Индекс базы данных ИРБИС

Материал из Wikipedia

Индекс базы данных ИРБИС – специальная структура, являющаяся частью базы данных, которая обеспечивает быстрый поиск.

Термин *словарь* получил широкое распространение и фактически стал в ИРБИС-сообществе своего рода заменой понятию *индекс* (см. подраздел *Словарь базы данных*).

В базах данных ИРБИС используется *инвертированный индекс* (http://ru.wikipedia.org/wiki/%D0%98%D0%BD%D0%B2%D0%B5%D1%80%D1%82%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D1%
, который принято называть *инвертированный файл* (также используют термин *инверсный файл*).

Индекс создаётся для каждой базы данных ИРБИС.

Для формулировки запросов, с помощью которых выполняется быстрый поиск, служит язык запросов ИРБИС.

Содержание

Словарь базы данных

Часто *словарь* используют как синоним понятия *индекс* базы данных. Хотя, строго говоря, *словарь* – лишь часть *индекса*, наряду со *списком индексных ссылок* (см. подраздел *Структура индекса базы данных ИРБИС*).

Примечание: пользователям хорошо знаком словарь, представленный непосредственно в графическом интерфейсе в ряде АРМ системы ИРБИС; возможно, поэтому термин "словарь" получил широкое распространение и фактически стал в ИРБИС-сообществе своего рода заменой понятию "индекс".

Определение *индекса* базы данных ИРБИС

Для построения *индекса* (и поддержания в актуальном состоянии) требуется его определение. Определением служат: *ТВП для инвертированного файла* и *таблица актуализации*.

ТВП для инвертированного файла

ТВП для инвертированного файла – ТВП, которая задаёт правила выбора элементов для инвертированного индекса базы данных ИРБИС.

В файловой структуре каждой базы предусмотрена одна *ТВП для инвертированного файла*. Подробнее см. в подразделе *ТВП для инвертированного файла* статьи *Файлы ИРБИС*.

Типовые базы данных (БД электронного каталога, полнотекстовая БД и т.д.) содержат в себе соответствующие типовые *ТВП для инвертированного файла*.

Также см.:

- подраздел *Внесение изменений в ТВП для инвертированного файла* статьи *Рекомендации по обслуживанию баз данных ИРБИС*.

Структура *индекса* базы данных ИРБИС

Индекс представляет собой совокупность двух структур: *словарь поисковых терминов* в структуре бинарного дерева и список *индексных ссылок*, соответствующих каждому термину.

Словарь поисковых терминов

Элементы, созданные посредством *ТВП для инвертированного файла*, составляют словарь поисковых терминов для базы данных.

Список индексных ссылок

Система связывает с каждым *поисковым термином* список *индексных ссылок*, обеспечивающих требуемую связь с записями. Каждый термин имеет столько *индексных ссылок*, сколько раз он встречается в базе данных.

Структура индексной ссылки

Для поддержки развитых средств поиска, имеющихся в языке поиска, таких, например, как поиск по ключевым словам в определенных элементах описания, каждая индексная ссылка содержит помимо MFN записи некоторую дополнительную информацию, имеющую отношение к расположению термина в записи.

Индексная ссылка имеет следующие 4 компоненты:

1. MFN записи, содержащей термин. Эта компонента вводится в состав индексной ссылки при актуализации/формировании словаря автоматически.
2. Идентификатор поля, используемый в процессе поиска при указании квалификатора. Эта компонента вводится в состав индексной ссылки на основе ТВП. Обратите внимание на то, что один и тот же идентификатор поля может быть присвоен различным полям, указанным в формате выборки.
3. Номер экземпляра (повторения) повторяющегося поля, необходимый для осуществления поиска на уровне поля и операторов близости расположения терминов в повторяющихся полях (в АРМах ИРБИС это используется при поиске по логике «И (в поле)»). Для того, чтобы можно было использовать указанный метод поиска (обычно для этого необходим метод индексирования 4 или 8), необходимо определить формат в ТВП таким, чтобы в его выходных данных между экземплярами повторяющегося поля располагался знак процента (%), для чего нужно задать его в качестве повторяющегося суффикс-литерала. Например, строка ТВП для инвертирования повторяющегося поля 10 должна содержать формат v10|%|. Система перед обработкой каждой строки ТВП устанавливает номер повторения в 1 и затем увеличивает его на 1 всякий раз, когда в созданном форматом тексте встречается символ %.

4. Последовательный номер термина, необходимый для осуществления поиска по близости расположения терминов (в АРМах ИРБИС это используется при поиске по логике «И (фраза)»). Управление присвоением данного номера происходит следующим образом: он устанавливается в 1 перед обработкой каждой строки ТВП и при изменении номера повторения и увеличивается на 1 для каждого элемента, созданного указанным методом индексирования. Например, предположим, что в повторяющемся поле 331 содержится краткое содержание литературного источника, причем каждое повторение состоит из одного абзаца. Пусть данное поле проиндексировано методом 4. Если определить формат выборки данных mdl,v331|%|, то начиная с каждого абзаца краткого содержания словам будет присваиваться последовательный номер, начиная с 1 в каждом абзаце, а если бы формат выборки был равным mdl,v331, то словам присваивался бы сквозной последовательный номер по всему краткому содержанию, например, первое слово второго абзаца имело бы последовательный номер на 1 больше номера последнего слова первого абзаца.

Обслуживание индекса

Инвертированный индекс в ИРБИС не является полностью автоматизированным, и в определённых ситуациях может потребоваться вмешательство администратора баз данных ИРБИС.

Могут возникнуть следующие ситуации, требующие обслуживания индекса:

- Новые записи, введенные в файл документов, недоступны при поиске.
- Записи, которые подвергались модификации, доступными при поиске, но под старыми элементами доступа.
- Удаленные записи все еще зарегистрированы под их элементами доступа, однако сами записи не отображаются.
- *ТВП для инвертированного файла* была изменена (результаты поиска остались прежними).

В системе ИРБИС имеются две операции, которые приводят индекс базы данных в актуальное состояние, соответствующее *ТВП для инвертированного файла* и содержимому базы данных: это операции *создания словаря* и *актуализации словаря*. Их отличие заключается в алгоритме и особенностях применения.

Флаг актуализации

Флаг актуализации позволяет отмечать каждую запись базы данных как *актуализированную* или *неактуализированную*, при этом считается, что:

- запись *актуализирована* – значит инвертированный индекс отражает её содержимое;
- запись *неактуализирована* – значит инвертированный индекс НЕ отражает её содержимое.

Благодаря использованию в системе ИРБИС данного флага возможно:

- установить факт наличия *неактуализированных* записей и, соответственно, сделать вывод о необходимости привести *инвертированный индекс* в актуальное состояние;
- посчитать соотношение *актуализированных* и *неактуализированных* записей, в соответствии с которым принимать решение о приведении *инвертированного индекса* в актуальное состояние с помощью *создания словаря* или *актуализации*.

Создание словаря

Создание словаря – это создание инвертированного индекса с использованием *ТВП для инвертированного файла* на основе всех *записей* (*документов*) базы данных.

Типичные примеры ситуаций, в которых выполняют создание словаря:

- имеется значительное количество неактуализированных записей по сравнению с общим количеством записей в базе данных;
- было добавлено значительное количество текстов в полнотекстовую базу данных;
- была изменена *ТВП для инвертированного файла*.

Создание словаря осуществляется с помощью АРМ Администратор. Ознакомьтесь с рекомендациями по созданию словаря.

Алгоритм создания словаря предусматривает три этапа, которые могут быть выполнены по отдельности:

- отбор
- сортировка
- загрузка.

Актуализация словаря

Актуализация словаря – это приведение инвертированного индекса в актуальное состояние на основании документов, для которых по каким-либо причинам (авария, глобальная корректировка, импорт и копирование через АРМ Администратор) не выполнялась автоматическая актуализация при их вводе/корректировке.

Типичные примеры ситуаций, в которых выполняют создание словаря:

- количество неактуализированных записей невелико по сравнению с общим количеством записей в базе данных.

Актуализация осуществляется с помощью АРМ Администратор. Ознакомьтесь с рекомендациями по актуализации словаря.

Описание механизма актуализации инвертированного файла в связи с изменением отдельной записи см. в статье Механизм актуализации записи.

Реорганизация словаря

Реорганизация словаря представляет собой структурное перестроение *инвертированного файла* с целью уменьшения размера файла и повышения быстродействия работы с ним.

Возникновение необходимости реорганизации словаря связано с тем, что в результате выполнения актуализации словаря может происходить усложнение структуры инвертированного файла и появление «пустот», которые реорганизация устраняет.

Реорганизация словаря осуществляется с помощью АРМ Администратор. Ознакомьтесь с рекомендациями по реорганизации словаря.

Файлы *индекса* базы данных ИРБИС

Индекс базы данных ИРБИС хранится на файловой системе в виде трёх файлов: *словарь поисковых терминов* в файлах `.n01` и `.l01`; список *индексных ссылок* в файле `.ifp`.

В бинарном дереве файл с расширением `.n01` содержит узлы дерева и файл с расширением `.l01` – листья. Записи с листьями указывают на файл ссылок `.ifp`.

Об особенностях размещения файлов `.n01`, `.l01` и `.ifp` см. подраздел *Файлы баз данных ИРБИС* статьи *Файлы ИРБИС*.

Взаимосвязи между файлами `.n01` и `.l01` обеспечиваются ссылками, которые представляют собой относительные адреса соответствующих записей. Относительный адрес это порядковый номер записи в данном файле.

Структура записи одинакова для `.n01` и `.l01` файлов. Размер (длина) записи зависит от реализации (512, 1024, 2048, 4096). Таким образом, максимальный размер файлов `.l01` и `.n01` определяется как 2 Гб * размер записи. В данной реализации размер записи 2048.

Адрес корневой записи файла `.n01` сохраняется как номер первой записи.

Смещение на запись в файле `.ifp` сохраняется в файле `.l01` и имеет длину 64 байта (в данной реализации используется только младшее слово этого смещения).

Формат файлов `.n01` и `.l01`

Файлы состоят из записей (блоков) постоянной длины. Записи состоят из трех частей: лидера, справочника и ключей переменной длины.

Формат лидера записи:

Число бит	Параметр	Описание
32	NUMBER	номер записи (начиная с 1; в `.n01` файле номер первой записи равен номеру корневой записи дерева)
32	PREV	номер предыдущей записи (если нет = -1)
32	NEXT	номер следующей записи (если нет = -1)
16	TERMS	число ключей в записи
16	OFFSET_FREE	смещение на свободную позицию в записи (от начала записи)

Справочник это таблица, определяющая поисковый термин. Каждый ключ переменной длины, который есть в записи, представлен в справочнике одним вхождением следующего формата:

Число бит	Параметр	Описание
16	LEN	длина ключа
16	OFFSET_KEY	смещение на ключ (от начала записи)
32	LOW	В `.n01` файле: ссылка на запись файла `.n01` (если LOW > 0) или файла `.l01` (если LOW < 0), у которых 1-й ключ равен данному. Положительное значение LOW определяет ветку индекса иерархически более низкого уровня. Самый низкий уровень индекса (LOW < 0) соответствует ссылкам на записи (листья) файла `.l01`. В `.l01` файле: младшее слово 8-байтового смещения на ссылочную запись в `.ifp`.
32	HIGH	В `.n01` файле: всегда 0. В `.l01` файле: старшее слово 8 байтового смещения на ссылочную запись в `.ifp`.

Ключи переменной длины записываются начиная с конца записи, так что порядок входов, соответствующих им, определяется алфавитным порядком ключей. Сами ключи располагаются вплотную друг к другу без разделителей в порядке поступления на запись.

- **Длина справочника** 12 * TERMS.
- **Длина ключей** = [Размер записи] – OFSET_FREE.
- **Размер свободного места в записи** = 16 + 12 * TERMS - [длина ключей].
- **Размер записи** зависит от реализации и может быть равен в байтах: 512, 1024, 2048, 4096.

Формат файла `.ifp`

Файл содержит список ссылок для каждого термина словаря.

Список ссылок может быть представлен в 2-х различных форматах. Выбор формата размещения ссылок осуществляется при загрузке словаря из файла `.lk1` (этот файл формируется после отбора и сортировки терминов) в зависимости от общего числа ссылок для данного термина. Обыкновенный формат – это заголовок блока и набор упорядоченных ссылок. По превышении определенного числа ссылок (MIN_POSTINGS_IN_BLOCK – в данной реализации 256) формат включает специальный блок и набор блоков обыкновенного формата размер которых определяется по следующей схеме: блоки 4, 8, 16, 32 Кб для общего числа ссылок соответственно 256-32000, 32000-64000, 64000-128000, 128000 и более.

Такая схема оптимизирует работу с диском в процессе инвертирования записи в базах данных, характеризующихся большим количеством ссылок на термин.

Обыкновенный формат записи `.ifp`

Запись состоит из заголовка и упорядоченного набора ссылок.

Ссылка имеет следующий формат:

Число бит	Параметр	Описание
32	PMFN	номер записи
32	PTAG	идентификатор поля, назначенный при отборе терминов в словарь
32	POCC	номер повторения
32	PCNT	номер термина в поле

Заголовок имеет следующий формат:

Число бит	Параметр	Описание
32	LOW	младшее слово смещения на следующую запись (если нет 0)
32	HIGH	старшее слово смещения на следующую запись (если нет 0)
32	TOTP	общее число ссылок для данного термина (только в первой записи); число ссылок в данном блоке (в следующих записях)
32	SEGP	число ссылок в данном блоке
32	SEGC	вместимость записи в ссылках

Признак последнего блока – LOW = HIGH = -1

Специальный формат записи `.ifp`

В этом случае первой записью является специальный блок, который представляет собой заголовок (обыкновенного формата), в котором смещения имеют специальные значения = -1001, и набор вхождений следующего формата:

Число бит	Параметр	Описание
32	POSTING	1-я ссылка из записи обыкновенного формата
32	LOW	младшее слово смещения на следующую запись (если нет 0)
32	HIGH	старшее слово смещения на следующую запись (если нет 0)

Число вхождений кратно 4. Записи, на которые ссылается специальный блок связаны между собой как описано выше. Общее количество ссылок для данного термина сохраняется только в специальном блоке.

Модификация записей файла `.ifp`

При выполнении актуализации инвертированного файла могут создаваться новые дополнительные записи при добавлении новых ссылок. В этом случае создается новая запись размером, равным общему количеству ссылок, если нет специального блока, и размером, равным количеству ссылок в данной записи, если есть. Новая запись создается таким образом, чтобы не нарушалась возрастающая последовательность следования ссылок. Новая запись связывается с существующими через поле NXT_, ссылки распределяются равномерно между старой и новой записью.

Ссылки

См. также:

- Базы данных ИРБИС
- Язык запросов ИРБИС
- Таблица выбора полей
- Механизм актуализации записи
- Механизм полнотекстового поиска
- Файлы ИРБИС
- The UNESCO micro CDS/ISIS Software. Приложение 7. Структура инвертированного файла и форматы записей (http://nbuv.gov.ua/books/19/isis/24.htm)
- Сценарии поиска

Источники информации:

- Общее описание системы ИРБИС64
- ИРБИС 64/128. Общее описание системы (http://lib.omgtu.ru/irb/doc/irbis128.doc)