Представление текстовых документов в виде наборов ключевых предложений

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 сокращения числа запросов к поисковым системам при поиске нечетких дубликатов в сети Интернет. Деление документа на части и создание для каждой из них поисковых запросов позволяет существенно снизить нагрузку на поисковые системы и увеличить скорость нахождения нечетких дубликатов. Лучше всего для создания запросов подходит нахождение ключевых предложений. .
The work is devoted to the problem of reducing the number of queries to search engines when searching for duplicates on the Internet. Dividing the source document into parts and generating queries for each of them can significantly reduce the load on search engines and increase the speed of finding duplicates. The most effective way to generate requests is to identify key sentences.

Введение

Для поиска дубликатов документов в сети Интернет необходимо сравнить содержимое анализируемого документа со всеми документами, размещенными в Интернет. Так как объемы информации в интернете огромны, невозможно за сравнительно небольшие промежутки времени скачать и проанализировать все доступные страницы (документы). Для ускорения анализа содержимое сайтов скачивается на локальные диски и индексируется. Но количество сайтов огромно. В настоящее время существует более 1,7 миллиардов сайтов. Для хранения такого количества информации потребуются сотни терабайт. Только для того, чтобы прочитать такое количество информации с жесткого диска необходимо огромное количество времени.

По этой причине чаще всего процесс поиска дубликатов производится с использованием уже готовых решений – поисковых систем [1, 2]. Поисковые системы уже имеют свои поисковые индексы, оптимизированные для быстрой работы с огромными массивами информации [3, 4].

Системы поиска дубликатов посылают поисковым системам запросы в виде кусков проверяемого текста и анализируют полученные от них результаты. В идеале в качестве запроса должен выступать полный текст анализируемого документа. Но поисковые системы (Google, Яндекс) ограничивают размер запроса несколькими десятками слов. Например, Яндекс накладывает следующие ограничения на запросы: максимальная длина запроса - 400 символов; максимальное количество слов – 40.

По этой причине системы поиска дубликатов должны произвести разбивку проверяемого документа на небольшие части и проводить поиск по каждой из них. Но и здесь возникает несколько проблем. Во-первых, для больших документов число частей будет достаточно большое, и их проверка также требует немало времени. Во-вторых, поисковые системы ограничивают число обрабатываемых от пользователей запросов. Другими словами, они не позволяют формировать большое количество запросов от пользователей в течении дня.

Модель представления документа набором ключевых предложений

Одной из основных проблем поиска дубликатов тексов в сети Интернет является большое число запросов к поисковым системам. Число запросов прямо

пропорционально размеру проверяемого документа. При решении задачи в лоб, число запросов равно количеству предложений в документе.

Естественно, возникает задача сокращения числа запросов до приемлемого значения N. Одним из вариантов решения проблемы является представление документа набором ключевых предложений. Каждое из таких предложений будет служить отдельным запросом к поисковой системе. Таким образом, выбрав N ключевых предложений, можно сократить число запросов к поисковым системам до N.

Основная задача ключевых предложений – служить лучшим представлением содержания документа. Другими словами, использование ключевого предложения в качестве запроса должно обеспечивать выборку наиболее близких по содержания документов поисковыми системами. Ключевое предложение должно быть наиболее оригинально. Оно не должно состоять из наиболее употребимых фраз, типа «Сегодня прекрасное утро» или «Добрый день уважаемые коллеги». Если предложение будет слишком распространенным, то результатом выполнения запроса поисковой системы будет большой список найденных документов, среди которых действительно похожий документ можно и пропустить. Дело в том, что при поиске дубликатов текстов в сети Интернет для сокращения времени работы обычно выбирают несколько первых результатов поисковой выдачи (например, 10 ссылок с первой страницы). Соответственно, если нужный документ будет располагаться не на первой странице результатов поиска, то вероятнее всего он не будет рассматриваться системой проверки. Из этого следует, что ключевое предложение должно быть, во-первых, достаточно длинным, во-вторых, наиболее уникальным, чтобы обеспечить нахождение дубликатов (при их наличии) в первых строках поисковой выдачи.

Предлагается следующий подход для определения ключевых предложений:

1. Весь документ d разбивается на N частей.

$$d = \{p_1, p_2, p_3, ..., p_N\}$$

где p_i – i-я часть документа.

Части могут формироваться постранично (например, одна или две страницы), по определенному количеству символов (например, 2000 символов), по определенному числу предложений (например, 10 предложений) или путем деления документа на заданное количество частей. При этом границы частей устанавливаются по границам предложений (предложения не разрываются на части).

2. Для каждой части формируется список входящих в нее предложений

$$p_i = \{s_1^i, s_2^i, s_3^i, ..., s_m^i\}$$

где s_k^i – k-е предложение в i-й части документа.

3. Из списка предложений для каждой части по определенному алгоритму выявляется самое значимое (ключевое) предложение.

$$ks_i = \max f(s_1^i, s_2^i, s_3^i, ..., s_m^i)$$

где $f(s_k^i)$ – функция, вычисляющая вес k-го предложения.

4. Формируется набор ключевых предложений.

$$KS = \{ks_1, ks_2, ks_3, ..., ks_N\}$$

5. Ключевые предложения из набора в виде запросов направляются в поисковые системы для поиска похожих документов в сети Интернет.

Методы выявления ключевых предложений

Ключевые предложения можно выбирать по различным признакам. Можно выделить следующие из них:

- Самые длинные предложения;
- Предложения с самой высокой суммой весов слов (самые «тяжелые» предложения);
- Предложения с самым высоким средним весом слов;
- Предложения с самым большим количеством весомых слов.

Вес слов может определяться по-разному. Часто используется метод TF*IDF в различных модификациях [5].

Величина инверсной частоты IDF может подсчитываться в рамках коллекции документов, одного документа или одной части документа [6].

В задаче выбора ключевых предложений, каждое предложение фактически является отдельной единицей. Так как выбор ключевых предложений осуществляется в каждой части по отдельности, то на выбор оказывают влияния слова только из рассматриваемой части. Соответственно, достаточно использовать только частоты слов в каждой части документа. Недостатком такого подхода может являться завышенный вес общеупотребимых слов, особенно в небольших по объему частях текста. Возможна ситуация, когда все слова в части будут встречаться по одному разу, и, соответственно, иметь одинаковый вес.

В другой стороны, все части представляют собой один документ, то есть между ними существует смысловая связь. По этой причине, встречаемость слов в других частях может косвенно свидетельствовать о важности слова в рассматриваемой части. Тогда, подсчет частот слов по всему документу тоже может быть оправдана.

Учет встречаемости слов по всей коллекции также имеет определенный смысл. Это позволяет выявить высокочастотные, общеупотребимые слова. Такие слова в небольших текстах могут получить неоправданно высокий (завышенный) вес. По этой причине имеет смысл либо учитывать встречаемость слов во всей коллекции документов, либо проводить фильтрацию высокочастотных слов (например, на основе закона Ципфа или имеющейся статистики).

Заключение

Предложенные в работе подходы по поиску дубликатов документов с помощью набора ключевых предложений используются в системе Автор.НЕТ [7, 8]. Исследования показали, что использование ключевых предложений позволяет в десять раз сократить число запросов к поисковым системам. При этом сохраняются хорошие результаты обнаружения дубликатов текстов.

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

Литература

1. Brin S., Davis J., Garcia-Molina H. Copy detection mechanisms for digital documents // Proc. ACM SIGMOD Annual Conference, pp. 398–409, 1995.
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. Henzinger M. R. Finding near-duplicate web pages: a large-scale evaluation of algorithms // SIGIR 2006, pp. 284–291, 2006.

4. 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.
5. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management, vol. 24, n. 5, pp. 513–523, 1988.
6. Wu H.C., Luk R.W.P., Wong K.F., Kwok K.L. Interpreting TF-IDF term weights as making relevance decisions // ACM Transactions on Information Systems, vol. 26 (3), 2008.
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.
8. Шарапова Е.В., Шарапов Р.В. Обнаружение нечетких дубликатов текстов в больших массивах информации // Проблемы управления и моделирования в сложных системах. Труды XXI Международной конференции. В 2-х томах. Под редакцией С.А. Никитова, Д.Е. Быкова, С.Ю. Боровика, Ю.Э. Плешивцевой. 2019. С. 335-339.