

Организация БЗ в современных СУБД

Б.И. Рабинович

Институт проблем информатики РАН

1. Введение

Основным требованием к интеллектуальным системам является наличие в них Баз Знаний (БЗ). Для управления БЗ используются специальные языки программирования. Для представления знаний используются специального вида семантические сети, а управления - системы продукций. Основная проблема, с которой сталкиваются разработчики подобного рода систем, связана со способом хранения знаний. В технологии, которая предусмотрена в языке ДЕКЛ [1], знания имеют вид «плоских файлов» (flat-file databases). Такой способ хранения обладает рядом недостатков: медленная обработка при больших объемах данных, высокая трудоемкость удаления, сложность поиска. Помимо этого возникают проблемы защиты данных и управления. В то же время в современных СУБД, таких как Oracle, MSSQL, MySQL и т.п. эти проблемы уже решены.

В данном докладе описывается модель, позволяющая организовать БЗ в одной из перечисленных выше СУБД. В качестве исходных данных рассматривается система, ядром которой является БЗ на «плоских файлах», управление знаниями в которой осуществляется при помощи языка ДЕКЛ.

2. Структура существующей Базы Знаний

БЗ представляет собой набор плоских файлов, в которых хранятся следующие данные: тексты загруженных в систему документов, семантические сети, каталоги и индексы, которые автоматически строятся на их основе. Кроме этого, есть еще два типизированных файла, в которых хранится информация о том, на какой строке файлов БЗ с текстами и семантическими сетями начинаются документы, и какая у них длина. Это сделано для организации более быстрого доступа к данным[2].

В БЗ есть два индексных файла. Первый представляет собой перечень ключевых слов, найденных в семантических сетях, частоту их появления во всех документах и адрес (смещение) списка документов во втором файле, в которых были найдены эти слова. Ключевыми словами являются имена и существительные, такие как: адрес, факс, номер счета и т.п.

Также на этапе загрузки строятся файлы с каталогами основных объектов, по которым проводится быстрый поиск по БЗ. Такими объектами могут быть, например, адрес, телефон, ФИО.

Работа с БЗ осуществляется с помощью стандартных функций языка ДЕКЛ. Вот перечень основных функций и процедур: запись семантической сети, запись текста загруженного файла, запись индексов и т.д.

3. Структура БЗ в реляционной СУБД

Теперь попытаемся спроектировать структуру БЗ нашей системы, учитывая то, что она будет реализована в современной реляционной СУБД (InterBase, MsSQL, MySQL, Oracle и т.п.).

Поскольку быстрый поиск является основным требованием и основным преимуществом СУБД, то можно отказаться от типизированных файлов db_txt.dbi и

db_zzz.dbi. Данные, которые хранятся в файлах db_txt.db и db_zzz.db можно объединить, поскольку они являются характеристикой одной и той же сущности «Документ». Так же, для ускорения работы можно отдельно хранить заголовок документа, поскольку в чаще всего отображаются не документы, а именно заголовки, которые представляют собой первые несколько слов документов. Файлы net2.slv и net2.ind представляют собой единую сущность «Слово», которая характеризуется самим словом, частотой и списком документов, в котором оно встретилось. Характеристика «частота» используется для того, чтобы показать сколько раз то или иное слово встречается во всей БЗ, соответственно, когда в результате удаления какого-то документа эта характеристика становится равной нулю, это слово можно удалить. Для создания экземпляра сущности «Слово», необходимо при загрузке каждого документа разбирать его, преобразовывать каждое слово в нормальную форму (единственное число, мужской род, именительный падеж для существительных), и уже после этого заполнять соответствующие таблицы. Кроме этого, для примера, создадим сущности «Каталог ФИО» и «Каталог паспортов». Схема такой базы данных представлена на рисунке 1:



Рис. 1. Схема Базы Знаний.

Как видно из схемы, все данные необходимые для того, чтобы система работала так же, как и с БЗ на плоских фалах сохранены. Более того, после переноса БЗ в СУБД возникнет возможность подключения внешних баз. Теперь обратим внимание на преимущества, которые дает нам применение новой технологии хранения знаний: решается проблема удаления и изменения документов, значительно возрастает скорость поиска, значительно увеличивается уровень безопасности, появляется возможность восстановления данных и введения RAID массивов, кроме этого существенно повышается надежность системы в целом.

Литература

1. Кузнецов И.П. Продукционный язык программирования ДЕКЛ. Система обработки декларативных структур знаний Деклар-2 ИПИРАН 1988 г.
2. Рабинович Б.И. Аналитическая система обработки и управления структурированной информацией. Интеллектуальные технологии и система. Выпуск 5. Издательство ООО «Эликс+». Москва 2003, стр. 284-296.