К проблеме обнаружения нечетких дубликатов в потоках сообщений

Е.В. Шарапова

Федеральное государственное бюджетного образовательного учреждения высшего образования "Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых", E-mail: info@vanta.ru

Рассматриваются подходы к обнаружению похожих сообщений, передаваемых в непрерывных потоках. Анализируется трудоемкость различных методов. Наибольшие перспективы с точки зрения производительности имеют сигнатурные методы, представляющие сообщение числовым кодом. Сравнение сообщений с помощью сигнатур требует достаточно небольшое количество вычислительных ресурсов.

Approaches to the detection of similar messages transmitted in continuous streams are considered. Analyzed the complexity of various methods. The most promising in terms of performance are signature methods that represent a message with a numeric code. Comparison of messages using signatures requires a rather small amount of computational resources.

Введение

В настоящее время в связи с развитием средств телекоммуникаций наблюдается активный поток сообщений различного вида, используемых пользователями для обмена информацией и общения. Сюда можно отнести электронные письма, SMS сообщения, новостные посты, сообщения в приложениях Viber, WhatsApp и т.д. Несмотря на явное удобство подобных средств общения, с их использованием возникает ряд проблем. Во-первых, появляются нежелательные сообщения, рекламирующие какие-либо продукты, услуги и т.д. Чаще всего такие сообщения называют спамом. Во-вторых, при работе с новостными потоками, возникает задача фильтрации новостей, похожих друг на друга и поступающих из различных источников. В обоих случаях сообщения могут иметь небольшие отличия по содержанию, т.е. являются не полными копиями, а нечеткими дубликатами друг друга [1].

При значительных потоках информации, на задачу обнаружения похожих сообщений налагаются ограничения по вычислительным ресурсам и, соответственно, на время обработки одного сообщения. Так, при потоке в 100 электронных писем в секунду, время обработки одного сообщения не должно превышать 1/100 секунды.

Таким образом, возникает задача экспресс-оценки сообщений на подобие, позволяющая за короткое время обнаруживать нечеткие дубликаты сообщений.

Методы обнаружения нечетких дубликатов

Существует множество подходов к обнаружению похожих сообщений. Метод n-грамм предполагает разбиение текста на элементы одинаковой длины, например, по 3 символа (триграммы) [2]. Тогда, чем больше n-грамм в двух сообщениях совпадают, тем более похожи документы. Преимущество метода заключается в его простоте и возможности искать похожие документы даже на незнакомых языках. Недостатком является его трудоемкость – число n-грамм будет сравнимо с длиной сообщения, а число сравнений – произведению длин обоих сообщений [3].

В методе шинглов предлагается сцеплять в строку цепочки из нескольких слов (обычно 5), для которой вычисляется хэш-код [4]. Полученный таким образом набор хэш-кодов представляет собой описание содержания документа. Сравнение наборов хэш-кодов для различных документов позволяет оценить меру их близости – чем

большее количество хэш-кодов совпадает, тем больше документы похожи друг на друга. Основным недостатком метода является прямая зависимость числа сравнению групп хэш-кодов от их количества. Например, если первое сообщение представлено m хэш-кодами, а второе n, то число операций сравнения будет лежать в диапазоне от m до m×n.

Большую популярность получила группа методов, основанная на оценке частоты слов в сообщениях (например, TF*IDF). В этом случае сообщение представляется с помощью вектора, содержащего значения TF*IDF каждого терма из словаря [5, 6]. Проблема заключается в том, что словарь, построенный по коллекции, может содержать более ста тысяч термов, и размер вектора будет достаточно большим. В отличие от метода щинглов, размеры TF*IDF вектора для каждого документа будет неизменным, а сам вектор будет сильно разреженным (содержать 0 для слов, не входящих в документ). Тогда число операций сравнения будет равно размеру словаря.

Сигнатурные методы обнаружения нечетких дубликатов

Сигнатурные методы заключаются в представлении содержания сообщений неким кодом (хэш-функцией, числом, контрольной суммой), позволяющим с высокой степенью вероятности выявить одинаковые (или практически одинаковые) сообщения. Фактически, сравнение сообщений сводится к сравнению нескольких чисел – сигнатур. 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 сообщения (отдельные слова или взаимосвязанные цепочки слов). Можно выделить следующие сигнатуры [7]:

- контрольная сумма документа (CRC, MD5),
- сигнатуры, вычисленные по набору наиболее частотных слов (на основе формул TF, TF*IDF, TF*RIDF);
- сигнатуры на основе самых длинных или значимых предложений;
- сигнатуры, рассчитанные на основе I-Match функции [8];
- сигнатуры, базирующиеся на методе шинглов и его модификациях (мегашинглы, супершинглы);
- сигнатуры, построенные на основе словаря опорных слов [9];
- 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.

Выводы

Сигнатурные методы являются достаточно перспективным решением для определения похожих сообщений в непрерывных потоках. Они требуют небольших вычислительных ресурсов и занимают достаточно немного памяти. Конечно, сигнатуры не позволяют оценить степень подобия сообщений, но достаточно неплохо справляются с обнаружением слегка измененных данных. Кроме того, сигнатуры без

затруднений могут использоваться в системах фильтрации почтового спама, новостных агрегаторах и т.д.

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. Sidorov G., Velazquez F., Stamatatos E., Gelbukh A., Chanona-Hernández L. Syntactic Dependency-based n-grams as Classification Features // Lecture Notes in Artificial Intelligence, 2012, vol. 7630, p. 1–11.
3. Гастфилд Д. Строки, деревья и последовательности в алгоритмах. - СПб.: Невский диалект, 2003.
4. Broder A., Glassman S., Manasse M., Zweig G. Syntactic clustering of the web // Computer Networks and ISDN Systems, 1997, vol. 29, n. 8, p. 1157–1166.
5. Salton G., McGill M. J. Introduction to modern information retrieval. -McGraw-Hill, 1983. 448 p.
6. Salton G., Wong A., Yang C. S. A vector space model for automatic indexing // Communications of the ACM, 1975, vol.18, n.11, p.613-620.
7. 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.
8. Chowdhury A., Frieder O., Grossman D., McCabe M. Collection statistics for fast duplicate document detection // ACM Transactions on Information Systems (TOIS), 2002, vol. 20, issue 2, p. 171-191
9. 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.