ABSTRACT

Theme of the Graduation Thesis:” Modelling an intelligent terminological information extraction system”

Author: Pogosyan Gevorg Aleksandrovich

Thesis tutor: Candidate of Technical Sciences, assistant professor of department of Information and Communication Technologies, Mathematics and Information Security Timchenko Olga Viktorovna.

Information about the contracting authority: Scientific and Educational Innovation Complex «Information and Communication and Mathematical Technologies» of Pyatigorsk State University.

Relevance of the research topic: Information processed by computer systems represents texts in a human language (HL). And over time, the share of such information only increases in the total volume. As a rule, when solving problems associated with automatic word processing, it is assumed to search and extract the given units (words, phrases). While working with scientific and technical texts (STT), these units are terms (words, phrases associated with a specific subject area), which, among the most frequent units, carry and convey meaning.

Objective: to search and extract automatically terminological information from a single text including terms and keywords; to determine the frequency (quantity) of the use of recognized terms and options in the text, to calculate their information content weights which allow us to evaluate their significance in relation to each other in the document.

Tasks:

1. To look at modern methods of extracting terms and existing means of formal representation of structures of a human language, to study their applicability for automatic recognition of terms, their options and constructions of their use;

2. To examine the procedures for extracting (based on partial parsing) of various terminological information from a single text;

3. To make proposals for the software implementation of the procedures for extracting terminological information mentioned above;

4. To write a software programme to implement the procedures for extracting terminological information with automatic calculation of indicators of information content of the found keywords.

Theoretical and practical significance of the research: The speed and accuracy of information analysis is gaining particular importance these days. An increase in the volume of processed information, the possibility of making adjustments to the work of computing systems, the extraction of predetermined terms automatically are only some of the characteristics and functional opportunities which, when properly configured, provide high efficiency for such systems. The function of automatic search and selection of terms is widely used in technical translation, creating dictionaries on various subjects.

Results of the research: as a result of the work done, a software system has been developed in the Delphi integrated development environment, which performs the following tasks:

- search for exact matches of terms (the number of terms entered for search can be several) with terms in the search text;
- highlighting the terms found for further processing;
- calculation of the amount of the use of the selected term in the search text;
- calculation of TF, IDF and TF-IDF coefficients;
- output of 10 main keywords with the highest TF-IDF coefficients in the form of a table in alphabetical order.

Recommendations: The obtained results and the developed software product are recommended for use when working with scientific and technical text.