А.А. Спрысков, Ю.В. Бидуля
Тюменский государственный университет, г.Тюмень
УДК 532.546.2

АГРЕГАЦИЯ И СУММАРИЗАЦИЯ ТЕКСТОВ ПОИСКОВОЙ ВЫДАЧИ ПО ЗАПРОСУ ПОЛЬЗОВАТЕЛЯ

Аннотация. Цель работы: описать подход к разработке программного приложения для компактного представления полезной информации, полученной в результате поискового запроса.

Ключевые слова: автоматическое реферирование, суммаризация, метрика TF-IDF, симметричное реферирование.

Автоматическое реферирование (или «суммаризация» от английского “summarization”) представляет одну из актуальных задач современности. Рост публикуемых в Сети данных вынуждает каждого человека затрачивать все больше времени на ознакомление и поиск полезной и необходимой информации.

Агрегация текстов поисковой выдачи по запросу пользователя подразумевает под собой получение ссылок на веб-страницы с контентом, соответствующим запросу пользователя, и получение и разбор содержимого HTML-файла веб-страницы с целью извлечения несущего смысл текста. Задачу можно разделить на две части: агрегация и суммаризация.

Получение ссылок предлагается осуществлять через взаимодействие с существующими поисковыми системами посредством их API (в реализованном приложении используется API Yandex.XML) [1]. Для получения HTML-файла веб-страницы используется возможности библиотеки libcurl [2], а для разбора содержимого HTML-файла по результатам сравнения производительности используется преобразование файла в объект класса DOMDocument и использование функционала генерируемого на его основе объекта класса DOMXPath, предоставляющего быстрый и удобный интерфейс для поиска содержимого тегов документа [3]. Результатом выполнения этого блока

операций является преобразование строки с запросом пользователя в набор текстов, тематика которых пересекается с текстом запроса.

Задача суммаризации может быть разбита на подготовку текста, определение ключевых слов текста и построение суммирующего текста.

Подготовка текста декомпозируется на следующие подзадачи: выделение в каждом тексте предложений; выделение в каждом предложении знаменательных слов; стеммирование каждого слова для каждого предложения каждого текста. Для выделения предложений текст разбивается по каждому символу, являющемся маркером конца предложения. Во избежание ошибок при разбиении в тексте предварительно «глушатся» символы-разделители, находящиеся в блоках (под блоком подразумеваются конструкции, заключенные в скобки или кавычки. Разбивать их не следует поскольку семантической единицей в данном случае является весь блок, а не одно предложение.), а также точки в аббревиатурах и адресах веб-сайтов. Для выделения в предложении знаменательных слов используется словарь служебных слов. Для стеммирования реализован алгоритм стеммера Портера для русского языка [4].

Для определения ключевых слов текста используется модифицированная метрика TF-IDF [5]. Формула расчета метрики TF-IDF для слова w документа d_w из корпуса документов C:

$$TFIDF(w) = \sum_{d=1}^{D} \left(\frac{S(t,d)}{W_d} * \sqrt{\frac{W_c}{S(t,c) * W_d}} * A \right) * B,$$

где S(t,d) – количество слов t во множестве d, W(d) – суммарное количество всех слов в множестве d, A и B – модификаторы, меняющие своё значение в зависимости от условий.

Чем выше текст в поисковой выдаче, тем сильнее (по мнению поисковика) его содержимое соответствует теме запроса, и модификатор A

отражает эту зависимость, с установленным шагом увеличивая метрику слова в документах по мере роста их позиции в выдаче. Модификатор B служит для изменения суммарного рейтинга для входящих в поисковой запрос слов, а также для слов, которые встречаются только в одном документе (такие слова могут попасть в документ при неправильном разборе сайта или из-за изначального несоответствия содержимого сайта теме запроса). В общем случае модификатор A служит для корректировки метрики слова в конкретном документе и зависит от документа, а модификатор B – для корректировки суммарной метрики слова и зависит от самого слова.

Важно отметить, что операция расчёта метрики применяется для стеммированных слов.

После определения метрики TF-IDF для всех слов корпуса выбирается некоторое количество слов, чьи метрики максимальны. Эти слова и будут являться ключевыми словами при построении суммирующего документа.

Для построения суммирующего документа используется модификация алгоритма под названием «Симметричное реферирование» [6,7]. Оригинальный механизм подразумевает его применения для построения реферата по единственному тексту путем подсчета для каждого предложения количества связей с другими документами. Связью в данном контексте является отношение, в котором два предложения имеют в себе одно и то же ключевое слово. В суммирующий документ попадают предложения с наибольшим числом связей. В случае, если два предложения с одинаковым количеством связей конкурируют за место в суммирующем документе, выбирается предложение, количество связей соседей которого выше.

Модифицированный алгоритм использует модификатор для изменения веса связи между предложениями из разных текстов. Кроме того, соседние предложения влияют на веса друг друга, делясь устанавливаемой модификатором частью своего рейтинга. Первый механизм позволяет

регулировать количество заимствований из разных текстов – при высоком значении модификатора большее значение будет иметь смысловая важность предложения, но суммирующий текст будет более «рваным». Второй механизм регулирует долю веса, которую предложение отдает своим соседям, что позволяет предложениям одного текста объединяться в кластеры, что повысит связность текста, но может привести к потере значащих предложений.

Формула для расчета рейтинга предложения s_i текста d_s имеет следующий вид:

$$R(s_i) = \left(\sum_{k=1}^{K}\left(\sum_{d=1}^{D} o(k,d) * C\right)\right) * D + (R(s_{i-1}) + R(s_{i+1}))*E$$

, где o(k, d) – количество содержащих ключевое слово k предложений текста d, C, D и E – модификаторы веса связей между предложениями разных текстов, влияния положения предложения в тексте и влияния веса соседнего элемента соответственно. В случае, если d_w=d, модификатор C считается равным единице. Модификатор D не равен единице только для первых предложений текстов (повышение веса первых предложений текста требуется для компенсации отсутствия левого соседа и для отражения наблюдения, по которому зачастую наибольшую семантическую ценность несут первые предложения текстов).

После расчета рейтинга для каждого предложения они сортируются в порядке убывания ценности. Затем итеративно выбираются предложения с наибольшими весами, проверяются на совпадение слов с другими выбранными предложениями (так, если в выбранном предложении содержится как минимум 75% слов одного из выбранных предложений, оно отбрасывается) и добавляется в список выбранных. Выбор предложений происходит до тех пор, пока количество выбранных предложений не достигнет установленной длины реферата в предложениях. После этого предложения объединяются в группы по принципу принадлежности к одному тексту, затем элементы групп объединяются в подгруппы по принципу соседства в исходном тексте.

Полученные подгруппы представляют собой абзацы суммирующего предложения; они выводятся в порядке нахождения первого предложения абзаца в исходном тексте; при совпадении положений первых предложений абзацев разных текстов первым выводится та группа, исходный документ которой ближе к первому.

СПИСОК ЛИТЕРАТУРЫ

1. Яндекс.XML Документация: [Электронный ресурс]: URL: https://tech.yandex.ru/xml/doc/dg/concepts/about-docpage/ (Дата доступа 19.03.2017).

2. Curl manual: [Электронный ресурс]: URL: http://curl.haxx.se/docs/manpage.html (Дата доступа 14.03.2017).

3. Класс DOMDocument: [Электронный ресурс]: URL: https://php.ru/manual/class.domdocument.html (Дата доступа 19.03.2017).

4. Russian stemming algorithm: [Электронный ресурс]: URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (Дата доступа 14.03.2017).

5. Lukas Havrlant, Vladik Kreinovich. A Simple Probabilistic Explanation of Term Frequency-Inverse Document Frequency (tf-idf) Heuristic (and Variations Motivated by This Explanation). International Journal of General Systems, 2017

6. Яцко, В.А. Симметричное реферирование: теоретические основы и методика НТИ. Сер. 2. - 2002.

7. Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4 (NAACL-ANLP-AutoSum '00), Vol. 4. Association for Computational Linguistics, Stroudsburg, PA, USA, 21-30.