

Тема выпускной квалификационной работы: «Моделирование интеллектуальной системы извлечения терминологической информации».

Автор ВКР: Погосян Геворг Александрович

Научный руководитель ВКР: канд. экон. наук, доц. кафедры информационно-коммуникационных технологий, математики и информационной безопасности Тимченко Ольга Викторовна

Сведения об организации-заказчике: НОИК ИКИМТ ПГУ.

Актуальность темы исследования: информация, обрабатываемая вычислительными системами, представляет собой тексты на естественном языке (ЕЯ). И со временем, доля подобной информации в общем объеме только увеличивается. Как правило, при решении задач, связанных с автоматической обработкой текстов, предполагается поиск и извлечение заданных единиц (слов, словосочетаний). В работе с научно-техническими текстами (НТ) указанные единицы являют собой термины (слова, словосочетания связанные с определенной предметной областью), которые в числе наиболее частотных единиц несут и передают смысловую нагрузку.

Цель работы: поиск и автоматическое извлечение из отдельно взятого текста терминологической информации, включающей термины и ключевые слова, определение частоты (количества) употребления в тексте распознанных терминов и вариантов, расчет их весов информативности, который позволяет оценить их значимость по отношению друг к другу в документе.

Задачи:

1. Рассмотреть современные методы извлечения терминов и существующие средства формального представления конструкций ЕЯ, исследовать их применимость для автоматического распознавания терминов, их вариантов и конструкций их употребления;

2. Рассмотреть процедуры извлечения (на базе частичного синтаксического анализа) различной терминологической информации из отдельно взятого текста;

3. Составить предложения по программной реализации рассмотренных процедур извлечения терминологической информации;

4. Программно реализовать процедуры извлечения терминологической информации с автоматическим расчетом показателей информативности найденных ключевых слов.

Теоретическая и практическая значимость исследования: Скорость и точность анализа информации приобретают в наши дни особое значение. Увеличении объемов перерабатываемой информации, возможность внесения корректив в работу вычислительных систем, извлечение заданных терминов автоматически - только некоторые характеристики и функциональные возможности, которые при правильной настройке обеспечивают высокий КПД подобных систем. Функция автоматического поиска и выделения терминов активно используется при техническом переводе, создании словарей различных тематик.

Результаты исследования: в результате проделанной работы разработана программная система в среде Delphi, которая выполняет следующие задачи:

- поиск точных совпадений терминов (количество введенных для поиска терминов может быть несколько) с терминами в искомом тексте;
- выделение найденных терминов для дальнейшей обработки;
- подсчет количества использования выделенного термина в искомом тексте;
- подсчёт коэффициентов TF, IDF и TF-IDF;
- вывод 10 основных ключевых слов с наибольшим коэффициентом TF-IDF в виде таблицы в алфавитном порядке.

Рекомендации: полученные результаты и разработанный программный продукт рекомендуется использовать при работе с научно-техническим текстом.