Использование текстового индекса при работе с документами в универсальной базе данных

Пржиялковский В. В.
prz@ccas.ru
Москва, Вычислительный центр РАН

Современные универсальные базы данных изначально создавались как «фактографические», т. е. допускающие хранение скалярных данных, организованных в таблицы. Хранение и обработка «сложно устроенных» данных традиционно выполнялась специализированными системами. Это порождало проблемы интеграции и синхронизации логически единых, но разнородных данных, вынуждено разнесённых по разным системам, а также препятствовало использованию таких возможностей универсальных баз данных, как хранение особо больших объёмов, восстановимость после потерь, поддержание целостности данных, защита. Однако, начиная с первой половины 90-х годов, эти проблемы стали постепенно разрешаться вследствие того, что основные производители универсальных СУБД начали добавлять в свои системы возможности работы со сложно устроенными данными.

Характерной в этом отношении является СУБД Oracle. Сейчас она дает возможность хранить в базе данных и обрабатывать (в разной степени) следующие категории нескалярных данных:

1. «Большие неструктурированные объекты» (LOB, large objects). БД относится к ним как к строкам байтов или текстовых символов с разрешённой длиною до 128 терабайт. СУБД обеспечивает только хранение (обеспечивается степень сжатия больше, чем в файловых системах), а обработка и интерпретация возложена на прикладные программы.
2. Структурированные объекты, построенные по принципам объектно-ориентированного подхода. СУБД полностью знает структуру хранимых объектов и использует её в запросах и в хранении.
3. Частично структурированные объекты. Объекты этого рода как правило не имеют жёсткой структуры и устроены по-разному соответственно разным предметным областям.

Для работы с объектами последней категории Oracle использует так называемые «предметные индексы» (domain indexes). Предоставляются средства программирования такого рода индексов, но одновременно имеется и несколько встроенных видов, готовых к употреблению. Например, пространственный индекс позволяет хранить в базе данных и обрабатывать с помощью СУБД пространственные данные; текстовый индекс, вместе со встроенной в СУБД поисковой текстовой машиной (Oracle

Text), позволяет расширить возможности базы данных традиционным инструментарием информационно-поисковых систем (ИПС).

Замечательно, что запросы к базе данных могут быть комбинированными, например, в одном запросе может содержаться обращение как к скалярным таблично-организованным данным, так и к текстовым документам.

Возможности текстового индекса

В Oracle Text имеется три разновидности текстового индекса применительно к трем случаям текстовой обработки:

— `CTXSYS.CONTEXT` — для выполнения полнотекстового поиска по текстовым документам как внутреннего хранения, так и внешнего (файловая система, интернет);
— `CTXSYS.CTXCAT` — для выполнения упрощенного и ускоренного поиска в каталогах с краткими описаниями, например в лентах новостей;
— `CTXSYS.CTXRULE` — для построения классификаций, или рубрикаций, документов при том, что признаки классифицирования описываются набором характерных запросов.

Документы для индексирования могут находиться в базе данных, в интернете, в файлах или же в произвольном месте, обращение к которому можно оформить программой. Форматы документов могут быть самых разных видов, включая простой текст, PDF, RTF, MS Word, XML, HTML и прочие.

Наибольшие возможности имеет полнотекстовый индекс. Содержательно он хранит тройки ⟨документ, словоместо, индексируемое слово или несколько слов⟩, и позволяет по предъявленному поисковому слову получить список пар ⟨документ, словоместо⟩. В сочетании с имеющейся программной логикой он позволяет осуществлять поиск по документам со следующими свойствами:

— Точный поиск. В документе ищется в точности указанное слово, например `'Java'`.
— Позиционный поиск группы слов. Может выполняться поиск комбинации слов, например `'Java Development Kit'`. Может искаться близкое расположение слов, например слово `'Java'`, расположенное на расстоянии не более чем N слов от слова `'development'`. Может искаться встреча двух слов, например, `'Java'` и `'development'`, в одном предложении или же параграфе.
— Нечеткий поиск. Могут искаться слова, похожие по звучанию или похожие по написанию (последствия опечаток). Могут искаться слова одного корня. Могут искаться слова с указанными подстроками.

Могут искаться слова с привлечением тезауруса, построенного в соответствие со стандартом ISO-2788.

Результаты поиска ранжируются по встроенному правилу, или по запрограммированному самостоятельно.

Продвинутые возможности Oracle Text позволяют отойти от поиска по словам и обеспечивают следующее:

— Тематический поиск. Поиск документов по темам, а не отдельным словам.
— Рубрицирование предъявленного документа по темам.
— Автоматическое формирование резюме документа.
— Автоматическое формирование набора классификационных правил на основе «обучения» предъявленными документами.
— Кластеризация. Группирование документов по близости содержания.

Некоторые возможности текстового индекса в Oracle Text не имеют готовых реализаций для русского языка. Например, это касается поиска по словам, близким по произношению, или же морфологического поиска. Однако имеется программный инструментарий для восполнения некоторых подобного рода пробелов.

Количественные характеристики текстового индекса

Практичность информационной системы часто напрямую зависит от её технических характеристик. Ниже приводятся характерные затраты ресурсов компьютера при работе Oracle Text, доступные для простой самостоятельной проверки.

В документации Oracle имеется два файла по Oracle Text в формате PDF: `b14217.pdf` и `b14218.pdf`. Их общий объем примерно 4,6 Мб или 742 страницы. Полученные характеристики индексирования: время построения индекса — 77 сек.; объем всех выше перечисленных структур индекса — 8 Мб; количество лексем в индексе — 40546; характерное время выполнения запроса сочетания `'oracle text'` по индексу — 1,28 сек.

Замеры приблизительные и соответствуют процессору Celeron с тактовой частотой 1 Ггц и оперативной памяти 512 Мб. Формат PDF — не самый экономный; индексирование документа формата HTML даст меньшее количество лексем, меньший объём индекса, более быстрое его построение и более быстрые ответы на запросы.

Более мощные вычислительные возможности способны дать лучшие характеристики. В отечественной практике есть случай использования Oracle Text для индексации и семантического анализа в базе данных электронных писем размером в несколько терабайт (без учёта вспомогательных структур).

Работа выполнена при поддержке РФФИ, проект №07-07-00181.

Литература

[1] Text Application Developer's Guide, 10g Release 2 (10.2). — Oracle Corp., Part Number B14217–01.

[2] Text Reference, 10g Release 2 (10.2). — Oracle Corp., Part Number B14218–01.

[3] *Плешко В. В.* Поиск с учетом словоформ русского языка. — Oracle Magazine, июнь/июль 2003. `www.oracle.com/ru/oramag/june2003/index.html?russia\_rco3.html`.