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 распределенной системы поиска текстовых заимствований *

Исследования показали, что для обеспечения достаточной производительности системы проверки оригинальности текстов необходимо использовать распределенную модульную структура [1, 2, 3]. При этом каждый модуль призван выполнять определенный круг задач и в значительной степени быть независимым от других.

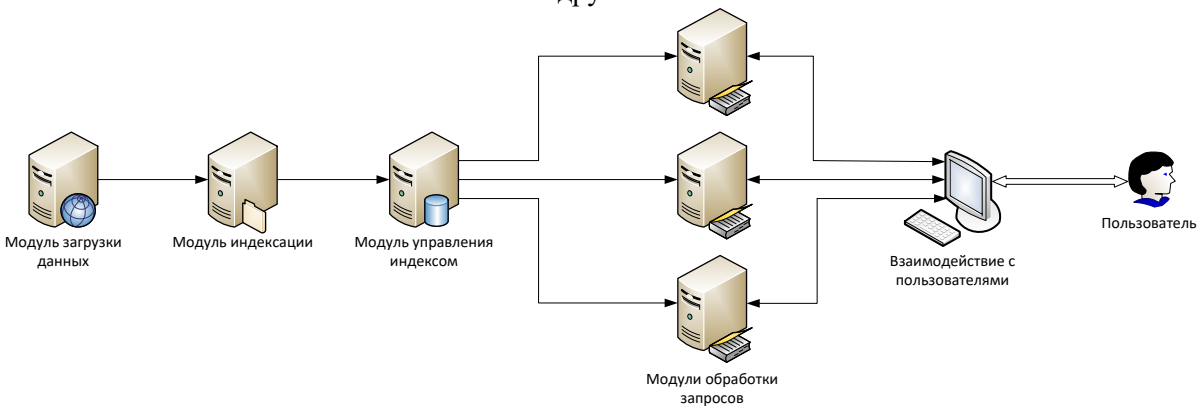

Рис. 1. Структура распределенной системы поиска заимствований.

1. Модуль загрузки данных

Основное назначение модуля – сбор данных (документов) для наполнения поискового индекса. Основным источником данных является сети Интернет. Модуль выполняет следующие действия:

- Загружает документы с вебсайтов
- Преобразует их в машиночитаемый формат (в нашем случае в обычный текст без разметки)
- При необходимости производит перевод текста в единую кодировку (UTF-8)
- Производит очистку полученного текста от элементов разметки, спецсимволов, выполняет слияние слов с переносами, разорванных переходами на новую страницу предложений

2. Модуль индексации

Модуль получает загруженные документы и производит их индексацию. При этом осуществляется подсчет сигнатур, вычисление TF*IDF индекса, набора шинглов и т.д. Модуль содержит средства нормализации входящих в документ слов, фильтрации стоп-слов (наиболее часто встречающихся слов во всей коллекции документов).

Модуль позволяет искать уже имеющиеся копии добавляемого документа в поисковом индексе, а также обновлять поисковый индекс более свежими версиями документов.

3. Модуль управления индексом

Пополнение поискового индекса в реальном времени чаще всего представляет собой сложную задачу, требующую больших вычислительных затрат, что оказывает влияние на производительность всей системы. В первую очередь это связано с необходимостью перестройки упорядоченных индексов, размеры которых могут составлять сотни гигабайт [4].

По этой причине чаще всего поисковый индекс обновляется пакетно. Иногда, для ускорения обновления индекс разбивается на две части – основной статический (большой) индекс, и дополнительный динамический (обновляемый) индекс, в который записываются все вновь добавленные документы. При очередном цикле обновления данные из дополнительного индекса переносятся в основной индекс, а дополнительный индекс очищается в ожидании новых данных.

Всеми операциями создания и модификации основного и дополнительного индексов занимается модуль управления индексом. Он также производит проверку индексов, удаления из них дублирующихся данных, производит подсчет статистики и т.д. Итогом работы модуля являются готовые индексные файлы, передаваемые системе обработки запросов.

4. Модули обработки запросов

Для повышения производительности системы предлагается вместо единого хранилища индексов использовать распределенное хранение. При этом, в зависимости от нагрузки к системе можно производить работу с поисковым индексом либо целиком в каждом модуле, либо разбив его на части (каждый модуль работает со своей частью индекса – шинглами, сигнатурами, TF*IDF и т.д.) [5].

Число модулей обработки запросов зависит от количества обращений к системе проверки. Это позволяет добиться масштабируемости системы без снижения ее производительности.

5. Интерфейс взаимодействия с пользователями

Интерфейс взаимодействия с пользователями предполагает получение документов для проверки от пользователей, их предварительную обработку, отправку запросов на поиск похожих документов модулям обработки запросов, получение результатов их работы и выдачу отчетов пользователям.

Литература

1. Sharapova E.V. Computational load reduction of fuzzy duplicate detection in large amounts of information // IOP Conference Series: Materials Science and Engineering, vol.734, 2020. 012119.

2. Sharapova E. Increase the Speed of Search Index in the Duplicate Text Detection Systems // 2020 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), Svetlogorsk, Russia, 2020, pp. 1-5.

3. Sharapova E.V. One way to solve the problem of fuzzy text duplicates detection // Materials of the International Conference “Scientific research of the SCO countries: synergy and integration” - Reports in English (July 12, 2019). Beijing, PRC, 2019. P.59-64.

4. Шарапова Е.В. Структура поискового индекса системы обнаружения нечетких дубликатов текстов // Наука и образование в развитии промышленной, социальной и экономической сфер регионов России. XII Всероссийские научные Зворыкинские чтения: сб. тез. докл. Всероссийской межвузовской научной конференции. Муром, 7 февр. 2020 г.– Муром: Изд.-полиграфический центр МИ ВлГУ, 2020. – [Электронный ресурс]: 1 электрон. опт. диск (CD-ROM). С128-129.

5. Шарапова Е.В. Обнаружение нечетких дубликатов текстов в больших массивах информации с помощью сигнатур содержания // Управление развитием крупномасштабных систем MLSD'2019 Материалы двенадцатой международной конференции (1–3 октября 2019 г., Москва). – М. Институт проблем управления им. В.А. Трапезникова, 2019. С. 1009-1011.

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