УДК 004.622:004.654

АЛГОРИТМЫ И ПРОГРАММНЫЕ СРЕДСТВА АВТОМАТИЗАЦИИ НОРМАЛИЗАЦИИ ТАБЛИЦ В ХРАНИЛИЩАХ ДАННЫХ

Л. С. Лазута, А. А. Карпук

Белорусская государственная академия связи,
Минск, Беларусь, lenya.lazuta@mail.ru, a_karpuk@mail.ru

Приведен обзор известных алгоритмов приведения таблиц хранилищ данных к третьей нормальной форме (3НФ). Показаны недостатки существующих подходов к автоматизации приведения таблиц к 3НФ. Представлено разработанное программное средство автоматизации приведения таблиц хранилищ данных к 3НФ.

Ключевые слова: хранилище данных; функциональная зависимость; третья нормальная форма; алгоритмы нормализации данных; автоматизация приведения к третьей нормальной форме.

ALGORITHMS AND SOFTWARE FOR AUTOMATION OF DATA NORMALIZATION IN DATA WAREHOUSING

L. S. Lazuta, A. A. Karpuk

Belarusian State Academy of Communications,
Minsk, Belarus, lenya.lazuta@mail.ru, a_karpuk@mail.ru

The article provides an overview of known algorithms for reducing data tables to the third normal form (3NF). The shortcomings of existing approaches to automating the reduction of tables to 3NF are shown. The developed software for automating the reduction of data warehouse tables to 3NF is presented.

Keywords: data warehouse; functional dependency; third normal form; data normalization algorithms; automation of conversion to third normal form.

1. Введение

Современные системы Big Data могут иметь архитектуру реляционного хранилища данных (Relational Data Warehouse, RDW), озера данных (Lake Data), современного хранилища данных (Modern Data Warehouse, MDW), фабрики данных (Data Fabric), озерного хранилища данных (Data Lakehouse), сетки данных (Data Mesh) [1]. Во всех перечисленных архитектурах, кроме архитектуры Lake Data, на одном или нескольких этапах

обработки данных решаются задачи очистки данных, предварительной обработки данных и нормализации данных в виде приведения таблиц данных к третьей нормальной форме (3НФ). Для решения задачи нормализации данных требуется знание системы образующих структуры функциональных зависимостей (ФЗ) между данными. В настоящей работе рассматриваются алгоритмы и программные средства автоматизации приведения таблиц хранилищ данных к 3НФ.

2. Алгоритмы приведения таблиц данных к 3НФ

Теория приведения таблиц данных к 3НФ бурно развивалась в последние десятилетия прошлого века в рамках теории проектирования реляционных баз данных. В настоящее время актуальной является задача приведения к 3НФ таблиц хранилищ данных, которые отличаются от таблиц баз данных значительным увеличением количества атрибутов в таблицах (в десятки раз), количества строк в таблицах (в тысячи раз) и количества выявленных ФЗ между атрибутами таблиц (в десятки раз). В этих условиях приведение таблиц хранилищ данных к 3НФ вручную невозможно, поэтому требуется разработать алгоритмы и программные средства автоматизации приведения таблиц хранилищ данных к 3НФ.

Известны два подхода к приведению таблиц данных к 3НФ с использованием ФЗ между атрибутами таблиц: анализ таблиц на удовлетворение требованиям первой нормальной формы (1НФ), второй нормальной формы (2НФ) и 3НФ, и выделение в отдельные таблицы тех атрибутов, которые нарушают эти требования; синтез таблиц в 3НФ на основе известной системы образующих структуры ФЗ между атрибутами. При первом подходе используются определения 1НФ, 2НФ и 3НФ, данные Э. Коддом. А именно, таблица находится в 1НФ, если все ее атрибуты являются атомарными (неделимыми) с точки зрения их дальнейшей обработки. Таблица находится во 2НФ, если для любого потенциального ключа таблицы любой атрибут таблицы, не входящий в состав этого потенциального ключа, функционально полно зависит от этого ключа, т. е. функционально зависит от всего ключа, но не зависит от никакого подмножества этого ключа. Таблица находится в 3НФ, если она находится во 2НФ, и в ней отсутствуют такие ФЗ между атрибутами, в левой и правой части которых одновременно присутствуют атрибуты, не входящие ни в один из потенциальных ключей таблицы (неключевые атрибуты). Алгоритм приведения таблицы к 3НФ сначала приводит таблицу к 1НФ, а затем приводит к 2НФ и 3НФ путем проверки каждой известной ФЗ таблицы на соответствие требованиям 2НФ и 3НФ. При этом подходе требуется решить задачу поиска всех потенциальных ключей таблицы, которая в общем случае является NP-трудной.

Все алгоритмы синтеза таблиц в 3НФ на основе известной системы образующих структуры ФЗ между атрибутами используют понятие замыкания множества атрибутов относительно структуры ФЗ. Пусть в результате выделения и поиска ФЗ между атрибутами множества атрибутов A получена система образующих структуры ФЗ между атрибутами

$$P = \{X_j \to Y_j \mid X_j \subset A, Y_j \subseteq A, j = \overline{1,m}\}.$$

Структуру ФЗ, заданную системой образующих P, будем обозначать $S(P)$. Замыканием множества атрибутов $X \subset A$ относительно структуры ФЗ $S(P)$ называется множество $X^+(P) \subseteq A$, такое, что для любого $Y \subseteq A$ из $X \to Y$ следует $Y \subseteq X^+(P)$. Система образующих структуры ФЗ $E = \{H_j \to T_j \mid H_j \subset A, T_j \subseteq A, j = \overline{1,n}\}$ называется элементарным базисом структуры ФЗ $S(E)$, если удаление любого атрибута из левой или правой части любой ФЗ из E приводит к структуре ФЗ, не эквивалентной $S(E)$. Алгоритмы построения замыкания заданного множества атрибутов относительно заданной системы, образующих структуры ФЗ и построения элементарного базиса структуры ФЗ, можно найти в работе [2].

Алгоритм приведения таблиц к 3НФ К. Делобеля – Р. Кейси [3] находит элементарный базис структуры ФЗ E и представляет таблицу в виде естественного соединения проекций на атрибуты каждой ФЗ элементарного базиса, если хотя бы одна из проекций содержит один из потенциальных ключей таблицы. Для проверки этого условия достаточно построить замыкание левой части каждой ФЗ элементарного базиса относительно этого элементарного базиса. Если хотя бы для одной ФЗ в построенное замыкание войдут все атрибуты таблицы, то условие выполняется. Если условие не выполняется, то к полученным проекциям добавляется проекция на любой потенциальный ключ таблицы. Алгоритм поиска одного потенциального ключа таблицы сводится к последовательному удалению из списка атрибутов таблицы тех атрибутов, которые принадлежат замыканию оставшихся атрибутов относительно элементарного базиса структуры ФЗ.

Алгоритм приведения таблиц к 3НФ П. Бернштейна [4] на первом шаге также находит элементарный базисе E структуры ФЗ. Затем производится попарное объединение тех ФЗ из E, для которых биекция левых частей принадлежит $S(E)$. Таблица представляется в виде естественного соединения проекций по атрибутам, вошедшим в каждое из попарных объединений, и проекций по атрибутам каждой из ФЗ, оставшихся в элементарном базисе E. Если ни одна из полученных проекций не содержит

ни одного из потенциальных ключей таблицы, то к полученным проекциям необходимо добавить проекцию по атрибутам любого потенциального ключа таблицы. С. Ислуром [5] была предложена процедура, позволяющая сразу производить групповое, а не попарное объединение тех ФЗ из E, биекции левых частей которых принадлежат $S(E)$. Действительно, необходимым и достаточным условием наличия биекции левых частей некоторой группы ФЗ из E является равенство замыканий левых частей этой группы относительно $S(E)$, что используется в алгоритме С. Ислура.

На первом шаге алгоритма приведения таблиц к 3НФ Е. А. Неклюдовой – М. Ш. Цаленко [6] также строится элементарный базис E, находятся все потенциальные ключи таблицы, и множество атрибутов, вошедших в потенциальные ключи, рассматривается как подструктура структуры ФЗ $S(E)$. Из элементарного базиса E удаляются все ФЗ, вошедшие в эту подструктуру. Производится расширение подструктуры путем добавления атрибутов, не вошедших в состав потенциальных ключей и функционально полно зависящих от потенциальных ключей. Если все атрибуты таблицы войдут в формируемую подструктуру, то она находится в 3НФ. В противном случае анализируются ФЗ, оставшиеся в элементарном базисе. Если в них входят не все атрибуты из A, то множество атрибутов этих ФЗ рассматривается как подструктура структуры $S(E)$, для которой все шаги алгоритма повторяются, иначе применяется алгоритм П. Бернштейна. Авторы [6] утверждали, что их алгоритм приводит к минимальному количеству получаемых таблиц в 3НФ. Однако в работе [7] А. А. Карпук показал, что в общем случае алгоритм Е. А. Неклюдовой – М. Ш. Цаленко не дает оптимального решения по количеству получаемых таблиц в 3НФ и получил необходимые и достаточные условия оптимальности алгоритмов Делобеля – Кейси и Бернштейна по количеству получаемых таблиц в 3НФ.

3. Автоматизация приведения таблиц данных к 3НФ

До настоящего времени ни в одном из имеющихся CASE-средств проектирования баз данных нет средств автоматического приведения таблиц к 3НФ [8]. Возможно, это объясняется тем, что все перечисленные алгоритмы, кроме алгоритма Э. Кодда, начинают работу с построения элементарного базиса структуры ФЗ на множестве атрибутов (в англоязычной литературе – с построения неизбыточного покрытия структуры ФЗ). К. Делобель и Р. Кейси изложили алгоритм построения элементарного базиса структуры ФЗ в терминах задачи декомпозиции булевых функций, затем Д. Мейер [9] описал алгоритм построения элементарного базиса структуры ФЗ в виде, требующем построения замыкания структуры ФЗ.

Обе эти задачи являются NP-трудными. Некоторые авторы утверждают, что для реализации алгоритма К. Делобеля – Р. Кейси требуется найти все потенциальные ключи структуры ФЗ, а это тоже NP-трудная задача. Естественно, что разработчики CASE-средств проектирования баз данных не включали в состав своих систем NP-трудные задачи, которые за приемлемое время можно решить только приближенными методами.

Известно несколько подходов к автоматизации приведения таблиц данных к 3НФ. И. А. Зорин [10] предложил подход, который он назвал теоретико-графовым. Работа И. А. Зорина имеет терминологические неточности и фактические ошибки. То, что в его работе называется «отношение», на самом деле должно называться «функциональная зависимость». Построение графа «отношений» между полями универсальной таблицы (атрибутами) начинается с вершин, соответствующих ключу всей системы «отношений». Но в общем случае вся система ФЗ между атрибутами имеет более одного ключа, что не предусмотрено в алгоритме И. А. Зорина. Система образующих структуры ФЗ между атрибутами представляется в виде графа, вершины которого соответствуют атрибутам, а дуги – ФЗ между атрибутами. При таком представлении ФЗ с двумя и более атрибутами в левой части не отличаются от ФЗ с одним атрибутов в левой части. На рис. 1 из работы [10] совершенно непонятно, показаны там три ФЗ $a \to i$, $e \to i$, $h \to i$ или одна ФЗ $aeh \to i$. Такой подход возможен только в случае, когда в левой части любой ФЗ системы образующих содержится ровно один атрибут. Для правильного представления системы образующих структуры ФЗ между атрибутами в виде графа следует использовать двудольные графы или гиперграфы

Проверка выполнения условий нахождения таблицы во 2НФ производится примитивным путем построения множества всех подмножеств единственного ключа таблицы и проверки наличия в системе образующих структуры ФЗ соответствующих ФЗ. На самом деле, условие нахождения таблицы во 2НФ состоит не в отсутствии соответствующей ФЗ в системе образующих, а в отсутствии этой зависимости в замыкании множества ФЗ относительно структуры ФЗ. Т. е. в общем случае искомой ФЗ может не быть в системе образующих структуры ФЗ, но таблица не будет находиться во 2НФ. При приведении таблиц к 3НФ не учитывается ситуация, когда таблица имеет более одного ключа, а также ситуация, когда в таблице имеется ФЗ неключевого атрибута от двух или более атрибутов, в состав которых входят ключевые и неключевые атрибуты. Таким образом, алгоритм И. А. Зорина может дать правильное решение только в случае, когда система ФЗ между атрибутами имеет единственный ключ, и все ФЗ системы образующих структуры ФЗ содержат один атрибут в левой части.

Ошибку И. А. Зорина при представлении системы образующих структуры ФЗ в виде ориентированного графа повторяет И. В. Клименко [11], который все свои рассуждения о нормализации таблиц реляционной базы данных строит на представлении системы образующих структуры ФЗ в виде графа, в котором вершины соответствуют множеству атрибутов, а дуги соответствуют множеству ФЗ. Еще раз отметим, что такое представление возможно только в тех случаях, когда в левых частях ФЗ содержится один атрибут. Далее И. В. Клименко вводит понятие максимального транзитивно независимого множества атрибутов как максимального подмножества вершин графа, из которых транзитивно не достигаются другие вершины графа и сводит задачу нормализации таблиц базы данных к задаче поиска и анализа всех максимальных транзитивно независимых множеств атрибутов, которая является NP-трудной.

Возникает вопрос, можно ли решить задачу приведения таблиц данных к 3НФ без решения NP-трудных задач. Ответ на этот вопрос дал А. А. Карпук [7], который преобразовал алгоритм К. Делобеля – Р. Кейси к виду, не требующему построения замыкания структуры ФЗ. В этой модификации алгоритма К. Делобеля – Р. Кейси требуется неоднократно строить замыкания множества атрибутов относительно структуры ФЗ, а также находить один любой ключ этой структуры, но эти задачи имеют полиномиальную сложность. Это дает возможность разработать программные средства для автоматической нормализации таблиц хранилища данных, которые можно применять для таблиц данных с любым количеством атрибутов и любым количеством известных ФЗ между атрибутами. Попытка разработать программное средство для автоматизации приведения таблиц данных к 3НФ была предпринята А. И. Говоровым и Р. И. Масленниковым под руководством М. М. Говоровой [12], однако в их программе был реализован алгоритм приведения таблиц к 3НФ И. А. Зорина, недостатки которого указаны выше.

Авторами разработано экспериментальное программное средство ReductionTo3NF для приведения таблиц хранилища данных к 3НФ. Программное средство реализовано в виде хранимых процедур и функций на языке Transact-SQL для СУБД Microsoft SQL Server. Исходными данными для программного средства являются данные о составе атрибутов таблицы, хранящиеся в таблице базы данных (БД) Attributes и данные о системе образующих структуры ФЗ между атрибутами, хранящиеся в таблицах БД FD, Left_Attributes_FD и Right_Attributes_FD. Результатом работы программного средства являются данные о таблицах в 3НФ, записанные в таблицы БД Tables3NF, Attributes_Tables, Keys_Tables, Attributes_Keys. В состав программного средства входят хранимые процедуры и функции построения замыкания множества атрибутов; построения элементарного базиса структуры ФЗ между

атрибутами путем удаления «посторонних» атрибутов в левых и правых частях ФЗ системы образующих; преобразования элементарного базиса структуры ФЗ в таблицы в 3НФ по алгоритму К. Делобеля – Р. Кейси, включая функции проверки наличия потенциального ключа таблицы в одной из полученных таблиц в 3НФ и поиска одного потенциального ключа таблицы; построения таблиц в 3НФ по алгоритмам П. Бернштейна и С. Ислура; проверки оптимальности полученного решения по количеству полученных таблиц в 3НФ.

Разработанное экспериментальное программное средство для автоматического приведения таблиц хранилища данных к 3НФ может непосредственно использоваться в системах Big Data, а также может служить основой для разработки новых платформенно независимых масштабируемых систем нормализации данных в системах Big Data.

Библиографические ссылки

1. *Serra J.* Deciphering Data Architectures. Choosing Between a Modern Data Warehouse, Data Fabric, Data Lakehouse and Data Mesh. O'Reilly, 2024.

2. *Карпук А. А*. О построении элементарного базиса системы функциональных зависимостей в базе данных // Информационные технологии и программные средства: проектирование, разработка и применение: сб. научн. статей / Под ред. Л. В. Рудиковой. Гродно : ГрГУ, 2011. С. 185–190.

3. *Delobel C., Casey R. G*. Decomposition of a Data Base and the Theory of Boolean Switching Functions // IBM J. Res. And Dev. 1973. Vol. 17, no 5. P. 374–386.

4. *Bernstein P. A*. Synthesizing third normal form relations from functional dependencies // ASM Transactions on Database Systems. 1976. Vol. l, no 4. P. 277–298.

5. *Isloor S. S.* An algorithm with logical simplicity for designing third normal form relational database schema from functional dependencies // Proc. of Int. Conf. on DBMSs (ICMOD 78). Fast Milan, Italy. 1978. P. 31–50.

6. *Неклюдова Е. А., Цаленко М. Ш.* Синтез логической схемы реляционной базы данных // Программирование. 1979. № 6. С. 58–68.

7. *Карпук А. А.* Алгоритмы нормализации таблиц реляционной базы данных // Системы управления и информационные технологии. 2017. № 2(68). С. 53–67.

8. *Анисимова Н. С., Назарова О. Б.* CASE-средства для проектирования баз данных: обзор и краткая характеристика // Наука, Информатизация, Технологии, Образование : материалы XI международной научно-практической конференции, ФГАОУ ВО «Российский государственный профессионально-педагогический университет», Екатеринбург, 26 февраля – 2 марта 2018 г. С. 472–480.

9. *Мейер Д.* Теория реляционных баз данных: Пер. с англ. М. : Мир, 1987.

10. *Зорин И. А.* Теоретико – графовое приведение реляционной базы данных к третьей нормальной форме Э. Кодда // Устойчивое инновационное развитие, проектирование и управление. 2009. Т. 5. С. 51–59.

11. *Клименко И. В.* Метод формальной нормализации отношений реляционной базы данных // Вестн. Ростовского гос. ун-та путей сообщения. 2012. № 2. С. 79–87.

12. *Говорова М. М., Говоров А. И., Масленников Р. И.* Программа нормализации реляционных баз данных как основа предметно-ориентированной интеллектуальной обучающей системы // Образовательные технологии и общество. 2015. Т. 18, № 1. С. 505–518.