АЛГОРИТМ ПРЕДВАРИТЕЛЬНОГО ОТБОРА ПОХОЖИХ ДОКУМЕНТОВ С УЧЁТОМ ДУБЛИКАТОВ И НЕЕСТЕСТВЕННЫХ ТЕКСТОВ

Дудник А. В. [1)], **Безрук В. И.** [1)], **Бульдяк Я. А.** [1)], **Белоткач А. А.** [1)]
[1)] ***Национальный технический университет «ХПИ», г. Харьков.***

Нынешняя эпоха заслуженно названа информационной: современного человека буквально захлёстывают информационные потоки самого различного качества и тематики, нередко несущие противоречивые и даже взаимоисключающие сведения. Сказанное в полной мере относится и к сугубо научным источникам, а недавний скандал [1] в очередной раз показал, что даже высокостатусные научные издания не гарантируют публикацию проверенных материалов. Таким образом, поиск адекватных опубликованных материалов становится сложной задачей, при решении которой приходится опираться на технические интеллектуальные средства. В [2] было рассмотрено одно из таких средств, приведена его структура и алгоритм работы. В данном докладе рассматриваются методы предварительного отбора текстов.

Предположим, что для подтверждения или опровержения некоторой гипотезы следует опереться на опубликованные материалы. После того, как запрос на поиск текстов определённого содержания сформирован, начинает выполняться соответствующий поисковый алгоритм. В данной работе за основу выбран алгоритм латентно-семантического анализа (LSA). Алгоритмы этого типа находят сходство или различия между текстами, основываясь на том, какие слова и в каком количестве входят в каждый из них.

Предварительно тексты преобразовывают по правилу «мешка слов»: слова берутся в исходной морфологической форме, каждому слову указан вес в форме tf-idf сообразно числу вхождений в текст. Получаем векторную модель текста. Далее алгоритм работает не с исходным документом, а с таблицей, столбцы которой привязаны к анализируемым текстам, а строки — к словам. Очевидно, что с этого момента смысловое содержание текста игнорируется.

Следует отметить, что если один векторизированный текст относительно невелик, то таблица, соответствующая некоторой группе документов обладает большой размерностью, что существенно замедляет работу алгоритма. Также таблица будет разреженной, поскольку многие ячейки будут нулевыми.

С другой стороны, поскольку в данном случае важным является тематическое соответствие документа запросу, следует вести анализ не в плоскости документ-слово, а в плоскости документ-тема. С этой целью из векторизированных документов формируется матрица, которая

подвергается сингулярному разложению методом SVD. Разложение выполняется по формуле:

$$A = U \times S \times W^T,$$

где A – исходная матрица слово-документ, U – матрица слово-тема, S – сингулярная матрица тема-тема, W^T – транспонированная матрица тема-документ.

Матрица S позволяет оценить темы, затронутые в исходном наборе документов, их количество и мощность. В данном случае очевидно ограничение: один документ – одна тема. Однако если предварительно отобраны документы одного направления и их количество более 5, то данное ограничение преодолевается.

Темы с малыми сингулярными числами игнорируются, что снижает количество анализируемых исходных данных и повышает скорость работы алгоритма.

Однако рассмотренный алгоритм абсолютно не затрагивает смысл текстов, и если они дублируются или должным образом синтезированы (т.н. неестественные тексты), то гарантированно будут отобраны. Чтобы не допустить этого, вносится ряд дополнений в исходный алгоритм.

Так, для выбраковывания неестественных текстов следует учесть, что для естественных текстов характерна глобальная тематическая связность. Как правило, они содержат одну основную тему и несколько второстепенных. Синтезированные тексты характеризуются отсутствием единой тематики. Поэтому, после того, как тематическая мощность будет выявлена для всей группы текстов, следует её уточнить для каждого текста. В этом случае было решено дополнить алгоритм скрытым распределением Дирихле (LDA), как предложено в [3].

Предварительно осуществляется распознавания дубликатов. Для этого пары векторизованных текстов соотносятся с центроидами кластеров «дубликаты» и «не дубликаты», настроенных при обучении алгоритма.

Список литературы

1. Научный скандал года: ученые писали фейк-исследования, чтобы разоблачить лженауку [Электронный ресурс]. – Режим доступа: https://www.bbc.com/russian/features-45751968

2. Дудник А.В. Система анализа сегмента информационного пространства / А.В. Дудник, Н.О. Артюхов, В.С. Багнюк, Д.А. Микитюк, Т.А. Обухова // Актуальні проблеми автоматики та приладобудування: матеріали І Міжнародної науково-технічної конференції, 2017. С. 21–22.

3. Павлов А.С. Методы обнаружения массово порождаемых неестественных текстов на основе анализа разнообразия тематической структуры текстов / А.С. Павлов // Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL, 2011. С. 195–200.