Методы и алгоритмы формирования многомерных данных с использованием промежуточных представлений

05.13.11 —«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»
Диссертация на соискание учёной степени кандидата физико-математических наук

С.В. Мосин

Научный руководитель (консультант):
Зыкин Сергей Владимирович
профессор, доктор технических наук

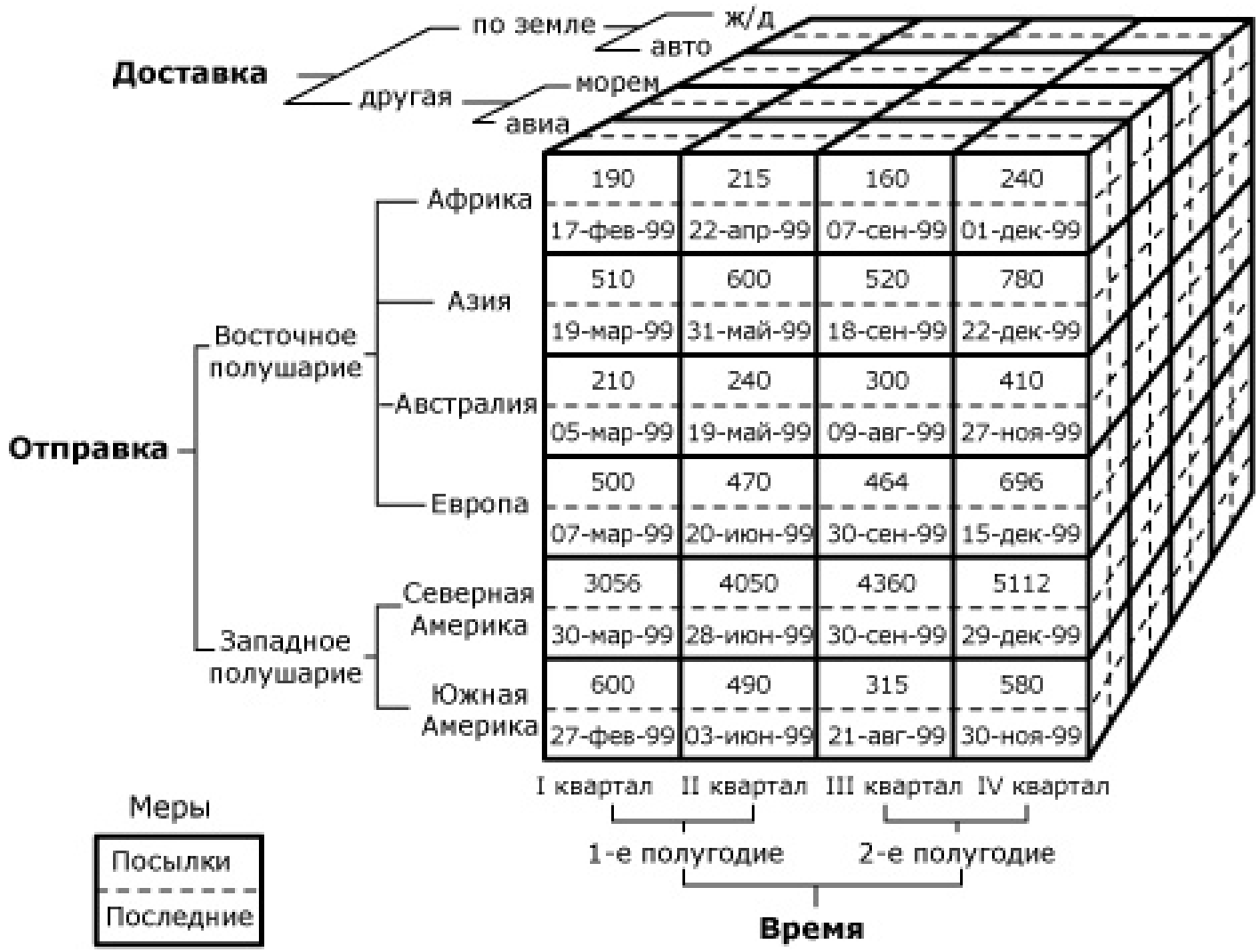
Цель работы

Разработка методов и алгоритмов формирования многомерного представления данных из реляционного представления при наложении логических ограничений на размерности и при использовании сохраненных данных.

Основные задачи

- ✓ Разработать алгоритм направленного перебора для формирования контекстов.
- ✓ Разработать оптимизированный алгоритм формирования представления данных «Таблица Соединений».
- ✓ Исследовать метод сравнения областей истинности логических ограничений при анализе сохраненных (кэшированных) данных.
- ✓ Разработать алгоритм повторного использования сохраненных данных и вычисления недостающих данных на основе сравнения областей истинности.
- ✓ Разработать алгоритм идентификации необходимых данных при анализе сохраненных данных.
- ✓ Реализовать программное обеспечение, формирующее гиперкубическое представление из исходного реляционного представления

Технология OLAP



Формализация задачи

$$RM \Rightarrow TJ \Rightarrow GC$$

где RM — реляционное представление данных, TJ — «Таблица Соединений», GC — гиперкуб

Формализация задачи

Исходные данные:

- Схема БД $\Re = \{R_1, R_2, \ldots, R_k\}$, находится в 5 НФ.
- Множество атрибутов $U = \{A_1, A_2, \ldots, A_k\}$.
- $\forall i, j : i \neq j \quad \langle R_i \rangle \not\subseteq \langle R_j \rangle$
- Неопределенное значение $NULL$ атрибута A_j в кортеже t не равно любому другому значению, в том числе другому неопределенному значению.

Многомерное представление:

- Совокупность размерностей: $\{D_1, D_2, \ldots, D_k\}$.
- D_l — множество расширенных имен атрибутов: $R_i.A_j$, $A_j \in \langle R_i \rangle$.
- M — множество мер, также заданных в виде расширенных имен атрибутов.
- Логическая формула F_l задана в ДНФ.
- НЕТ зависимости $D_1 D_2 \ldots D_d \to M \to$ в одной ячейке гиперкуба несколько значений (список) атрибута $R_i.A_j \in M$

Формирование контекстов

Пусть $C = R_1, R_2, \ldots, R_q$ — произвольное подмножество отношений реляционной БД.

Определение 1

Зависимость $dep_j \in DEP$ будем считать ***реализованной*** *на C, если операция дополнения, удаления или модификации кортежа в произвольном отношении $R_i \in C$ будет заблокирована, если при этом нарушается зависимость dep_j.*

Определение 2

Множество C будем называть ***контекстом****, если оно удовлетворяет свойству соединения без потери информации на зависимостях DEP, реализованных в C.*

Формирование контекстов

Теорема 1

Множество отношений C обладает свойством СБПИ, если существует отношение $R_i \in C$, замыкание первичного ключа которого совпадает со всем множеством атрибутов отношений множества C.

Формирование контекстов

Теорема 1

Множество отношений C обладает свойством СБПИ, если существует отношение $R_i \in C$, замыкание первичного ключа которого совпадает со всем множеством атрибутов отношений множества C.

Пусть $m = \{R_1, R_2, \ldots, R_q\}$ — произвольное множество отношений и $\langle C_m \rangle = \langle R_1 \rangle \cup \langle R_2 \rangle \cup \cdots \cup \langle R_q \rangle$.

Теорема 2

Множество отношений $C_{m+1} = \{R_1, R_2, \ldots, R_q, R_{q+1}\}$ не обладает свойством СБПИ на DEP, если зависимость $Z \twoheadrightarrow X(Y)$ не выводима из DEP, где $X \subseteq \langle C_m \rangle$, $Y \subseteq \langle R_{q+1} \rangle$ и $\langle C_m \rangle \cap \langle R_{q+1} \rangle \subseteq Z$.

Формирование контекстов

Определение 3

Существующее соединение Выражение $R_1 \bowtie R_2 \bowtie \cdots \bowtie R_q$ *будем называть существующим соединением, если для совокупности отношений* $R_i, i = 1, \ldots, q$ *существует хотя бы одна перестановка* $V_1, V_2, \ldots, V_q$ *отношений* $R_1, R_2, \ldots, R_q$ *такая, что* $([V_1] \cup [V_2] \cup \cdots \cup [V_j]) \cap [V_{j+1}] \neq \emptyset, j = 1, \ldots, q-1.$

Теорема 3

Если множество отношений $C = R_1, R_2, \ldots, R_q$ *не образует существующее соединение, то оно не обладает свойством СБПИ на множестве функциональных зависимостей* FD.

Формирование контекстов

1. Замыкание первичного ключа нового отношения R_i совпадает со всем множеством атрибутов в выбранных отношениях. Такое отношение получает **приоритет 3**.

Формирование контекстов

1. Замыкание первичного ключа нового отношения R_i совпадает со всем множеством атрибутов в выбранных отношениях. Такое отношение получает **приоритет 3**.
2. Для отношения R_i выполнено условие существования связи, соответствующей ЗВ $R_i[X] \subseteq R_j[X]$ с уже выбранными отношениями R_j, где множество атрибутов X является первичным ключом отношения R_j. Такое отношение получает **приоритет 2**.

Формирование контекстов

1. Замыкание первичного ключа нового отношения R_i совпадает со всем множеством атрибутов в выбранных отношениях. Такое отношение получает **приоритет 3**.
2. Для отношения R_i выполнено условие существования связи, соответствующей ЗВ $R_i[X] \subseteq R_j[X]$ с уже выбранными отношениями R_j, где множество атрибутов X является первичным ключом отношения R_j. Такое отношение получает **приоритет 2**.
3. Если дополняемое отношение R_i не удовлетворяет условиям теорем 2 и 3, то такое отношение получает **приоритет 1**.

Формирование контекстов

1. Замыкание первичного ключа нового отношения R_i совпадает со всем множеством атрибутов в выбранных отношениях. Такое отношение получает **приоритет 3**.
2. Для отношения R_i выполнено условие существования связи, соответствующей ЗВ $R_i[X] \subseteq R_j[X]$ с уже выбранными отношениями R_j, где множество атрибутов X является первичным ключом отношения R_j. Такое отношение получает **приоритет 2**.
3. Если дополняемое отношение R_i не удовлетворяет условиям теорем 2 и 3, то такое отношение получает **приоритет 1**.
4. Остальные отношения получают **приоритет 0**.

Формирование контекстов

Алгоритм формирования контекстов.

1. Подсчет весов для отношений R^1, и их упорядочение по убыванию весов.

Формирование контекстов

Алгоритм формирования контекстов.

1. Подсчет весов для отношений R^1, и их упорядочение по убыванию весов.
2. Формирование сочетаний без повторений из отношений R^1:
   $(1), (2), \ldots, (1,2), (1,3), \ldots, (2,3) \ldots$
   СБПИ выполнено $\rightarrow$ дополняем R^1 к контексту.

Формирование контекстов

Алгоритм формирования контекстов.

1. Подсчет весов для отношений R^1, и их упорядочение по убыванию весов.
2. Формирование сочетаний без повторений из отношений R^1:
   $(1), (2), \ldots, (1,2), (1,3), \ldots, (2,3) \ldots$
   СБПИ выполнено $\rightarrow$ дополняем R^1 к контексту.
3. В процессе выполнения алгоритма пользователю предлагается выбрать нужный контекст.

Формирование контекстов

Алгоритм формирования контекстов.

1. Подсчет весов для отношений R^1, и их упорядочение по убыванию весов.
2. Формирование сочетаний без повторений из отношений R^1:
   $(1), (2), \ldots, (1,2), (1,3), \ldots, (2,3) \ldots$
   СБПИ выполнено $\rightarrow$ дополняем R^1 к контексту.
3. В процессе выполнения алгоритма пользователю предлагается выбрать нужный контекст.

Формирование контекстов

$R_1 =$ Специальности

№ специальности	Специальность
...	...

$R_2 =$ Предметы

№ предмета	Предмет
...	...

Формирование контекстов

$R_1 =$ Специальности

№ специальности	Специальность
...	...

$R_2 =$ Предметы

№ предмета	Предмет
...	...

$R_3 =$ Контроль

№ специальности	№ предмета	Семестр	Вид занятия	Количество часов
...	...	...	...	...

Определение 4

Множество атрибутов KM_j будем называть ключом атрибута меры $R_i.A_j \in M$ в контексте C, если $KM_j \subseteq \langle C \rangle$, зависимость $KM_j \to R_i.A_j$ выводима на множестве функциональных зависимостей, и не существует выводимой зависимости $Y \to R_i.A_j$, где $Y \subset KMj$.

Обозначим $KM = KM_1 \cup KM_2 \cup \cdots \cup KM_h$, где $h = |M|$, — общий ключ для всех мер гиперкуба.

Промежуточное представление данных

Определения

Определение 5

Реализацией контекста C называется множество кортежей c, определяемое по следующему правилу:

$$c = \sigma_F(R_{mas(1)}[W_{mas(1)}] \bowtie R_{mas(2)}[W_{mas(2)}] \bowtie \cdots \bowtie R_{mas(p)}[W_{mas(p)}]),$$

где $V_{mas(i)} \subseteq W_{mas(i)}, i = 1, 2, \ldots, p$, V_i — множество атрибутов $A_j \in \langle R_i \rangle$:

1. $\exists D_l : R_i.A_j \in D_l$
2. $R_i.A_j \in M$
3. $R_i.A_j \in KM$
4. $\exists R_v \in C : A_j \in \langle R_v \rangle, i \neq v$
5. $\exists F_l : A_j \in \langle F_l \rangle$

Промежуточное представление данных

Правила задания

Построение «Таблицы Соединений»:

- Находим все контексты из набора $R_1, R_2, \ldots, R_k$. (подмножества данного множества отношений, удовлетворяющие свойству СБПИ)

Промежуточное представление данных

Правила задания

Построение «Таблицы Соединений»:

- Находим все контексты из набора $R_1, R_2, \ldots, R_k$. (подмножества данного множества отношений, удовлетворяющие свойству СБПИ)
- Пусть c — реализация контекста C. $\forall u \in c$ формируем кортеж t «Таблицы Соединений»: $t[A_j] = u[A_j]$, если $A_j \in \bigcup\limits_{R \in C} \langle R \rangle$, иначе $t[A_j] = emp$, где emp — пустое значение.

Промежуточное представление данных

Правила задания

Построение «Таблицы Соединений»:

- Находим все контексты из набора $R_1, R_2, \ldots, R_k$. (подмножества данного множества отношений, удовлетворяющие свойству СБПИ)
- Пусть c — реализация контекста C. $\forall u \in c$ формируем кортеж t «Таблицы Соединений»: $t[A_j] = u[A_j]$, если $A_j \in \bigcup\limits_{R \in C} \langle R \rangle$, иначе $t[A_j] = emp$, где emp — пустое значение.
- Каждому кортежу поставим в соответствие битовый вектор $g(t) = (g_1(t), g_2(t), \ldots, g_k(t))$, где $g_j(t) = 1$, если отношение R_j участвует в текущем соединении, и $g_j(t) = 0$ в противном случае.

Промежуточное представление данных

Правила задания

- Удаляем все подчиненные кортежи.

Определение 6

Кортеж $t \in s$ *является менее определенным или равным кортежу* $t' \in s$, *когда для любого атрибута* A_i *выполнено условие: если* $t[A_i] \neq t'[A_i]$, *то* $t[A_i] = emp$ *и* $g_j(t') \geq g_j(t), j = 1, \ldots, k$, *причем* $t[A_i] = t'[A_i]$, *если* A_i *принимает значение* $NULL$ *в обоих кортежах. В этом случае будем писать* $t \prec t'$, *назовем кортеж* t *подчиненным кортежу* t' *и оба этих кортежа будем считать сравнимыми.*

Промежуточное представление данных

Правила задания

- Удаляем все подчиненные кортежи.

Определение 6

Кортеж $t \in s$ является менее определенным или равным кортежу $t' \in s$, когда для любого атрибута A_i выполнено условие: если $t[A_i] \neq t'[A_i]$, то $t[A_i] = emp$ и $g_j(t') \geq g_j(t), j = 1, \ldots, k$, причем $t[A_i] = t'[A_i]$, если A_i принимает значение $NULL$ в обоих кортежах. В этом случае будем писать $t \prec t'$, назовем кортеж t подчиненным кортежу t' и оба этих кортежа будем считать сравнимыми.

- Удаляем кортежи $t : F(t) = FALSE$.

Промежуточное представление данных

Свойства

Определение 7

Проекция $\pi_{R(L)}(s)$ *есть совокупность кортежей* $u[R(L)]$, *определенных на множестве всех атрибутов отношений* $R(L)$, *где для каждого* $u[R(L)]$ *существует кортеж* $t \in s$, *такой, что* $u[R(L)] = t[R(L)]$ *и* $g_{mas(i)}(t) = 1, i = 1, 2, \ldots, p$.

Теорема 4

Для любого множества отношений $R^* = \{R_1^*, R_2^*, \ldots, R_q^*\} \subseteq C'$, *удовлетворяющего свойству СБПИ, где* C' *— контекст и* s *— таблица соединения, соответствующая* C', *выполнено:*

$$\pi_{R^*(L)}(s)[Z] = \sigma_F(R_1^*[Z_1] \bowtie R_2^*[Z_2] \bowtie \cdots \bowtie R_q^*[Z_q]),$$

где $V_i \subseteq Z_i \subseteq W_i$, $Z = Z_1 \cup Z_2 \cup \cdots \cup Z_q$, L *— вектор номеров отношений в* R^*.

Промежуточное представление данных

Пример

Таблица: Пример «Таблицы Соединений».

№ студ.	**№ группы**	**ФИО**	**Код группы**	**№ специальности**	**№ курса**	g_1	g_2
1	2	Алексенко С. В.	М-220	5	2	1	1
2	2	Белоусов П. О.	М-220	5	2	1	1
3	2	Бессараб О. П.	М-220	5	2	1	1
4	2	Вяткин М. С.	М-220	5	2	1	1
5	2	Драница А. А.	М-220	5	2	1	1
6	2	Ефимов Е. С.	М-220	5	2	1	1

Промежуточное представление данных

Оптимизированный алгоритм формирования. Обозначения.

Вход:

- $C = \{R_1, R_2, \ldots, R_p\}$ — контекст.
- DEP — реализованные зависимости.
- F — логическая формула.

Выход: s — «Таблица Соединений»

Промежуточное представление данных

Оптимизированный алгоритм формирования. Обозначения.

Вход:

- $C = \{R_1, R_2, \ldots, R_p\}$ — контекст.
- DEP — реализованные зависимости.
- F — логическая формула.

Выход: s — «Таблица Соединений»

Обозначения:

- $Comb(i, p, mas)$ — процедура формирования сочетаний из p по i.
- $Transform(R)$ — процедура преобразования кортежей: добавление вектора g.

Промежуточное представление данных

Алгоритм формирования Таблицы Соединений

$s = \emptyset$
for $i = 1, \ldots, p$ **do**
    $mas = (0, 0, \ldots, 0)$
    while $Comb(i, p, mas)$ **do**
        if $Llj(C, mas, DEP)$ **then**
            $s = AddContext(s, C, mas)$
        end if
    end while
end for

Промежуточное представление данных

Алгоритм формирования Таблицы Соединений

---

function AddContext(s, C, mas)

$R = R_{mas(1)}[W_{mas(1)}] \bowtie R_{mas(2)}[W_{mas(2)}] \bowtie \cdots \bowtie R_{mas(i)}[W_{mas(i)}]$

$s' = Transform(R)$

for all $t_i \in s$ **do**

for all $t_j \in s'$ **do**

if $t_i \prec t_j$ **then**

$s = s - t_i$

end if

end for

end for

return $s = s \cup \sigma_F(s')$

end function

Гиперкубическое представление данных

$$TJ \Rightarrow GC$$

$GC = GC(T_0, T_1, T_2, \ldots, T_d)$, где T_i — преобразованные «Таблицы Соединений»

Гиперкубическое представление данных

$$TJ \Rightarrow GC$$

$GC = GC(T_0, T_1, T_2, \ldots, T_d)$, где T_i — преобразованные «Таблицы Соединений»
Таблица T_0 соответствует контексту приложения и определена на множестве атрибутов $D_0 = D_1 \cup D_2 \cup \cdots \cup D_d \cup M$. При этом:

$$T_0 = s_0[D_0]\,, \tag{1}$$

где s_0 — таблица соединения для контекста приложения C_0, таблица T_0 содержит остатки соединения и сопоставленные значения мер значениям размерностей.

Гиперкубическое представление данных

Для формирования каждой размерности $(1 \leq i \leq d)$ используется одна из трех последующих формул:

- $T_i = s_i[D_i]$, где s_i — таблица соединения для контекста C_i

Гиперкубическое представление данных

Для формирования каждой размерности ($1 \leq i \leq d$) используется одна из трех последующих формул:

- $T_i = s_i[D_i]$, где s_i — таблица соединения для контекста C_i
- $T_i = \pi_{R^*(L)}(s_0)[D_i]$, где $C_i = \{R'_1, R'_2, \ldots, R'_q\}$ — пустой контекст, L — вектор номеров отношений пустого контекста

Гиперкубическое представление данных

Для формирования каждой размерности ($1 \leq i \leq d$) используется одна из трех последующих формул:

- $T_i = s_i[D_i]$, где s_i — таблица соединения для контекста C_i
- $T_i = \pi_{R^*(L)}(s_0)[D_i]$, где $C_i = \{R'_1, R'_2, \ldots, R'_q\}$ — пустой контекст, L — вектор номеров отношений пустого контекста
- $T_i = \sigma_{F_i}(R'_1[W_1] \ltimes R'_2[W_2] \ltimes \cdots \ltimes R'_p[W_p])[D_i]$

Гиперкубическое представление данных

Таблица: Фрагмент учебного плана.

	Семестр	2			
	Предмет	Математика			
	Вид занятия	Лекции	Практика	Лаб. раб.	*Контроль успеваемости*
Специальность		*Часы*	*Часы*	*Часы*	
История		36	36	18	зач., экз.
Филология		18	18	18	зач.
Правоведение		48	48	24	зач., экз.

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД
Возможные значения T_j^i:

- $Expr_1 \; \theta \; Expr_2$, $(\theta \in \{=, \neq, >, <, \leq, \geq\})$, $Expr_i$ — атрибуты/константы, согласованные по типам.

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД
Возможные значения T_j^i:

- $Expr_1 \; \theta \; Expr_2$, $(\theta \in \{=, \neq, >, <, \leq, \geq\})$, $Expr_i$ — атрибуты/константы, согласованные по типам.
- $Expr_1 \; [NOT] \; BETWEEN \; Expr_2 \; AND \; Expr_3$ (содержимое в прямоугольных скобках $[*]$ для предиката не является обязательным при написании)

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД
Возможные значения T_j^i:

- $Expr_1 \; \theta \; Expr_2$, $(\theta \in \{=, \neq, >, <, \leq, \geq\})$, $Expr_i$ — атрибуты/константы, согласованные по типам.
- $Expr_1 \; [NOT] \; BETWEEN \; Expr_2 \; AND \; Expr_3$ (содержимое в прямоугольных скобках $[*]$ для предиката не является обязательным при написании)
- $Expr \; [NOT] \; IN \; S$, где S — список значений либо подзапрос.

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД
Возможные значения T_j^i:

- $Expr_1 \; \theta \; Expr_2$, $(\theta \in \{=, \neq, >, <, \leq, \geq\})$, $Expr_i$ — атрибуты/константы, согласованные по типам.
- $Expr_1 \; [NOT] \; BETWEEN \; Expr_2 \; AND \; Expr_3$ (содержимое в прямоугольных скобках $[*]$ для предиката не является обязательным при написании)
- $Expr \; [NOT] \; IN \; S$, где S — список значений либо подзапрос.
- $Str_1 \; [NOT] \; LIKE \; Str_2$, где Str_i — строки;

Логические ограничения

Общий вид логической формулы:

$$F = K_1 \vee K_2 \vee \cdots \vee K_m, \tag{2}$$

$$K_i = T_1 \& \; T_2 \ldots \& \; T_n, i = 1, \ldots, m, \tag{3}$$

здесь $T_j, j = 1, \ldots, n$ - предикаты на множестве атрибутов БД
Возможные значения T_j^i:

- $Expr_1 \; \theta \; Expr_2$, $(\theta \in \{=, \neq, >, <, \leq, \geq\})$, $Expr_i$ — атрибуты/константы, согласованные по типам.
- $Expr_1 \; [NOT] \; BETWEEN \; Expr_2 \; AND \; Expr_3$ (содержимое в прямоугольных скобках $[*]$ для предиката не является обязательным при написании)
- $Expr \; [NOT] \; IN \; S$, где S — список значений либо подзапрос.
- $Str_1 \; [NOT] \; LIKE \; Str_2$, где Str_i — строки;
- $Expr \; \theta \; ALL/ANY \; S$.

Логические ограничения

Область истинности

$\mathcal{A} = \{(a_1, \ldots, a_n) \mid a_i \in Dom(A_i), i = 1, \ldots, n\}$, где $Dom(A_i)$ - множество всех допустимых значений атрибута A_i.

$\mathcal{A} = Dom(A_1) \times Dom(A_2) \times \cdots \times Dom(A_n)$ - n-мерное пространство значений всех атрибутов базы данных. Текущее состояние базы данных соответствует подмножеству этого пространства.

Логические ограничения

Область истинности

$\mathcal{A} = \{(a_1, \ldots, a_n) \mid a_i \in Dom(A_i), i = 1, \ldots, n\}$, где $Dom(A_i)$ - множество всех допустимых значений атрибута A_i.

$\mathcal{A} = Dom(A_1) \times Dom(A_2) \times \cdots \times Dom(A_n)$ - n-мерное пространство значений всех атрибутов базы данных. Текущее состояние базы данных соответствует подмножеству этого пространства.

Определение 8

Областью истинности формулы F, определенной (2), (3), называется множество $M(F) = \{a \in \mathcal{A} \mid F(a) = TRUE\}$.

Логические ограничения

Проекция

Определение 9

Проекцией логической формулы F, заданной (2), (3), на множество атрибутов X называется логическая формула $F[X], \langle F[X] \rangle = X$, в которой все термы, содержащие атрибуты $R_i^F.A_i^F \notin X$, заменены на тривиальный терм $TRUE$.

Логические ограничения

Проекция

Определение 9

Проекцией логической формулы F, заданной (2), (3), на множество атрибутов X называется логическая формула $F[X], \langle F[X] \rangle = X$, в которой все термы, содержащие атрибуты $R_i^F.A_i^F \notin X$, заменены на тривиальный терм $TRUE$.

Утверждение 1 (Свойство включения)

$\forall X \subseteq \langle F \rangle \quad M(F) \subseteq M(F[X])$

Кэширование данных

Обозначения

P=$\{P_1, P_2, \ldots, P_m\}$ - сохраненные данные

$$P_v = \pi_{X_v}(\sigma_{F_v}(R_1^v \bowtie R_2^v \bowtie \cdots \bowtie R_{s(v)}^v))$$

$s(v)$ – количество отношений БД, использованных при формировании P_v,
π_{X_v} - операция проекции по множеству атрибутов X_v,
σ_{F_v} - операция селекции с логическим ограничением на кортежи F_v.

Целевое выражение, которое надо будет получить из представлений P:

$$P^* = \pi_{X^*}(\sigma_{F^*}(R_1^* \bowtie R_2^* \bowtie \cdots \bowtie R_l^*)\,.$$

Кэширование данных

Случай одного представления данных

Теорема 5

$P^* \subseteq \pi_{X^*}(\sigma_{F^*[X_v]}(P_v))$, *если:*
а) $X^* \subseteq X_v$
б) $\{R_1^v, \ldots, R_{s(v)}^v\} \subseteq \{R_1^*, \ldots, R_l^*\}$
в) $M(F^*) \subseteq M(F_v)$.

Теорема 6

$P^* = \pi_{X^*}(P_v)$, *если:*
а) $X^* \subseteq X_v$
б) $\{R_1^v, \ldots, R_{s(v)}^v\} = \{R_1^*, \ldots, R_l^*\}$
в) $M(F^*) = M(F_v)$.

Кэширование данных

Случай нескольких представлений данных

Теорема 7

$P^* \subseteq \pi_{X^*}(\sigma_{F^*[X]}(P_1 \bowtie \cdots \bowtie P_n))$, *где* $X = \bigcup\limits_{v=1}^{n} X_v$ *если:*

а) $X^* \subseteq X$

б) $\bigcup\limits_{v=1}^{n} \{R_1^v, \ldots, R_{s(v)}^v\} = \{R'_1, \ldots, R'_{s'}\} \subseteq \{R_1^*, \ldots, R_l^*\}$

в) $M(F^*) \subseteq M(F_v), v = 1, \ldots, n.$

Кэширование данных

Случай нескольких представлений данных

Теорема 8

$P^* = \pi_{X^*}(P_1 \bowtie \cdots \bowtie P_n)$, *где* $X = \bigcup\limits_{v=1}^{n} X_v$ *если:*

а) $X^* \subseteq X$, $X_v \supseteq \langle \bowtie_{i=1}^{s(v)} R_i^v \rangle \cap (\bigcup\limits_{\substack{w=1 \\ w \neq v}}^{n} \langle \bowtie_{i=1}^{s(w)} R_i^w \rangle), v = 1, \ldots, n$

б) $\bigcup\limits_{v=1}^{n} \{R_1^v, \ldots, R_{s(v)}^v\} = \{R'_1, \ldots, R'_{s'}\} = \{R_1^*, \ldots, R_l^*\}$

в) $M(F^*) = M(F_1 \& \ldots \& F_n)$.

Кэширование данных

Случай нескольких представлений данных

Теорема 8

$P^* = \pi_{X^*}(P_1 \bowtie \cdots \bowtie P_n)$, *где* $X = \bigcup\limits_{v=1}^{n} X_v$ *если:*

а) $X^* \subseteq X$, $X_v \supseteq \langle \bowtie_{i=1}^{s(v)} R_i^v \rangle \cap (\bigcup\limits_{\substack{w=1 \\ w \neq v}}^{n} \langle \bowtie_{i=1}^{s(w)} R_i^w \rangle), v = 1, \ldots, n$

б) $\bigcup\limits_{v=1}^{n} \{R_1^v, \ldots, R_{s(v)}^v\} = \{R'_1, \ldots, R'_{s'}\} = \{R_1^*, \ldots, R_l^*\}$

в) $M(F^*) = M(F_1 \& \ldots \& F_n)$.

Кэширование данных

Сравнение областей истинности

$F^*, F_1, F_2, \ldots, F_n$ заданы (2), (3).

- Применимость кэша
  $M(F^*) \cap (M(F_1) \cup M(F_2) \cup \cdots \cup M(F_n)) = \emptyset \iff M(F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n)) = \emptyset \iff \boldsymbol{F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n) \equiv FALSE}.$

Кэширование данных

Сравнение областей истинности

$F^*, F_1, F_2, \ldots, F_n$ заданы (2), (3).

- Применимость кэша
  $M(F^*) \cap (M(F_1) \cup M(F_2) \cup \cdots \cup M(F_n)) = \emptyset \iff M(F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n)) = \emptyset \iff \boldsymbol{F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n) \equiv FALSE}.$
- Проверка условий теорем
  $M(F^*) \subseteq M(F_v), v = 1, \ldots, n \iff M(F^*) \subseteq M(F_1) \,\&\, M(F^*) \subseteq M(F_2) \,\&\, \ldots \,\&\, M(F^*) \subseteq M(F_n) \iff M(F^*) \subseteq M(F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n) \iff F^* \to F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n \equiv TRUE \iff \boldsymbol{F_1 \,\&\, F_2 \,\&\, \ldots \& F_n \vee \neg F^* \equiv TRUE}.$

Кэширование данных

Сравнение областей истинности

$F^*, F_1, F_2, \ldots, F_n$ заданы (2), (3).

- Применимость кэша
  $M(F^*) \cap (M(F_1) \cup M(F_2) \cup \cdots \cup M(F_n)) = \emptyset \iff M(F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n)) = \emptyset \iff \boldsymbol{F^* \,\&\, (F_1 \vee F_2 \vee \cdots \vee F_n) \equiv FALSE}.$
- Проверка условий теорем
  $M(F^*) \subseteq M(F_v), v = 1, \ldots, n \iff M(F^*) \subseteq M(F_1) \,\&\, M(F^*) \subseteq M(F_2) \,\&\, \ldots \,\&\, M(F^*) \subseteq M(F_n) \iff M(F^*) \subseteq M(F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n) \iff F^* \to F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n \equiv TRUE \iff \boldsymbol{F_1 \,\&\, F_2 \,\&\, \ldots \& F_n \vee \neg F^* \equiv TRUE}.$
- Формула для получения недостающих данных
  $M(F^*) \setminus M(F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n) = M(F_{n+1}).$
  $\boldsymbol{F_{n+1} = F^* \,\&\, \neg(F_1 \,\&\, F_2 \,\&\, \ldots \,\&\, F_n)}.$

Кэширование данных

Алгоритм использования кэша. Обозначения.

Вход:

- $P = \{P_1, \ldots, P_n\}$ — сохраненные данные.
- $R^* = \{R_1^*, \ldots, R_l^*\}$ — отношения нового пользовательского запроса.
- F^* — логическая формула нового запроса.
- X^* — множество атрибутов нового запроса.

Выход: P^* — данные пользовательского запроса

Кэширование данных

Алгоритм использования кэша. Обозначения.

Вход:

- $P = \{P_1, \ldots, P_n\}$ — сохраненные данные.
- $R^* = \{R_1^*, \ldots, R_l^*\}$ — отношения нового пользовательского запроса.
- F^* — логическая формула нового запроса.
- X^* — множество атрибутов нового запроса.

Выход: P^* — данные пользовательского запроса

Обозначения:

- Функция *get_ data_ from_ server()* запрашивает требуемые данные с сервера БД, минуя кэш.
- Функция *get_ data_ from_ cache()* получает данные из кэша, отсеивая лишние данные.

Кэширование данных

Алгоритм использования кэша

---

if $M(F^*) \cap (M(F_1) \cup M(F_2) \cup \cdots \cup M(F_n)) = \emptyset$ **then**
    return $get\_data\_from\_server()$
end if
if $M(F^*) \subseteq (M(F_1) \cap M(F_2) \cap \cdots \cap M(F_n))$ **then**
    return $get\_data\_from\_cache()$
end if
$F_{n+1} = F^* \ \& \ \neg(F_1 \ \& \ F_2 \ \& \ \ldots \ \& \ F_n)$
$P_{n+1} = \pi_{X^*}(\sigma_{F_{n+1}}(R_1 \bowtie \cdots \bowtie R_l))$
return $get\_data\_from\_cache(F^*) \cup P_{n+1}$

Кэширование данных

Пример

Рассмотрим фрагмент схемы БД, представляющий учебный план в университете:

$$R_1 = \textit{Студенты}\ (\textbf{№ студента}, \text{ФИО}, \text{Группа})$$

$$R_2 = \textit{Группы}\ (\textbf{Группа}, \text{Факультет})$$

$$R_3 = \textit{Успеваемость}\ (\textbf{№ студента}, \textbf{Дисциплина}, \text{Оценка})$$

Имена отношений выделены *курсивом*, их первичные ключи – **жирным** шрифтом.

Кэширование данных

Пример

Сохраненные в кэше запросы:

- Список студентов математического факультета:

$$P_1 = \pi_{X_1}(\sigma_{F_1}(R_1 \bowtie R_2)),$$

где $X_1 = \{\text{ФИО, Группа}\}$,
$F_1 = (\text{Факультет} = \text{"Математический"})$.

- Список студентов, сдавших физику:

$$P_2 = \pi_{X_2}(\sigma_{F_2}(R_1 \bowtie R_3)),$$

где $X_2 = \{\text{ФИО}, \text{Группа}, \text{Оценка}\}$,
$F_2 = (\text{Дисциплина} = \text{"Физика"} \ \& \ \text{Оценка} \geq 3)$.

Кэширование данных

Целевой запрос:

Отчет успеваемости по физике математического факультета:

$$P^* = \pi_{X^*}(\sigma_{F^*}(R_1 \bowtie R_2 \bowtie R_3)),$$

где $X^* = \{\text{ФИО}, \text{Группа}, \text{Оценка}\}$, F^* — логическая формула:
$F^* = (\text{Факультет} = \text{"Математический"} \ \& \ \text{Дисциплина} = \text{"Физика"})$.

Кэширование данных

Целевой запрос:

Отчет успеваемости по физике математического факультета:

$$P^* = \pi_{X^*}(\sigma_{F^*}(R_1 \bowtie R_2 \bowtie R_3)),$$

где $X^* = \{\text{ФИО}, \text{Группа}, \text{Оценка}\}$, F^* — логическая формула:
$F^* = (\text{Факультет} = \text{"Математический"} \ \& \ \text{Дисциплина} = \text{"Физика"})$.

$$P^* \subseteq \{P_1, P_2\}?$$

Кэширование данных

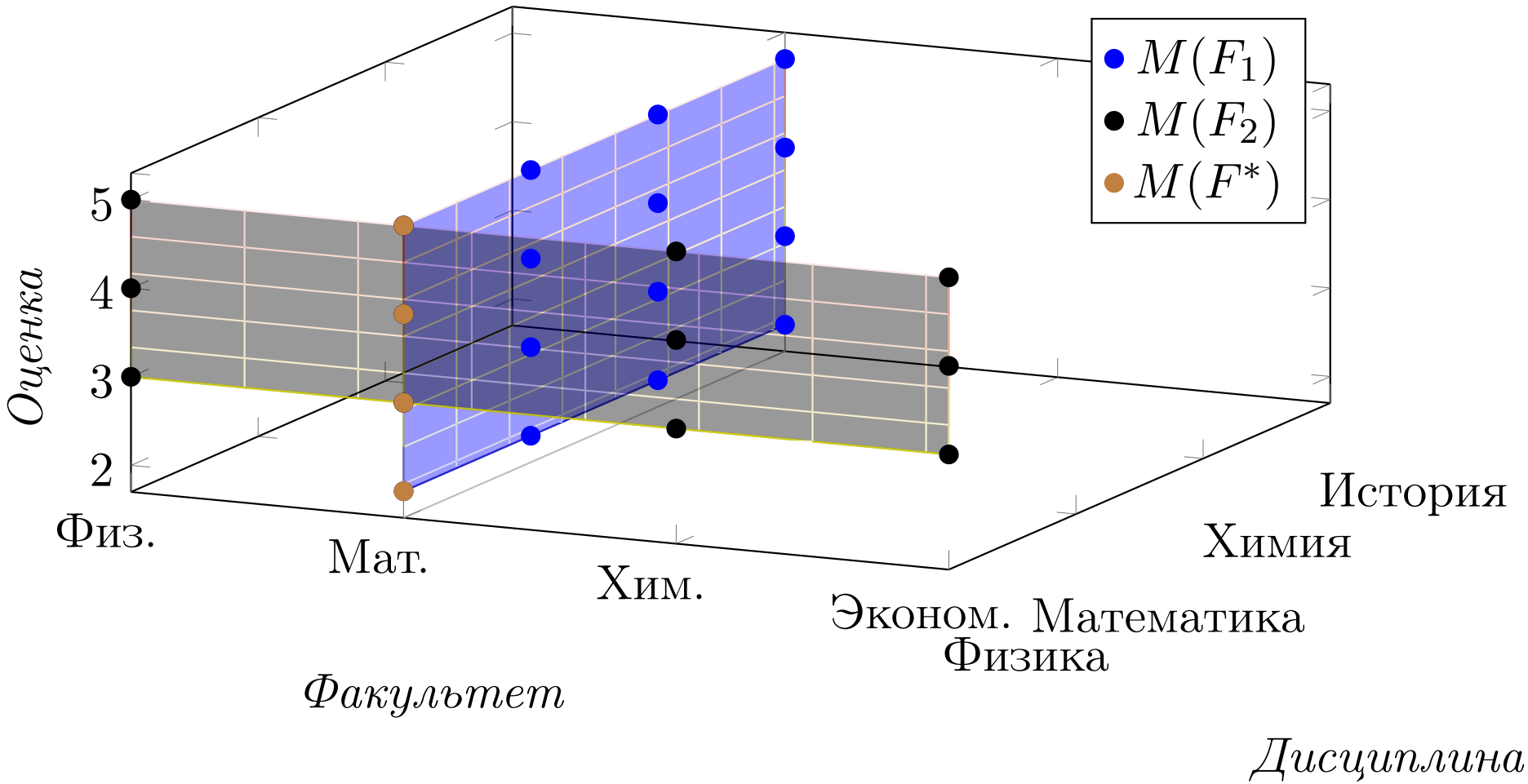
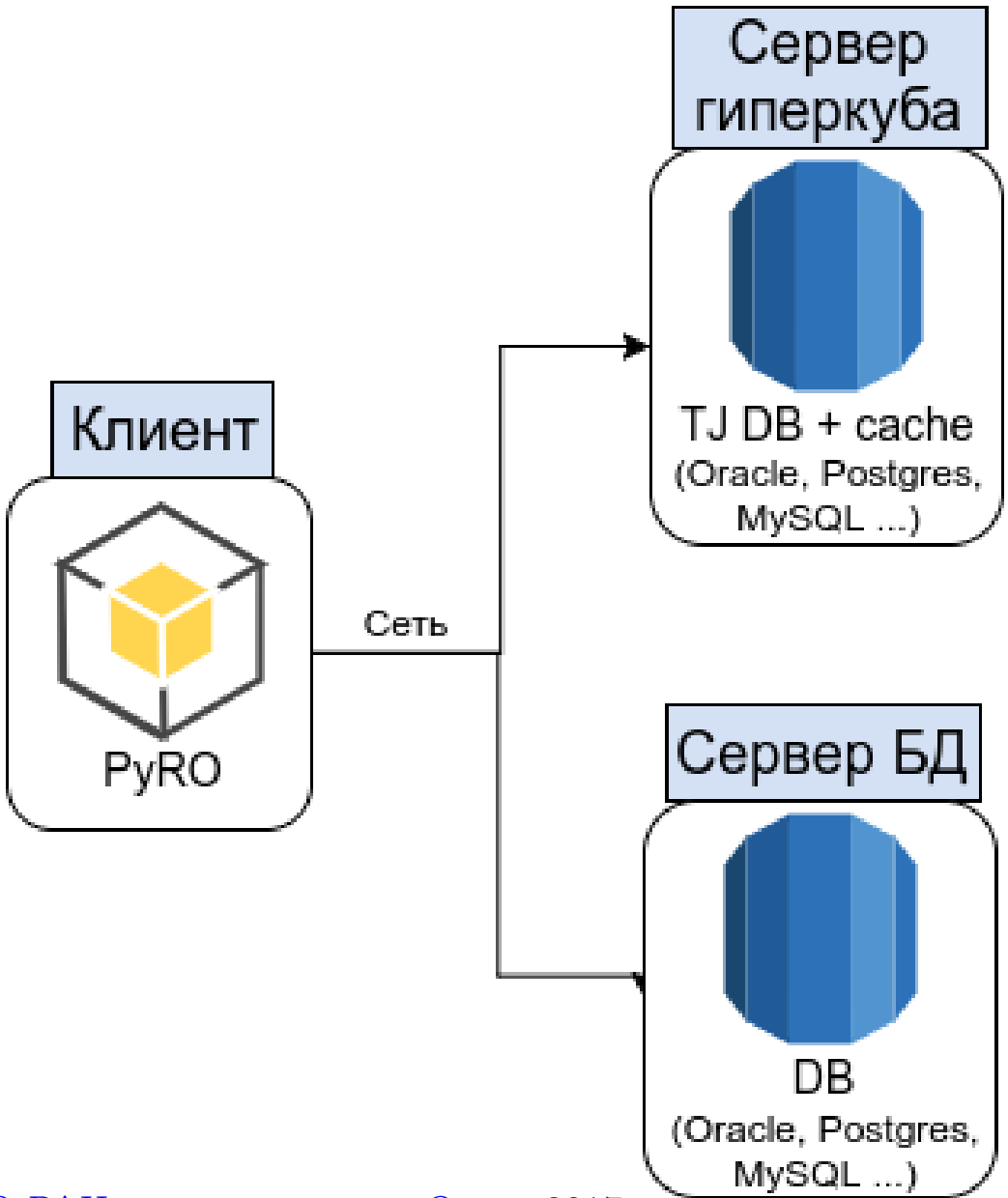


Рис.: Области истинности логических формул

Программное обеспечение

Архитектура



Программное обеспечение

Сравнение производительности

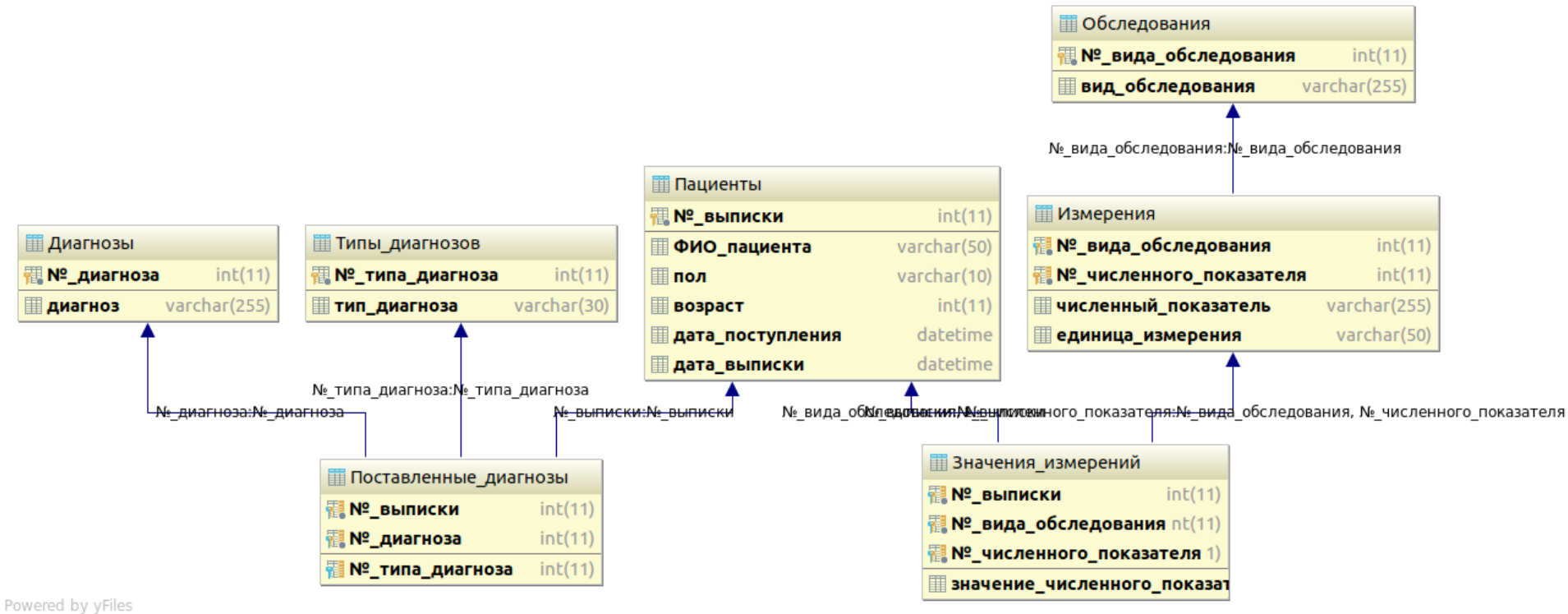


Рис.: Связи между отношениями.

Программное обеспечение

Сравнение производительности

Атрибуты размерностей:

1. Размерность "Показатели": Измерения.численный показатель, Пациенты.пол
2. Размерность "Диагнозы": Типы диагнозов.тип диагноза, Диагнозы.диагноз

Мера: Значения измерений.Значение численного показателя.

Мощности отношений (количество кортежей):

- Пациенты: 1443
- Значения измерений: 23342
- Поставленные диагнозы: 13391
- Типы диагнозов: 3
- Диагнозы: 200
- Измерения: 69
- Обследования: 67

Программное обеспечение

Сравнение производительности

Эксперименты:

1. $F_1^1 = \text{True}$, $F_2^1 = \text{True}$, никаких ограничений на данные
2. $F_1^2 = \{\text{численный показатель} = Hb \vee \text{численный показатель} = L \vee \text{численный показатель} = \textit{Белок}\}$, $F_2^2 = \{\text{диагноз} <> \textit{ИБС} \wedge \text{диагноз} <> \textit{абдоминальное ожирение}\}$
3. $F_1^3 = \{(\text{численный показатель} = Hb \wedge \text{пол} = \textit{мужской}) \vee \text{численный показатель} = L \vee \text{численный показатель} = \textit{Белок}\}$, $F_2^3 = \{\text{диагноз} = \textit{адгезивный перикардит} \vee \text{диагноз} = \textit{ФН}\}$

$$M(F_1^2) \subseteq M(F_1^1), M(F_2^2) \subseteq M(F_2^1) \tag{4}$$

$$M(F_1^3) \subseteq M(F_1^2), M(F_2^3) \subseteq M(F_2^2) \tag{5}$$

Программное обеспечение

Сравнение производительности

$$F_1^1 = \text{True}, F_2^1 = \text{True}$$

Таблица: Сравнительный анализ 1

ПО	Время (секунды)
PyRO	$3,02 * 10^5$
MS Analyses Services	2,84
Oracle OLAP	1,49

Программное обеспечение

Сравнение производительности

$F_1^2 = \{$численный показатель $= Hb \vee$ численный показатель $=$
$L \vee$ численный показатель $=$ *Белок*$\}$
$F_2^2 = \{$диагноз $<>$ *ИБС* $\wedge$ диагноз $<>$ *абдоминальное ожирение*$\}$

Таблица: Сравнительный анализ 2

ПО	Время (секунды)
PyRO	$2,6 * 10^{-1}$
MS Analyses Services	2,92
Oracle OLAP	1,41

Программное обеспечение

Сравнение производительности

$F_1^3 = \{(\text{численный показатель} = Hb \wedge \text{пол} = \textit{мужской}) \vee$
$\text{численный показатель} = L \vee \text{численный показатель} = \textit{Белок}\}$
$F_2^3 = \{\text{диагноз} = \textit{адгезивный перикардит} \vee \text{диагноз} = \textit{ФН}\}$

Таблица: Сравнительный анализ 3

ПО	Время (секунды)
PyRO	$2,4 * 10^{-1}$
MS Analyses Services	2,8
Oracle OLAP	1,39

Основные результаты, выносимые на защиту

- ✓ Разработан алгоритм направленного перебора для формирования контекстов.
- ✓ Разработан оптимизированный алгоритм формирования представления данных «Таблица Соединений».
- ✓ Предложен и исследован оригинальный метод сравнения областей истинности логических ограничений при анализе сохраненных (кэшированных) данных.
- ✓ Разработан алгоритм повторного использования сохраненных данных и вычисления недостающих данных на основе сравнения областей истинности.
- ✓ Разработан алгоритм идентификации необходимых данных при анализе сохраненных данных.
- ✓ Реализовано программное обеспечение, формирующее гиперкубическое представление из исходного реляционного представления.

Основные публикации

Публикации из перечня ВАК

Mosin S., *Zykin S.* — Truth space method for caching database queries. — // Modeling and Analysis of Information Systems. — 2015. — т. 22, № 2. — с. 248—258. — индексирована в MathSciNet.

Зыкин С. В., *Мосин С. В.*, *Полуянов А. Н.* — Технология раздельного формирования многомерных данных. — // Вестник Донского государственного технического университета. — 2016. — т. 85, № 2. — с. 114—129.

Мосин С. В. — Сравнение областей истинности запросов к реляционной базе данных. — // Вестник ЮУрГУ. — 2016. — т. 5, № 1. — с. 85—99. — (Вычислительная математика и информатика).

Мосин С. В. — Алгоритм использования кэша запросов к реляционной базе данных. — // Вестник СибГУТИ. — 2017. — № 1. — с. 47—57.

Основные публикации

Публикации в изданиях, индексируемых в Scopus и Web of Science

Mosin S. V., *Zykin S. V.* — Using logical formulas for caching uniform RDB queries. — // 2015 International Siberian Conference on Control and Communications (SIBCON). — май 2015. — с. 1—5.

Zykin S., *Mosin S.*, *Poluyanov A.* — Technology of separate formation of multidimensional data with lists of measure values. — // 2015 International Siberian Conference on Control and Communications (SIBCON). — май 2015. — с. 1—11.

Апробация

1. XVI Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL'2014)» (13 – 16 октября 2014 г., Дубна).
2. V Международная молодежная научно-практическая конференция с элементами научной школы «Прикладная математика и фундаментальная информатика (ПмиФИ'2015)» (23 – 30 апреля 2015 г., Омск).
3. Международная IEEE Сибирская конференция по управлению и связи (SibCon'2015) (21 – 23 мая 2015 г., Омск).
4. Международная IEEE Сибирская конференция по управлению и связи (SibCon'2015) (21 – 23 мая 2015 г., Омск).
5. VI Всероссийская научно-техническая конференция «Россия молодая: передовые технологии— в промышленность!» (10–11 ноября 2015 г., Омск)