Таблица выбора полей

Материал из Wikipedia

Таблица выбора полей (ТВП) определяет правила выбора одного или нескольких ***элементов*** в контексте записи базы данных. ***Элемент*** может быть в общем виде определён как фрагмент записи, выделенный в результате определённой обработки. Хотя во многих случаях *элементами* будут *элементы данных* (поля или подполя), могут также употребляться слова, фразы или другие фрагменты данных, которые имеют определённое значение в специальных приложениях.

Содержание

Применение ТВП

Выбранные элементы применяются в зависимости от конкретного варианта использования.

ТВП может использоваться для переформатирования записей во время операций импорта, экспорта или копирования.

ТВП для инвертированного файла

ТВП для инвертированного файла – специальная ТВП, которая определяет содержимое *индекса* (*словаря*) базы данных ИРБИС.

Общие сведения

Таблица выбора полей как отдельная структура сохраняется в текстовом файле с расширением `.fst`.

Файл ТВП представляет собой набор строк, каждая из которых содержит следующие три элемента, разделенные знаком пробел:

- ***идентификатор поля*** (ИП);
- ***метод индексирования*** (МИ);
- ***формат выборки данных***, представленный на языке форматирования системы ИРБИС.

Когда появляется необходимость в выборке элементов с использованием ТВП, система производит чтение требуемых записей файла документов, и выполняет следующие действия по каждой записи и каждой строке ТВП:

1. использует *формат выборки данных* для извлечения и форматирования соответствующих данных из записи;
2. применяет указанный *метод индексирования* к данным, полученным с помощью предыдущего шага;
3. присваивает каждому элементу, полученному подобным образом, указанный *идентификатор поля*.

Процесс выборки элементов с использованием ТВП является чисто механическим, описанные шаги связаны только лишь данными, которые создаются при их выполнении. Например, тот факт, что на шаге 1 произошла выборка данных из конкретного поля, является несущественным на шаге 2. На шаге 1 могут использоваться все возможности языка форматирования для создания строки символов, которая затем поступает в распоряжение шага 2. На шаге 2 поступившие строки символов обрабатываются в соответствии с указанным методом индексирования. Методы индексирования представляют собой операции со строками символов, а не с записями или полями. Именно благодаря такому универсальному пониманию сути ТВП, предоставляется возможность использовать их для таких, на первый взгляд совершенно не связанных целей, как определение содержимого инвертированного файла и способ преобразования данных при импорте документов.

В самом общем смысле ТВП представляет собой механизм порождения элементов данных на основе имеющихся, применяемый для выполнения определенных задач.

Параметры ТВП

Ниже описаны параметры строк ТВП в том порядке, в каком они обрабатываются (в строке ТВП они расположены в обратном порядке).

Формат выборки данных

Формат выборки данных создается с использованием средств языка форматирования.

Такие средства языка форматирования как средства RTF и HTML и переменные метки нельзя применять в форматах выборки ТВП.

Для некоторых методов индексирования существенной является концепция *строки*. В этом случае весьма ответственным является употребление команд перехода на новую строку.

Также весьма важным является употребление команд *режима вывода*, так как некоторые методы индексирования фактически требуют наличия конкретного режима вывода. В связи с этим, пользователь сам должен побеспокоиться, чтобы тот или иной формат выборки данных содержал необходимые команды режима вывода.

Следует также отметить, что использование преобразования букв в прописные может повлиять на дальнейшую обработку данных, создаваемых с помощью ТВП. Как правило, не нужно использовать такое преобразование, то есть надо использовать режимы mpl, mhl, mdl, а не mpu, mhu, mdu. Система автоматически производит преобразование букв в прописные, когда в этом появляется необходимость. Например, все элементы, создаваемые ТВП для инвертированного файла, переводятся в прописные буквы до их размещения в словаре, даже если ТВП порождает элементы в виде строчных букв.

Методы индексирования

Метод индексирования определяет специфическую обработку данных, созданных форматом. Имеется девять методов индексирования. Они идентифицируются числовыми кодами от 0 до 8.

Метод индексирования 0

Создаёт элемент из каждой строки, сформированной в соответствии с форматом. Этот метод обычно используется для индексирования в целом всего поля или подполя. Следует обратить особое внимание, что система в данном случае строит элементы из строк, а не из полей. В качестве выходного результата форматирования выступает строка символов, в которой нет никакого указания на ее принадлежность (или принадлежность части строки) тому или иному полю или подполю. Поэтому следует быть особенно внимательным, чтобы формат порождал корректные данные, особенно в тех случаях, когда индексируются повторяющиеся поля и/или более одного поля. Другими словами, при использовании данного метода, выводимые в соответствии с форматом отбора данные должны быть представлены отдельной строкой для каждого индексируемого элемента.

Метод индексирования 1

Создаёт элемент из каждого подполя или строки, созданных форматом. Так как в этом случае система будет производить поиск кодов разделителей подполей в строке, созданной форматом, то для обеспечения правильной работы метода в формате должен быть указан режим проверки mpl (или вообще не указан никакой режим, так как режим проверки выбирается по умолчанию), который обеспечивает сохранность разделителей подполей в выходном результате формата. Напомним, что режимы заголовка и данных заменяют разделители подполей на знаки пунктуации. Отметим, что метод индексирования 1 позволяет сделать описание более коротким, чем метод индексирования 0.

Ниже приведены примеры использования данного метода. Показано, как работает метод в режиме данных.

Содержание поля 26 используемой в примерах записи:

```
26 ^aParis^bUnesco^c1965
```

ТВП	Результат форматирования	Порождённые ТВП элементы
`1 1 mpl,v26`	`^aParis^bUnesco^c1965`	Paris Unesco 1965
`1 0 mhl,v26^a/v26^b/v26^c`	Paris Unesco 1965	Paris Unesco 1965
`1 1 mdl,v26`	`Paris, Unesco, 1965`	`Paris, Unesco, 1965`

Метод индексирования 2

Создаёт элемент из каждого термина или фразы, заключенных в угловые скобки (<…>). Любой текст, расположенный вне скобок, не индексируется. Заметим, что данный метод требует, чтобы в формате указывался режим проверки, так как любой другой режим вывода удаляет угловые скобки. Например, текст

```
<Отчет> по использованию <информатики> и <программирования> в <средней школе>
```

приведет к порождению следующих элементов:

```
отчет
информатики
программирования
средней школе
```

Метод индексирования 3

Создаёт элемент из каждого термина или фразы, заключенных в косые черты (/…/). Во всём остальном он работает точно так же, как и метод индексирования 2. Например, текст

```
/Отчет/ по использованию /информатики/ и /программирования/ в /средней школе/
```

приведет к порождению следующих элементов:

```
отчет
информатики
программирования
средней школе
```

Метод индексирования 4

Создаёт элемент из каждого слова в тексте, созданном форматом.

Подробнее о выборе слов см. подраздел *Алгоритм выбора слов*.

При использовании данного метода для индексации поля, содержащего разделители подполей, в формате выборки данных необходимо указать режимы заголовка или данных (mhl или mdl) с тем, чтобы замена разделителей подполей произошла до индексации, так как в противном случае буква разделителя подполей будет рассматриваться как составная часть слова.

Методы индексирования 5, 6, 7, 8

Методы индексирования 5, 6, 7 и 8 аналогичны соответственно методам 1, 2, 3, 4 за исключением того, что они дополнительно предоставляют возможность присоединять к индексируемым терминам префиксы. Присоединяемый префикс определяется в формате выборки данных в виде безусловного литерала и имеет следующий вид:

```
'dp...pd', [format]
```

где:

- `d` – выбранный по усмотрению пользователя ограничитель, который не попадает в текст префикса;
- `p..p` – собственно префикс.

Например, строка ТВП

```
1  8  '/K=/',v200^a
```

приведет к индексированию каждого слова подполя `a` поля 200 с предварительным присоединением к каждому термину префикса `K=`.

Эти методы широко применяются в системе ИРБИС для определения принадлежности терминов к определенным элементам описания. Именно на основе этих методов создается модель словарей по различным элементам данных ("Авторы", "Заглавие" и т.д.). При этом при показе словарей соответствующие префиксы опускаются.

Метод индексирования 9

Метод индексирования 9 был разработан в качестве расширения концепции ТВП. Реализован в рамках технологии полнотекстовых баз данных ИРБИС, и работает только для функций создания и актуализации словаря.

Подробнее см. подразделе *Метод индексирования 9* статьи *Механизм полнотекстового поиска*.

Идентификатор поля

Идентификатор поля – это число, которое назначается каждому элементу, создаваемому на шаге индексирования. Смысл идентификатора поля зависит от целей последующего использования ТВП. В случае использования ТВП для создания инвертированного файла идентификатор поля представляет собой вторую компоненту индексной ссылки для соответствующего термина. При использовании ТВП для переформатирования (импорта, экспорта, копирования), идентификатор поля представляет собой метку, присваиваемую экспортируемому, импортируемому или копируемому полю.

Алгоритм выбора слов

Выбор слов происходит при использовании методов индексирования 4, 8 и 9.

Выбор слов необходим при индексировании базы данных.

Примечание: см. также рекомендации по настройке процесса выбора слов в словарь базы данных.

Выбор слов из текста

Разбиение некоторого текста на слова осуществляется на основе разделения символов на ***алфавитные*** (являющиеся частью слова) и ***неалфавитные***.

Разбиение текста на слова осуществляется в соответствии со следующими правилами:

- словом считается непрерывная последовательность ***алфавитных символов***;
- ***неалфавитные*** символы отделяют одно слово от другого.

Символ считается алфавитным, если код символа присутствует в так называемой ***таблице алфавитных символов***. ***Таблица алфавитных символов*** присутствует в системе ИРБИС в виде текстового файла `ISISACW.TAB`.

Отбрасывание стоп-слов

Система отбрасывает стоп-слова.

Список стоп-слов определён в ***файле стоп-слов***.

Ссылки

См. также:

- Базы данных ИРБИС
- Язык форматирования системы ИРБИС
- Язык запросов ИРБИС
- Индекс базы данных ИРБИС
- Механизм полнотекстового поиска
- Механизм актуализации записи

Источники информации:

- Общее описание системы ИРБИС64