ОБНАРУЖЕНИЕ НЕЧЕТКИХ ДУБЛИКАТОВ ТЕКСТОВ В БОЛЬШИХ МАССИВАХ ИНФОРМАЦИИ С ПОМОЩЬЮ СИГНАТУР СОДЕРЖАНИЯ

Шарапова Е.В., Шарапов Р.В.

Владимирский государственный университет, Россия, г. Муром, ул. Орловская д.23
mivlgu@mail.ru, info@vanta.ru

Аннотация: В работе рассматриваются вопросы применения сигнатур содержания для обнаружения похожих документов в больших массивах информации. Суть сигнатурных методов сводится к представлению документа неким кодом, позволяющим с высокой степенью вероятности выявить одинаковые документы.

Ключевые слова: нечеткий дубликат, сигнатура, текст, содержание.

Введение

Нечеткие дубликаты представляют собой тексты, подвергшиеся некоторой корректировке и видоизменению, но сохранившие смысловое сходство с оригиналами [1].

Существует большой круг задач, требующих обнаружения текстов с похожим содержанием. Системы поиска плагиата активно используются в издательствах и учебных учреждениях. Системы проверки уникальности текстов используются при заказе услуг копирайтеров. Новостные агрегаторы определяют похожие новости, группируют их в единые цепочки и удаляют дубли. Системы борьбы с почтовым спамом отслеживают появление большого количества электронных писем со схожим содержанием. По этой причине задача обнаружения текстов, значительно похожих друг на друга (нечетких дубликатов) является достаточно актуальной.

Интерес представляет использование подходов, позволяющих осуществлять поиск подобных документов с минимальным числом операций сравнения. Этого можно достичь путем использования различного рода сигнатур, описывающих содержание документа в виде одного или группы чисел.

1 Сигнатуры содержания

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.

Сигнатуры могут строиться по различным принципам. Чаще всего за основы берется содержание документа (отдельные слова или взаимосвязанные цепочки слов). Можно выделить следующие сигнатуры [2]:

1) контрольная сумма документа (CRC, MD5),
2) сигнатуры, вычисленные по набору наиболее частотных слов (на основе формул TF, TF*IDF, TF*RIDF);
3) сигнатуры на основе самых длинных или значимых предложений;
4) сигнатуры, рассчитанные на основе I-Match функции;
5) сигнатуры, базирующиеся на методе шинглов и его модификациях (мегашинглы, супершинглы);
6) сигнатуры, построенные на основе словаря опорных слов;
7) сигнатура Рабина для подсчета нечетких контрольных сумм документов;
   сигнатура Winnowing для получения «отпечатков пальцев» документов и т.д.

В связи с тем, что большинство сигнатур описывает содержание документов, то изменение текста может существенно снизить возможность их использования, так как вероятность того, что изменения повлияют на значение сигнатуры – достаточно велика. По этой причине, для обнаружения нечетких дубликатов, лучше всего подходят составные и нечеткие сигнатуры, позволяющие определить схожие документы по частичному совпадению значений сигнатур.

В ходе экспериментов был предложен следующий метод. По аналогии с [3] 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 качестве меры схожести может быть использовано количество совпадающих ненулевых значений в векторах документов, соотнесенное к общему количеству ненулевых элементов в этих векторах. Мера близости двух документов может быть представлена в следующем виде:

$$w(d_1, d_2) = 2\frac{\left|\{t_i^1 \in d_1, t_i^2 \in d_2 \,|\, t_i^1 = 1, t_i^2 = 1\}\right|}{\left|\{t_i^1 \in d_1 \,|\, t_i^1 = 1\}\right| + \left|\{t_i^2 \in d_2 \,|\, t_i^2 = 1\}\right|}. \quad (1)$$

где d_1, d_2 – бинарные векторы для документов 1 и 2, t_i^1, t_i^2 – значение i-го элемента бинарного вектора для документов 1 и 2 соответственно.

Чем ближе значение меры близости $w(d_1, d_2)$ к единице, тем больше документы похоже друг на друга.

2 Результаты исследования

Для проведения тестирования были собраны базы рефератов из нескольких источников:

1) Коллекция 15000 Рефератов – 18587 документов;
2) Сайт studentbank.ru – 71193 документов;
3) Сайт bestreferat.ru – 360534 документов.

Итого 450314 документов с общим размером текстовой части 30 Гб. Размер файла сигнатур на основе словаря опорных слов составил 225 Мбайт. Файлы сигнатур, вычисленных по набору наиболее частотных слов, самых длинных и значимых предложений имеют размер по 3 Мбайта каждый.

Для оценки качества работы использовались следующие метрики: полнота, точность, F-мера.

Для тестирования метода было выполнено два прогона с различными значениями порогового значения меры сходства. Для обнаружения полных дубликатов использовалось пороговое значение 0.95, для нечетких дубликатов со значительными изменениями 0.5. С обнаружением полных дубликатов справились все сигнатуры. Результаты обнаружения нечетких дубликатов приведены в таблице 1.

Таблица 1. Результаты тестирования для нечетких дубликатов

Сигнатура	**Полнота**	**Точность**	**F-мера**
CRC	0,11	1	0,19
TF	0,74	0,78	0,76
TF*IDF	0,70	0,82	0,75
TF*RDF	0,71	0,84	0,77
Самые длинные предложения	0,69	0,79	0,74
Самые значимые предложения	0,63	0,81	0,70
Словарь опорных слов	0,86	0,72	0,78

Эффективное обнаружение идентичных по содержанию документы возможно путем использования практически любых сигнатур. Сложности возникают при поиске частично похожих документов. Различные сигнатуры позволяют либо добиться высокой точности поиска, либо высокой полноты. Так, сигнатура контрольной суммы документа обеспечивает точность 1 при полноте всего в 0,11.

Сигнатуры, вычисленные по набору наиболее частотных слов (TF, TF*IDF, TF*RIDF) дают достаточно сбалансированное соотношение полноты и точности поиска. Лучше всего показала себя сигнатура, построенная на основе комбинации частоты слов TF и остаточной обратной частоты документов RIDF. При точности 0,84 она обеспечивает полноту в 0,71. Недостатком указанной группы сигнатур является ограниченное число учитываемых слов (чаще всего 6 слов). Это приводит к тому, что документы, посвященные одной тематике, могут иметь одинаковые сигнатуры и ошибочно признаваться нечеткими дубликатами. Кроме того, даже незначительные изменения

документов могут привести к перестановке в сигнатуре слов с близким расположением частот, что приведет к ошибочному признанию документов различными.

Сигнатуры, основанные на цепочках самых длинных или наиболее значимых предложений (с точки зрения веса входящих в них слов по формуле TD*IDF) показывают неплохую точность, но при изменениях и компиляциях текстов часто не позволяют обнаруживать похожие документы.

Сигнатура, построенная на основе словаря опорных слов, показала наиболее сбалансированное значение полноты и точности (о чем свидетельствует наибольшее значение F-меры). По своей сущности она похожа на сигнатуры наиболее частотных слов, но при этом рассматривается существенно больший набор из нескольких тысяч слов. Этим достигается исключение влияния изменения частот на качество обнаружения дубликатов.

Заключение

Использование таких популярных методов, как n-граммы и шинглы для больших объемов данных приводит к значительным вычислительным затратам. Так, число операций сравнения для двух документов в тысячу слов будет исчисляться миллионами. В связи с тем, что современные коллекции содержат в себе сотни тысяч, а иногда миллионы документов, число операций сравнения одного документа с текстовой коллекцией будет исчисляться триллионами операций! Это слишком много для того, чтобы обеспечить высокое быстродействие. Несмотря на то, что существуют подходы, позволяющие сократить число операций сравнения на несколько, время сравнения все равно остается слишком большим.

С этой точки зрения, сигнатуры содержания обеспечивают значительно лучшую производительность. Например, для текстовой коллекции в 1 Тбайт, размер большинства сигнатур составит около 100 Мбайт, на обработку которых уйдет менее 1 секунды. Размер файла сигнатуры, построенной на основе словаря опорных слов, в таком случае составит около 7 Гбайт. При современном развитии компьютерной техники такой объем можно разместить в оперативной памяти, что позволит существенно повысить скорость поиска по сигнатуре.

Таким образом, для больших объемов данных, использование сигнатур содержания является основным средством быстрого поиска нечетких дубликатов.

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

Литература

1. 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.

2. Зеленков Ю. Г., Сегалович И.В. Сравнительный анализ методов определения нечетких дубликатов для Web–документов // Тр. 9-й Всеросс. научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Переславль-Залесский: Изд-во ИПС РАН, 2007. – С. 166-174.

3. Ilyinsky S., Kuzmin M., Melkov A., Segalovich I. An efficient method to detect duplicates of Web documents with the use of inverted index // Proc. of the 11th International World Wide Web Conference, WWW 2002, Honolulu, Hawaii, USA, ACM, New York.