О ПЛАТФОРМЕ ДЛЯ ЭЛЕКТРОННОЙ ТЕКСТОВОЙ БИБЛИОТЕКИ

Веретенников А. Б.

Уральский государственный университет им. А. М. Горького

alexander@veretennikov.ru

Автором разработана новая структура данных CLB-дерево [5,7] предназначенная для создания индекса для поиска в большом массиве текстовых документов. В частности подобные задачи решаются поисковыми системами в сети Интернет. Например, указав несколько слов, пользователь получает набор документов, которые содержат искомые слова. Часто использующиеся для решения подобных задач инвертированные файлы трудно обновлять. В отличие от них CLB-дерево обладает возможностью быстрого добавления в индекс новых данных, что весьма актуально при работе с часто обновляющимися коллекциями текстовых документов.

При разработке автор применяет подход «модель — алгоритм — программный комплекс». На основании созданной структуры данных и алгоритмов разработан программный комплекс, предназначенный для поиска в большом массиве текстовых документов. Также разработан алгоритм поиска похожих документов [8].

Далее приведены результаты одного из экспериментов, показывающие возможность быстрого обновления индекса. Структура эксперимента:

1) Создание индекса для базового массива документов.
2) Добавление в индекс небольшого файла.
3) Добавление в индекс 1, 2, 3, ..., 9 Мб данных.
4) Добавление в индекс 10, 20, 30, ..., 90 Мб данных.
5) Добавление в индекс 100, 200, 300, ..., 900 Мб данных.
6) Добавление в индекс 1000, 2000, 3000, ..., 10000 Мб данных.

После добавления в индекс каждой порции данных сохраняется весь кеш и программа создания индекса завершает работу, затем она снова запускается и добавляет следующую порцию данных. При работе с файлом, содержащим кластеры, кеширование операционной системы отключается. Таким образом мы проверяем скорость обновления индекса для данных разного размера.

Эксперименты проводились на следующей конфигурации:

Процессор: Intel Core i7 920, 2.66 GHz, оперативная память: 12 Гб, DDR3, жесткий диск: Seagate Barracuda ST32000641AS, 7200.12, 7200 RPM, кэш 64 Мб, объем 2 Тб, ОС: Microsoft Windows 2008 Enterprise x64 Edition with Service Pack 2.

Все файлы представляли собой обычный текст. В следующей таблице в первой колонке указан размер добавляемых данных, далее указано полное время, затраченное на

добавление данных, которое складывается из времени чтения файлов и времени создания индекса. Далее указан размер прочитанных и записанных кластеров (см. описание алгоритма в [5,7]). Время указано в формате час:мин:сек. Общее затраченное время 10 час. 9 мин.

Размер	Время	Время создания индекса	Чтение кластеров	Запись кластеров
85.23 Гб	03:29:31	01:28:32	7.14 Гб	56.62 Гб
1.41 Кб	00:01:08	00:01:04	4.46 Гб	4.46 Мб
1.08 Мб	00:02:03	00:01:58	4.46 Гб	321.58 Мб
2.17 Мб	00:02:20	00:02:15	4.46 Гб	458.27 Мб
3.09 Мб	00:02:41	00:02:36	4.47 Гб	575.57 Мб
4.03 Мб	00:02:46	00:02:41	4.46 Гб	664.48 Мб
5.03 Мб	00:02:53	00:02:48	4.46 Гб	721.57 Мб
6.25 Мб	00:03:01	00:02:56	4.47 Гб	790.28 Мб
7.31 Мб	00:03:06	00:03:01	4.47 Гб	826.96 Мб
8.21 Мб	00:03:05	00:03:00	4.47 Гб	845.46 Мб
9.19 Мб	00:03:12	00:03:06	4.47 Гб	900.57 Мб
10.02 Мб	00:03:18	00:03:12	4.47 Гб	931.06 Мб
10.02 Мб	00:03:17	00:03:12	4.47 Гб	929.92 Мб
20.03 Мб	00:03:32	00:03:26	4.47 Гб	1.10 Гб
30.37 Мб	00:03:51	00:03:44	4.48 Гб	1.21 Гб
40.19 Мб	00:04:05	00:03:58	4.48 Гб	1.30 Гб
50.60 Мб	00:04:18	00:04:09	4.49 Гб	1.35 Гб
60.17 Мб	00:04:23	00:04:14	4.49 Гб	1.38 Гб
70.26 Мб	00:04:27	00:04:17	4.51 Гб	1.41 Гб
80.46 Мб	00:04:37	00:04:27	4.50 Гб	1.44 Гб
90.01 Мб	00:04:39	00:04:29	4.50 Гб	1.46 Гб
100.09 Мб	00:04:43	00:04:31	4.51 Гб	1.49 Гб
200.00 Мб	00:05:43	00:05:09	4.60 Гб	1.73 Гб
300.00 Мб	00:06:25	00:05:39	4.66 Гб	1.84 Гб
400.44 Мб	00:07:22	00:06:24	4.74 Гб	2.02 Гб
500.13 Мб	00:07:51	00:06:43	4.80 Гб	2.20 Гб
600.01 Мб	00:08:13	00:06:59	4.85 Гб	2.22 Гб
700.90 Мб	00:08:49	00:07:26	4.99 Гб	2.41 Гб
800.62 Мб	00:09:12	00:07:43	4.95 Гб	2.44 Гб

900.24 Мб	00:09:39	00:07:56	5.07 Гб	2.54 Гб
1000.54 Мб	00:09:56	00:08:10	5.12 Гб	2.71 Гб
1.95 Гб	00:13:37	00:10:27	5.55 Гб	3.72 Гб
2.93 Гб	00:17:17	00:12:25	5.97 Гб	4.89 Гб
3.90 Гб	00:21:02	00:14:32	6.52 Гб	6.11 Гб
4.88 Гб	00:24:23	00:16:23	6.89 Гб	7.00 Гб
5.85 Гб	00:29:08	00:19:52	7.41 Гб	8.09 Гб
6.83 Гб	00:32:36	00:21:29	7.75 Гб	8.95 Гб
7.81 Гб	00:35:49	00:23:21	8.26 Гб	9.92 Гб
8.78 Гб	00:38:37	00:24:48	8.24 Гб	10.63 Гб
9.76 Гб	00:42:26	00:27:06	8.74 Гб	11.71 Гб

С помощью дополнительных модулей реализована обработка файлов в различных форматах, например RTF, PDF, CHM, HTML, DJVU и кодировках, например UNICODE, UTF8, CP1251, ASCII, KOI8. Поддерживается обработка архивов форматов ZIP, CAB, RAR, 7Z, ARJ, TAR, и др. Разработан API для доступа к системе на базе COM.

В дальнейшем планируется на основании созданного программного комплекса разработать универсальную платформу для обработки электронных текстовых библиотек, включающую в себя различные компоненты, такие как:

- WEB сервер для доступа к содержимому библиотеки, позволяющий осуществлять поиск в библиотеке и использовать другие компоненты (предполагается использование современных технологий AJAX, HTML5).
- Распознавание текста в графических документах.
- Поиск похожих документов.
- Эффективная обработка XML файлов с применением XML схем и XSLT.
- Автоматическая классификация документов.
- Извлечение текстовой информации из мультимедийных данных (аудио, видео).
- Трансформация документов в другие форматы по запросу (например PDF, RTF).
- Средства для предоставления доступа к библиотеке по различным протоколам (FTP, BitTorrent, SMB).
- Модуль индексации баз данных.
- Модуль индексации документов в сети Интернет.
- Модули индексации документов, используя как источник FTP.
- Дополнительные API для доступа к системе, на базе WEB сервисов и JAVA.
- SQL подобный язык запросов.

Отдельной задачей является разработка распределенной системы, которая может быть развернута на произвольном количестве вычислительных узлов.

Список литературы:

1. Веретенников А. Б., Лукач Ю. С. CLB-деревья: новый способ индексации больших массивов текстов. Международная алгебраическая конференция: К 100-летию со дня рождения П. Г. Конторовича и 70-летию Л. Н. Шеврина. Тез. докл. Екатеринбург: Изд-во Урал. ун-та, 2005, с. 173-175.
2. Веретенников А. Б., Лукач Ю. С. Еще один способ индексации больших массивов текстов. Известия Уральского государственного университета. Серия «Компьютерные науки», 2006. №43. с. 103-122.
3. Веретенников А. Б. Новый подход к быстрому выделению памяти в программах на C++. Проблемы теоретической и прикладной математики: Труды 37-й Региональной молодежной конференции. Екатеринбург: УрО РАН, 2006, с. 413-417.
4. Веретенников А. Б. Эффективное создание текстовых индексов. Проблемы теоретической и прикладной математики: Труды 39-й Всероссийской молодежной конференции. Екатеринбург: УрО РАН, 2008. с. 348-350.
5. Веретенников А. Б. Создание легко обновляемых текстовых индексов. Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008». Дубна: ОИЯИ, 2008. с. 149-154.
6. Веретенников А. Б. Библиотека для создания оконных интерфейсов на любых скриптовых языках в операционной системе Windows. Информационно-математические технологии в экономике, технике и образовании: Тезисы докладов Третьей международной научной конференции. Екатеринбург: УГТУ-УПИ, 2008. с. 220-221.
7. Веретенников А. Б. Эффективная индексация текстовых документов с использованием CLB-деревьев. Системы управления и информационные технологии, 2009, 1.1(35). - С. 134-139.
8. Веретенников А. Б. Гибкий подход к проблеме поиска похожих документов. Проблемы теоретической и прикладной математики: Труды 40-й Всероссийской молодежной конференции. Екатеринбург: УрО РАН, 2009. с. 392-396.
9. Веретенников, А. Б. Сравнение эффективности CLB-дерева в 32-битных и 64-битных архитектурах. Материалы межвузовской научной конференции по проблемам информатики «СПИСОК 2009». Екатеринбург. 2009. с. 7-13.