Секция «Вычислительная математика и кибернетика»

Технология выявления, хранения и анализа первичных и вторичных понятий в корпусе текстов

Сабурова Мария Ивановна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: masha-saburova@yandex.ru

Содержательной задачей работы, является задача структурированного представления текста в виде перечня ключевых понятий, их вхождений в текст и связей между понятиями и частями текста. Первоначальная идея, из которой развилась эта задача, заключается в том, чтобы создать аналог поисковой системы, на вход которой подавались бы два понятия, и в результате система показывала бы, как эти понятия связываются в корпусе текстов. Вообще же, структурированное представление текста помогает извлекать и визуализировать информацию из больших коллекций документов, что весьма востребовано в современном мире.

В настоящее время ведется работа по созданию программного обеспечения, которое помогает человеку создавать разметку текста, т.е. подсказывает ему понятия и их вхождения. Также разрабатываются инструменты для изучения накопленных разметок. Задача создания такой программной системы нетривиальна, поэтому на первом этапе было создано ПО, которое позволяет аналитику удобно размечать текст вручную.

Процесс разработки программы включал в себя дизайн хранилищ данных. При формулировке требований производилось итеративное уточнение. Анализ сценариев использования ПО позволил оценить частотность выполнения операций, в соответствии с чем развивался пользовательский интерфейс программы. Текущий вариант программы был сочтен удобным.

С помощью программы были собраны данные. Входные данные – 18 текстов, объемом 2-3 абзаца каждый. В результате для каждого текста была получена разметка. Также был получен словарь, состоящий из 37 понятий. Получив эти данные, появилась возможность начать этап исследований.

Визуализация полученных данных была выполнена в виде графа, где вершины – понятия из словаря, ребром соединялись вершины, если эти понятия встречались в одном тексте. Вес ребра определялся разными способами. Во-первых, просто число общих текстов для двух понятий. Также был рассчитан индекс парной зависимости понятий [2]. Показаны самые зависимые и наименее зависимые пары понятий, определяющие новые связи, ранее не отмеченные экспертами.

В рамках анализа формальных понятий построена диаграмма Хассе для понятий и текстов, из которой были получены новые связи между объектами анализа [3].

Ближайшие цели: продолжать анализ полученных в эксперименте данных, кластеризовать тексты, визуализировать получаемые данные удобным образом, ввести функционал выделения связей в тексте и меду текстами.

Литература

1. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск.—ООО "И.Д.Вильямс 2011.—528 с.

2. Ивченко Г. И., Медведев Ю.И. Введение в математическую статистику.— Издательство ЛКИ, 2009

3. http://www.upriss.org.uk/fca/ – Formal Concept Analysis Homepage