МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. № 7 Москва 2019

ИНФОРМАЦИОННЫЙ ПОИСК

УДК 519.876.5:004.65

Е.М. Бениаминов

Алгебра скрытых отношений как средство моделирования статистических баз данных

Рассматривается аналог реляционной алгебры, используемый для моделирования баз данных в случае, когда пользователю значения атрибутов видны не полностью, а только с точностью до некоторого отношения эквивалентности на доменах. Этот случай важен для так называемых статистических баз данных, когда нужно выдавать и всесторонне анализировать сводную информацию по данным, хранящимся в некоторой базе данных, но доступ пользователям к персональным данным в базе данных закрыт. По аналогии с реляционным подходом в этом случае вводятся понятие скрытого отношения и система операций над скрытыми отношениями. Показаны некоторые соотношения между этими операциями и язык запросов.

Ключевые слова: *реляционная база данных, статистическая база данных, скрытые домены, операции над отношениями, группа симметрий, k-анонимность*

МОТИВАЦИЯ И НЕФОРМАЛЬНОЕ ВВЕДЕНИЕ

Проблема, которой посвящена настоящая работа не нова. В литературе давно обсуждается вопрос, как организовать доступ пользователям к базам данных для решения задач статистического анализа, не раскрывая при этом персональных данных, которые могут храниться в этих БД [1]. Эта тема не устарела и сейчас. Международные конференции "*Privacy in Statistical Databases*" («Конфиденциальность в статистических базах данных») проводятся уже в течение многих лет [2].

Один из подходов к статистическим базам данных состоит в том, что обычную реляционную БД, которая описывает классификацию объектов предметной

области, их свойства и связи между объектами, дополняют числовыми признаками-показателями, заданными на строках некоторых отношений. Запрос к такой базе данных состоит в построении таблиц агрегированных и статистических показателей на основе хранящихся в ней показателей и отношений. Такие статистические базы данных могут быть использованы для научных целей или для онлайн аналитических приложений (*OLAP)*. В настоящей работе предлагается развитие именно такого подхода.

Так же, как и в реляционном подходе к базам данных [3, 4], в этом подходе важными являются понятия отношения и атрибута отношения. Атрибут отношения задается именем *a* атрибута и областью значений (доменом) этого атрибута D_a. В отличие от стандартного подхода здесь будет предполагаться, что с одним и тем же именем атрибута могут связываться разные домены. При этом значения некоторых доменов могут наблюдаться полностью и применяться пользователями статистической базы данных в запросах. Такие домены D_a иногда будут называться наблюдаемыми. Предполагается также, что с именем атрибута *a* могут быть связаны и другие домены D'_a. Значения атрибутов в этих доменах могут не полностью наблюдаться пользователем в отношениях. Соответствие между ненаблюдаемыми и наблюдаемыми значениями задается отображением $p': D'_a \to D_a$. При этом предполагается, что каждое значение $d' \in D'_a$ пользователь воспринимает как значение $p'(d') \in D_a$ в области наблюдаемых значений D_a атрибута *a*.

Для примера в качестве атрибута рассмотрим атрибут с именем возраст *age* с наблюдаемыми значениями D_{age} в виде множества интервалов, которые именуются, как "младенческий", "младший дошкольный" и т. д. В качестве домена реальных значений для атрибута "возраст" может рассматриваться целочисленный тип данных $D'_{age} = \text{int}$. Отображение $p': D'_{age} \to D_{age}$ соответствует принадлежности числа к соответствующему интервалу.

Отношением *r* с набором атрибутов $\{a_1,...,a_n\}$ называется конечное подмножество $r \subset D'_{a_1} \times ... \times D'_{a_n}$ в декартовом произведении доменов соответствующих атрибутов. В нашем случае предполагается, что для статистических исследований строки отношений могут рассматриваться как в расширенных (скрытых) доменах D'_{a_i}, так и в наблюдаемых доменах. Отношения, в которых используются скрытые домены, естественно назвать скрытыми отношениями. Статистические показатели такие, как доход, количество штук, численность населения и т. д. задают веса на строках таких отношений.

Соединяя эти понятия, мы приходим к понятию показателя на скрытых отношениях, который представляется функцией (показателем), заданной на множестве строк отношения, где есть атрибуты с расширенными (скрытыми) доменами, и со значениями функции в множестве действительных чисел. Носителем такого показателя является множество строк, на которых этот показатель имеет ненулевое значение.

В настоящей работе мы попытаемся проанализировать ситуацию, в которой выполняется следующее основное предположение: пользователь знает имена отношений из схемы базы данных, может обратиться к скрытому отношению по имени, но видеть он может только суммарные значения весов строк по скрытым доменам самого отношения для строк наблюдаемой части носителя скрытого отношения. То же относится к скрытым отношениям, полученным в результате операций над ними, т. е. ответы на запросы пользователей к базе данных со скрытыми отношениями должны быть отношениями с наблюдаемыми доменами.

Чтобы прояснить эту ситуацию следует ввести систему операций над скрытыми отношениями и формально точно определить, что может наблюдать пользователь. Это и есть цель настоящей статьи.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Введем основные определения понятий, которые нам потребуются для моделирования баз данных со скрытыми отношениями: основное – это понятие скрытого отношения, которое является обобщением обычного понятия отношения.

Так же, как и в реляционном подходе к базам данных, атрибут отношения — это пара *a* и D_a, где *a* – имя атрибута, D_a – область значений (домен) этого атрибута.

Набором атрибутов $A=\{(a_1, D_{a_1}),\ldots,(a_n, D_{a_n})\}$ называется последовательность атрибутов с различными именами.

Отношением *r* с набором атрибутов $A=\{(a_1, D_{a_1}),\ldots,(a_n, D_{a_n})\}$ называется произвольное подмножество в декартовом произведении доменов атрибутов, то есть $r \subseteq D_{a_1} \times ... \times D_{a_n}$.

Для приложений баз данных к статистическим исследованиям или для задач аналитического анализа понятие отношения нужно немного модернизировать.

<u>Определение 1.</u> *Взвешенным отношением (показателем) φ с набором атрибутов* $A=\{(a_1, D_{a_1}), \ldots, (a_n, D_{a_n})\}$ *называется произвольная функция* $\phi: D_{a_1} \times ... \times D_{a_n} \to R$, *где R – множество действительных чисел (с указанием единицы измерения). Через V(A) мы будем обозначать множество всех взвешенных отношений с набором атрибутов A, т.е.* $V(A) \stackrel{def}{=} \{\phi \mid \phi: D_{a_1} \times ... \times D_{a_n} \to R\}$. *Если* $(d_1,...,d_n) \in D_{a_1} \times ... \times D_{a_n}$ *– произвольная строка из декартового произведения доменов атрибутов, то значение функции* $\phi(d_1,...,d_n)$ *на этой строке называется весом строки* $(d_1,...,d_n)$ *в отношении φ.*

Множество $r_\phi \overset{def}{=} \{(d_1,...,d_n) : \phi(d_1,...,d_n) \neq 0\}$ *строк с ненулевым весом называется носителем взвешенного отношения φ.*

Нас, как правило, будут интересовать взвешенные отношения с конечным носителем.

Далее, по аналогии с реляционной алгеброй, мы должны определить операции над взвешенными отношениями.

Аналогами теоретико-множественных операций над отношениями являются операции сложения и умножения взвешенных отношений.

Определение 2. *Если* $\phi_1, \phi_2 : D_{a_1} \times ... \times D_{a_n} \to R$ *– два взвешенных отношения с одинаковыми наборами атрибутов* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$, *то их суммой* $\phi_1 + \phi_2$, *разностью* $\phi_1 - \phi_2$ *и произведением* $\phi_1 \cdot \phi_2$ *называются взвешенные отношения с тем же набором атрибутов, веса которых на строках* $(d_1,...,d_n) \in D_{a_1} \times ... \times D_{a_n}$ *задаются, соответственно, выражениями:*

$$(\phi_1 + \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n) + \phi_2(d_1,...,d_n),$$
$$(\phi_1 - \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n) - \phi_2(d_1,...,d_n) \text{ и}$$
$$(\phi_1 \cdot \phi_2)(d_1,...,d_n) \overset{def}{=} \phi_1(d_1,...,d_n)\phi_2(d_1,...,d_n),$$

т. е. $\phi_1 + \phi_2$, $\phi_1 - \phi_2$ *и* $\phi_1 \cdot \phi_2$ *как функции являются суммой, разностью и произведением функций* φ_1 *и* φ_2.

Перейдем к определениям для построения операций, связывающих взвешенные отношения с разными наборами атрибутов.

Определение 3. *Пусть* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$ *и* $B=\{(b_1, D_{b_1}),...,(b_k, D_{b_k})\}$ *– два набора атрибутов. Морфизмом* $\gamma : A \to B$ *из набора атрибутов A в набор атрибутов B называется отображение множеств имен атрибутов* $\gamma : \{a_1,...a_n\} \to \{b_1,...,b_k\}$, *которое мы будем обозначать той же буквой* γ, *вместе с отображениями доменов* $\gamma_{a_i} : D_{\gamma(a_i)} \to D_{a_i}$, *для каждого* $a_i \in \{a_1,...,a_n\}$.

Заметим, что, если $\gamma : A \to B$ морфизм наборов атрибутов, то по нему естественным образом строится отображение декартовых произведений

$$\Gamma_\gamma : D_{b_1} \times ... \times D_{b_k} \to D_{a_1} \times ... \times D_{a_n} \qquad (1)$$

по правилу: если $(h_{b_1},...,h_{b_k}) \in D_{b_1} \times ... \times D_{b_k}$, то

$$\Gamma_\gamma(h_{b_1},...,h_{b_k}) \overset{def}{=} (d_{a_1},...,d_{a_n}) \in D_{a_1} \times ... \times D_{a_n}, \qquad (2)$$

где $d_{a_i} = \gamma_{a_i}(h_{\gamma(a_i)})$

для $a_i \in \{a_1,...,a_n\}$.

Определение 4. *Если* $\phi : D_{a_1} \times ... \times D_{a_n} \to R$ *– произвольное взвешенное отношение с набором атрибутов* $A=\{(a_1, D_{a_1}),...,(a_n, D_{a_n})\}$, *и* $\gamma : A \to B$ *– морфизм атрибутов, где* $B=\{(b_1, D_{b_1}),...,(b_k, D_{b_k})\}$, *то через* $\gamma_*(\phi)$ *обозначим взвешенное отношение с набором атрибутов B, которое как функция* $\gamma_*(\phi) : D_{b_1} \times ... \times D_{b_k} \to R$ *является композицией функции* Γ_γ, *заданой выражениями (1) и (2), и функции φ; т. е.* $\gamma_*(\phi) \overset{def}{=} \phi \circ \Gamma_\gamma$, *где* $\circ$ *– знак композиции функций. Взвешенное отношение* $\gamma_*(\phi)$ *называется прямым образом взвешенного отношения φ относительно морфизма γ.*

Определение 5. *Если* $\psi : D_{b_1} \times ... \times D_{b_k} \to R$ *взвешенное отношение с набором атрибутов B, и* $\gamma : A \to B$ *– морфизм атрибутов, то через* $\gamma^*(\psi)$ *обозначим взвешенное отношение с набором атрибутов A, которое как функция* $\gamma^*(\psi) : D_{a_1} \times ... \times D_{a_n} \to R$ *на любой строке* $(d_{a_1},...,d_{a_n}) \in D_{a_1} \times ... \times D_{a_n}$ *задается выражением:*

$$\gamma^*(\psi)(d_{a_1},...,d_{a_n}) \overset{def}{=}$$
$$\overset{def}{=} 0 + \sum_{\substack{(h_{b_1},...,h_{b_k}) \in D_{b_1} \times ... \times D_{b_k} \\ where \quad \Gamma_\gamma(h_{b_1},...,h_{b_k}) = (d_{a_1},...,d_{a_n})}} \psi(h_{b_1},...,h_{b_k}). \qquad (3)$$

Взвешенное отношение $\gamma^*(\psi)$ *называется обратным образом взвешенного отношения φ относительно морфизма* γ.

Особенностью операции прямого образа γ_* в случае, когда морфизм атрибутов $\gamma : A \to B$ не сюръективен и домены D_{b_i} бесконечны, является то, что взвешенные отношения $\phi : D_{a_1} \times ... \times D_{a_n} \to R$ с конечным носителем могут переводиться этой операцией в скрытые отношения с бесконечным носителем. Для сглаживания этого недостатка операция прямого образа будет применяться в сочетании с операцией умножения на взвешенное отношение с конечным носителем, т. е. в виде $\psi\gamma_*(\phi)$, где $\psi : D_{b_1} \times ... \times D_{b_k} \to R$ – взвешенное отношение с конечным носителем.

Следующие свойства операций прямого и обратного образа, перечисленные в Утверждении 1, легко проверяются.

Утверждение 1. *Операции прямого и обратного образа над взвешенными отношениями удовлетворяют следующим соотношениям.*

1. Операции прямого и обратного образа функториальны. Это значит, что:

если $1_A : A \to A$ *— тождественное согласование атрибутов, то операции* $(1_A)^*$ *и* $(1_A)_*$ *являются тождественными отображениями;*

если $\gamma_1 : A \to B$, $\gamma_2 : B \to C$ *— морфизмы наборов атрибутов и* $\gamma_2 \circ \gamma_1 : A \to C$ *— композиция этих морфизмов, то выполняются следующие равенства* $(\gamma_2 \circ \gamma_1)_* = (\gamma_2)_* \circ (\gamma_1)_*$ *и* $(\gamma_2 \circ \gamma_1)^* = \gamma_1^* \circ \gamma_2^*$.

2. Для любого морфизма наборов атрибутов $\gamma : A \to B$ *операция прямого образа* $\gamma_* : V(A) \to V(B)$ *является гомоморфизмом алгебр взвешенных отношений относительно операций сложения и умножения. То есть* $\gamma_*(\phi_1 + \phi_2) = \gamma_*(\phi_1) + \gamma_*(\phi_2)$ *и* $\gamma_*(\phi_1\phi_2) = \gamma_*(\phi_1)\gamma_*(\phi_2)$, *для любых взвешенных отношений* $\phi_1, \phi_2 \in V(A)$ *с набором атрибутов* $A = \{a_1, ..., a_n\}$.

3. Для любого морфизма наборов атрибутов $\gamma : A \to B$ *операция прямого образа* $\gamma^* : V(B) \to V(A)$ *перестановочна с операцией суммы, т. е.* $\gamma^*(\psi_1 + \psi_2) = \gamma^*(\psi_1) + \gamma^*(\psi_2)$ *для любых взвешенных отношений* $\psi_1, \psi_2 \in V(B)$ *с набором атрибутов B, но не сохраняет произведения, т. е., в общем случае* $\gamma^*(\psi_1\psi_2) \neq \gamma^*(\psi_1)\gamma^*(\psi_2)$.

Утверждение 2. *Для любого морфизма наборов атрибутов* $\gamma : A \to B$ *и любых взвешенных отношений* $\psi \in V(B)$ *и* $\phi \in V(A)$ *с конечными носителями и с наборами атрибутов* $B = \{b_1, ..., b_k\}$ *и* $A = \{a_1, ..., a_n\}$, *соответственно, выполняется следующее равенство:* $\gamma^*(\gamma_*(\psi)\phi) = \psi\gamma^*(\phi)$.

Доказательство Утверждения 2 проводится непосредственной проверкой равенства с использованием определений операций. Пусть $d_B : \{b_1, ..., b_k\} \to D_{b_1} \cup ... \cup D_{b_k}$ – произвольная строка из декартового произведения $D_{b_1} \times ... \times D_{b_k}$, тогда по Определениям 4 и 5 операций прямого и обратного образа взвешенного отношения, а также из свойства дистрибутивности умножения относительно сложения, из левой части равенства Утверждения 2 на этой строке получаем:

$$
\begin{aligned}
(\gamma^*(\gamma_*(\psi))\phi)(d_B) &\overset{def}{=} 0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} (\gamma_*(\psi)\phi)(d_A) = \\
&= 0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} \psi(\Gamma(d_A))\phi(d_A) = \\
&= \psi(d_B)\left(0 + \sum_{\substack{d_A \in D_{a_1} \times ... \times D_{a_n} \\ where \Gamma_\gamma(d_A) = d_B}} \phi(d_A)\right) = \\
&= \psi(d_B)(\gamma^*(\phi)(d_B)).
\end{aligned}
\tag{4}
$$

В результате значение правой части равенства Утверждения 2 располагается на строке $d_B \in D_{b_1} \times ... \times D_{b_k}$.

Определение 6. *Если* $r \subseteq D_{a_1} \times ... \times D_{a_n}$ *– отношение, в котором один из атрибутов (например,* a_n*) функционально зависит от остальных атрибутов* $\{a_1, ..., a_{n-1}\}$ *и домен* D_{a_n} *этого атрибута является числовым, то через* $\psi(r, a_n)$ *обозначается показатель* $\psi(r, a_n) : pr(r \mid a_1, ..., a_{n-1}) \to D_{a_n}$, *заданный на проекции* $pr(r \mid a_1, ..., a_{n-1})$ *отношения r на множество атрибутов* $\{a_1, ..., a_{n-1}\}$, *который каждой строке* $(d_1, ..., d_{n-1})$ *из проекции* $pr(r \mid a_1, ..., a_{n-1})$ *выдает значение* $\psi(r, a_n)(d_1, ..., d_{n-1}) = d_n$ *строки* $(d_1, ..., d_{n-1}, d_n)$ *отношения r.*

Определим операцию, обратную только что определенной.

Определение 7. *Если* $\psi : D_{a_1} \times ... \times D_{a_{n-1}} \to R$ *– взвешенное отношение с носителем* $r \subseteq D_{a_1} \times ... \times D_{a_{n-1}}$, и a_n *– имя атрибута, которого нет среди имен атрибутов отношения r, то через* (r, a_n, ψ) *обозначается отношение* $(r, a_n, \psi) \subset D_{a_1} \times ... \times D_{a_{n-1}} \times R$, *состоящее из строк вида* $(d_1, ..., d_{n-1}, \psi(d_1, ..., d_{n-1}))$ *для всех строк* $(d_1 ..., d_{n-1}) \in r$.

Следующая операция важна для составления по некоторому отношению *r* сводных отчетов-отношений с использованием стандартного набора операторов, принимающих числовые значения, и операторов над числами.

Обычно, как, например, в языке запросов SQL, используются следующие стандартные операторы:

count(| [distinct] имя_атрибута)* – вычисления числа строк или числа различных значений в выделенном атрибуте;

max(имя_атрибута), min(имя_атрибута), sum([distinct] имя_атрибута), avg([distinct] имя_атрибута) – вычисления, соответственно, минимального, максимального, суммарного или среднего значения по некоторому атрибуту, принимающему, числовое значение (для всех или только различных значений этого атрибута в зависимости от присутствия ключевого слова *distinct* в соответствующем операторе).

Для применения операции *GROUP BY* к отношению *r* с набором атрибутов *A* нужно указать подмножество атрибутов $B \subseteq A$, по которому следует произвести группирование (разбиение множества строк исходного отношения на подмножества с равными значениями в подстроках с атрибутами *B*), и некоторый оператор, из перечисленных выше, с указанием атрибута из *A*, не входящего в *B*. Результатом такой операции является взвешенное отношение-показатель с набором атрибутов *B*, функция весов строк которого вычисляется указанным оператором по атрибуту в каждой группе.

Помимо приведенных выше операций допускаются также все операции Кодда над отношениями, включая теоретико-множественные операции и операции проекции, соединения и переименования атрибутов в отношениях.

Рассмотрим далее базу данных с отношениями $r_1, ..., r_n$, с именами этих отношений $R_1, ..., R_n$ и доменами атрибутов отношений $D_1, ..., D_k, D'_1, ..., D'_s$. Домены $D_1, ..., D_k$ называются наблюдаемыми, а до-

мены $D'_1,...,D'_s$ – скрытыми. К наблюдаемым относятся также некоторые числовые домены.

Определение 8. *Скрытым отношением r называется отношение, содержащее атрибуты со скрытыми доменами. Запросом к базе данных называется правильно построенное выражение из $R_1,...,R_n$ и имен операций над отношениями, задающее композицию из определенных выше операций с этими отношениями. Ответом на запрос к базе данных в текущем состоянии $r_1,...,r_n$ называется результат вычисления выражения запроса после подстановки в запрос вместо имен отношений $R_1,...,R_n$ их состояний $r_1,...,r_n$..*

Определение 9. *Запрос к базе данных со скрытыми отношениями называется внешним, если ответом является отношение с наблюдаемыми доменами атрибутов.*

СИММЕТРИЯ СКРЫТЫХ ОТНОШЕНИЙ

Пусть, как и ранее, база данных задается отношениями $r_1,...,r_n$ с именами этих отношений $R_1,...,R_n$ и доменами атрибутов отношений $D_{t_1},...,D_{t_k}, D'_{s_1},...,D'_{s_m}$, где $T=\{t_1,...,t_k\}$ – множество всех типов наблюдаемых доменов базы данных, и $S=\{s_1,...,s_m\}$ – множество всех типов скрытых доменов базы данных. Обозначим через $D=D_1\cup...\cup D_k$ разъединенное объединение всех наблюдаемых доменов, через $D'=D'_1\cup...\cup D'_s$ – разъединенное объединение всех скрытых доменов базы данных, а через $D^{all}=D\cup D'$ – объединение всех доменов. Естественно, что каждый элемент $d\in D^{all}$ имеет свой тип $type(d)\in T\cup S$, заданный отображением $type: D^{all}\to T\cup S$, которое определяется равенством $type(d)=t_j$, если $d\in D_{t_j}$, для $j=1,...,k$ или $type(d)=s_i$, если $d\in D'_{s_i}$, для $i=1,...m$. Через D_c^{all} обозначим подмножество в D^{all}, состоящее из всех элементов, входящих в строки отношений $r_1,...,r_n$. Множество D_c^{all} естественно назвать доменом – носителем базы данных $r_1,...,r_n$.

Предположим также, что среди отношений $r_1,...,r_n$ схемы базы данных нет отношений порядков на скрытых доменах.

Определение 10. *Биективное отображение $h: D_c^{all}\to D_c^{all}$, оставляющее на месте элементы наблюдаемых доменов и переставляющее элементы скрытых доменов, но сохраняющее типы элементов, называется симметрией базы данных $r_1,...,r_n$, если для любой строки $(d_1,...,d_q)\in r_i$ любого отношения $r_i\in r_1,...,r_n$ строка $(h(d_1),...,h(d_q))$ также принадлежит этому же отношению r_i, т. е. $(h(d_1),...,h(d_q))\in r_i$.*

Так как отображение симметрии h – биективно, то оно, по определению, и инъективно, и, следовательно, различные строки отношения r_i переводятся в различные строки этого же отношения. Так как отношение r_i конечно, то из инъективности h на строках следует, что h биективно действует на строках каждого отношения $r_i\in r_1,...,r_n$.

Легко проверить, что каждая симметрия базы данных есть и симметрия отношения, которое является ответом на запрос к этой базе данных (предполагается, что в запросах не используются константы из скрытых доменов и отношения, отличные от $r_1,...,r_n$, содержащие атрибуты со скрытыми доменами). С этой целью нужно проконтролировать выполнение этого свойства для всех перечисленных операций над отношениями (так как каждый запрос – это композиция элементарных операций над отношениями).

Нетрудно доказать также, что,

если $h: D_c^{all}\to D_c^{all}$ является симметрией базы данных, то и обратное отображение $h^{-1}: D_c^{all}\to D_c^{all}$ – это симметрия;

если $h_1,h_2: D_c^{all}\to D_c^{all}$ – две симметрии базы данных, то композиция этих отображений $h_1\circ h_2: D_c^{all}\to D_c^{all}$ также является симметрией базы данных;

тождественное отображение $1_{D_c^{all}}: D_c^{all}\to D_c^{all}$, очевидно, тоже симметрия базы данных.

Определение 11. *Множество всех симметрий состояния r_1, ..., r_n базы данных вместе с этими операциями над симметриями образуют группу, которая называется группой симметрий базы данных (в данном состоянии).*

Обозначим эту группу через *H*.

Определение 12. *Если H – группа симметрии базы данных со скрытыми отношениями и $d'\in D'_j$ – элемент скрытого домена, то множество $H(d')\overset{def}{=}\{d''\in D'_j \mid d''=h(d') \text{ для } h\in H\}$ называется траекторией элемента d' относительно группы H. Элемент $d'\in D'_j$ называется неподвижным, если множество H(d') состоит из одного элемента.*

Если группа *H* симметрий базы данных со скрытыми отношениями оставляет неподвижными элементы наблюдаемых доменов (т. е. $h(d)=d$ для любых $d\in D$ и любого $h\in H$), то она оставляет неподвижными и строки отношений, которые являются ответами на внешние запросы к базе данных. Это следует из того, что такие отношения по определению не содержат атрибутов со скрытыми доменами.

Пусть *r* – отношение, которое является ответом на некоторый запрос к состоянию базы данных $r_1,...,r_n$. Допустим также, что отношение *r* имеет атрибуты со значениями в скрытых доменах. Если группа *H* – группа симметрий этой базы данных, то все элементы группы *H,* как было отмечено, являются симметриями отношения *r*, т. е. группа *H* действует на строках отношения *r*. Более того, легко видеть: если группа *H* действует на элементы скрытых доменов без неподвижных элементов, то и на строках отношения *r* она действует без неподвижных элементов. Отсюда следует важное для баз данных утверждение.

Утверждение 3. *Если у группы H симметрий базы данных нет неподвижных элементов в скрытых доменах, то по внешним запросам невозможно однозначно определить строку отношения со скрытыми доменами, являющегося ответом на запрос к этой базе данных, т. е.: если отношение r – ответ на запрос к базе данных и строка* $\bar{d}$ *принадлежит r, то множество строк* $H(\bar{d}) = \{h(\bar{d}) \mid h \in H\}$, *полученных действием симметрий из группы H на строку* $\bar{d}$, *является подмножеством в r, т.е.* $H(\bar{d}) \subseteq r$.

Таким образом, если отношение *r* является ответом на запрос к базе данных и содержит атрибуты со скрытыми доменами, то число строк в $H(\bar{d})$ для любой строки $\bar{d} \in r$ не меньше, чем число $k = \min_{d' \in D'} |H(d')|$ – минимальное число элементов в траекториях элементов скрытых доменов. Отсюда следует, что при построении баз данных со скрытыми отношениями для увеличения степени неопределенности распознавания скрытых элементов отношений необходимо увеличивать число *k,* которое в публикациях по безопасности баз данных и социальных сетей получило название *k*-анонимность (*k-anonymity*) [5, 6].

Теперь рассмотрим, в какой степени множества элементов наблюдаемых доменов определяют множества элементов скрытых отношений в базе данных со скрытыми доменами атрибутов.

Пусть, как и ранее, $r_1,...,r_n$ – состояние базы данных со скрытыми отношениями и $M \subseteq D$ – подмножество во множестве *D* элементов наблюдаемых доменов. Заменим все элементы наблюдаемых доменов, не входящих в подмножество *M*, в строках отношений $r_1,...,r_n$ на новый элемент *NULL.* Получим отношения $r_1^M,...,r_n^M$, имеющие атрибуты, значения которых лежат во множестве $D^M = M \cup \{NULL\} \cup D'$, где *D'*, как и ранее, – объединение всех скрытых доменов.

Пусть H^M – группа симметрии состояния $r_1^M,...,r_n^M$ базы данных с доменом D^M и скрытым доменом *D'* (см. Определение 11), т. е. H^M – множество всех биекций $h: D^M \to D^M$, сохраняющих типы элементов и оставляющих элементы множества $D^M \cup \{NULL\}$ на месте, которые переводят отношения $r_1^M,...,r_n^M$ в себя. В частном случае, когда *M=D,* группа $H^D = H$.

Соответственно, $H^M(d')$, где $d' \in D' \subset D^M$, – это траектория элемента *d'* скрытого домена относительно группы H^M, действующей на множестве D^M. Аналогично $k = \min_{d' \in D'} |H(d')|$ – минимальному числу элементов в траекториях элементов скрытых доменов относительно группы *H*, определим $k^M = \min_{d' \in D'} |H^M(d')|$ – минимальное число элементов в траекториях элементов скрытых доменов относительно группы H^M.

Число k^M – это минимальный размер подмножества в скрытых доменах, определяемого подмножеством наблюдаемых элементов $M \subseteq D$ в состоянии базы данных $r_1,...,r_n$.

БЛИЗКИЕ РАБОТЫ

Проблема, как организовать доступ пользователям к базам данных для задач статистического анализа, не раскрывая персональных сведений, которые могут храниться в этих базах данных, актуальна и сегодня. Первая статья по *k*-анонимности появилась в 1998 г. [5], более подробно этот подход изложен в [6]. Затем поток публикаций на тему *k*-анонимности стал значительно расти. Связь *k*-анонимности с симметрией данных в социальных сетях обсуждалась в [7]. Современное состояние исследований, связанных с анонимностью в базах данных и сетях, *k*-анонимностью, а также с некоторыми проблемами конфиденциальности, возникающими в результате анализа Больших Данных, показано в [8].

ЗАКЛЮЧЕНИЕ

Объектом исследования в настоящей работе являются реляционные базы данных, в которых некоторые атрибуты принимают значения в скрытых доменах. Предполагается, что пользователи таких баз данных могут обращаться к ним с произвольными запросами, но значения атрибутов в скрытых доменах не выдаются в качестве ответов на запросы к таким базам данных.

Нами рассмотрены взвешенные отношения в реляционных базах данных (показатели) и операции над такими отношениями, а также приведены некоторые соотношения между этими операциями.

Основная проблема, которая возникает при работе с базами данных, содержащими скрытые отношения, – это как обеспечить невозможность вывода значений атрибутов в скрытых доменах этой базы по любым допустимым запросам к ней. Мы предлагаем подход к этой проблеме, основанный на применении понятия группы симметрий состояния базы данных. Этот подход позволил найти условия, при которых элементы некоторых множеств в скрытых доменах невозможно различить с помощью внешних запросов к этой базе данных.

СПИСОК ЛИТЕРАТУРЫ

1. Jaideep Srivastava, Hung Q. Ngo. Statistical databases // In Wiley Encyclopedia of Electrical and Electronics Engineering. – Hoboken, NY: Wiley, 1999. – URL: https://onlinelibrary.wiley.com/doi/book/10.1002/047134608X
2. Privacy in Statistical Databases “UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018” // Proceedings, Springer, LNCS, 2018. – Vol. 11126. – URL: https://www.springer.com/gp/book/9783319997704

3. Мейер Д. Теория реляционных баз данных. – М. : Мир, 1987. – 608 с.
4. Бениаминов Е.М. Алгебраические методы в теории баз данных и представлении знаний. – М.: Научный мир, 2003. – 184 с.
5. Samarati P., Sweeney L. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression // Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International, 1998. – URL: http://www.csl.sri.com/papers/sritr-98-04/
6. Sweeney L. *K*-anonymity: a model for protecting privacy // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. – 2002. – Vol. 10(5). – P. 557-570.
7. Wu W., Xiao Y., Wang W., He Z., Wang Z. *K*-symmetry model for identity anonymization in social networks // Advances in Database Technology – EDBT 2010 Proceedings of the 13th International Conference on Extending Database Technology Lausanne, Switzerland, March 22–26, 2010. – P. 111–122.
8. Salas J., Domingo-Ferrer J. Some Basics on Privacy Techniques, Anonymization and their Big Data Challenges // Math. Comput. Sci. – 2018, – Vol.12(3). – P. 263–274.

Материал поступил в редакцию 22.04.19

Сведения об авторе

БЕНИАМИНОВ Евгений Михайлович – доктор физико-математических наук, профессор, заведующий кафедрой Российского государственного гуманитарного университета (РГГУ), Москва.
e-mail: ebeniamin@yandex.ru