

Тема выпускной квалификационной работы: «Разработка интеллектуальной системы кластеризации текстовой информации (на материале русскоязычных текстов)»

Автор ВКР: Гатиева Анжела Владимировна

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: В условиях стремительного развития web-пространства и увеличение объема хранимой информации появляются новые проблемы, связанные с обработкой такого большого количества данных. В частности, это проблема релевантного поиска информации в Web; категоризация и рубрикация документов (для каталоговых служб). Решение таких задач возможно с помощью методов кластеризации, позволяющих разбивать множество web-страниц на категории в зависимости от характеристик данных документов.

Цель работы: разработать интеллектуальную систему кластеризации текстовых документов. Система должна автоматически определять теги документов из коллекции, что в дальнейшем и позволяет распределять тексты по кластерам.

Задачи:

1. Обзор и анализ современных методов кластеризации текстов;
2. Постановка задачи автоматического тегирования текста как задачи кластеризации;
3. Разработка архитектуры интеллектуальной системы;
4. Разработка программного продукта, реализующего задачу автоматического тегирования текстового документа.

Теоретическая и практическая значимость исследования: состоит в разработке и апробации предложенной в рамках выпускной квалификационной работы интеллектуальной технологии кластеризации и

тегирования текстов.

Результаты исследования: результатом стала интеллектуальная система кластеризации и тегирования текстовых документов, обязательным элементом которой является система взаимодействия с пользователем.

Рекомендации: алгоритм автоматического извлечения тегов должен содержать более глубокую предварительную обработку текста, чем тот, который реализован в рамках ВКР. Вероятно, это позволит повысить точность тегирования.