УДК 004.912

МЕТОДЫ ИНФОРМАЦИОННОГО ПОИСКА ДЛЯ ОПРЕДЕЛЕНИЯ СМЫСЛОВОЙ СХОЖЕСТИ ДОКУМЕНТОВ

Маркелова Ю.В. (Университет ИТМО)
Научный руководитель – профессор, д.т.н., Бесмертный И.А.
(Университет ИТМО)

В современном мире жизнь человека трудно представить без сети Интернет и возможности в любой момент найти нужную информацию, всего лишь используя поисковый запрос в строке браузера. При этом, в большинстве случаев предложенные результаты будут удовлетворять ожиданиям пользователей и соответствовать поисковым намерениям, заложенным в запросе. Это достижимо благодаря различным методам информационного поиска – действиям и процедурам, позволяющим осуществлять отбор определенной информации из массива данных.

Введение.
Далеко не всегда достаточно короткого, состоящего из нескольких ключевых слов и/или словосочетаний запроса для поиска ресурсов. В некоторых случаях требуется определить релевантность содержимого документа не только поисковому запросу, но и другому документу, что представляет собой задачу сравнения различных текстов и определения степени их смысловой схожести. В качестве примеров таких задач можно привести следующие:

- подбор подходящих материалов для научных статей;
- оценка соответствия написанной работы поставленным целям и теме;
- выявление плагиата;
- выбор из тысяч резюме нескольких, подходящих конкретной вакансии и др.

Вышеперечисленные задачи относятся к анализу текстов и определению степени их смысловой схожести и, во многих случаях, решаются человеком. Однако для прочтения и последующего анализа большого количества литературы требуются значительные временные ресурсы и постоянная внимательность специалиста.

Основная часть.
Существует несколько основных методов для обнаружения сходства текстов. К ним можно отнести сравнение на основе текстового содержимого документов (сравнивая слово за словом с целью найти одинаковые слова, словосочетания и фразы); поиск по ключевым словам с помощью современных информационно-поисковых систем, таких как Google; стилистический и семантический анализ текстов; вероятностные тематические модели.

Важным этапом перед программным анализом документов является их предварительная обработка, которая состоит из ряда преобразований над исходным текстом. Она может включать в себя лемматизацию, стемминг, исключение шумовых и редких слов, выделение ключевых фраз и именованных сущностей, а также учет опечаток и ошибок в текстах.

Наиболее популярным методом для определения смысловой схожести различных текстов является их семантический анализ, основанный на дистрибутивной гипотезе и векторном представлении документов. Для определение меры сходства текстов вычисляется косинусное расстояние между векторами, соответствующими векторным представлениям документов.

Выводы.
Существуют различные подходы для программной обработки текстов на естественном языке и дальнейшего их сравнения между собой, которые позволят значительно сэкономить время, а также существенно увеличить количество обрабатываемых документов в единицу времени по сравнению с человеком.

Маркелова Ю.В. (автор)

Бессмертный И.А. (научный руководитель)