Информационно-поисковая система МОДИС в виде картриджа СУБД ORACLE

Л.М.Зуева, Е.В.Козырева, В.В.Комоско,
П.А.Самойлова

РФЯЦ-ВНИИТФ,456770, Снежинск, Челябинской области, а/я 245, Россия

Аннотация

Описан картридж СУБД Oracle, включающий в себя алгоритм автоиндексирования текстов на естественном языке, реализованный в системе МОДИС.

ВВЕДЕНИЕ

Информационно-поисковая система МОДИС разработана во Всероссийском научно-исследовательском институте технической физики (ВНИИТФ).

Работы по созданию информационно-поисковых систем ведутся во ВНИИТФ с 1968 года - это системы АСИОР (1968-1975 гг), МОДИС-БЭСМ (1974-1982) [1], МОДИС-ЕС (1981-1990) [2-3], МОДИС-РС (1989-1999)[4]. Все эти системы успешно эксплуатировались и эксплуатируются по сей день во многих организациях страны, где накоплены большие массивы информации. Алгоритм автоиндексирования системы МОДИС зарегистрирован в международном каталоге "Автоматизированная обработка текстов на естественном языке".

Основные характеристики МОДИС:

1. Уникальный алгоритм автоиндексирования, т.е. глубокого семантического и синтаксического анализа текстов с учетом грамматики естественного русского и английского языков;

2. Алгоритм обработки параметрической информации, т.е. приведение единиц размерностей, встречающихся в тексте, к единицам Международной системы единиц СИ;

3. Язык запросов, моделирующий фразы естественных русского и английского языков.

©Вторая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
26-28 сентября 2000г., Протвино

При анализе текста входной информации учитываются парадигматические отношения для улучшения полноты поиска и текстуальные связи для улучшения точности поиска, а при морфологическом анализе используются различные вспомогательные словари и таблицы (словарь основ - для контроля входной лексики, словарь неинформативных терминов - для выделения значащих терминов, словари суффиксов, окончаний и префиксов - для усечения слов и определения их грамматических классов, словарь синонимов - для устранения неоднозначности).

ИНДЕКСИРОВАНИЕ

Индексирование - это процесс компактного описания содержимого документов в форме, удобной для последующего поиска. В ИПС МОДИС реализованы различные алгоритмы обработки индексируемых полей, зависящие от их типа - "ДАТА", "КОДЫ", "РАЗМЕРНОСТЬ", "ЧИСЛО", "ТЕКСТ", "ФАМИЛИЯ". В картридж МОДИС включен только алгоритм индексирования текстовой информации.

Алгоритм автоматического индексирования основывается на использовании сведений о внешней (синтаксической) структуре текста и дополнительных данных, извлекаемых из вспомогательных списков и таблиц. Данный алгоритм создан кандидатом физико-математических наук В.С. Авраменко. При разработке алгоритма были проведены необходимые исследования, на основе которых составлены следующие вспомогательные списки и таблицы:

- списки (словари) неинформативных терминов (слов, словосочетаний, сокращений и т.п.) русского и английского языков;
- таблица признаков для определения грамматических классов русских слов;
- таблица признаков для определения отглагольных существительных;
- список окончаний русского языка;
- список суффиксов русского и английского языков;
- список префиксоидов русского языка;

- список синонимических основ русского языка;
- список количественных и порядковых числительных русского языка;
- список единиц размерностей физических величин.

В словари неинформативных терминов русского (800 единиц) и английского (100 единиц) языков вошли служебные термины (союзы, предлоги, артикли и т.д.), общеупотребительные термины, а также термины, являющиеся неинформативными только в определенных словосочетаниях. При этом словари являются общими для любой предметной области. Пользователь может расширять и модифицировать эти словари с учетом конкретной проблемной ориентации.

Таблица признаков для определения грамматических классов русских слов включает около 250 буквосочетаний (от 1 до 5 символов), каждое из которых взаимосвязано с определенным набором окончаний из таблицы окончаний. Определение грамматического класса английских слов осуществляется только посредством анализа их суффиксальной последовательности.

Первый этап индексирования состоит в представлении входного текста в виде набора индексационных терминов. Термины текста последовательно подвергаются семантическому и морфологическому анализу, которым после приведения к каноническому виду (индексам) присваивается признак одного из трех типов:

- незначащий;
- значащий;
- полуинформативный.

На стадии предварительного семантического анализа термины всех трех типов разделяются на слова естественного языка (русские, английские), числовую информацию и термины, состоящие из смешанных символов. Числовая информация (числа, числительные) и единицы размерностей приводятся к каноническому виду с учетом соответствующих коэффициентов пропорциональности. Смешанным и числовым терминам после приведения их к каноническому виду присваиваются соответствующие семантические признаки, и в таком виде они попадают в инвертированный файл.

Все ”слова”естественного языка подвергаются морфологическому анализу, задачей которого является отделение у каждого слова основы и приписывание этой основе морфологической и семантической информации, необходимой для установления текстуальных отношений между словами на этапе синтаксического анализа.

Морфологический анализ выполняется по следующей схеме:

- определяется грамматический класс слова по его буквенному составу (только для русских слов);
- существительные подвергаются анализу на определение отглагольных существительных;
- сложные слова разбиваются на самостоятельные единицы: префиксоид и основу (только для русских слов);
- различные словоформы одного слова приводятся к одному и тому же парадигматическому коду (основе).

На заключительной стадии семантического анализа каждому индексу присваивается семантико-грамматическая категория (предмет, объект, прибор, свойство, качество, процесс, действие), уменьшающая объем обозначаемых им понятий.

Целью синтаксического анализа является разбиение текста на определенные конструкции (предложения, сегменты, индексы) и установление соответствующих связей между элементами этих конструкций.

Индекс - это набор символов, полученных из терминов текста на стадии лексического анализа.

Сегмент - последовательность индексов, ограниченная знаками препинания или неинформативными терминами.

Предложение - последовательность сегментов, ограниченная точкой с запятой, восклицательным знаком или точкой, обозначающей конец предложения.

Внутри каждого сегмента устанавливаются следующие связи межу элементами:

- атрибутивная связь (”префиксоид + основа”, ”существительное + существительное”, ”число + размерность физической величины”);
- дефисная связь (”прилагательное + существительное”, ”существительное + существительное”, ”число + число”и т.п.);
- грамматическая связь (попарное согласование существительного и прилагательного).

В результате автоиндексирования текста из каждого значащего термина формируется 6-символьный индекс, который характеризуется грамматической категорией, номерами предложения, сегмента и слова в сегменте, признаком префиксоида или основы, признаком связи и, если это число, диапазоном значений.

Скорость индексирования около 0,5 Мб в минуту на ЭВМ Pentium-1 (133 Мгц).

ЗАДАНИЕ ЗАПРОСА

Запросы на поиск подвергаются точно такой же обработке, как и тексты. Благодаря наличию алгоритма автоиндексирования и обработке данным алгоритмом как документов, так и запросов, поиск производится практически на естественном языке и позволяет получать информацию с большой полнотой и точностью. В процессе составления запроса можно использовать логические операторы ”И”, ”ИЛИ”, ”НЕ”, а также применять контекстуальные операторы и операторы отношения.

Включение в запрос контекстуальных операторов позволяет более детально указать местоположение и связь терминов в тексте. К таким операторам относятся:

- встречаемость в предложении;
- встречаемость в сегменте;
- согласование существительного и его определения;

- согласование числа и характеризующей его размерности;
- дефисная связь двух слов;
- префиксоидная связь;
- принудительная связь.

Если между двумя словами в запросе встречается знак или какое-либо неинформативное слово, то считается, что данные слова должны встретиться в одном предложении.

Если между двумя словами запроса нет знаков или неинформативных терминов, то система воспринимает это как указание на встречаемость слов в одном сегменте.

Если в запросе присутствует существительное и относящееся к нему определение, согласованное в числе и падеже, то будет вестись поиск именно такого словосочетания.

Дефисная связь требует наличия двух слов, написанных через дефис. Задав в запросе два слова через дефис, можно потребовать взаимное расположение двух терминов, стоящих рядом в строго определенном порядке.

Префиксоидная связь позволяет находить информацию по одной из частей сложного слова, т.е. по префиксоиду или основе.

Для числовых данных могут быть заданы операторы отношения:

- больше или равно;
- меньше или равно;
- встречаемость в диапазоне.

В запросах при индексировании каждому термину также приписывается признак семантико-грамматической категории, указывающий на логическую роль данного термина в тексте. Учитывая или не учитывая эти признаки, можно регулировать процесс поиска.

Примеры запросов:

1. Химический лазер на смеси серного ангидрида.
2. Термоупругие волны напряжений; электромагнитный импульс.
3. Электропарамагнитный резонанс.
4. Картографирование с разрешающей способностью до 100 м.
5. Поршневые двигатели мощностью 200-300 л.с.
6. Жидководородная мишень.

КАРТРИДЖ МОДИС

СУБД Oracle имеет определенные возможности поиска по текстовым неструктурированным полям, но при этом не учитывается семантика и синтаксис естественных русского и английского языков. В СУБД Oracle эти возможности реализованы в виде картриджа ConText Option [5], но, на наш взгляд, алгоритмы поиска системы МОДИС по русскоязычным текстам являются более мощными, и формулирование запросов на поиск более приближено к естественному языку, хотя в МОДИС, в то же время, отсутствуют некоторые возможности ConText Option. К примеру, в МОДИС возможен поиск:

- с учетом грамматической категории слова;
- с учетом местоположения слова в предложении;
- с учетом связи существительного и прилагательного;
- по запросам, включающим одновременно тексты и параметрические данные;
- по части сложного слова .

Механизм картриджей, реализованный в СУБД Oracle, делает систему открытой и позволяет дополнять ее внешними модулями, поэтому алгоритмы системы МОДИС реализованы в виде картриджа.

На данный момент текст в полях, обрабатываемых МОДИС, должен быть в формате ASCII.

Картридж МОДИС для Oracle представляет собой:

- два процессора в виде динамически загружаемых библиотек - один для автоматического индексирования, другой для поиска информации,
- OLE-элемент ModisViewer - для просмотра результатов поиска,
- несколько дополнительных таблиц и триггеров форм и базы данных.

Никаких ограничений на таблицы в базах данных не накладывается. Единственное требование - наличие уникального числового поля (номера документа). В Oracle уникальное поле организовано с помощью числовой последовательности.

Для инициализации картриджа необходимо в каждой базе данных создать ряд служебных таблиц:

- MODIS - содержит данные о месторасположении вспомогательных списков системы МОДИС и временных файлов;
- MODIS_USER_POLICY - содержит все необходимые для индексирования алгоритмами МОДИС какого-либо поля таблицы базы данных сведения. Это имя таблицы, имя данного поля, имя ключевого поля, месторасположение индексов, количество заиндексированных записей и ссылка на фильтр для конвертирования информации из других форматов в формат ASCII. Для каждого индексируемого поля таблицы индексы хранятся в отдельных файлах;
- MODIS_QUERY - для временного хранения результатов поиска;
- MODIS_DELETE - для хранения номеров уничтоженных записей.

К каждой таблице, имеющей поля, индексируемые алгоритмами МОДИС, добавляется триггер, позволяющий сохранять номера уничтожаемых записей в служебной таблице MODIS_DELETE. Эти записи не индексируются и не участвуют в поиске информации.

Так как Oracle не позволяет подключать в триггере базы данных программы из динамически загружаемых библиотек, то автоиндексирование производится не в момент записи информации в базу данных, а отдельной программой. При этом отслеживается количество заиндексированных записей и выдается вся информация на текущий момент. Индексы хранятся в инвертированных файлах.

При построении формы Oracle необходимо подключить четыре триггера: два на блок (PRE_QUERY, PRE_SELECT) и два на форму (WHEN_NEW_FORM_INSTANCE, POST_FORM). Триггеры полностью независимые, т.е. подключаются к любой форме для любой таблицы. Обращение к библиотекам МОДИС происходит через пакет ORA_FFI.

При выполнении запроса отрабатывает триггер PRE_QUERY, в котором опознаются поля, обрабатываемые алгоритмами МОДИС, производится поиск по данным полям и в таблице MODIS_QUERY запоминаются уникальные номера релевантных документов. Затем запрос на поиск корректируется, т.е. формируется вложенный запрос, учитывающий наличие отобранных системой МОДИС записей, и передается Oracle на выполнение.

Для просмотра результатов поиска используется OLE-элемент ModisViewer, позволяющий просматривать отобранную информацию (в том числе и поля LONG ROW) с подсветкой релевантных терминов. В полях, заиндексированных МОДИС, запрос должен и формироваться по правилам МОДИС, т.е. можно использовать логические и контекстуальные операторы, осуществлять поиск по смешанному запросу (текст и числа), искать по числам с размерностями физических величин. Поэтому пользователю, осуществляющему ретроспективный поиск в форме Oracle, желательно знать поля, обрабатываемые системой МОДИС, чтобы сформировать более грамотный запрос и отобрать информацию с учетом семантики и синтаксиса русского языка.

При формировании очередного запроса или закрытии формы информация в таблице MODIS_QUERY, относящаяся к последнему обработанному запросу, уничтожается.

ЗАКЛЮЧЕНИЕ

В качестве одной из областей применения алгоритма Автоиндексирования возможно создание информационно-поискового узла Internet, обеспечивающего контекстно-свободный поиск литературных источников и другой информации на русском и английском языках.

Просматривается возможность создания "Русской поисковой машины МОДИС"в сети Internet на базе СУБД Oracle 8.0, алгоритмов автоиндексирования и поиска информации системы МОДИС. В этом случае можно использовать Oracle как среду хранения URL-адресов документов, а процессоры МОДИС - для индексирования WEB-страниц.

Список литературы

[1] В.Р.Хисамутдинов, В.С.Авраменко, В.И.Легоньков. Автоматизированная система информационного обеспечения разработок. Москва, "Наука", 1980 г.

[2] В.С.Авраменко, Л.М.Зуева. Математическое обеспечение диалоговых информационных систем для ЕС ЭВМ (МОДИС-ЕС). Сборник научных трудов "Автоматизированные библиотечно-информационные системы", Новосибирск, изд. ГКНТБ СОАН СССР, 1985 г.

[3] В.С.Авраменко, В.И.Легоньков, В.Р.Хисамутдинов. Математическое обеспечение диалоговых информационных систем. Москва, "Наука", 1990 г.

[4] Л.М.Зуева, Т.В.Ермолаева, В.И.Легоньков, Е.Ю.Лукьянова, Л.А.Стракевич. МОДИС - математическое обеспечение диалоговых информационных систем. Сборник научных трудов "Автоматизированные библиотечно-информационные системы", Новосибирск, изд. ГКНТБ СОАН СССР, 1991 г.

[5] Oracle ConText Option V2.0, An Oracle Q&A. Oracle Corporation 1997.