Шарапова Е.В.
Муромский институт (филиал) федерального государственного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»
602264, г. Муром, Владимирская обл., ул. Орловская, 23
E-mail: sharapovamivlgu@gmail.com

Обнаружение полных дубликатов текстов

Полный дубликат документа — это экземпляр текстового документа на естественном языке, без учета формы представления (оформления) и формы хранения (формата файла).

Примерами полных дубликатов являются произведения одних и тех же авторов, напечатанные в разных издательствах, тексты, сохраненные в разных текстовых форматах (DOC, DOCX, RTF, TXT), различных кодировках (cp866, windows 1251, KOI8-R, Unicode).

Для обнаружения полных дубликатов наиболее подходит использование сигнатурных методов. Суть сигнатурных методов сводится к представлению документа неким кодом (хэш-функцией, числом, контрольной суммой), позволяющим с высокой степенью вероятности выявить одинаковые (или практически одинаковые) документы. Фактически, сравнение документов сводится к сравнению нескольких чисел — сигнатур документов. Это существенно сокращает потребности памяти и вычислительные затраты. Отличительная особенность сигнатур — это возможность подсчета их в любое время, в том числе заранее, а не в момент проверки.

Перед подсчетом сигнатур предлагается проводить предварительную обработку текстов: приведение к единой кодировке, удаление переносов и разрывов страниц, склейка разорванных при этом слов, удаление номеров страниц, очистку форматирования и представление документов в виде обычных текстов.

Простейшим примером сигнатуры является контрольная сумма документа, например, CRC32 или MD5.

Один из способов подсчета сигнатур [1, 2] основан на подсчете частоты встречаемости слов в документе TF (term frequency). Сигнатура строится по нескольким (5 или 6) наиболее частотным словам. В качестве сигнатуры используется контрольная сумма CRC32 строки, состоящей из выбранных слов, упорядоченных по алфавиту.

Немного более сложная версия сигнатуры предполагает подсчет веса слов не по формуле TF, а по формуле TF*IDF. В данном случае учитывается не только частота слов в документе (TF), но и общая встречаемость слов во всех документах коллекции (IDF). Для подсчета сигнатуры, подсчитывается вес каждого слова, путем вычисления tf(d)×idf(t) и выбираются несколько слов с наибольшим весом. Сигнатура вычисляется как контрольная сумма CRC32 строки, состоящей из выбранных слов, упорядоченных по алфавиту.

Хорошо зарекомендовала себя сигнатура, построенная на основе комбинации частоты слов TF и остаточной обратной частоты документов RIDF. Для подсчета сигнатуры, аналогично предыдущему способу, подсчитывается вес каждого слова по формуле tf(d)×ridf(t), выбираются несколько слов с наибольшим весом, слова упорядочиваются по алфавиту и сцепляются в строку, для которой вычисляется контрольная сумма CRC32.

Сигнатура $TF*IDF_{Opt}$ является модифицированной версией сигнатуры TF*IDF. Модификация заключается в изменении принципа подсчета значения IDF на основе так называемой «оптимальной частоты».

Сигнатура, построенная на основе двух самых длинных предложений, позволяет довольно хорошо находить похожие документы [1]. Для этого в тексте находятся два самых длинных предложения и сцепляются в одну строку в алфавитном порядке. Для строки подсчитывается контрольный код CRC32, который и является сигнатурой. Надо заметить, что при неправильном определении границ предложений (например, при переносах на другую страницу или для больших списков) сигнатура может не определить похожие документы.

Сигнатура двух самых «тяжелых» предложений строится по аналогичному принципу [1]. Из текста выбираются два предложения. Но предложения выбираются на основе суммы весов

(рассчитываемых по формуле tf(d)×idf(t)) входящих в нее слов. Два предложения с наибольшей суммой весов упорядочиваются по алфавиту, сцепляются в одну строку, для которой подсчитывается контрольный код CRC32.

Сигнатура I-Match строится на основе вычисления значения функции, предложенной в [3]. Для всей коллекции документов составляется словарь (список) слов, имеющих среднее значение IDF (слова со слишком большими или маленькими значениями IDF в список не включаются). Для каждого документа формируется множество слов и определяется его пересечение со словарем. При превышении пересечением некоторого порога, для множества слов вычисляется хеш-функция SHA1 (I-Match сигнатура).

Рассмотренные сигнатуры позволяют представить документ одним числовым значением. Понятно, что чем меньше сигнатура, тем менее она подходит для определения нечетких дубликатов текстов. Причина проста — если изменения в тексте сказываются на результирующем значении сигнатуры, то оценить меру похожести текстов уже не удастся. Решением проблемы может являться расширение сигнатуры, то есть включение в нее ограниченного ряда значений неких показателей. В этом случае, похожесть документов может определяться как некая мера (функция) близости двух сигнатур. Чаще всего это совпадение части значений сигнатур похожих документов [4].

Надо заметить, что длина сигнатур должна быть, по возможности, фиксированной и не превышать нескольких килобайт. Дело в том, что при различной длине сигнатур усложняется функция вычисления их близости, а при большой длине – повышается трудоемкость вычислений.

Для вычисления сигнатуры MegaShingles для всего множества шинглов документа производится вычисление 84 различных хэш-функций. Далее по критерию максимума или минимума каждой функции выбираются 84 шингла, которые разбиваются на 6 групп, по каждой из которых строятся 6 супершинглов. Сигнатура состоит из 15 чисел (мегашинглов), представляющих собой все возможные парные сочетания указанных 6 супершинглов.

Неплохие результаты показала сигнатура, построенная на основе словаря опорных слов [5]. Суть ее сводится к следующему. Из всего словаря, построенного по коллекции, выбираются несколько тысяч слов, наиболее качественно описывающих документы. Для каждого документа высчитывается бинарный вектор, длина которого равна количеству опорных слов. Для каждого опорного слова подсчитывается значение TF. В случае, если значение превышает некоторый порог, в соответствующую позицию вектора записывается 1, в противном случае 0. Сигнатура документа представляет собой указанный бинарный вектор. В отличие от ранее описанных сигнатур, мера схожести документов вычисляется как мера совпадения бинарных векторов (при этом совпадение не обязательно должно быть полным). Сигнатура отличается большой длиной. Так, для 4000 опорных слов длина сигнатуры составляет 500 байт. Фактически сигнатура является результатом фиксации векторной модели по длине словаря. Тем не менее, она сохраняет суть сигнатур — представление документа числовым значением фиксированной длины

Так как различные документы могут содержать сноски, комментарии, колонтитулы и т.д. использование одной сигнатуры, например, контрольной суммы, может не дать должного результата. По этой причине предлагается использовать наборы сигнатур, включающие контрольные суммы документов, сигнатуры на основе наиболее частотных слов, сигнатуры на основе самых длинных или самых значимых предложений.

Такое сочетание позволяет нивелировать влияние различных факторов и обеспечивает хорошую полноту поиска полных дубликатов.

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

Литература

1. Зеленков Ю.Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для WEB-документов // Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL’2007: сб. работ участников конкурса –Переславль-Залесский, 2007. – Т. 1. – С. 166-174.

2. 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 (3-6 сентября 2019 г. Самара, Россия). – Самара: ООО “Офорт”, 2019. Том 2. С. 335-339.

3. Kolcz A., Chowdhury A., Alspector J. Improved Robustness of Signature-Based Near-Replica Detection via Lexicon Randomization // KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004. – P. 605-610.

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) [Электронный ресурс] : материалы Двенадцатой междунар. конфер, 1–3 окт. 2019 г., Москва. – М.: ИПУ РАН, 2019. С. 1009-1011.

5. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // WWW Conference 2002.