ПРИМЕНЕНИЕ МЕТОДА ПОИСКА КЛЮЧЕВЫХ СЛОВ ДЛЯ ЗАДАЧИ ИНДЕКСИРОВАНИЯ ДОКУМЕНТОВ

Вабищевич Д.П.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Парамонов А.И. – канд. техн. наук, доцент

В работе предлагается подход к организации индексирования документов, для организации поиска по ним с ранжированием результатов. В основе положено применение алгоритма TextRank и дополнительная предобработка текстовых документов. В качестве индекса предлагается использовать вектор ключевых слов документной базы.

Значительное увеличение цифровых документов требует новых решений целого ряда задач, связанных с организацией документов и в том числе задачи поиска информации. Для её решения были разработаны алгоритмы от простейшего поиска по ключу (прямое нахождение слов запроса в документах) до построения моделей документов [1] и последующего сравнения их с запросом. В данной работы предлагается решение задачи индексирования документов, которое даст возможность поиска по документной базе с ранжированием результатов и сопоставления документов на схожесть.

В качестве результата операции индексирования должно получаться некоторое представление документа, которое будет отражать его основное содержание. В качестве такого представления документа предлагается использовать вектор по универсуму терминов документной базы. Терминами выступают извлеченные из документов ключевые слова и фразы. Размерность вектора соответственно равна общему числу терминов, которые получены из документов на данный момент.

Для процесса ранжирования необходимо определить алгоритм вычисления меры схожести документов (а точнее их представлений) друг с другом и с поисковым запросом. Для идентичности операции сравнения с представлениями документов, запрос также преобразуются в аналогичный вектор терминов. Он будет содержать информацию о присутствии ключевых слов в запросе. Вектора проиндексированных текстов помещаются в k-мерное дерево, так как данная структура данных обеспечивает быстрый поиск и ранжирование по набору k-мерных векторов. Для задачи ранжирования в качестве меры схожести запроса и документа предлагается использовать метрику Минковского с p=1 (расстояние городских кварталов, L1 норма). В нашем случае оно будет равно количеству терминов, относящихся только к запросу и только к документу. Таким образом, тексты с меньшим значением расстояния более похожи на запрос по смыслу. Следует учесть, что при добавлении новых текстов k-мерное дерево разбалансируется, что уменьшает скорость поиска и ранжирования. Также с каждым добавлением нового документа может изменяться и размерность вектора индексирования. Все это приводит к необходимости перестроения дерева в течение работы модуля, при накоплении определенного количества новых документов. Вопрос частоты балансировки дерева требует дополнительного исследования и будет рассмотрен в дальнейшем.

Извлечение ключевых слов предлагается выполнять с помощью метода поиска ключевых слов TextRank [2]. Принцип его работы заключается в построении графа со словами в вершинах и подсчете на его основе для каждого слова числового рейтинга. Он показывает, насколько слово часто встречается в этом тексте в разных контекстах. Чем больше это число, тем вероятнее, что слово важно для текста и отражает суть его содержания. Основной плюс алгоритма в том, что для извлечения ключевых слов он использует только статистики слов в текущем документе. Это значит, что он не требует хранения никакой дополнительной информации.

Для улучшения качества результатов работы алгоритма к тексту применяется предобработка. Во-первых, из текста убирается пунктуация и служебные слова, которые сами по себе не несут смысловой нагрузки (предлоги, местоимения и т.п.). Затем применяется алгоритм нахождения устойчивых словосочетаний, который основывается на подсчёте того, как чаще слова встречаются вместе или по отдельности [2]. В результате работы этого алгоритма часть ключевых слов (словосочетания) объединяются в один токен, и TextRank работает с ними уже как с одним термином.

Предложенный подход обеспечивает хорошее качество индексирования документной базы и как следствие быстрого поиска с ранжированием по ней. При этом для начала работы не требуется предварительно научения и наборов доменно-специфичных текстов, что делает качество работы независимым от доменной области и языка, а также от количества хранимых в базе текстов.

Список использованных источников:

1. Парамонов А.И. Представление знаний гибридной моделью для систем интеллектуального поиска / Вестник Донецкого национального университета – 2005. – Серия А, №1. – С. 404-409.

2. R. Mihalcea and P. Tarau. TextRank: Bringing Order into Texts// Proc. of the 9th Conf. on Empirical Methods in Natural Language Processing. – 2004. – С. 404–411.

3. Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean // In Advances in Neural Information Processing Systems 26, pages 3111–3119, 2013. – 9 с.