ТЕМАТИЧЕСКИЙ АНАЛИЗ НЕСТРУКТУРИРОВАННОЙ ТЕКСТОВОЙ ИНФОРМАЦИИ

Гонтарь Ю.Н., Кустов А.А., Баранова Ю.Н., Чередниченко О.Ю.
Национальный технический университет «Харьковский политехнический институт», г. Харьков

Накопленные объемы информации и постоянно увеличивающиеся темпы ее роста определяют актуальность и значимость исследований в области информационного поиска. Развитие сетевых технологий, в том числе и Интернет, способствуют значительному увеличению доступных информационных ресурсов и объемов передаваемой информации. Как правило, это разнородная, динамическая, слабо структурированная и избыточная информация.

Во многом поиск определяется слабо формализуемыми и нечеткими условиями, в значительной степени зависящими от опыта и предпочтений человека. Далеко не всегда пользователь информационно-поисковой системы может четко и однозначно сформулировать именно тот набор ключевых слов, который и приведет его к искомому результату. Представленные на сегодняшний день в большинстве популярных поисковых систем способы организации полнотекстового поиска и методы анализа документов не учитывают в достаточной мере человеческий фактор. Все это обуславливает актуальность и значимость исследований, направленных на решение проблемы адекватного отображения информационных потребностей пользователей.

Одним из вариантов решения этой проблемы является поиск документов по образцу, когда человек задает некоторый документ в качестве образца, а система, реализующая данный вариант поиска подбирает документы подобные заданному (подобные по содержанию, тематике). Анализ существующих исследований, посвященных решению задач поиска документов по образцу, выявил отсутствие достаточно проработанной теории и практики решения задач тематического анализа неструктурированной, естественно-языковой текстовой информации произвольного содержания.

В работе предлагается модель структурного представления текста в виде ориентированного мультиграфа, а также рассматриваются вопросы анализа такой модели применительно к решению задач поиска документов по образцу. Реализовать поиск документов по образцу предлагается на основе решения двух основных задач: 1. Выделение тематики документа, которая отражает содержание документа через множество ключевых слов, находящихся в некоторой зависимости друг от друга. 2. Вычисление тематической близости документов, что дает возможность проранжировать документы по степени значимости.

Разработка моделей тематического обобщения набора документов и оценки тематической близости документов является предметом дальнейших исследований.