Критерии выбора ключевых предложений при поиске нечетких дубликатов

Е.В. Шарапова

Муромский институт (филиал) ФГБОУВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирской обл., ул. Орловская, 23
E-mail: info@vanta.ru

Работа посвящена вопросам определения методики выбора ключевых предложений, обеспечивающей наиболее быстрое нахождение дубликатов текстов при минимальном числе запросов к поисковым системам. На основе критерия «Ценность ответа» проведена оценка различных подходов к определению ключевых предложений. Лучшим критерием является выбор предложений с наибольшей суммой весов слов, входящих в предложения с учетом глобальной частоты слов.
The work is devoted to determining the methodology for choosing key sentences that ensures the most rapid finding of duplicate texts with a minimum number of queries to search engines. Based on the criterion "Response Value", various approaches to determining key sentences were evaluated. As a criterion for choosing key sentences, the largest sum of the weights of the words included in the sentence is calculated, taking into account the global frequency of words.

Введение

При разработке систем обнаружения нечетких дубликатов текстов возникает задача выявления предложений, наиболее хорошо характеризующих анализируемый текст. Такие предложения будем называть ключевыми предложениями.

Необходимость определения ключевых предложений состоит в потребности снижения вычислительной нагрузки на системы обнаружения дубликатов и уменьшении числа запросов к поисковым системам, являющимся поставщиками контента для поиска дубликатов. Выбор ключевых предложений может осуществляться по различным методикам. Цель работы – определить методику выбора ключевых предложений, обеспечивающую наиболее быстрое нахождение дубликатов при минимальном числе запросов к поисковым системам [1, 2].

Оценка влияния выбранных ключевых предложений на качество поиска

Для оценки влияния выбранных ключевых предложений на качество поиска дубликатов в сети Интернет было проведено исследование, суть которого заключалась в оценке результатов поиска.

Чем выше найденный дубликат находится в поисковой выдаче (в идеале – на первом месте), тем лучше ключевое предложение представляет анализируемый текст.

Надо заметить, что найденные поисковыми системами документы практически всегда будут соответствовать поисковому запросу. Проблема заключается в том, что найденные документы могут соответствовать запросу, но не быть дубликатами анализируемого документа. Другими словами, ключевые предложения могут встречаться в других документах, мало связанных по содержанию с анализируемым. Поэтому, чем более уникальным будет предложение, тем больше шансов, что найденные поисковыми системами документы действительно будут являться дубликатами.

Для оценки использовалась метрика «Ценность ответа» (ReciprocalRank) [3, 4, 5]. Она позволяет оценить, сколько усилий требуется программе, чтобы найти первый ответ на свой вопрос, или какова вероятность того, что программа досмотрит результаты до позиции, где находится первый правильный ответ. Формально «ценность» ответа на конкретное задание вычисляется как:

$$ReciprocalRank = rank(pos),$$

где pos – это минимальная позиция, на которой находится релевантный ответ.

Если правильных ответов в ответе нет, то «ценность» равна 0.

Функция rank(pos) обычно задается некоторой линейкой значений для нескольких первых позиций и считается равной 0 для всех остальных.

$$rank(pos)=1/pos/$$

ReciprocalRank обратно пропорционален позиции первого релевантного ответа системы.

Для тестирования были выбраны несколько вариантов формирования ключевых предложений.

Для оценки влияния длинных последовательностей слов в запросах были выбраны предложения с самой большой длинной [6]. Для оценки влияния весов слов, подсчитанных по разным методикам, были выбраны предложения с подсчетом IDF по анализируемой части документа, всего документа и глобальной коллекции документов. Предложения с самым большим количеством весомых слов по принципу выбора похожи на самые длинные предложения, но в данном случае не учитываются общеупотребимые, высокочастотные слова. За счет этого последовательность слов в выбранных предложениях является более редкой, и, как следствие, более уникальной [7].

Предложения с самым высоким средним весом слов представляют собой предложения, насыщенные ключевыми словами. При этом, длина таких предложений не всегда бывает большой, и может составлять всего несколько слов, что не всегда достаточно.

Для сравнения качества выбора ключевых предложений в исследование были добавлены варианты формирования запросов из случайных предложений анализируемого текста.

В противоположность длинным и тяжелым предложениям, в качестве сравнения были выбраны также короткие предложения, предложения с наименьшим весом входящих в них слов.

Для исследования были выбраны 10 документов из разных коллекций рефератов, объемом 30-50 страниц каждой. Документы разбивались на части размером по 2000 символов. Результаты тестирования приведены в таблице 1.

Таблица 1 – Результаты исследования

Текст запроса	**ReciprocalRank**
Короткие предложения	0,23
Самые «легкие» предложения	0,26
Случайное предложение	0,64
Длинные предложения	0,81
Самые «тяжелые» предложения (часть)	0,62
Самые «тяжелые» предложения (документ)	0,77
Самые «тяжелые» предложения (коллекция)	0,84
Предложения с самым высоким средним весом слов	0,58
Предложения с самым большим количеством весомых слов	0,82

Заключение

Как можно заметить, короткие предложения и самые «легкие» предложения дают очень маленькое значение ReciprocalRank. Это означает, что по таким запросам находится большое количество документов, которые содержат эти предложения, но не являются дубликатами исходного текста. Оба вида стремятся к коротким предложениям с общеупотребительными словами.

Наилучший результат дают самые «тяжелые» предложения (коллекция) и предложения с самым большим количеством весомых слов. Немного уступают им самые длинные предложения. Все три вида стремятся к наиболее длинным предложениям, содержащим значимые, не часто употребимые слова.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-07-00692.

Литература

1. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации с помощью сигнатур содержания // Управление развитием крупномасштабных систем MLSD'2019. Материалы двенадцатой международной конференции Научное электронное издание. Под общей ред. С.Н. Васильева, А.Д. Цвиркуна. 2019. С. 1009-1011.
2. Sharapova E. One way to fuzzy duplicates detection // International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM. 14. 2014. С. 273-278.
3. Ageev M., Kuralenok I., Nekrestyanov I. Official ROMIP 2010 metrics // Russian Workshop on Evaluating Information Search Methods. Proceedings of ROMIP 2010, pp.172-187, 2010.
4. Voorhees E.M. Proceedings of the 8th Text Retrieval Conference // TREC-8 Question Answering Track Report, pp. 77–82, 1999.
5. Chapelle O., Metlzer D., Zhang Y., Grinspan P. Expected reciprocal rank for graded relevance // Proceeding of the 18th ACM Conference on information and Knowledge Management CIKM '09, pp. 621-630, 2009.
6. Zelenkov Y., Segalovich I. Comparative analysis of methods for fuzzy duplicate detection for Web-documents // Proceeding of 9-th Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» RCDL2007, pp. 166-174, 2007.
7. Sharapova E.V., Sharapov R.V. The problem of fuzzy duplicate detection of large texts // Proceedings of the International Conference Information Technology and Nanotechnology. Session Data Science. Samara, Russia, 24-27 April, 2018. CEUR Workshop Proceedings. Vol. 2212, pp. 270-277.