DATA MINING FROM TEXT DOCUMENTS

Vovnjanka Roman Volodymyrovych[1], Dosyn Dmytro Grygorovych[2], Kovalevych Vira Mykhailivna[2]

[1] Lviv Polytechnic National University, S. Bandery Str., 12, Lviv, 79013, UKRAINE,

[2]Karpenko Physico-Mechanical Institute of the National Academy of Science of Ukraine

The approach to creating a computer system of an automated development of basic ontology is described in the article. This approach is based on natural language text documents analysis. The method for semantic analysis of documents using the software Link Grammar Parser and machine learning techniques is discussed. Machine learning tools operate along with OWL-ontology. Ontology provides grammatical and semantic structure patterns for the statements recognition (logic predicates of 1st order) in researched and / or educational texts. As a result of such recognition, new items are added to the ontology.

The text corpus consists of different text documents, each of them containing from 1 to 10-20 sentences. These sentences are in sequential logical connection. The text is divided into an ordered set of sentences. The sentences consecutively undergo the basic procedure of recognition. Complex sentences are divided to simple sentences by means of parsing. Substitution of pronouns by nouns of the first part of the sentence to which these pronouns refer to is performed in the process of separation. The preparation of sentence proposals is carried out so that the algorithm can clearly identify all the concepts involved in the formulated sentences in future. There are also differentiated generalized concepts (classes) and specific concepts (instances of the corresponding classes).

The architecture of ontology synthesis system is designed. The basic modules of the system and their purposes are described. The choice of software tools for the practical implementation of the system is justified and practical implementation of the proposed method is done. The functionality of the developed system has been tested. The system allows in automatic mode to fill in the ontology of a domain.

It is proposed to use the intelligent system for OODA loop simulation. Ontology is the core of knowledge base in the intelligent system. It consists of the domain ontology and the ontology of applications. The content ontology directly affects the second and third stages of the OODA cycle, and the structure and content of the ontology depends on the 1st and 2nd stages.

Keywords – ontology, learning ontologies, intelligent agent, knowledge base, text document.