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

Структура поискового индекса системы обнаружения нечетких дубликатов текстов

При построении систем обнаружения дубликатов текстов одним из ключевых параметров, влияющим на производительность, является организация поискового индекса. Значение имеют такие параметры, как размеры индекса и скорость доступа к данным.

Исследования показали, что использование в качестве поискового индекса различного рода СУБД (MySQL, Oracle, MSSQL Server и т.д.) для больших коллекций документов не обеспечивает требуемого уровня производительности - время доступа к данным достаточно большое.

Для хранения поисковых индексов могут использоваться текстовые или структурированные бинарные файлы. При хранении индекса в виде текстовых файлов с индексом легко работать, но обработка проводится построчно. Это замедляет скорость работы с индексом. Во втором случае можно добиться высокой скорости операций чтения индекса и перемещения по нему. По этой причине использование структурированных бинарных файлов для хранения поискового индекса кажется наиболее привлекательным вариантом.

Рассмотрим структуру поискового индекса, применяемого в система Автор.NET более подробно. Поисковый индекс можно разделить на несколько групп файлов:

1. Термы.

В группу входит словарь термов (слов, встречающихся в документах), список стоп слов (наиболее часто встречающихся слов из словаря, которые являются не информативными и не рассматриваются при поиске), IDF индекс.

IDF индекс представляет собой значения инверсных частот документов, подсчитанных для каждого терма (слова). Инверсных частот вычисляется по формуле [1]:

$$\mathrm{idf}(t) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

где $|D|$ – число документов в коллекции, $|\{d_i \in D | t \in d_i\}|$ – число документов, в которых встречается слово t.

2. Документы.

Группа включает в себя список всех внесенных в систему документов, с указанием их уникальных идентификаторов, расположением в системе и сети Интернет, языка документа, числа слов в нем, принадлежность к той или иной коллекции и т.д., а также полные тексты документов в текстовом формате в кодировке Unicode.

3. TF*IDF индекс.

Группа включает в себя индексы значений частоты термов TF и TF*IDF, подсчитанные для каждого слова в каждом документе. Значение TF вычисляется по формуле:

$$\mathrm{tf}(t) = \frac{n_t}{\sum_k n_k}$$

где n_t – число вхождений слова t в документ, $\sum_k n_k$ – общее количество слов в документе.

Значение TF*IDF представляет собой произведение значений tf(d)*idf(t).

4. Сигнатуры.

В группу входят сигнатуры, подсчитанные для каждого документа – контрольная сумма CRC32 и MD5, сигнатуры шести самых значимых слов по значениям TF, TF*IDF, TF*RIDF [2, 3], сигнатуры самых длинных и самых значимых предложений, а также «длинная» сигнатура опорных слов [4].

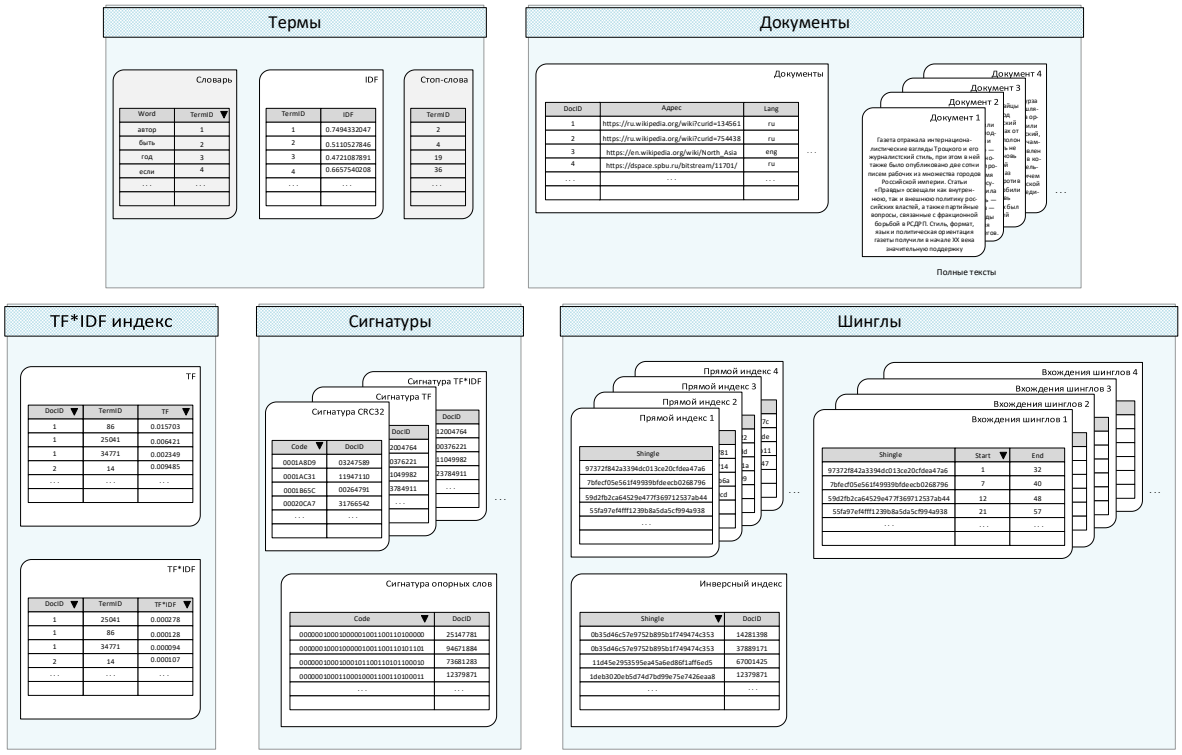

Рис. 1. Структура поискового индекса.

5. Шинглы.

Группа включает в себя наборы прямых индексов шинглов [5] (построенная по порядку следования щинглов в документе) для каждого документа, индексы вхождения шинглов в каждом документе (начальная и конечная позиция в тексте), а также инверсный индекс, сортированный по значениям шинглов.

Предложенная структура позволяет разбить индекс на независимые части и работать с каждой из их по отдельности, в зависимости от выполняемых задач. Кроме того, размеры индексов сигнатур достаточно малы, что позволяет хранить их в оперативной памяти для обеспечения быстрой обработки.

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

Литература

1. Sharapova E.V., Sharapov R.V. Detection of Fuzzy Duplicate Texts in News Feeds // 2019 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Russia, 2019, pp. 1-5.

2. 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.

3. 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.

4. 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.

5. 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.