll for extraction of text from pdf files was used for implementation. PDFBox is a library that allows creating and manipulating PDF files and also allows for extracting its content.

Library dotnetRdf.dll is used for parsing of rdf-file. This library provides for work of Rdf in .Net. The analysis results are recorded in file report „Log.txt" and are acceptable for review.

3. Conclusions

The algorithm developed for electronic resource text analysis allows to form a protocol, which gives information about document, evaluation, conclusion and recommendation. It should be underlined that the final decision on of document positioning is made by an expert.

References

[1] Андреев А.М., Березкин Д.В., Рымарь В.С., Симаков К.В.: *Использование технологии Semantic Web в системе поиска несоответствий в текстовых документах*, 6-я Всероссийская научная конференция RCDL'2004 , 1-2.

[2] Гаврилова Т.А.: *Базы знаний интеллектуальных систем*. Питер, 2001.

[3] Тузовский А.Ф., Ямпольский В.З.: *Системы управления знаниями в образовании*. Современные средства и системы автоматизации. Изд-во Том. ун-та, 2002, с. 295-299.

Tatyana Balova
e-mail: TBalova@ektu.kz

Doctor of science, associate professor of academic department "Information systems", Information Systems Technologies and Energy Faculty.

Natalya Rokhas Kriulko
e-mail: NRohas-Kriulko@ektu.kz

Senior professor of academic department "Information systems", Information Systems Technologies and Energy Faculty

Andrzej Kotyra
e-mail: a.kotyra@pollub.pl

Associate professor of Institute of Electronic and Information Technology, Faculty of Electrical Engineering and Computer Science at Lublin University of Technology.