УДК 681.3.07

А. В. Брешенков, Мин Тхет Тин

ПРЕОБРАЗОВАНИЕ НЕРЕЛЯЦИОННЫХ ТАБЛИЦ К РЕЛЯЦИОННОМУ ВИДУ БЕЗ ИСПОЛЬЗОВАНИЯ СЛОЖНЫХ АТРИБУТОВ

Рассмотрены проблемы избавления от сложных атрибутов и исключения внутренних подзаголовков при преобразовании нереляционных таблиц к реляционному виду, а также задачи преобразования нереляционных таблиц, в частности задача избавления от подзаголовков таблиц.

E-mail: minthettin85@gmail.com

Ключевые слова: *информация табличного вида, реляционная таблица, однотипность, преобразование, внешние и внутренние подзаголовки, сложные атрибуты.*

Во всех областях человеческой деятельности накопился большой объем информации, представленной в табличной форме. В качестве примеров информации такого рода служат справочники, каталоги, ведомости, прайс-листы и др. Причем данные представлены в различных формах: в видах электронных таблиц, текстовых редакторов, текстовых процессоров и др. Опираясь на опыт работы в организациях, занимающихся хранением, использованием и обработкой информации, можно сказать, что, как правило, существует потребность использования средств современных баз данных для обработки такой информации. Введем понятие информации табличного вида, которое определяет все указанные представления документов. Опыт разработок баз данных показывает, что существует проблема преобразования информации табличного вида к реляционному виду, т.е. к виду, приемлемому для использования в реляционных системах управления базами данных.

Реляционные таблицы — это таблицы, имеющие следующие свойства:

- в таблице не должно быть подзаголовков [1];
- для каждого столбца должно выполняться требование однотипности элементов;
- каждая строка должна быть уникальной;
- порядок строк и столбцов может быть произвольным;
- каждая ячейка не должна содержать в себе других ячеек.

В работе [2] рассмотрены задачи автоматизированного проектирования реляционных баз данных на основе использования существующей информации табличного вида. В [2] проанализированы и решены следующие задачи:

- приведение табличной информации к реляционному виду;

• назначение первичных ключей в заполненных реляционных таблицах;

• нормализация заполненных реляционных таблиц;

• формирование связей между заполненными реляционными таблицами.

В настоящей работе рассматриваются вопросы, относящиеся к первой задаче, причем только в части последнего требования к реляционным таблицам (в ней не должно быть подзаголовков). При этом анализируются вопросы, которые не нашли своего отражения в работе [2].

В нереляционных таблицах могут встречаться подзаголовки трех типов: внешние, внутренние и подзаголовки-столбцы. В реляционных таблицах подзаголовки недопустимы. Рассмотрим примеры подзаголовков различных типов.

Внешние подзаголовки. Такого рода подзаголовки часто встречаются в таблицах-справочниках. В качестве примера в табл. 1 приведем фрагмент каталога электрооборудования автомобилей.

Таблица 1

Указатели				Системы зажигания			
Габаритные указатели		Поворотные указатели		Контактные		Бесконтактные	
№	Марка	№	Марка	№	Марка	№	Марка

Как следует из структуры таблицы, она имеет заголовки трех уровней — основной заголовок, подзаголовоки первого и второго уровней. В реляционных таблицах допустимы заголовки только одного уровня. Для информации, представленной в табличной форме для общего случая, справедливы следующие утверждения:

T={З1,З2,. . . ,ЗК,. . . .,ЗN}, ЗК={П1ЗК,П2ЗК,. . . ,ПLЗК,. . . .,ПFЗК}

ПLЗК={П1ПLЗК, П2ПLЗК,. . . ., ПQПLЗК,. . . ., ПРПLЗК},

где Т — схема таблицы; ЗК — схема K-го заголовка таблицы Т; ПLЗК — L-й подзаголовок первого уровня ЗК-го заголовка таблицы Т; ПQПLЗК — Q-й подзаголовок второго уровня ПLЗК-го подзаголовока первого уровня ЗК-го заголовки таблицы Т.

Следует обратить внимание на то, что число уровней подзаголовков может быть существенно бóльшим, но правила формирования их псевдонимов аналогичны приведенным. Здесь же из соображений ограничения объема работы принято ограничиться тремя уровнями заголовков и подзаголовков. Как показывает анализ доступных документов различных предметных областей, обычно задействуют два уровня

подзаголовков. Это следует из примера, приведенного во фрагменте таблицы крепежных деталей (табл. 2).

Таблица 2

Болты				Шайбы				Винты								Штифты				Гайки				Шплинты	
6-гранные		Цилиндрические с внутренним 6-гранником		Плоские		Гравер		с чечевичной головкой		с полусферической головкой		с потайной головкой		с цилиндрической головкой		Простые		Конусообразные		Основные		Коронки			
N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T	N	T

В этой таблице могут быть задействованы заголовки более высокого уровня. Их схема следующая:
нормаль = {гайки, болты, винты, шайбы};
фиксирующие детали = {штифты, шплинты, клинья}.

В обозначениях столбцов типа ПQПLЗК заложены индексы, с помощью которых могут быть построены циклы, содержащиеся в алгоритмах исключения внешних заголовков. Например, для основного цикла могут быть задействованы индексы K и N. Для 1-го внутреннего цикла индексы L и K. Для 2-го внутреннего индексы Q и P. Чтобы исключить сложные заголовки и привести таблицу к реляционному виду, необходимо организовать сканирование подзаголовков самого нижнего уровня и для каждого подзаголовка осуществить сбор всей относящейся к нему информации. Затем собранную информацию необходимо использовать в качестве неделимого (атомарного) заголовка.

Например, 1-й полученный таким образом заголовок табл. 2 будет выглядеть следующим образом: N 6-гранные болты или болты 6-гранные N. Второй заголовок может быть таким: болты 6-гранные тип. Далеко не всегда заголовки, полученные таким образом, могут быть воспринимаемы потенциальным пользователем базы данных. Более того, их длина может превышать максимально допустимую длину атрибутов используемых инструментальных средств. В связи с этим процедура формирования атомарных столбцов должна быть не автоматической, а автоматизированной, т.е. пользователь средств исключения внешних подзаголовков должен иметь возможность вмешаться в процесс формирования заголовков столбцов в целях присвоения атомарным столбцам приемлемых имен. Один из возможных алгоритмов исключения внешних подзаголовков приведен в работе [3].

Следует обратить внимание на то, что исходная информация может быть представлена различными способами: на бумаге, в виде текстовых файлов, в формате электронных таблиц и др. Удобнее всего для разработчика средств преобразования было бы использование в качестве исходного единый формат данных. Например, в качестве основного — использование формата Excel, а все другие представления данных преобразовывать в данный формат. Но, к сожалению, не всегда это возможно. Например, табличные данные. представленные на бумаге, далеко не всегда удается отсканировать в формат электронных таблиц.

Поэтому необходимо предусмотреть реализацию алгоритма для информации табличного вида, представленной не только в формате .xls, но и в формате .txt.

Внутренние подзаголовки. Подзаголовки такого рода часто встречаются в прайс-листах, в отчетах по покупке и продаже товаров, т.е. в тех случаях, когда структуры нескольких таблиц совпадают, но их данные относятся к различным группам. Группировка может осуществляться, например, по датам, категориям товаров, регионам. В качестве примера рассмотрим список проданных за день товаров магазина “Соки–воды” (табл. 3).

Таблица 3

Название	Объем	Цена	Количество
Газированная вода			
Тархун	1 л	30 р	7
Байкал	2 л	60 р	5
Колокольчик	1,5л	40р	10
Соки			
Вишневый	1 л	45 р	3
Ананасовый	1 л	45 р	15
Яблочный	1 л	45 р	8
Морсы			
Клюквенный	1,5 л	60 р	7
Малиновый	1,5 л	60 р	7
Рябиновый	1 л	45 р	7
. . .			

Как следует из примера, таблица легко воспринимается визуально. Однако в таком виде она неприменима в составе баз данных. Такая таблица с небольшим числом строк без особых усилий может быть

преобразована в реляционную таблицу вручную. Однако в реальных таблицах может быть тысячи или более строк. В этом случае их преобразование не очевидно. Необходимы специальные автоматизированные средства.

В случае если табл. 3 будет импортирована в какую-либо систему управления базами данных она примет вид табл. 4.

Таблица 4

Название	Объем	Цена	Количество
Газированная вода			
Тархун	1 л	30 р	7
Байкал	2 л	60 р	5
Колокольчик	1,5 л	40 р	10
Соки			
Вишневый	1 л	45 р	3
Ананасовый	1 л	45 р	15
Яблочный	1 л	45 р	8
Морсы			
Клюквенный	1,5 л	60 р	7
Малиновый	1,5 л	60 р	7
Рябиновый	1 л	45 р	7

Как следует из табл. 4, смысл импортированной таблицы утрачен. В частности, из таблицы следует, что напиток “Газированная вода” не имеет объема, цены и не продавался. Хотя на самом деле все обстоит по-другому. Необходимо избавиться от противоречий такого рода. Для этого

- нужно выявить внутренние подзаголовки;
- исключить соответствующие строки;
- сформировать реляционную таблицу или таблицы таким образом, чтобы сохранить смысл исходных таблиц.

Информация табличного вида нередко представлена таким образом, что в области заголовков имеют место заголовки, которые включают в себя несколько позиций. Например, какой-либо заголовок может быть представлен следующим образом: континент, часть света, страна. Таблица такого рода не является реляционной, и преобразование ее к реляционному виду — задача нетривиальная. Более того, проблема преобразования может существенно усложниться в связи с тем, что такие заголовки влекут за собой необходимость использования сложных заголовков. Например, рассмотрим табл. 5.

Таблица 5

Континент, часть света, страна		Количество крупных городов	Количество крупных рек
Европа			
Восток	Запад		
Россия		33	26
	Испания	8	14

И в первом и втором случаях нарушена первая нормальная форма — атрибуты реляционных таблиц должны быть неделимы. Кроме того, такого рода подзаголовки могут встречаться внутри таблицы. Приведем пример такого рода таблицы (табл. 6).

Таблица 6

Континент, часть света, страна		Количество крупных городов	Количество крупных рек
Европа			
Восток	Запад		
Россия		33	26
	Испания	8	14
Азия			
Восток	Запад		
Мьянма		14	4
	Индия	26	8

Такую таблицу невозможно обрабатывать с помощью языка запросов. В связи с этим ее можно представить в виде двух связанных реляционных таблиц: “Части света. Континенты” и “Страны”. Однако в этой таблице существуют сложные атрибуты, поэтому надо избавляться от сложных атрибутов, избавляться от подзаголовков, которые попали в значения атрибутов. В табл. 7 показаны эти изменения.

Таблица 7

Континент, часть света, страна		Количество крупных городов	Количество крупных рек
Восточная Европа	Западная Европа		
Россия		33	26
	Испания	8	14
Восточная Азия	Западная Азия		
Мьянма		14	4
	Индия	26	8

Предлагается следующая последовательность действий. Формируется новый столбец с номерами частей света, континентов. Сканируется преобразованная таблица; очередной части света, континенту присваивается номер и этот номер распространяется на страны. Но

эта таблица не очень удачна для двух связанных реляционных таблиц: “Части света. Континенты” и “Страны”. Поэтому предлагается следующий вариант — табл. 8.

Таблица 8

Континент, часть света, страна	Количество крупных городов	Количество крупных рек
Восточная Европа		
Россия	33	26
Западная Европа		
Испания	8	14
Восточная Азия		
Мьянма	14	4
Западная Азия		
Индия	26	8

Результат формирования нового столбца с номерами частей света, континентов и заполнением столбца приведен в табл. 9.

Таблица 9

№	Континент, часть света, страна	Количество крупных городов	Количество крупных рек
1	Восточная Европа		
1	Россия	33	26
2	Западная Европа		
2	Испания	8	14
3	Восточная Азия		
3	Мьянма	14	4
4	Западная Азия		
4	Индия	26	8

Результаты формирования новой таблицы “Части света. Континенты” и исключения записей с “частью света, континентом” из исходной таблицы приведены соответственно в табл. 10. В табл. 11 приведены страны.

Таблица 10

№	Континент, часть света
1	Восточная Европа
2	Западная Европа
3	Восточная Азия
4	Западная Азия

Таблица 11

№	Страна	Количество крупных городов	Количество крупных рек
1	Россия	33	26
2	Испания	8	14
3	Мьянма	14	4
4	Индия	26	8

Для таблиц данного вида можно строить реляционные запросы. Затем из записей с “частью света, континентом” формируется новая таблица, соответствующие записи из исходной преобразованной таблицы удаляются. После этих преобразований будут сформированы две таблицы, связанные между собой связью типа $1{:}\infty$.

Описанные манипуляции вполне можно выполнить вручную для небольшого рассмотренного примера, но не для реальных таблиц мощностью десятки тысячи записей. Поэтому необходима разработка алгоритма автоматизированного избавления от сложных атрибутов при преобразовании заполненных нереляционных таблиц к реляционному виду.

Предварительно представим таблицу (отношение) рассматриваемого типа в общем виде (табл. 12).

Таблица 12

A_1	A_2	.....	A_i	.....	A_k
a_{11}	NULL	.....	NULL	.....	NULL
a_{21}	a_{22}	.....	NULL	.....	NULL
a_{31}	NULL	.....	a_{3i}	.....	a_{3k}
NULL	a_{42}	.....	a_{4i}	.....	a_{4k}
a_{j1}	NULL	.....	NULL	.....	NULL
.....	.....	.....	.....	.....	.....
a_{f1}	NULL	.....	NULL	.....	NULL
a_{m1}	a_{m2}	.....	a_{mi}	.....	a_{mk}

Особенность таблицы такого рода состоит в том, что в некоторых ее строках значение имеет один или два атрибута. Принимается, что такой атрибут является внутренним сложным подзаголовком таблицы.

Предлагается укрупненный алгоритм исключения сложных подзаголовков:

П1: Выполняется сканирование всех записей отношения R. Каждая запись проверяется на наличие в ней только одного значения атрибута. Записи такого рода подсчитываются. Если таких записей несколько, то подзаголовки в отношениях R присутствуют и выполняется переход к следующему пункту (П2). В противном случае алгоритм завершает работу.

П2: Избавление от сложных атрибутов:

```
r=1;
FOR r=1 то m
    i=1;
    FOR i=1 то 2
      WHILE a1i<>a2i;
      a1i:=Concat (a1i,' ',a2i);
    END;
END;
```

П3: Для двух связанных реляционных таблиц:

```
k=0;
i=1;
j=1;
r=1;
WHILE arj<>END FILE
    WHILE aij <> Подзаголовок
      IF k=0 THEN r=1;
      k=1;
      a1ij:=aij;
      i:=i+1;
    END WHILE;
    k:=0;
    IF j=1 THEN j:=2;
    ELSE j:=1;
END WHILE;
```

П4: К отношению R приписывается дополнительный атрибут KR с типом “числовой”.

```
      COUNTER:=0;
```

П5: Выполняется сканирование всех записей отношения R’. Если в записи имеется только одно заполненное значение атрибута, то счетчик подзаголовков COUNTER увеличивается на единицу. Атрибуту KR присваивается значение COUNTER.

П6: Создается новое отношение R2, включающее в себя два атрибута NR и атрибут с подзаголовком. Выполняется сканирование всех записей отношений R’. Записи, которые имеют только одно (кроме ключевого) заполненное значение атрибута, перемещаются в отношение R2.

В результате выполнения алгоритма сформируются отношение R2 (с подзаголовками исходной таблицы) и отношение R1 без подзаголовков. Связи между таблицами обеспечиваются посредством ключевых атрибутов KR, которые присутствуют в обоих отношениях.

Формализованный алгоритм исключения внутренних подзаголовков и избавления от сложных атрибутов при преобразовании нереляционных таблиц к реляционному виду выглядит следующим образом:

```
k=0;
j=1;
r=1;
    FOR r=1 то m
       i=1;
       FOR i=1 то 2
          WHILE a1i<>a2i;
          a1i:=Concat (a1i,' ',a2i);
END
WHILE arj<>END FILE
    WHILE aij <> Подзаголовок
       IF k=0 THEN r=1; k=1;
       a1ij:=aij;
       i:=i+1;
    END WHILE;
    k:=0;
    IF j=1 THEN j:=2;
END WHILE;
COUNTER=0;
FOR r=1 то m
    COUNTER1=0;
    FOR f=1 то p
       IF arp=NULL THEN COUNTER1=COUNTER1+1;
    NEXT f
    IF COUNTER1=p-1 THEN COUNTER=COUNTER+1;
NEXT r
IF COUNTER < 2 THEN EXIT
REM Формирование двух отношений
R'=R(A1,. . . . . ,Ai,. . . . . ,Ap)+R(KR)
COUNTER=0
FOR r=1 то m
COUNTER1=0;
FOR f=1 то p
    IF arp=NULL THEN COUNTER1=COUNTER1+1;
    NEXT f
IF COUNTER1 =p-1 THEN
    COUNTER=COUNTER+1;
    Z(R2COUNTER,1)=COUNTER;
    Z(R2COUNTER,2)=arp;
    DELETE * FROM R' WHERE (A1=arp);
ELSE
    Z(R'r,1)=COUNTER;
```

END IF
NEXT r;

Здесь: m — мощность отношения R; P — степень отношения R; END FILE — конец таблицы; подзаголовок — заголовок 2-го уровня; выражение R'=R+R(KR) означает добавление к R атрибута с именем KR; выражение $Z(R2_{COUNTER,1})$ означает значение элемента R2 в строке COUNTER и первом столбце; выражение Z(R'r,1) означает значение элемента R' в строке r и первом столбце.

Заключение. Исключение подзаголовков в таблицах небольшой размерности может быть выполнено вручную в соответствии с рекомендациями авторов.

В отношениях с числом столбцов несколько десятков и числом строк несколько тысяч предложенный метод автоматизированного преобразования заполненных таблиц к первой нормальной форме позволяет избавиться от сложных атрибутов, исключить подзаголовки внутри таблиц.

СПИСОК ЛИТЕРАТУРЫ

1. Дейт К., Дж. Введение в системы баз данных: Пер. с англ. – М.: Вильямс, 2005. – 1328 с.
2. Балдин А. В., Брешенков А. В. Анализ проблемы проектирования реляционных баз данных на основе использования существующей информации табличного вида // Вестник МГТУ им. Н.Э. Баумана. Сер. Приборостроение. – 2007. – № 4. – С. 43–52.
3. Брешенков А. В., Гудзенко Д. Ю., Казаков Г. И. Проектирование реляционных баз данных на основе информации табличного типа: Учеб. пособие. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2009. – 150 с.

Статья поступила в редакцию 6.06.2011

Александр Владимирович Брешенков родился в 1955 г., окончил МВТУ им. Н.Э. Баумана в 1982 г. Д-р техн. наук, профессор кафедры "Компьютерные системы и сети" МГТУ им. Н.Э. Баумана. Автор 110 научных работ в области САПР ЭВМ и баз данных.

A.V. Breshenkov (b. 1955) graduated from the Bauman Moscow Higher Technical School in 1982. D. Sc. (Eng.), professor of "Computer Systems and Networks" department of the Bauman Moscow State Technical University. Author of 110 publications in the field of systems of automated design and data bases.

Мин Тхет Тин родился в 1985 г., окончил МГУ им. М.В. Ломоносова в 2010 г. Стажер кафедры "Компьютерные системы и сети" МГТУ им. Н.Э. Баумана.

Min Thet Tin (b. 1985) graduated from Moscow State University in 2010. The trainee of computer System and Network department of the Moscow State Technical University.