ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 201 за 2018 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

Трифонова Е.Е.

Восстановления баз данных с ограничениями из одного класса

Рекомендуемая форма библиографической ссылки: Трифонова Е.Е. Восстановления баз данных с ограничениями из одного класса // Препринты ИПМ им. М.В.Келдыша. 2018. № 201. 31 с. doi:10.20948/prepr-2018-201
URL: http://library.keldysh.ru/preprint.asp?id=2018-201

Ордена Ленина
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Российской академии наук

Е. Е. Трифонова

Восстановления баз данных с ограничениями из одного класса

Москва — 2018

Трифонова Е.Е.

Восстановления баз данных с ограничениями из одного класса

В данной работе рассматриваются вопросы построения наилучших восстановлений баз данных, содержащих противоречия, с точки зрения формальной логики. Выделяется класс формул-ограничений, позволяющий выражать на языке исчисления предикатов формулы-ограничения, записанные на языке SQL, использующемся в реальных базах данных. Для этого класса рассматриваются различные варианты построения наилучших восстановлений в зависимости от вида формулы.

Ключевые слова: формулы-ограничения, восстановление базы данных, противоречие.

Ekaterina Evgen'evna Trifonova

Database repairs for a class of constraints

We consider the use of formal logic for eliminating inconsistencies in database content. We define a class of formulas in predicate calculus that allows expressing constraints defined in SQL for real databases. We consider different constructions of database repairs depending on the structure of formulas from this class.

Key words: database repair, constraints, inconsistency.

Работа подготовлена при поддержке программы Президиума РАН №01 «Фундаментальная математика и ее приложения» (грант PRAS-18-01).

Оглавление

Введение

В существующих информационных системах одной из проблем является обеспечение целостности и непротиворечивости данных. Особенно острой она становится при работе с распределенными базами данных и информационными системами со множественными источниками информации. Для преодоления возникающих сложностей A. Motro [4] предложил группу методов, которая состоит в классификации неопределенностей и противоречий, возникающих в информационных системах, разработке системы мер и правил, позволяющих корректировать описание реального мира; посредством подобных мер предполагается исправлять имеющиеся проблемы с данными, а также работать в приложениях с неполными и неточными данными. Вторую группу методов предложил J. Chomiki [1–3] — они основаны на создании различных вспомогательных языков и программных средств, которые, с одной стороны, позволяли бы лучше описывать предметную область, а с другой — являлись бы сами средством для работы с противоречиями. Среди предложенных мер: механизм для создания политик, которые бы логически описывали работу системы, чтобы избежать конфликтов между данными [2]; методика выбора предпочтительных кортежей для ответа на некоторые виды запросов (содержащих операторы отсеивания) [3]. Им предложен один из способов построения восстановления, который называется «поддержанием минимально изменяемой целостности», состоящий в том, что при построении восстановления необходимо оставить как можно больше кортежей [1]. Целью настоящей работы является рассмотрение вопроса о том, как будут выглядеть наилучшие восстановления для некоторых классов формул с точки зрения формальной логики при поддержании минимально изменяемой целостности.

В настоящий момент наблюдается большой интерес к базам данных с большими объемами информации, поступающей из множественных источников. Для таких баз данных часта ситуация, когда возникают противоречия, т.е. об одних и тех же вещах содержится разная (противоречащая друг другу) информация. Существуют различные способы работы с возникающими противоречиями. Рассматриваемая в настоящей работе модель наиболее близка к модели, рассматриваемой в [1].

Основные определения

Рассмотрим основные понятия теории баз данных с точки зрения исчисления предикатов.

Схемой базы данных будем называть некое конечное множество предикатов $\mathbb{P}$ и множество ограничений, записанных в виде формулы Φ. *Базой данных* D будем называть совокупность таблиц $\mathbb{T}$ и формулы Φ: $D = \langle \mathbb{T}, \Phi \rangle$. При этом каждому предикату P из множества предикатов $\mathbb{P}$ ставится в соответствие таблица T из $\mathbb{T}$, которая определяет значение истинности предиката. Элементом таблицы является *кортеж*.

Будем обозначать множество кортежей, которые присутствуют в таблице T, как X_T, а множество кортежей, которые присутствуют в базе D, как $X = X(D) = \bigcup\limits_{T \in \mathbb{T}} X_T$. Определим множество функций $F : X \to \mathbb{N}$, где $\mathbb{N}$ — множество натуральных чисел. *Вспомогательным предикатом* от двух переменных $g(x_1, x_2)$ будем называть выражение вида $(f_i(x_1) \mathcal{R} f_j(x_2))$, где $f_i, f_j \in F$, а $\mathcal{R}$ — одно из отношений $<, >, \leqslant, \geqslant, \neq, =$. Пусть G — множество всевозможных вспомогательных предикатов.

Для сокращения записи будем называть *вспомогательным условием* и обозначать как $h(x_1, \ldots, x_n)$ формулу, образованную по следующей схеме:

$$h(x_1, \ldots, x_n) = g_1(x_{i_1}, x_{j_1}) \Diamond g_2(x_{i_2}, x_{j_2}) \Diamond \ldots \Diamond g_k(x_{i_k}, x_{j_k}),$$

где $i_1, \ldots, i_k, j_1, \ldots, j_k \in \{1, \ldots, n\}, g_1, g_2, \ldots, g_k \in G$, $\Diamond$ — места размещения логических операторов $\&$ и $\vee$.

В качестве формулы-ограничения Φ будем рассматривать замкнутые выполнимые формулы на языке первого порядка, записанные с использованием связок $\vee$, $\&$, $\neg$, $\to$, предикатов из $\mathbb{P}$ и вспомогательных предикатов из G. Переменные принимают значения из множества кортежей X. Если формула Φ истинна на множестве X, то базу данных D будем называть *непротиворечивой*.

Если формула Φ ложна на множестве X, то будем говорить, что в базе данных D содержатся противоречия. Устранять противоречия будем посредством удаления кортежей из базы данных, при этом под удалением кортежа будем подразумевать удаление его из всех таблиц, в которых он содержится.

Восстановлением Q для базы данных D будем называть базу данных, для которой выполняется следующее:

1. Схемы баз данных Q и D совпадают.

2. Каждый кортеж из таблицы Q содержится в соответствующей таблице D.

3. База данных Q — непротиворечивая.

4. Добавление к Q любого кортежа из D, который в Q не содержится, в соответствующую таблицу приводит к тому, что в Q возникают противоречия.

Наилучшим восстановлением будем называть восстановление, содержащее наибольшее число кортежей среди всех восстановлений для рассматриваемой базы данных. Наилучших восстановлений может быть несколько.

Будем обозначать как Z_Q множество кортежей, удаленных из D для получения восстановления Q: $Z_Q = X(D) \setminus X(Q)$. Для каждой базы, содержащей противоречия, существует некоторое множество восстановлений. Те из них, которые содержат наибольшее число кортежей, будем называть *наилучшими восстановлениями*. Множество наилучших восстановлений для базы D будем обозначать как $\mathbb{Q}_{max} = \mathbb{Q}_{max}(D)$. Если в базе данных нет противоречий, то $\mathbb{Q}_{max} = \{D\}$. Если все таблицы базы данных пусты, то считаем, что в базе данных нет противоречий.

С логической точки зрения база данных D является интерпретацией формулы Φ. Восстановление является для формулы Φ моделью с универсумом, который есть подмножество множества $X(D)$. Наилучшее восстановление — это модель с максимально возможной мощностью универсума, являющегося подмножеством $X(D)$.

Под логической эквивалентностью формул будем понимать определение, используемое, в частности, в [5].

Утверждение 1. *Множества наилучших восстановлений логически эквивалентных формул совпадают.*

Доказательство. Рассмотрим логически эквивалентные формулы Φ_1 и Φ_2 и произвольную базу D. Допустим, что Q_1 — наилучшее восстановление базы данных D для формулы Φ_1. То есть в Q_1 содержится наибольшее число кортежей среди всех восстановлений. Предположим, что Q_1 не будет наилучшим восстановлением базы данных D для формулы Φ_2. Тогда есть некое восстановление Q_2 базы данных D, которое будет наилучшим для Φ_2. То есть в нем содержится больше кортежей, чем в Q_1. Формула Φ_2 истинна на Q_2, значит, на нем истинна и Φ_1. Следовательно, нашлось такое восстановление базы данных D для формулы Φ_1, а именно Q_2, в котором содержится больше кортежей, чем в Q_1. Это

противоречит тому, что Q_1 — наилучшее восстановление. Наше предположение неверно. Следовательно, наилучшее восстановление для Φ_1 будет наилучшим восстановлением и для Φ_2. □

Класс формул A

В качестве базисных формул возьмём формулы следующих трех видов:

$$\forall x(P(x) \to h(x)), \tag{I}$$

$$\forall x_1 \forall x_2(P(x_1)\&P(x_2) \to h(x_1, x_2)), \tag{II}$$

$$\forall x_1 \exists x_2(P_1(x_1) \to P_2(x_2)\&h_1(x_1, x_2) \vee h_2(x_1)). \tag{III}$$

Будем называть классом A множество конечных формул, построенных из формул вида (I), (II), (III), соединенных конъюнкцией, где $P, P_1, P_2 \in \mathbb{P}$ и h, h_1, h_2 — вспомогательные условия, построенные по приведенным выше правилам с использованием вспомогательных предикатов из G.

Будем говорить, что формула Φ может быть построена в классе A, если найдётся такая формула Φ_1, которая представляет собой конъюнкцию формул вида (I), (II) и (III) и логически эквивалентна формуле Φ. Класс всех формул исчисления предикатов не совпадает с классом A. Однако с помощью формул класса A возможно записать ограничения, которые могут быть выражены на языке SQL базовой версии (подробнее см. [6]). Поэтому рассмотрение способов построения восстановлений баз данных с формулами-ограничениями из класса A имеет особую, в том числе и практическую, значимость.

Подклассы формул-ограничений

Разобьём все формулы, выразимые в классе A, на следующие подклассы:

1. конъюнкция одного вида формул (первый подкласс) :

   (I)&(I)&...&(I); (II)&(II)&...&(II); (III)&(III)&...&(III);

2. конъюнкция двух видов формул (второй подкласс):

   (I)&...&(I)&(II)&...&(II); (II)&...&(II)&(III)&...&(III);
   (I)&...&(I)&(III)&...&(III);

3. конъюнкция трёх видов формул (третий подкласс):
   (I)&...&(I)&(II)&...&(II)&(III)&...&(III).

Запись формулы в виде (I)&...&(I)&(II)&...&(II) означает, что в развернутом виде ее можно представить как

$$\forall x_1(P_1(x_1) \to h_1(x_1))\& \ldots \&\forall x_k(P_k(x_k) \to h_1(x_k))\&$$
$$\&\forall x_{k+1}\forall x_{k+2}(P_{k+1}(x_{k+1})\&P_{k+1}(x_{k+2}) \to h_{k+1}(x_{k+1}, x_{k+2}))\& \ldots \&$$
$$\&\forall x_{k+2m-1}\forall x_{k+2m}(P_{k+m}(x_{k+2m-1})\&$$
$$\&P_{k+m}(x_{k+2m}) \to h_{k+m}(x_{k+2m-1}, x_{k+2m})). \quad (1)$$

Аналогично могут быть развернуты и формулы других подклассов.

Восстановления для первого подкласса

Пусть задана формула Φ и требуется построить наилучшее восстановление $Q_{max} \in \mathbb{Q}_{max}$. Рассмотрим его построение для некоторых классов формул.

Графом противоречий будем называть гиперграф (т.е. граф, у которого ребро может соединять более чем две вершины), вершинами которого являются кортежи, а ребрами соединены те кортежи, которые не могут совместно присутствовать ни в одном восстановлении, так как в каком бы подмножестве кортежей исходной базы данных они ни оказались совместно, формула будет обращаться в ложь на данном множестве, то есть будет присутствовать противоречие.

Пустым подграфом гиперграфа противоречий будем называть такое подмножество вершин гиперграфа, что ни одна пара вершин из данного подмножества не соединена ребром в гиперграфе. Максимальным пустым подграфом для гиперграфа будем называть такой пустой подграф, добавление к которому любой вершины из гиперграфа, не принадлежащей подграфу, приводит к нарушению свойства пустого подграфа.

Очевидно, что для любого гиперграфа может быть несколько максимальных пустых подграфов. Наибольшим максимальным пустым подграфом для данного гиперграфа будем называть такой максимальный пустой подграф, в котором наибольшее число вершин среди всех максимальных пустых подграфов данного гиперграфа.

Заметим, что восстановлению базы данных будет соответствовать максимальный пустой подграф гиперграфа противоречий. Наилучшему восстановлению будет соответствовать наибольший максимальный пустой подграф гиперграфа противоречий.

Будем обозначать как $\mathbb{P}_{\text{I}}$ множество предикатов, которые в формуле Φ участвуют в формировании формул вида (I); $\mathbb{P}_{\text{II}}$ — множество предикатов, которые в формуле Φ участвуют в формировании формул вида (II); $\mathbb{P}_{\text{III}}$ — множество предикатов, которые в формуле Φ участвуют в формировании формул вида (III); $\mathbb{P}_{\text{III}_1}$ — множество предикатов, которые стоят на месте P_1 (до импликации) в частях формулы Φ, имеющих вид (III), $\mathbb{P}_{\text{III}_2}$ — множество предикатов, которые стоят на месте P_2 (после импликации) в частях формулы Φ, имеющих вид (III). При этом очевидно, что $\mathbb{P}_{\text{III}} = \mathbb{P}_{\text{III}_1} \cup \mathbb{P}_{\text{III}_2}$

Утверждение 2. *Пусть $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ и $T_1, \ldots, T_n$ — соответствующие им таблицы,*

$$\Phi = (\text{I})(\text{I})\&(\text{I})\& \ldots \&(\text{I}),$$

и в базе данных D содержатся противоречия.

Тогда существует единственное восстановление для D. Оно будет являться максимальным. При этом $Z(Q_{max}) = \bigcup\limits_{i=1}^{n} \{x | (x \in T_i)\&(h_i(x) = 0)\}$.

Доказательство. В базе данных содержатся противоречия, следовательно, формула Φ ложна на $X(D)$. Поскольку Φ представляет собой конъюнкцию подформул, то для того, чтобы формула была истинна, истинной должна быть каждая из подформул. Таким образом, для каждой из ложных подформул устраняем те кортежи, которые обращают её в ложь. То есть в итоге необходимо удалить $Z(Q_{max}) = \bigcup\limits_{i=1}^{n} \{x | (x \in T_i)\&(h_i(x) = 0)\}$. Докажем, что восстановление Q, построенное подобным образом, будет максимальным и единственным.

Предположим, что построенное восстановление Q не является максимальным. Тогда есть восстановление Q', в котором содержится больше кортежей. То есть $X(Q') \setminus X(Q) \neq \varnothing$. Без ограничения общности будем считать, что $X(Q') \setminus X(Q) = \{t'\}$.

Тогда, чтобы в базе данных не было противоречий, $\Phi(t')$ должна быть истинной.

Рассмотрим t'. Так как этот кортеж не содержится в $X(Q)$, то найдется такая формула ϕ_m вида (I) ($m \in \{1, \ldots, n\}$), являющаяся частью формулы Φ, для которой $\phi_m(t') = 0$, следовательно, $\Phi(t') = 0$. Это противоречит

тому, что Q' — восстановление. Тогда наше предположение неверно, остается признать, что Q — максимальное восстановление.

Предположим, что наше восстановление Q не является единственным, тогда существует еще одно восстановление Q_1. Оно не может содержать больше кортежей, чем в Q, по только что доказанному. Однако если в нём будет содержаться меньше кортежей, то найдется такой кортеж t_0, что $Q \setminus Q_1 = \{t_0\}$. Тогда при добавлении кортежа t_0, содержащегося в базе данных, к Q_1 получаем другое восстановление, что противоречит определению восстановления базы данных.

Отсюда получаем, что наше предположение неверно, и восстановление Q, построенное по приведенным выше правилам, будет являться максимальным и единственным. Утверждение доказано. □

Утверждение 3. *Пусть $D = \langle \mathbb{T}, \Phi \rangle$, $\mathbb{P}$ — множество соответствующих предикатов, $T \in \mathbb{T}$ — таблица предиката P, Φ — формула вида* (II)*, а именно*

$$\Phi = \forall x_1 \forall x_2 (P(x_1) \& P(x_2) \to h(x_1, x_2)),$$

и в базе данных содержатся противоречия.

Тогда для базы данных D может существовать несколько восстановлений. Для каждого из них $Z(Q) \subseteq T$, при этом для гиперграфа противоречий вершины, соответствующие кортежам из множества $T \setminus Z(Q)$, образуют максимальный пустой подграф для гиперграфа противоречий.

Доказательство. В базе данных содержатся противоречия, следовательно, формула Φ ложна на $X(D)$. Это возможно в том случае, когда в базе данных содержится хотя бы одна такая пара кортежей (a_1, a_2), что $a_1, a_2 \in T$ и $h(a_1, a_2) = 0$. Для устранения противоречия достаточно удалить один кортеж из такой пары. Однако таких пар может быть несколько, и некоторые кортежи могут входить в несколько пар одновременно.

Рассмотрим гиперграф противоречий для данного случая. Вершинами этого графа являются все кортежи из $X(D)$, ребра соединяют, очевидно, кортежи t_1, t_2, для которых выполняется следующее:

1. $t_1, t_2 \in T$;
2. $h(t_1, t_2) = 0$.

Для того чтобы построить восстановление, из гиперграфа противоречий нужно удалить минимальное количество вершин таким образом, чтобы в нем не осталось ребер. То есть нужно удалить хотя бы одну

вершину для каждого ребра, присутствующего в гиперграфе противоречий. Это соответствует построению максимального пустого подграфа для гиперграфа противоречий. В восстановление будут включены те кортежи, которые соответствуют вершинам, входящим в максимальный пустой подграф. При этом максимальному восстановлению будет соответствовать наибольший максимальный пустой подграф. При этом очевидно, что любой кортеж $t \in X(D) \setminus T$ будет входить в каждое из восстановлений, поскольку он не будет инцидентен ни одному ребру из гиперграфа противоречий.

Докажем, что построенная таким образом база данных действительно будет являться восстановлением. По построению восстановления Q получаем, что выполняются пункты 1–3 определения восстановления базы данных. Предположим, что не выполняется пункт 4. Тогда если существует восстановление Q_1, которое может быть получено из Q добавлением кортежа t_1, то есть $X(Q_1) = X(Q) \cup \{t_1\}$. В $X(Q)$ содержатся все кортежи из $t_1 \in X(D) \setminus T$ и все кортежи из T, которые образуют максимальный пустой подграф. Таким образом, кортеж t_1 не может быть из этих множеств. Тогда остается признать, что $t_1 \in T$ и не принадлежит рассматриваемому максимальному пустому подграфу, это означает что в гиперграфе противоречий есть ребро между t_1 и неким кортежем $t_0 \in X(Q)$. Следовательно, $h(t_1, t_0) = 0$. Отсюда следует, что Φ ложна на $X(Q) \cup \{t_1\}$. А значит, Q_1 не является восстановлением. Таким образом, Q, построенное подобным образом, удовлетворяет всем пунктам определения восстановления.

Заметим, что подобных восстановлений может быть несколько, так как в графе противоречий может быть несколько максимальных пустых подграфов, каждый из них будет соответствовать своему восстановлению.

Теперь предположим, что существует какое-то другое восстановление Q_2, которое не может быть построено по приведенным выше правилам. Очевидно, что в него войдут все кортежи из $X(D) \setminus T$, поскольку на них формула Φ не является ложной. Далее, по нашему предположению, кортежи из T, входящие в Q_2, не будут образовывать максимальный пустой подграф для гиперграфе противоречий. Тогда возможны два варианта:

1. $\forall t_1 \forall t_2 h(t_1, t_2) = 1$, где $t_1, t_2 \in X(Q_2)$, $t_1, t_2 \in T$. Тогда, очевидно, множество кортежей из T, входящих в Q_2, будет содержаться в одном из максимальных пустых подграфов. Следовательно, Q не будет являться восстановлением.

2. Будут существовать некие кортежи $a, b \in X(Q_2)$, что для них

$h(a,b)=0$. Тогда Q также не будет являться восстановлением.

Таким образом, доказано, что для базы данных, в качестве формулы-ограничения содержащей одиночную формулу вида (II), восстановление может быть построено только по вышеприведенным правилам. □

Утверждение 4. *Пусть $D=\langle\mathbb{T},\Phi\rangle, \mathbb{P}$ — множество соответствующих предикатов, $T_1,\ldots,T_n\in\mathbb{T}$ — таблицы предикатов $P_1,\ldots,P_n$, Φ представляет собой конъюнкцию двух формул вида* (II)*, а именно*

$$\Phi=\text{(II)}\&\text{(II)},$$

и в базе данных содержатся противоречия.

Тогда для D может существовать несколько наилучших восстановлений. Задача поиска наилучшего восстановления сводится к поиску наибольшего максимального пустого подграфа для гиперграфа противоречий, и $Z(Q)\subseteq(T_1\cup T_2)$, где T_1 и T_2 —таблицы предикатов P_1, P_2, участвующих в образовании формулы Φ.

Доказательство. Для предикатов, составляющих формулу Φ, возможны два варианта.

1. $P_1=P_2$, то есть формула Φ может быть записана следующим образом:

$$\Phi=\forall x_1\forall x_2(P_1(x_1)\&P_1(x_2)\to h_1(x_1,x_2))\&$$
$$\&\forall x_3\forall x_4(P_1(x_3)\&P_1(x_4)\to h_2(x_3,x_4)).$$

В образовании данной формулы Φ участвует только один предикат – P_1. Тогда для этой формулы строим восстановление следующим образом: берем общий гиперграф противоречий, то есть вершинами будут кортежи из $X(D)$, а ребрами соединим те, которые вызывают противоречия либо в части формулы

$$\forall x_1\forall x_2(P_1(x_1)\&P_1(x_2)\to h_1(x_1,x_2)),$$

либо в части формулы

$$\forall x_3\forall x_4(P_1(x_3)\&P_1(x_4)\to h_2(x_3,x_4)).$$

Заметим, что ребра могут соединять только вершины, соответствующие кортежам из T_1.

Тогда наибольший максимальный пустой подграф для построенного таким образом гиперграфа противоречий будет максимальным восстановлением. Докажем это.

Добавление любого кортежа к построенному таким образом восстановлению приводит к тому, что возникает противоречие, так как очевидно, что можем добавить только кортеж из T_1, так как остальные кортежи в восстановлении присутствуют. Но добавление кортежа из E_1 нарушает свойство максимального пустого подграфа. Получаем противоречие. Доказано.

2. $P_1 \neq P_2$, то есть формула Φ будет выглядеть так:

$$\Phi = \forall x_1 \forall x_2 (P_1(x_1) \& P_1(x_2) \to h_1(x_1, x_2)) \&$$
$$\& \forall x_3 \forall x_4 (P_2(x_3) \& P_2(x_4) \to h_2(x_3, x_4)).$$

В данном случае имеем две независимые части формулы. С учетом того, что предикаты в подформулах не дублируются, получаем, что $X(Q) = X(Q_1) \cap X(Q_2)$, то есть восстановление для базы данных с формулой Φ может быть может быть построено с использованием восстановлений для каждой из частей формулы путем пересечения множеств кортежей данных восстановлений.

Для первой части формулы $\varphi_1 = \forall x_1 \forall x_2 (P_1(x_1) \& P_1(x_2) \to h_1(x_1, x_2))$ получаем восстановления с помощью построения максимального пустого подграфа для гиперграфа, в котором ребрами соединяются только кортежи из таблицы T_1.

Для второй части формулы $\varphi_2 = \forall x_3 \forall x_4 (P_2(x_3) \& P_2(x_4) \to h_2(x_3, x_4))$ получаем восстановления с помощью построения максимального пустого подграфа для гиперграфа, в котором ребрами соединяются только кортежи из таблицы T_2.

Возьмем в качестве восстановления для Φ пересечение множеств кортежей из восстановлений для T_1 и T_2 ($X(Q) = (X(Q_1) \cap X(Q_2))$). Докажем, что это будет восстановлением для Φ. Предположим, что это не будет восстановлением. Тогда к множеству $X(Q)$ мы можем добавить кортеж из D без возникновения противоречий. Этот кортеж не может быть не из таблицы T_1 или T_2, так как все такие кортежи уже содержатся в восстановлении по построению. Если добавляемый кортеж принадлежит таблице T_1, то при добавлении нарушится свойство максимального пустого подграфа для гиперграфа G_1, соответствующего формуле φ_1, если из T_2, то при добавлении нарушится свойство максимального пустого подграфа для гиперграфа G_2, соответствующего формуле φ_2. Следовательно, не можем добавить ни один кортеж, значит, нами построено восстановление.

В данном случае для построения максимального восстановления необходимо найти наибольшие максимальные пустые подграфы для каждого из гиперграфов G_1 и G_2. Поскольку в случае максимального

восстановления для первой части формулы и для второй части формулы получим в итоге максимальное восстановление для всей формулы.

□

Будем называть *ядром восстановления* множество кортежей из таблиц, которые не участвуют в образовании формулы Φ. Соответствующие им вершины никогда не могут быть соединены ребром в гиперграфе противоречий.

Здесь и далее, если не оговорено иное, будем рассматривать гиперграф противоречий и построение максимальных пустых подграфов в гиперграфе противоречий только для кортежей, которые не входят в ядро восстановления. Таким образом, получается, что все восстановления для какой-то одной конкретной базы данных имеют общую часть — ядро восстановления.

Утверждение 5. *Пусть $D = \langle \mathbb{T}, \Phi \rangle, \mathbb{P}$ — множество соответствующих предикатов, $T_1, \ldots, T_n \in \mathbb{T}$ — таблицы предикатов $P_1, \ldots, P_n$, Φ представляет собой конъюнкцию формул вида* (II)*, а именно*

$$\Phi = (\text{II})\&(\text{II})\& \ldots \&(\text{II}),$$

и в базе данных D содержатся противоречия.

Тогда для D может существовать несколько наилучших восстановлений. Задача поиска наилучшего восстановления сводится к поиску наибольшего максимального пустого подграфа для гиперграфа противоречий. Задача поиска множества восстановления сводится к поиску всех максимальных пустых подграфов для гиперграфа противоречий.

Доказательство. Рассмотрим формулу Φ. Без ограничения общности представим ее в следующем виде:

$$\Phi = \varphi_1 \& \varphi_2 \& \ldots \& \varphi_k,$$

где

$$\begin{array}{l} \varphi_i = \forall x_{i1} \forall x_{i2} (P_i(x_{i1}) \& P_i(x_{i2}) \to h_{i1}(x_{i1}, x_{i2})) \& \\ \&\forall x_{i3} \forall x_{i4} (P_i(x_{i3}) \& P_i(x_{i4}) \to h_{i2}(x_{i3}, x_{i4})) \& \ldots \\ \&\forall x_{i(m_i-1)} \forall x_{im_i} (P_i(x_{i(m_i-1)}) \& P_i(x_{im_i}) \to h_{i(m_i/2)}(x_{i(m_i-1)}, x_{im_i})) \end{array},$$

при $i \in \{1, \ldots, k\}$, $P_1, \ldots, P_k$ — предикаты, а $T_1, \ldots, T_k$ — их таблицы, которые участвуют в образовании формулы Φ.

То есть мы сгруппировали наши части формул по присутствующим в них предикатам. Тогда если рассмотреть формулу Φ с точки зрения построения восстановления, то задача распадется на k частей — по

числу различных предикатов, задействованных в формуле Φ. Это объясняется тем, что если будем рассматривать граф противоречий, то в его формировании будут участвовать кортежи из k таблиц (от 1 до k), при этом ребра гиперграфа будут возникать только между вершинами, соответствующими кортежам из одной таблицы. Фактически это соответствует разделению графа на несколько компонент связности. Таким образом, при построении наибольшего максимального пустого подграфа мы можем рассматривать подобное построение для каждой из k частей, а потом объединить результат. После этого полученный наибольший максимальный пустой граф для всего гиперграфа необходимо добавить к ядру восстановления, то есть к кортежам из таблиц, не входящих в $T_1, \ldots, T_k$. Полученное множество будет максимальным восстановлением по построению. Максимальных восстановлений может быть несколько, так как наибольших максимальных пустых подграфов для гиперграфа может быть несколько. Первая часть утверждения доказана.

Для построения множества всех восстановлений в данном случае может быть использован следующий алгоритм.

1. Добавляем в Q_0 все кортежи из ядра восстановления.
2. Группируем части (II) формулы Φ по содержащемуся в них предикату, как это было показано в доказательстве утверждения 5.
3. Строим гиперграф противоречий для формулы Φ и находим в нем все максимальные пустые подграфы. Обозначим соответствующие им множества кортежей как $q_1, q_2, \ldots, q_m$.
4. Построим следующим образом множество всех восстановлений: $\mathbb{Q} = \{Q_1, Q_2, \ldots, Q_m\}$, где $Q_i = Q_0 \cup q_i, i \in \{1, k\}$.

То, что в этом множестве содержатся все восстановления для данной базы данных, очевидно следует из доказанного ранее здесь и в утверждениях 4 и 3. □

Утверждение 6. *Пусть $D = \langle \mathbb{T}, \Phi \rangle$, $\mathbb{P}$ — множество соответствующих предикатов, $T_1, T_2 \in \mathbb{T}$ — таблицы предикатов P_1, P_2, $P_1 \neq P_2$, Φ — формула вида* (III)*, а именно*

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))),$$

и в базе данных содержатся противоречия.

Тогда существует единственное восстановление Q_{max}, и $Z(Q_{max}) \subseteq T_1$.

Доказательство. Противоречия возникают для данного вида формулы Φ в том случае, когда в таблице T_1 предиката P_1 найдется хотя бы один такой кортеж t, что для него $h_2(t) = 0$, при этом для любого кортежа x из таблицы T_2 предиката P_2 $h_1(t, x) = 0$.

Устранить противоречия можно только одним способом — удалить из таблицы T_1 все такие кортежи, для которых выполняется вышеизложенное. Если оставить хотя бы один кортеж, для которого выполняется вышеизложенное, то в базе данных останутся противоречия. Соответственно, в данном случае восстановление будет единственным и максимальным. И удалять кортежи необходимо только из таблицы T_1. Следовательно, $Z(Q_{max}) \subseteq T_1$. □

Утверждение 7. *Пусть $D = \langle \mathbb{T}, \Phi \rangle, \mathbb{P}$ — множество соответствующих предикатов, $T_1 \in \mathbb{T}$ — таблица предиката P_1, Φ — формула вида* (III)*, а именно*

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))),$$

и в базе данных содержатся противоречия.

Тогда существует единственное восстановление Q_{max}, и $Z(Q_{max}) \subseteq T_1$.

Доказательство. Для построения восстановления предлагается следующий алгоритм.

1. Определяем, для каких кортежей не выполняется условие h_2, помечаем их.

2. Среди помеченных кортежей находим те, для которых Φ обращается в ложь. Если их больше 0, то переходим к 3, иначе переходим к 4.

3. Удаляем найденные кортежи. Если множество помеченных кортежей не пусто, то возвращаемся к 2, иначе переходим к 4.

4. Строим восстановление Q следующим образом: добавляем к Q все кортежи из всех таблиц исходной базы данных, кроме T; из T добавляем те кортежи, которые остались после удаления.

Докажем, что построенное таким образом множество кортежей Q будет единственным восстановлением. Если к Q добавить хотя бы 1 кортеж из числа удаленных, то Φ обратится в ложь, следовательно, Q — восстановление. Пусть Q' — какое-то другое восстановление, т.е. рассмотрим вариант, когда Q не является единственным восстановлением. Тогда либо в Q' содержится хотя бы 1 удаленный кортеж из T, либо в Q'

нет какого-либо кортежа, который есть в Q. И в том и в другом случае Q' не будет восстановлением по определению. Остается признать, что Q — единственное восстановление, следовательно, оно же является максимальным. Утверждение доказано. □

Утверждение 8. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $\mathbb{P}$ *— множество соответствующих предикатов,* Φ *представляет собой конъюнкцию двух формул вида* (III)*, внутри каждой из которых предикаты различны:*

$$\Phi = (\text{III})\&(\text{III}),$$

и в базе данных содержатся противоречия.

Тогда существует единственное восстановление Q_{max}*, соответственно, оно же будет максимальным. Удаление кортежей для построения восстановления будет производиться из обеих таблиц предикатов, стоящих на первом месте в каждой из частей* (III) *формулы* Φ.

Доказательство. В данном случае для каждой из частей формулы Φ возможны различная их взаимосвязь между собой. Рассмотрим возможные варианты.

I. Части формулы Φ могут быть рассмотрены как независимые друг от друга, это возможно, когда Φ представима в виде:

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_3(x_3) \to (P_4(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))),$$

или

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& (x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_1(x_3) \to (P_3(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))),$$

или

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_1(x_3) \to (P_2(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))).$$

Построение восстановления для всей формулы Φ целиком сводится к следующим действиям:

1. Добавляем в Q все кортежи, которые не принадлежат таблицам предикатов, стоящих на первом месте (T_1 и T_2 для первой формулы, T_1 — для второй и третьей).

2. Для первой части формулы производим удаление из таблицы стоящего на первом месте предиката (T_1 для всех формул) всех вызывающих противоречие кортежей, как описано в утверждении 6.

3. Добавляем к Q те кортежи, которые не были удалены на предыдущем шаге.

4. Для второй части формулы производим удаление из таблицы стоящего на первом месте предиката (T_3 — для первой формулы, модифицированная после предыдущих шагов таблицы T_1 — для второй и третьей формулы) всех вызывающих противоречие кортежей, как описано в утверждении 6.

5. Добавляем к Q те кортежи, которые остались после удаления на предыдущем шаге.

Построенное таким образом множество Q будет являться единственным и максимальным восстановлением. Действительно, добавление любого кортежа из исходной базы данных, не добавленного к Q, приводит к возникновению противоречия в одной из частей формулы, следовательно, и во всей формуле, так как обе части формулы связаны конъюнкцией. Следовательно, Q — восстановление. Рассмотрим произвольное подмножество Q_0 множества кортежей базы данных. Докажем, что оно не может быть восстановлением ни в каком случае, кроме того, когда оно совпадает с Q. Действительно, раз оно не совпадает с Q, следовательно, либо является подмножеством Q, но в этом случае оно не будет восстановлением, либо содержит хотя бы один кортеж из удаленных. Однако если Q_0 содержит удаленный кортеж, то оно опять-таки не будет являться восстановлением из-за возникновения противоречия. Следовательно, остается признать, что Q — единственное восстановление, следовательно, оно же будет являться максимальным.

II. Части Φ не могут быть рассмотрены как независимые друг от друга, однако не возикает «цикла». Φ может быть записана как

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \& \\ \& \forall x_3 \exists x_4 (P_2(x_3) \to (P_3(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))).$$

Восстановление Q будем строить следующим образом:

1. Добавляем в Q все кортежи, не принадлежащие таблице T_1 и таблице T_2.

2. Удаляем из таблицы T_2 кортежи, которые вызывают противоречия во второй части формулы, как описано в утверждении 6.

3. Удаляем из таблицы T_1 кортежи, которые вызывают противоречия в первой части формулы, как описано в утверждении 6.

4. Добавляем к Q все кортежи из T_1 и T_2, которые мы не удалили.

Докажем, что полученное таким образом Q будет единственным восстановлением. Предположим, Q не является восстановлением. Тогда может добавить хотя бы еще один кортеж, так чтобы не возникло противоречий. Однако этого сделать нельзя, так как все кортежи — удаленные, а значит, их добавление вызывает противоречие. Следовательно, Q является восстановлением. Предположим, что восстановление Q не является единственным. Тогда существует другое восстановление Q', отличающееся от Q. Тогда в нем кроме кортежей из Q должны содержаться кортежи из тех, которые были удалены для получения восстановления Q, иначе оно не будет восстановлением, а будет лишь подмножеством Q. Однако добавление кортежа из удаленных приводит к возникновению противоречия, следовательно, Q будет единственным восстановлением. Раз оно единственно, то оно же является и максимальным.

III. Части Φ образуют «цикл». Это возможно только в том случае, когда Φ может быть записана как:

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_2(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_2(x_3) \to (P_1(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))).$$

Сложность данного варианта в том, что при попытке устранения противоречий в каждой из частей независимо от другой (как делается в утверждении 6) возможно возникновение новых противоречий во второй части. То есть на самом деле эти части не являются независимыми, поэтому обычный подход, который был использован в первом случае, не применим. Однако и алгоритм, использовавшийся для построения восстановления во втором случае, не подходит, так как там присутствовала одиночная прямая зависимость, а здесь зависимость "циклическая". Соответственно, для полного устранения противоречий действия по устранению противоречий в этом случае также будут циклическими.

Предлагаемая последовательность действий для построения восстановления будет такой:

1. Добавляем в Q все кортежи из базы данных, которые не принадлежат таблицам T_1 и T_2.

2. В первой части формулы:

   - удаляем те кортежи из таблицы T_1 (соответствующей предикату P_1), которые обращают в ложь первую часть формулы;

- помечаем те кортежи из таблицы T_1, для которых условие h_2 не выполняется, а выполняется условие h_1 (благодаря каким-то кортежам из T_2).

3. Во второй части формулы:

   - удаляем все кортежи из T_2, которые обращают в ложь вторую часть формулы;
   - помечаем те кортежи из T_2, для которых условие h_4 не выполняется, а h_3 выполняется (благодаря каким-то кортежам из T_1).

4. Проверяем первую часть на наличие противоречий, рассматривая только помеченные нами кортежи из T_1. Удаляем те из помеченных кортежей, для которых h_1 стало ложью.

5. Проверяем вторую часть формулы на наличие противоречий, рассматривая только помеченные нами кортежи из T_2. Удаляем те из помеченных кортежей, для которых h_3 стало ложью.

6. Если при выполнении 4 и 5 был удален хотя бы один кортеж, то возвращаемся к выполнению 4. Если же при последнем выполнении 4 и 5 не было удалено ни одного кортежа, то добавляем к Q все кортежи, которые остались после удалений в таблицах T_1 и T_2.

Построенное Q будет единственным и максимальным восстановлением. Докажем это. Предположим, что Q не будет восстановлением. Тогда можем добавить какой-то кортеж к Q, при этом формула останется непротиворечивой. Очевидно, что добавить может только один из удаленных кортежей из T_1 или из T_2. Однако добавление такого кортежа приводит к возникновению противоречий, то есть это не будет восстановлением. Таким образом, показали, что Q является восстановлением. Теперь предположим, что существует другое восстановление Q', то есть Q не будет являться единственным. Тогда Q' должно отличаться от Q, построенного нами, хотя бы на один кортеж, и при этом быть восстановлением. Очевидно, что в Q' должны войти все кортежи, которые содержались в исходной базе данных и не принадлежат T_1 и T_2. Далее, если в Q' не будет входить ни один кортеж из T_1 и T_2, то Q' не будет являться восстановлением. Если же мы включим в Q' другое подмножество кортежей из $X(T_1) \cup X(T_2)$, то возможны следующие варианты:

- туда попали кортежи, удаленные в ходе выполнения приведенного выше алгоритма; в этом случае Q' не будет восстановлением, так как в формуле Φ возникнут противоречия;

- туда не попали удаленные кортежи, а попали только те кортежи из T_1 и T_2, которые есть в Q; тогда Q' совпадет с Q или будет его подмножеством.

Следовательно, Q будет единственным восстановлением, а значит, оно также будет максимальным.

Таким образом, мы разобрали все варианты, утверждение доказано. □

Утверждение 9. *Пусть $D = \langle \mathbb{T}, \Phi \rangle, \mathbb{P}$ — множество соответствующих предикатов, Φ представляет собой конъюнкцию двух формул вида* (III)*:*

$$\Phi = \text{(III)}\&\text{(III)},$$

в которой хотя бы в одной из частей вида (III) *один и тот же предикат стоит и на первом и на втором месте, и в базе данных содержатся противоречия.*

Тогда существует единственное восстановление Q_{max}, соответственно, оно же будет максимальным. Удаление кортежей для построения восстановления будет производиться из таблиц предикатов, стоящих на первом месте в каждой из частей (III) *формулы* Φ.

Доказательство. Рассмотрим варианты различной взаимосвязи предикатов каждой из двух частей между собой. В общем случае их можно разбить на две большие группы.

I. В этом случае части формулы Φ можно считать независимыми друг от друга. Φ может быть записана в виде:

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_2(x_3) \to (P_3(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))),$$

или

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_2(x_3) \to (P_2(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))).$$

Устранение противоречий и построение восстановления для данной формулы может быть проведено следующим образом:

1. Устраняем противоречия для первой части формулы путем удаления кортежей из таблицы T_1, как описано в утверждении 7.

2. Устраняем противоречия для второй части формулы путем удаления кортежей из таблицы T_2, как описано в утверждении 6.

3. Формируем Q следующим образом: добавляем все кортежи из таблиц, не совпадающих с T_1 и T_2, затем добавляем все кортежи, оставшиеся в T_1 и T_2 после удаления кортежей при устранении противоречий.

Докажем, что построенное Q действительно восстановление. При добавлении кортежа из базы данных, не включенного в Q, а это неминуемо будет какой-либо удаленный кортеж либо из T_1, либо из T_2, возникнет противоречие. Следовательно, Q — восстановление. Покажем, что Q — единственное. Если возьмем любое другое подмножество кортежей из базы данных, то оно либо будет включено в Q, либо будет содержать удаленные из Q кортежи. В обоих случаях это другое подмножество не будет являться восстановлением. Следовательно, Q будет единственным, а значит, оно же будет максимальным восстановлением.

II. В этом случае части формулы Φ нельзя рассматривать независимо друг от друга. Φ может быть записана в виде:

$$\Phi = \forall x_3 \exists x_4 (P_1(x_3) \to (P_2(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))) \&$$
$$\& \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))),$$

или

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_2(x_3) \to (P_1(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))),$$

или

$$\Phi = \forall x_1 \exists x_2 (P_1(x_1) \to (P_1(x_2) \& h_1(x_1, x_2) \vee h_2(x_1))) \&$$
$$\& \forall x_3 \exists x_4 (P_1(x_3) \to (P_1(x_4) \& h_3(x_3, x_4) \vee h_4(x_3))).$$

Тогда для построения наилучшего восстановления Q может быть использован следующий алгоритм.

1. Устраняем противоречия для первой части формулы путем удаления кортежей из таблицы стоящего на первом месте предиката, как описано в утверждениях 7 или 6 (в зависимости от вида первой части).

2. Устраняем противоречия для второй части формулы путем удаления кортежей из таблицы стоящего на первом месте предиката, как описано в утверждениях 7 или 6 (в зависимости от вида второй части), используя для выявления противоречий модифицированную на предыдущем шаге таблицу.

3. Формируем Q следующим образом: добавляем все кортежи немодифицированных таблиц, затем добавляем все кортежи, оставшиеся в модифицированных таблицах после удаления кортежей при устранении противоречий.

Доказательство того, что построенное Q — действительно восстановление, и что оно единственно, аналогично доказанному ранее в этом утверждении и в утверждении 8.

Таким образом, построенное восстановление Q — единственное, следовательно, оно является максимальным. Утверждение доказано. □

Утверждение 10. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $\mathbb{P}$ *— множество соответствующих предикатов,* $T_1, T_2, \ldots, T_{n-1}, T_n \in \mathbb{T}$ *— таблицы предикатов* $P_1, P_2, \ldots, P_{n-1}, P_n$, Φ *представляет собой конъюнкцию формул вида* (III)*:*

$$\Phi = (\text{III})\&(\text{III})\& \ldots \&(\text{III}),$$

и в базе данных содержатся противоречия.

Тогда существует единственное восстановление Q_{max}*, и удаление кортежей производится из таблиц предикатов, стоящих на первых местах в соответствующих частях формулы.*

Доказательство. Для формулы Φ возможны два основных варианта.

I. $\mathbb{P}_{\text{III}_1} \cap \mathbb{P}_{\text{III}_2} = \varnothing$, то есть предикаты, стоящие на первых местах частей (III) формулы Φ, не совпадают с предикатами, стоящими на вторых местах, ни для какой пары частей. В этом случае восстановление Q может быть построено следующим образом:

1. К Q добавляем все кортежи из таблиц тех предикатов, которые не принадлежат множеству $\mathbb{P}_{\text{III}}$.

2. В Q добавляем кортежи из таблиц предикатов, которые принадлежат $\mathbb{P}_{\text{III}_2}$.

3. Последовательно перебирая части (III) формулы Φ, для каждой из них устраняем противоречия посредством удаления кортежей из таблицы предиката, принадлежащего множеству $\mathbb{P}_{\text{III}_1}$, то есть стоящего на первом месте в этой части, таким образом, как это описано в утверждении 6.

4. После того как устранили противоречия для всех частей формулы, добавляем кортежи из модифицированных таблиц предикатов к Q.

Докажем, что построенное Q будет действительно являться единственным восстановлением. Оно действительно является восстановлением, так как добавление к нему любого не содержащегося в нем кортежа из базы данных (а это может быть только кортеж, который был удален в процессе построения Q) приводит к тому, что формула Φ обращается в ложь, то есть возникают противоречия. Докажем, что Q — единственное. Предположим, что есть другое восстановление Q'. Очевидно, что в нем должны содержаться все кортежи из таблиц предикатов, принадлежащих $\mathbb{P}_{\text{III}_2}$. Далее, в нем должно содержаться какое-то подмножество множества кортежей из таблиц предикатов, принадлежащих $\mathbb{P}_{\text{III}_1}$. Очевидно, что в нем должен присутствовать хотя бы один кортеж из тех, которые мы удалили, чтобы Q' отличалось от Q и не было его подмножеством. Однако наличие такого кортежа в Q' приводит к тому, что формула Φ будет ложной на Q', а значит, Q' не будет восстановлением. Остается признать, что Q — единственное восстановление, а значит, максимальное.

II. $\mathbb{P}_{\text{III}_1} \cap \mathbb{P}_{\text{III}_2} \neq \varnothing$, то есть предикаты, стоящие на первых местах частей (III) формулы Φ, в одном или нескольких местах совпадают с предикатами, стоящими на вторых местах частей (III) формулы Φ. В этом случае восстановление Q может быть построено следующим образом:

1. В Q добавляем все кортежи из таблиц тех предикатов, которые не принадлежат множеству $\mathbb{P}_{\text{III}}$.

2. Для каждой из частей (III) формулы выполняем следующее:

   - устраняем кортежи из таблицы предиката, стоящего на первом месте, которые вызывают противоречия, как показано в утверждении 6;
   - помечаем кортежи из той же таблицы, которые не вызывают противоречия, однако для которых второе условие принимает значение «ложь».

3. Последовательно для каждой из частей (III) формулы Φ выполняем: устраняем кортежи из модифицированной (после удаления кортежей) таблицы предиката, стоящего на первом месте, которые вызывают противоречия, как показано в утверждении 6, при этом поиск производим только среди помеченных кортежей.

4. Если на предыдущем шаге ни для одной части (III) формулы Φ не произвели ни одного удаления, то тогда переходим к 5, иначе возвращаемся к 3.

5. К Q добавляем все кортежи из таблиц предикатов, принадлежащих $\mathbb{P}_{\text{III}}$, которые там остались после всех удалений. Восстановление построено.

Докажем, что построенное Q будет являться единственным восстановлением. Если попытаемся добавить к Q какой-либо кортеж, которого там нет, но который изначально присутствовал в базе данных, то неминуемо возникнет противоречие. Следовательно, Q — действительно, восстановление. Предположим, что существует другое восстановление Q', отличное от Q. Тогда туда войдут все кортежи из таблиц тех предикатов, которые не принадлежат множеству $\mathbb{P}_{\text{III}}$, иначе Q' не будет восстановлением. Из таблиц предикатов, которые не принадлежат множеству $\mathbb{P}_{\text{III}}$, в Q' должно входить другое подмножество, чем в Q. Однако в этом случае в Q' неминуемо попадет хотя бы один кортеж из удаленных, а это приведет к противоречию. Следовательно, Q' не может быть восстановлением, отличным от Q. Значит, Q — единственное восстановление, таким образом, оно же является максимальным. □

Восстановления для второго подкласса

Для второго подкласса формул для построения наилучших восстановлений справедливы следующие утверждения.

Утверждение 11. *Пусть $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ и $T_1, \ldots, T_n$ — соответствующие им таблицы,*

$$\Phi = \text{(I)}\& \ldots \&\text{(I)}\&\text{(II)}\& \ldots \&\text{(II)},$$

и в базе данных D содержатся противоречия. Тогда для D возможно построить наилучшее восстановление, при этом их может быть несколько.

Доказательство. Для построения восстановления будем использовать следующий алгоритм.

1. Строим наилучшее восстановление для части (I)$\& \ldots \&$(I) формулы (т.е. для базы данных $\langle \mathbb{T}, \text{(I)}\& \ldots \&\text{(I)} \rangle$) согласно алгоритму, приведенному в утверждении 2.

2. Находим множество наилучших восстановлений для (II)$\& \ldots \&$(II) (т.е. для базы данных $\langle \mathbb{T}', \text{(II)}\& \ldots \&\text{(II)} \rangle$, где $\mathbb{T}'$ — таблицы базы данных D после выполнения предыдущего действия) согласно алгоритму, приведенному в утверждении 5.

Построенное восстановление будет наилучшим, поскольку, даже если предикаты из обеих частей будут повторяться каким-либо образом, это не оказывает влияния на процесс восстановления, если удаление производить в указанном порядке. Сначала мы удаляем кортежи для устранения противоречий из части формулы, выраженной как (I)$\&\ldots\&$(I). Если при этом удалении возникнут противоречия, которых раньше не было, в части формулы, выраженной как (II)$\&\ldots\&$(II), то они будут устранены на втором шаге. При удалении каких-либо кортежей новые противоречия в части формулы, выраженной как (I)$\&\ldots\&$(I), возникнуть не могут, исходя из вида формулы. Соответственно, последовательное построение наилучших восстановлений приведет нас к требуемому результату. □

Утверждение 12. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ *и* $T_1, \ldots, T_n$ *— соответствующие им таблицы,*

$$\Phi = \text{(I)}\&\ldots\&\text{(I)}\&\text{(III)}\&\ldots\&\text{(III)},$$

и в базе данных D *содержатся противоречия.*

Тогда для D *существует только одно восстановление, наилучшее.*

Доказательство. Для построения восстановления будем использовать следующий алгоритм.

1. Сначала строим наилучшее восстановление для части (I)$\&\ldots\&$(I) формулы (т.е. для базы данных $\langle \mathbb{T}, \text{(I)}\&\ldots\&\text{(I)} \rangle$) согласно алгоритму, приведенному в утверждении 2.

2. Строим наилучшее восстановление для части (III)$\&\ldots\&$(III) (т.е. для базы данных $\langle \mathbb{T}', \text{(III)}\&\ldots\&\text{(III)} \rangle$, где $\mathbb{T}'$ — таблицы базы данных D после выполнения предыдущего действия) согласно алгоритму, приведенному в утверждении 10.

Построенное восстановление будет наилучшим, поскольку, даже если предикаты из обеих частей будут повторяться каким-либо образом, это не оказывает влияния на процесс восстановления, если удаление производить в указанном порядке. Сначала мы удаляем кортежи для устранения противоречий из части формулы, выраженной как (I)$\&\ldots\&$(I). Если при этом удалении возникнут противоречия, которых раньше не было, в части формулы, выраженной как (III)$\&\ldots\&$(III), то они будут устранены на втором шаге. При удалении каких-либо кортежей новые противоречия в части формулы, выраженной как (I)$\&\ldots\&$(I), возникнуть не могут, исходя из вида формулы. Соответственно, последовательное построение наилучших восстановлений приведет нас к требуемому результату. □

Утверждение 13. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ *и* $T_1, \ldots, T_n$ *— соответствующие им таблицы,*

$$\Phi = (\text{II})\& \ldots \&(\text{II})\&(\text{III})\& \ldots \&(\text{III}),$$

и в базе данных D *содержатся противоречия. Кроме этого, для формулы* Φ *выполняется следующее:* $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} = \varnothing$.

Тогда для D *может быть построено восстановление. При этом возможно существование нескольких наилучших восстановлений.*

Доказательство. Для построения наилучшего восстановления будем использовать следующий алгоритм.

1. Строим наилучшее восстановление для $(\text{III})\& \ldots \&(\text{III})$ (т.е. для базы данных $\langle \mathbb{T}, (\text{III})\& \ldots \&(\text{III}) \rangle$) согласно алгоритму, приведенному в утверждении 10.

2. Находим множество наилучших восстановлений для $(\text{II})\& \ldots \&(\text{II})$ (т.е. для базы данных $\langle \mathbb{T}', (\text{II})\& \ldots \&(\text{II}) \rangle$, где $\mathbb{T}'$ — таблицы базы данных D после выполнения предыдущего действия) согласно алгоритму, приведенному в утверждении 5.

Построенное восстановление будет наилучшим, поскольку предикаты из обеих частей не могут повторяться произвольным образом, так как $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} = \varnothing$.

Таким образом, после удаления кортежей из части формулы, выраженной как $(\text{III})\& \ldots \&(\text{III})$, возможно возникновение противоречий, которых раньше не было, в части формулы, выраженной как $(\text{II})\& \ldots \&(\text{II})$. Однако эти противоречия будут устранены вместе с остальными на втором шаге. На втором шаге при удалении каких-либо кортежей новые противоречия в части формулы, выраженной как $(\text{III})\& \ldots \&(\text{III})$, возникнуть не могут, поскольку $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} = \varnothing$. Соответственно, последовательное построение наилучших восстановлений в указанном выше порядке приведет нас к требуемому результату. □

Утверждение 14. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ *и* $T_1, \ldots, T_n$ *— соответствующие им таблицы,*

$$\Phi = (\text{II})\& \ldots \&(\text{II})\&(\text{III})\& \ldots \&(\text{III}),$$

и в базе данных D *содержатся противоречия. Кроме этого, для формулы* Φ *выполняется следующее:* $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} \neq \varnothing$.

Тогда для D *существует наилучшее восстановление, при этом оно может быть не одно.*

Доказательство. В данном случае для построения наилучшего восстановления может быть использован следующий алгоритм.

1. Строим наилучшее восстановление для (III)$\&\ldots\&$(III) (т.е. для базы данных $\langle\mathbb{T}$, (III)$\&\ldots\&$(III)$\rangle$) согласно алгоритму, приведенному в утверждении 10.

2. Находим множество $\mathbb{Q}_2$ всех восстановлений для (II)$\&\ldots\&$(II) (т.е. для базы данных $\langle\mathbb{T}'$, (II)$\&\ldots\&$(II)$\rangle$, где $\mathbb{T}'$ — таблицы базы данных D после удаления кортежей) согласно алгоритму, приведенному в утверждении 5.

3. Для каждого из восстановлений из множества $\mathbb{Q}_2$ производим поиск наилучшего восстановления для части (III)$\&\ldots\&$(III) (т.е. для базы данных $\langle\mathbb{T}''$, (III)$\&\ldots\&$(III)$\rangle$, где $\mathbb{T}''$ — таблицы базы данных D после выполнения предыдущего действия) согласно алгоритму, приведенному в утверждении 10.

4. Если на предыдущем шаге какое-либо восстановление из множества $\mathbb{Q}_2$ совпало с построенным для него наилучшим восстановлением для части (III)$\&\ldots\&$(III), то добавляем его к $\mathbb{Q}$ — множеству всех восстановлений формулы Φ. И далее это восстановление исключаем из $\mathbb{Q}_2$. Если после исключения всех совпавших восстановлений $\mathbb{Q}_2$ не пусто, то переходим к 5. Если же в нем не осталось восстановлений, то переходим к 6.

5. Сравниваем наилучшие восстановления, полученные на этапе 3, между собой, и удаляем те, которые совпадают (сравнение производим по множествам кортежей, содержащихся в таблицах предикатов). Из кортежей всех наилучших восстановлений базы данных формируем новое множество кортежей в таблицах базы данных — $\mathbb{T}'$. Кортежи добавляем только в те таблицы, в которых они содержались изначально, дублирующиеся кортежи не добавляем. Переходим к 2.

6. Процесс закончен, найдено множество всех восстановлений $\mathbb{Q}$. Среди всех построенных восстановлений находим те, у которых число кортежей наибольшее. Таких восстановлений может быть несколько. Эти восстановления образуют множество всех максимальных восстановлений для данной базы данных.

Введение дополнительных действий при поиске наилучших восстановлений вызвано тем, что $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} \neq \varnothing$. Следовательно, после того

как построим восстановление для части (II)&...&(II), возможно возникновение новых противоречий в части (III)&...&(III), и наоборот. Однако отсутствие удалений после выполнения 3 явно показывает, что восстановление построено.

Объединение множеств кортежей, производимое на этапе 5, обусловлено тем, что после удалений кортежей максимальные пустые подграфы для гиперграфа неминуемо изменятся и их надо находить заново.

Докажем, что при использовании приведенного алгоритма построим все восстановления приведенной базы данных.

Предположим, что найдется такое восстановление Q', которое не принадлежит $\mathbb{Q}$. Рассмотрим, какие кортежи могут входить в такое восстановление:

1. Очевидно, что ядро восстановления войдет в Q'.

2. Поскольку Q' не входит в $\mathbb{Q}$, то для него должно выполняться что-то из нижеперечисленного:

   - в Q' войдет один из кортежей, удаленный на этапе 1, однако тогда Q' не будет восстановлением, так как в Φ будут противоречия;
   - в Q не будет входить ни один максимальный пустой подграф, найденный на этапе 2, однако в этом случае либо в Q' войдет пустой подграф, не являющийся максимальным, однако тогда Q' не будет восстановлением (так как к нему можно будет добавить кортеж и противоречие не возникнет); либо войдет какое-либо другое подмножество, отличное от пустых подграфов, тогда Q' также не будет восстановлением, так как останутся противоречия;
   - аналогично для 3 — если оставим в Q' какой-то удаленный кортеж, то оно не будет восстановлением;
   - на этапе выполнения 5 ничего не удаляем и не строим восстановления, для вновь построенных таблиц баз данных все будет аналогично только что рассмотренному.

Соответственно, восстановления Q', которое бы не входило ни в одно множество $\mathbb{Q}$, построить не можем. Следовательно, в наше множество $\mathbb{Q}$ войдут все восстановления. И только они, по построению $\mathbb{Q}$ все входящие в него множества будут восстановлениями. □

Восстановления для третьего подкласса

Утверждение 15. *Пусть* $D = \langle \mathbb{T}, \Phi \rangle$, $P_1, \ldots, P_n \in \mathbb{P}$ *и* $T_1, \ldots, T_n$ *— соответствующие им таблицы,*

$$\Phi = \text{(I)}\& \ldots \&\text{(I)}\&\text{(II)}\& \ldots \&\text{(II)}\&\text{(III)}\& \ldots \&\text{(III)},$$

и в базе данных D *содержатся противоречия.*

Тогда для D *возможно построить наилучшее восстановление.*

Доказательство. Для построения наилучшего восстановления для данной формулы предлагается следующий алгоритм.

1. Строим наилучшее восстановление для части $\text{(I)}\& \ldots \&\text{(I)}$ формулы (т.е. для базы данных $\langle \mathbb{T}, \text{(I)}\& \ldots \&\text{(I)} \rangle$) согласно алгоритму, приведенному в утверждении 2.

2. Если $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} = \varnothing$, то для части формулы

   $$\text{(II)}\& \ldots \&\text{(II)}\&\text{(III)}\& \ldots \&\text{(III)}$$

   (т.е. для базы данных $\langle \mathbb{T}', \text{(II)}\& \ldots \&\text{(II)}\&\text{(III)}\& \ldots \&\text{(III)} \rangle$, где $\mathbb{T}'$ — таблицы базы данных D после удаления кортежей на предыдущем шаге) устраняем противоречия и строим наилучшее восстановление согласно алгоритму, сформулированному при доказательстве утверждения 13.

3. Если $\mathbb{P}_{\text{II}} \cap \mathbb{P}_{\text{III}_2} \neq \varnothing$, то для части формулы

   $$\text{(II)}\& \ldots \&\text{(II)}\&\text{(III)}\& \ldots \&\text{(III)}$$

   (т.е. для базы данных $\langle \mathbb{T}', \text{(II)}\& \ldots \&\text{(II)}\&\text{(III)}\& \ldots \&\text{(III)} \rangle$, где $\mathbb{T}'$ — таблицы базы данных D после удаления кортежей на предыдущем шаге) устраняем противоречия и строим наилучшее восстановление согласно алгоритму, сформулированному при доказательстве утверждения 14.

Доказательство того, что построенное в результате выполнения данного алгоритма восстановление будет максимальным, аналогично и базируется на соответствующих доказательствах, приведенных для утверждений 2, 13, 14. □

Следует заметить, что приведенный алгоритм может быть использован также для произвольной формулы из класса A. Действительно, предположим, что нам дана произвольная формула из класса A. Рассмотрим, каким образом мы можем построить для нее восстановление.

В общем случае, поскольку формула принадлежит классу A, мы можем выразить ее как конъюнкцию других подформул. Тогда в общем виде эта формула может быть записана как:

$$\Phi = (\mathrm{I})\& \ldots \&(\mathrm{I})\&(\mathrm{II})\& \ldots \&(\mathrm{II})\&(\mathrm{III})\& \ldots \&(\mathrm{III}).$$

В зависимости от того, какие из частей представлены в этой формуле, и какие взаимосвязи между предикатами внутри частей могут быть прослежены, для построения наилучшего восстановления может быть использован один из алгоритмов, сформулированных при доказательстве утверждений предыдущих и текущего разделов. При этом в самом худшем случае может быть использован алгоритм, предложенный при доказательстве утверждения 15.

Заключение

Таким образом, в данной работе был изучен класс формул A, соответствующий выразительным возможностям языка записи формул-ограничений SQL. Для данного класса формул рассмотрены возможности построения восстановлений для противоречивых баз данных и предложены соответствующие алгоритмы, исходя из концепции удаления кортежа.

Список литературы

[1] Chomicki J., Marchinkowski, J. Minimal-change integrity maintenance using tuple deletion. // Inf. Comput., 2005. 197(1-2). P.90-121.

[2] Chomiki J., Lobo J. and Naqvi S. Conflict Resolution Using Logic Programming // IEEE Transactions on Knowledge and Data Engineering, 2003. № 15(1). P. 244-249.

[3] Chomiki J. Semantic Optimization Techniques for Preference Queries // Information Systems, 2007. № 32(5). P. 670-684.

[4] Motro A. Sources of Uncertainty, Imprecision and Inconsistency in Information Systems. // Uncertainty Management in Information

Systems: From Needs to Solutions. Kluwer Academic Publishers, 1996. Ch. 2. P. 9-34.

[5] Мендельсон Э. Введение в математическую логику. М.:Наука, 1984.

[6] Трифонова Е.Е. О представлении ограничений, записанных с помощью SQl, для построения восстановлений баз данных // Материалы VIII молодежной школы по дискретной математике и её приложениям (Москва, 24-29 октября 2011 г.), 2011. ч.II. С.30-35.