РАНЖИРОВАНИЕ ТЕКСТОВЫХ ДОКУМЕНТОВ ПО РЕЛЕВАНТНОСТИ ЗАПРОСУ НА ОСНОВЕ ИХ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ

Чалая Л.Э., Лимаренко Д.В., Порчинский Э.В.
Харьковский национальный университет радиоэлектроники

В области информационного поиска и автоматической обработки электронных текстов можно выделить ряд относительно самостоятельных направлений: извлечение объектов и признаков, реферирование, классификация, кластеризация, интеллектуальный поиск, семантический анализ и т.п. [1]. В частности, задача организации эффективного доступа к неструктурированной тематической информации непосредственно связана с задачей ранжирования по релевантности запросу электронных текстов, извлекаемых из ресурсов сети Интернет или электронных библиотек. Для решения этой задачи разработано множество эффективных методов, позволяющих повысить качество поиска и количество обработанных источников. Для проведения кластерного анализа электронных текстовых документов достаточно сложно непосредственно получить сравнительные характеристики (кроме размера документов и их метаданных). Чтобы облегчить сравнение и ранжирование документов, необходимо использовать методы предварительной обработки текста для приведения его к виду, наиболее удобному для последующего анализа. Большинство современных поисковых систем используют векторную модель представления текста, представляющую собой таблицу частоты применения слов в документе. Такое представление позволяет но достаточно быстро определить ключевые слова документа и его тематику. В данной работе рассматриваются корпусы текстов технической тематики, особенность которых заключается в том, что их частотный словарь (по критерию $TF \cdot IDF$) определяет ключевые слова и тематическую направленность анализируемого текста. Современные методы векторного представления текстовой информации являются развитием моделей векторных пространств VSM (Vector Space Models). В этих моделях компоненты текстов (например, слова, словосочетания, фрагменты текстов, целые документы) представлены многомерными векторами, элементы которых определяются значениями некоторой заданной функции от совместной встречаемости текстов и их контекстов. Модель векторного пространства является алгебраической моделью представления текстовых документов в корпусе текстов.

Рассмотрим метод ранжирования документов по релевантности запросу на основании векторного представления текстовых документов. На начальном этапе метода создается линеаризованный словарь корпуса (коллекции) $T = \{t_1,..., t_n\}$ путем удаления стоп-слов, которые не несут семантической нагрузки, а затем осуществляется предварительная обработка текста, в процессе которой слова коллекции заменяются термами t_i, которые получаются в результате лемматизации (приведения к нормальной форме) либо стемминга (взятия общей основы для семейства слов). Слова в разных регистрах и в различных вариантах произношения, а также их аббревиатуры приводятся к

одной форме. Документы коллекции представляются набором входящих в них термов, а также частотами их вхождения. Каждый документ представляется мультимножеством слов. Под таким мультимножеством будем понимать неупорядоченную коллекцию, аналогичную обычному множеству, но допускающую наличие в коллекции одновременно двух и более одинаковых значений. Каждому терму соответствует координата векторного пространства, характеризующая количественно степень вхождения терма в документ. Таким образом, каждый документ характеризуется набором из n чисел. Определим матрицу M по формуле

$$M_{ij} = TF_{ij} \cdot IDF_i,$$

где M_{ij} – степень соответствия слова i документу j (каждый документ представляется в этой матрице в виде столбца (j фиксировано, i меняется)); TF_{ij} (Term Frequency, частота терма) – относительная доля слова i в документе j; IDF_i (Inversed Document Frequency) – величина, обратная количеству документов, содержащих слово i.

Если слово не встречается в данном документе, то значение соответствующей компоненты вектора равно 0. Меру релевантности векторного представления документа D_j вектору запроса Q будем задавать косинусом угла Q и D_j:

$$R\left(Q, D_j\right) = \cos\alpha = \frac{QD_j}{|Q||D_j|}.$$

Нормализация здесь позволят уравнять веса документов с различным количеством слов. Очевидно, что чем ближе $\cos\alpha$ к единице, тем больше оцениваемая мера релевантности (при $\cos\alpha = 0$ документ полностью не соответствует запросу. Вектор поискового запроса Q сравнивается со всеми документами, имеющимися в кэше, после чего с применением меры $R\left(Q, D_j\right)$ осуществляется ранжирование документов по релевантности: чем больше косинус, тем выше в списке результатов поиска документ [2]. Таким образом, результатом является ранжированный по релевантности перечень документов, соответствующих запросу.

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В./ – М.: МИЭМ, 2011. – 272 с. 2. Чалая, Л. Э. Оценивание пертинентности лингвистических дескрипторов в системах информационного поиска документов / Л.Э. Чалая, Ю.Ю. Харитонова // Восточно-европейский журнал передовых технологий. – 2015. – № 1/9(73). – С. 46–53.