УДК 004.657

А.П. ДИМИТРИЕВ, Т.А. ЛАВИНА
(Чебоксары)

СОПОСТАВЛЕНИЕ МЕХАНИЗМОВ ДОСТУПА К БАЗЕ ДАННЫХ ПРИ ОБРАБОТКЕ ТЕКСТОВЫХ ДАННЫХ

Обосновывается выбор подходящего механизма доступа к базе данных при обработке значительного объема текстовой информации. Производится сравнительный анализ форматов баз данных "dBase IV" и "Microsoft Access 2003" при прямом доступе к файлу и доступе с помощью SQL. Приложением является база данных автоматизированного перевода.

Ключевые слова: *база данных, SQL, студент, автоматизированный перевод, Microsoft Access.*

ALEXANDER DIMITRIEV, TATYANA LAVINA
(Cheboksary)

COMPARISON OF ACCESS MECHANISM TO DATABASE WHILE PROCESSING TEXT DATABASE

The article deals with the choice of the appropriate access mechanism to database while processing a reasonable amount of text database. There is given a comparison analysis of the format of the database "dBase IV" and "Microsoft Access 2003" with the direct access to the file and with the help of SQL. The application is the database of the computer-aided translation.

Key words: *database, SQL, student, computer-aided translation, Microsoft Access.*

Обучающиеся выпускных курсов в области информатики и вычислительной техники часто обладают недостаточным уровнем навыков программирования. Одной из причин является слабая мотивация изучения программирования [4]. Иногда мотивацией может послужить организация какого-либо соревнования. В данной статье проводится исследование по сопоставлению различных механизмов работы с базой данных, в том числе по скорости обработки, что может рассматриваться как соревнование. Такой подход реализует метод познания через сравнение как один из методов эмпирического уровня [6].

Для вышеуказанного исследования можно сгенерировать обезличенные тесты, которые будут нести не имеющую применения информацию узкому кругу учёных, и чтобы понимать, о чем речь, студенту потребуется подготовка. В уставе Чувашского государственного университета сообщается, что он создан для осуществления, наряду с образовательными, и научных функций [5]. Практика показывает, что из тысяч бывших студентов только единицы впоследствии имеют научные публикации, причем никто не планирует быть ученым. По результатам анкетирования 2015 г. в Чувашском государственном университете, только 25% выпускников планировали продолжить учебу в магистратуре и никто – в аспирантуре. Таким образом, в исследовании используем не абстрактные тесты, а знакомую большинству пользователей Интернета и широко востребованную предметную область – автоматизированный перевод.

В разработанном автором чувашско-русском автоматизированном переводчике применяется база данных (БД), содержащая сведения о частоте встречаемости русских словосочетаний [1, 3]. Для ее наполнения использованы тексты как из различных сфер жизни и областей деятельности: наука, футбол, право и т. д., так и из художественной литературы.

Процесс наполнения БД основан на использовании механизма BDE в Delphi. БД хранится в формате "dBase IV" (файл с расширением DBF с индексным файлом). В связи с тем, что этот формат

был выпущен в 1988 г., он морально устарел, и возникает вопрос, следует ли обучать студентов использованию этого формата. В настоящее время при выполнении лабораторных работ студенты занимаются импортом и экспортом в такой формат без объяснения, где он пригодится в будущей деятельности.

В данной работе предпринята попытка заменить формат “dBase IV” на формат “Microsoft Access 2003” (файл с расширением MDB). Как следует из года в названии, предлагаемый формат также не является новым. Однако на некоторых компьютерах успешно функционируют относительно старые операционные системы (ОС), такие как Windows XP, где нет встроенной поддержки еще более новых форматов “Microsoft Access”, таких как 2007 или 2010, но есть для Microsoft Access 2003.

Преимуществами применения файлов MDB по сравнению с DBF являются: поддержка более широкого спектра типов полей, автоматическая поддержка на всех ОС семейства Windows NT для персональных компьютеров, русификация, возможность парольной защиты.

С целью перехода к файлу MDB выполнен ряд следующих действий:

1. Произведен импорт файла DBF в Microsoft Access, необходимое поле проиндексировано.
2. В исходном тексте программы заменено имя компонента “Table1” на “ADOTable1”.
3. На форму вместо компонента типа “TTable” помещен компонент типа “TADOTable”, вместо “DataSource1” соответственно “ADOConnection1”.
4. В свойстве “ConnectionString” компонента “ADOConnection1” указан подключаемый файл MDB.

Итоговым преимуществом замены формата является уменьшение размера БД. Исходная БД данных занимала 109,5 Мб в трех файлах, а новая занимает 28,2 Мб в одном файле. Однако учитывая современные объемы жестких дисков, такие размеры БД не критичны.

Недостатком является многократное увеличение времени работы. Для тестирования использована часть первой главы диссертации автора в виде текстового файла размером 70,8 Кб, содержащий текст на русском языке с формулами.

В качестве аппаратной основы для исследования использованы компьютеры $П_1$ и $П_3$, аппаратные характеристики которых приводятся в работе [2], в которой, по аналогии с данной работой, производится их сопоставление при обработке некоторых задач. Время обработки текста с DBF-файлом составило 10 с на компьютере $П_1$, а с MDB-файлом – 177 с. На $П_3$ соответственно 14 с и 307 с.

Совокупный объем обработанного текста в работе [1] составлял более 72 Мб, что примерно в 1000 раз больше, чем вышеуказанный текст. Следовательно, в случае использования формата “Microsoft Access 2003” время обработки составило бы около $1000 \cdot 177$ с ≈ 2 сут. Время, затраченное при применении формата “dBase IV”, составляет $1000 \cdot 10$ с ≈ 2 ч, что говорит о колоссальной экономии машинного времени.

При обновлении формата увеличиваются и затраты оперативной памяти. Ранее при обработке указанного текста объем занимаемой процессом памяти увеличивался с 4,8 Мб до 22 Мб, а теперь до 130 Мб. Можно бороться с этим явлением, при обработке каждых 100 слов очищая память с помощью процедуры “TrimWorkingSet”. Тогда при работе приложения с новым форматом потребуется не более 18 Мб. Однако сначала требуется 54 Мб при загрузке вспомогательной БД, в то же время для формата dBase IV указанный объем составляет от 7 до 12 Мб.

Доступ к БД возможен не только напрямую. При работе с БД часто применяется клиент-серверная архитектура, основанная на запросах SQL. Данная архитектура обладает таким полезным в данном случае достоинством, как возможность многопользовательского режима. В маловероятном случае, если в перспективе задачей наполнения базы данных переводчика будут заниматься сотни добровольцев, то потребуется защита от злоумышленников, обеспечиваемая рядом SQL-серверов. Однако необходимо ориентироваться на факт, что такие запросы будут занимать больше времени. Для дока-

зательства данного утверждения использованы еще два файла, которые обладают размером меньшим, чем первый файл (см. табл.).

Показатели обработки файлов на компьютере $П_1$

Характеристика обработки	Файл 1	Файл 2	Файл 3
Число слов текста	10843	918	60
Объем, байт	72533	6410	490
Количество операций чтения БД	215662	21023	1301
Количество операций записи БД	189832	18464	1149
Количество операций позиционирования по закладкам	63877	6263	383
Количество операций индексного поиска	18969	1802	124
Время ЦП при использовании TADOQuery, с	–	–	32
Время ЦП при использовании TQuery, с.	–	11	1

Примечание. БД – база данных, ЦП – центральный процессор.

При этом не производились модификации данных, выполнялся лишь отбор (с помощью оператора “SELECT”), показывающий, что одного только замедления при чтении таблицы достаточно для отказа от запросов SQL. Так, при обработке только по чтению 72 Мб с использованием DBF-файла компонентом типа “TQuery” время обработки достигло бы

$$\frac{11\text{ с} \times 72\text{ Мб}}{6410\text{ б}} \approx 1{,}5$$

При использовании MDB-файла компонентом типа “TADOQuery” время должно достичь

$$\frac{32\text{ с} \times 72\text{ Мб}}{490\text{ б}} \approx 57\text{ сут}$$

Следует заметить, что использование флеш-накопителя с целью хранения базы данных недопустимо, во-первых, из-за многократного замедления при прочих равных условиях, а во-вторых, из-за быстрого износа. Одной из причин замедления при использовании компонента “TQuery” является создание в текущей папке временного файла для каждого запроса с именем “_QSQL000.DBF”, в данном случае размером 262 байта. При этом вместо подстроки “000” BDE может использовать другое трехзначное число, если уже есть файл с таким именем. При использовании “TADOQuery” подобного файла в текущей папке не создается, однако замедление настолько неприемлемое, что измерение производилось только для наименьшего из файлов.

Использовать приведенные преимущества Microsoft Access при обработке значительного объема текста нет необходимости. Совместимость с операционной системой наблюдается также и при использовании прежнего формата. Формат “dBase IV” обладает главным преимуществом – скоростью обработки данных. Использование запросов SQL приводит к многократному возрастанию времени обработки. Существуют и другие СУБД и использующие их среды разработки, которые могут применяться в данной предметной области. Однако некоторые из них имеют усложненную процедуру инсталляции и подключения в условиях затрудненной совместимости версий. Другие требуют изучения в условиях ограниченных часов учебного плана при том, что целесообразность комплексной организации такого обучения для востребованности выпускников не очевидна.

Литература

1. Димитриев А.П. База данных русско-чувашского переводчика // Компьютерные технологии и моделирование: сб. науч. тр. Вып. 9. Чебоксары, 2013. С. 67–75.

2. Димитриев А.П. Исследование эффективности методов оптимизации исходного кода на примере программы моделирования расписания занятий // Фундаментальные исследования. 2016. № 5-1. С. 28–32.

3. Димитриев А.П. Чувашско-русский переводчик: программная реализация // Прикладная информатика. 2011. № 6(36). С. 43–46.

4. Портнов М.С., Речнов А.В. О некоторых проблемах обучения программированию в средних специальных учебных заведениях // Состояние, направления и перспективы развития среднего профессионального образования: сб. материалов Междунар. заочной науч.-практ. конф. (г. Чебоксары, 24 марта 2017 г.). Чебоксары: ЧКИ РУК, 2017. С. 97–102.

5. Устав федерального государственного бюджетного образовательного учреждения высшего образования «Чувашский государственный университет имени И.Н. Ульянова» (новая редакция). [Электронный ресурс]. URL: http://www.chuvsu.ru/sveden/files/Ustav_obrazovatelynoy_organizacii.pdf (дата обращения: 16.04.2019).

6. Философия для аспирантов: учебное пособие. 2-е изд. / Кохановский В.П. [и др.]. Ростов н/Д.: Феникс, 2003.