УДК 004.056:519.115.1 DOI: 10.17587/it.27.314-321

Е. А. Саксонов, главный специалист, д-р техн. наук, проф., e-mail: saksmiem@mail.ru,
Московский технический университет связи и информатики, Москва

Способ обезличивания персональных данных

Предложен метод обезличивания больших объемов персональных данных. Метод сохраняет структуру и семантику данных, позволяет повысить безопасность обезличенных данных и обрабатывать персональные данные без предварительного обезличивания. Разработана математическая модель метода. Получены оценки защищенности обезличенных данных.

Ключевые слова: *персональные данные, обезличивание персональных данных, преобразование таблиц персональных данных, защищенность обезличенных персональных данных*

Введение

Обработка персональных данных (ПД) применяется для решения различных задач по обслуживанию клиентов, формированию статистических данных, проведению социальных и экономических исследований. Для этого обычно используются большие объемы различных типов ПД, хранение которых должно обеспечивать их защиту от неправомерного использования.

Одним из распространенных подходов к хранению ПД является их обезличивание, представляющее ПД в таком виде, когда без использования дополнительной информации невозможно определить принадлежность ПД конкретному субъекту ПД [1]. Обезличивание обеспечивает возможность многократной обработки ПД для решения различных задач.

Современные информационные технологии позволяют проводить обезличивание ПД различными способами, сохраняя их полноту, структуру и семантическую целостность для проведения обработки.

Способы различаются по применяемым алгоритмам, трудоемкости реализации, возможностям обработки запросов, стойкости к различным видам атак, направленных на деобезличивание [2—6].

Однако наличие большого числа разнообразных способов обезличивания приводит к сложностям в согласовании и обработке данных, обезличенных различными способами.

В данной статье предлагается способ обезличивания ПД, обеспечивающий возможность проведения процедуры деобезличивания, изменения параметров способа и проведения обработки ПД, внесения изменений в ПД без предварительного деобезличивания, повышающий стойкость к атакам, направленным на деобезличивание.

Рассматриваемый способ обеспечивает конфиденциальность, доступность и целостность обезличенных данных. Способ основан на приведении исходной таблицы ПД к такому виду, когда выделение ПД конкретных физических лиц становится практически невозможным из-за очень большой трудоемкости при неизвестных параметрах обезличивания. Способ реализует один из предлагаемых Роскомнадзором методов обезличивания и соответствует установленным требованиям [7, 8].

1. Описание способа обезличивания ПД

В предлагаемом способе обезличивания ПД используется исходная таблица ПД, которая может быть построена из имеющегося множества структурированных ПД, предварительно собранных в удаленных хранилищах в виде множеств файлов, таблиц баз данных. Для построения исходной таблицы требуется привести ПД каждого субъекта к единой структуре, которая будет представлена в строках исходной таблицы.

Таким образом, строка исходной таблицы представляет ПД одного субъекта, а каждый столбец (атрибут) содержит множество однотипных данных, соответствующих элементам структуры, по всем субъектам, занесенным в таблицу.

Допускается отсутствие данных о субъектах в отдельных элементах атрибутов, в этом случае элементом атрибута является специальный набор символов.

Способ основан на применении алгоритма многоэтапного преобразования множества данных к каждому атрибуту исходной таблицы ПД. На каждом этапе множество данных атрибута разбивается по заданному правилу на блоки, которые подвергаются операции перестановки, после чего этап повторяется. Число этапов, правила разбиения на блоки и перестановки блоков на этапе могут меняться для каждого атрибута.

Правила разбиения на блоки, параметры перестановки блоков и число этапов для каждого атрибута составляют ключ преобразования.

В результате преобразования меняется порядок данных во множестве элементов каждого атрибута и, соответственно, содержание строк в таблице, но сохраняется семантика каждого элемента. Предложенный способ позволяет рассеивать элементы атрибута в пределах всего множества элементов, устранить связи между структурными элементами каждой строки исходной таблицы ПД, что значительно усложняет возможность деперсонализации при неизвестном ключе преобразования, даже при наличии множества известных ПД, находящихся в исходной таблице.

Известны способы, применяющие разбиения и перестановки (перемешивание) [9—12]. Общими недостатками этих способов являются:

- сохранение состава подмножеств (блоков) ПД, получаемых после разбиения исходного множества, что приводит к слабому рассеиванию элементов в пределах множества всех элементов атрибута;
- циклические сдвиги, существенно ограничивающие число вариантов перестановок и не увеличивающие рассеивание;
- применение, как правило, одного этапа в преобразовании.

Все это значительно ослабляет защищенность этих способов.

2. Математическая модель преобразования

Пусть исходная таблица ПД $\mathbf{T}_0(N,M)$ содержит N записей (строк) и M атрибутов (столбцов). Все строки и атрибуты имеют уникальные номера от 1 до N и от 1 до M, соответственно.

Далее будем рассматривать атрибут номер j исходной таблицы, который содержит упорядоченное занумерованное множество из N элементов (данных о субъекте) — $C_{j0} = \{c_{j0}(m)\}$. Здесь $c_{j0}(m)$ — уникальный номер элемента, имеющего порядковый номер m (стоящего на месте m) во множестве C_{j0}. В исходной таблице $\mathbf{T}_0(N,M)$ порядковый номер элемента атрибута равен номеру строки, в которой этот элемент находится, и, в данном случае, уникальный номер элемента равен его порядковому номеру в атрибуте исходной таблицы: $c_{j0}(m) = m$, $(1 < c_{j0}(m) \leqslant N$; $1 \leqslant m \leqslant N)$. Таким образом, уникальные номера элементов во всех атрибутах, относящихся к одному субъекту (находящихся в одной строке исходной таблицы) совпадают. Уникальные номера являются постоянными и однозначно определяют элементы атрибутов ПД.

Преобразования приводят к несовпадению уникальных и порядковых номеров элементов.

Преобразование множества данных атрибута номер j ($j = 1, 2, ..., M$) происходит в несколько этапов, число которых R_j $(1 \leqslant R_j < \infty)$.

Каждый этап номер r ($r = 1, 2, ..., R_j$) содержит два шага.

На первом шаге происходит разбиение множества данных атрибута на непересекающиеся блоки данных. Число блоков равно Z_{jr}, где $1 < Z_{jr} \leqslant N$. Каждый блок номер i содержит $k_{jr}(i)$ последовательно занумерованных элементов с номерами от $\sum_{n=1}^{i-1} k_{jr}(n) + 1$ до $\sum_{n=1}^{i} k_{jr}(n)$ ($i = 1, 2, ..., Z_{jr}$; $0 < k_{jr}(i) < N$). Кроме того, $\sum_{n=1}^{Z_{jr}} k_{jr}(n) = N$, т. е. блоки разбиения содержат все элементы атрибута.

Число блоков Z_{jr} атрибута j на этапе r и вектор числа элементов в каждом блоке $\mathbf{k}_{jr} = (k_{jr}(1), k_{jr}(2), ..., k_{jr}(Z_{jr}))$ являются уникальными для каждого атрибута и каждого этапа.

На втором шаге этапа проводится перестановка блоков атрибута, задаваемая матрицей перестановки $\mathbf{X}_{jr} = \left\| x_{ij} \right\|$ $(i = 1, 2; j = 1, 2, ..., Z_{jr})$. Здесь x_{1n} _ порядковый номер блока при разбиении $(1 \leqslant x_{1n} \leqslant Z_{jr}$ и, как правило, $x_{1n} = n)$, а x_{2n} — номер блока разбиения, стоящего на месте номер n после перестановки $(1 \leqslant x_{2n} \leqslant Z_{jr})$. В результате получается преобразованное множество атрибута C^*_{jr}, где r — номер этапа ($r = 1, 2, ..., R_j$). После преобразования на этапе r изменяется расположение элементов атрибута и определяется $c_{jr}(m)$ — уникальный номер

элемента, имеющего после преобразования порядковый номер m ($m = 1, 2, ..., N$).

Исходным множеством для преобразования на первом этапе является множество данных атрибута номер j исходной таблицы — C_{j0}, исходными множествами для преобразований на этапах 2, 3, ..., R_j являются множества C^*_{j1}, C^*_{j2}, ..., $C^*_{j(R_j-1)}$. Множество $C^*_{jR_j}$ есть окончательный результат преобразования множества данных атрибута номер j. После окончания проведения преобразований всех атрибутов получается таблица обезличенных ПД — $\mathbf{T}^*(N, M)$.

Преобразование исходной таблицы ПД на каждом этапе изменяет порядковые номера элементов всех атрибутов, разрушая связи между элементами строки исходной таблицы.

Вычислим новый порядковый номер n элемента атрибута номер j после выполнения этапа r, если по окончании предыдущего этапа $(r-1)$ элемент имел порядковый номер d. Разбиение и перестановка задаются вектором $\mathbf{k}_{jr} = (k_{jr}(1), k_{jr}(2), ..., k_{jr}(Z_{jr}))$ и матрицей $\mathbf{X}_{jr} = \|x_{ij}\|$.

Определим сначала h — номер блока, куда после разбиения входит элемент номер d. Этот номер h вычисляется из условия

$$\sum_{i=1}^{h-1} k_{jr}(x_{1i}) \leqslant d \leqslant \sum_{i=1}^{h} k_{jr}(x_{1i}). \qquad (1)$$

Номер элемента внутри блока h вычисляется по формуле

$$b = d - \sum_{i=1}^{h-1} k_{jr}(x_{1i}). \qquad (2)$$

После перестановки блок номер h будет иметь порядковый номер z, который вычисляется из условия

$$x_{2z} = h. \qquad (3)$$

Номер элемента после перестановки вычисляется по формуле

$$n = b + \sum_{i=1}^{z-1} k_{jr}(x_{2i}). \qquad (4)$$

В результате определяется новый порядковый номер элемента после проведения этапа. Этот номер становится номером элемента перед следующим этапом преобразования. Уникальный номер элемента не меняется и всегда остается равным порядковому номеру этого элемента в атрибуте исходной таблицы $\mathbf{T}_0(N, M)$ и, таким образом, $c_{j(r-1)}(d) = c_{jr}(n)$.

При проведении деобезличивания для определенного субъекта ПД задается значение элемента атрибута этого субъекта, устанавливается его порядковый номер в соответствующем атрибуте таблицы $\mathbf{T}^*(N, M)$, после чего вычисляется значение уникального номера этого элемента. Затем разыскиваются все элементы других атрибутов, имеющие этот же уникальный номер.

Так, если после этапа r элемент атрибута j имел порядковый номер n, то вычислим порядковый номер этого элемента после этапа $(r-1)$. Используя матрицу перестановки $\mathbf{X}_{jr} = \|x_{ij}\|$, можно вычислить номер h блока разбиения, куда входит элемент с порядковым номером d. Этот номер h вычисляется из условия

$$\sum_{i=1}^{h-1} k_{jr}(x_{2i}) \leqslant n \leqslant \sum_{i=1}^{h} k_{jr}(x_{2i}). \qquad (5)$$

Далее номер элемента внутри блока с номером h вычисляется по формуле

$$b = n - \sum_{i=1}^{h-1} k_{jr}(x_{2i}). \qquad (6)$$

До перестановки блок имеет порядковый номер h, и порядковый номер элемента равен

$$d = b + \sum_{i=1}^{h-1} k_{jr}(x_{1i}). \qquad (7)$$

Таким образом, показано, как вычислить порядковые номера элементов после преобразований и как вычислить порядковые номера для деобезличивания (в обратном порядке).

Параметры преобразования атрибута j на этапе r задаются множеством $\mathbf{K}_{jr} = \{Z_{jr}, \mathbf{k}_{jr}, \mathbf{X}_{jr}\}$, которое является ключом преобразования атрибута на этапе. Вектор $\mathbf{K}_j = (R_j, \mathbf{K}_{j1}, \mathbf{K}_{j2}, ..., \mathbf{K}_{jRj})$ задает множество ключей для всех этапов преобразования атрибута j ($j = 1, 2, ..., M$).

Множество ключей для всех атрибутов и всех этапов преобразования исходной таблицы $\mathbf{K}(\mathbf{T}_0(N, M)) = \{\mathbf{K}_j\}$ ($j = 1, 2, ..., M$) называется ключом преобразования исходной таблицы.

Эффективность преобразования определяется рассеиванием элементов атрибута по всему множеству элементов. Рассеивание позволяет избежать устойчивых групп элементов, имеющихся в исходном множестве, наличие которых упрощает проведение атак на деобезличивание при отсутствии ключей преобразования.

Для оценки качества проводимых преобразований введем меру рассеивания элементов атрибута j на этапе r:

$$R_{jr} = \frac{1}{(N-1)} \sum_{i=1}^{N-1} \left| c_{jr}(i+1) - c_{jr}(i) \right|. \qquad (8)$$

Здесь величина $\left|c_{jr(i+1)} - c_{jri}\right|$ — это "расстояние" между соседними элементами с номерами $(i + 1)$ и i $(i = 1, 2, ..., N - 1)$ в таблице $\mathbf{T}_0(N, M)$ после проведения этапа r. Мера рассеивания для всей таблицы $\mathbf{T}^*(N, M)$ вычисляется по формуле

$$R(T^*) = \frac{1}{M}\sum_{j=1}^{M} R_{jR_j}. \qquad (9)$$

Предлагаемая мера дает возможность количественно оценить усредненное значение "расстояний" между соседними элементами исходной таблицы после деобезличивания.

Отметим, что циклическая перестановка незначительно влияет на значение меры рассеивания элементов атрибута.

Пример 1.

Пусть задана исходная таблица $\mathbf{T}_0(20,1)$, где атрибут номер 1 ($M = 1$) имеет 20 элементов с номерами от 1 до 20 ($N = 20$).

Рассмотрим первый этап преобразования. Число битов $Z_{11} = 5$. Разбиение задается вектором $\mathbf{k}_{11} = (k_{11}(1) = 5, k_{11}(2) = 3, k_{11}(3) = 4, k_{11}(4) = 2, k_{11}(5) = 6)$.

Матрица $\mathbf{X}_{11} = \|x_{ij}\|$, $(i = 1, 2; j = 1, 2, ..., 5)$ имеет вид: $\mathbf{X}_{11} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix}$.

До разбиения исходное множество элементов на этапе 0 C_{10} имеет вид (для простоты значения элементов совпадают с их уникальными номерами):

$$C_{10} = \{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12\ 13\ 14\ 15\ 16\ 17\ 18\ 19\ 20\}.$$

После разбиения на первом шаге первого этапа получены следующие блоки:

$$\{(1\ 2\ 3\ 4\ 5)\ (6\ 7\ 8)\ (9\ 10\ 11\ 12)\ (13\ 14)\ (15\ 16\ 17\ 18\ 19\ 20)\}.$$

После перестановки получено новое множество элементов C_{11}^* (показана перестановка блоков), где изменились порядковые номера элементов, но остались неизменными уникальные номера:

$$C_{11}^* = \{(9\ 10\ 11\ 12)\ (15\ 16\ 17\ 18\ 19\ 20)\ (1\ 2\ 3\ 4\ 5)\ (6\ 7\ 8)\ (13\ 14)\}.$$

Вычислим новый номер n элемента, который до преобразования имел, например, начальный номер $d = 9$. Используя соотношение (1), получим: $8 = \sum_{i=1}^{2} k_{11}(x_{1i}) \leqslant 9 \leqslant \sum_{i=1}^{3} k_{11}(x_{1i}) = 12$, откуда $h = 3$ и, используя (2), получим: $b = 9 - \sum_{i=1}^{2} k_{11}(x_{1i}) = 1$. Далее $z = 1$, поскольку из соотношения (3) будем иметь: $x_{21} = h = 3$.

Новый номер элемента, имевшего до преобразования номер 9, вычисляем, используя формулу (4):

$$n = 1 + \left(\sum_{i=1}^{0} k_{11}(x_{1i}) = 0\right) = 1.$$

Обратная процедура: пусть после этапа 1 порядковый номер элемента $n = 1$. Тогда из соотношения (5) получим

$$0 = \sum_{i=1}^{0} k_{11}(x_{2i}) \leqslant 1 \leqslant \sum_{i=1}^{1} k_{11}(x_{2i}),$$

Таблица 1

Преобразования атрибута

Этап r	Вектор разбиения	Матрица перестановки	Множество значений (уникальных номеров) элементов после преобразования на этапе $r - C_{1r}^*$
0			1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
1	(5,3,4,2,6)	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 5 & 1 & 2 & 4 \end{pmatrix}$	9 10 11 12 15 16 17 18 19 20 1 2 3 4 5 6 7 8 13 14
2	(3,4,3,5,2,3)	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 3 & 4 & 2 & 5 & 1 \end{pmatrix}$	8 13 14 18 19 20 1 2 3 4 5 12 15 16 17 6 7 9 10 11
3	(4,6,5,5)	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$	5 12 15 16 17 8 13 14 18 6 7 9 10 11 19 20 1 2 3 4
4	(4,6,5,5)	$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}$	7 9 10 11 19 5 12 15 16 20 1 2 3 4 17 8 13 14 18 6
5	(7,3,3,4,3)	$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 3 & 5 & 2 \end{pmatrix}$	4 17 8 13 7 9 10 11 19 5 12 1 2 3 14 18 6 15 16 20

откуда $x_{21} = 3$ и, следовательно, $h = 3$. Используя формулу (6), получим:

$$b = n - \left(\sum_{i=1}^{0} k_{jr}(x_{2i}) = 0 \right) = 1.$$

Далее из формулы (7) имеем

$$d = b + \sum_{i=1}^{h-1} k_{11}(x_{1i}) = 9.$$

В табл. 1 приведены результаты преобразований таблицы $\mathbf{T}_0(20,1)$ на пяти этапах. Исходное множество соответствует этапу номер 0.

Из табл. 1 видно, как меняется расстановка элементов атрибута в результате преобразований.

Значения меры рассеивания, вычисленные с применением формулы (5), по результатам преобразований, приведенных в табл. 1, равны: $R_{10} = 1$; $R_{11} = 2{,}26$; $R_{12} = 2{,}79$; $R_{13} = 4{,}16$; $R_{14} = 5{,}63$; $R_{15} = 6{,}31$, откуда видно, что рассеивание элементов увеличивается с каждым этапом.

Однако понятно, что мера рассеивания имеет верхнюю границу, после достижения которой, преобразования становятся неэффективными.

Применяя приведенные формулы для каждого элемента исходного множества и для всех этапов, можно определить новые номера элементов каждого атрибута номер j после окончания преобразования, т. е. после проведения R_j этапов ($j = 1,2,...,M$).

3. Оценка защищенности преобразования

Защищенность таблицы $\mathbf{T}^*(N, M)$ от атак, направленных на деобезличивание, при отсутствии данных о ключе преобразования, можно оценить числом возможных вариантов преобразования.

Сначала вычислим число вариантов разбиения на блоки. Если имеются N элементов, то между ними будет $(N - 1)$ промежутков, которые имеют порядковые номера от 1 до $(N - 1)$. Число разбиений N элементов на k блоков равно числу возможных сочетаний из $(N - 1)$ по $(k - 1)$, при этом наборы не будут повторяться. Таким образом, число возможных разбиений N элементов на k блоков равно

$$U_0(k, N) = C_{N-1}^{k-1} = \frac{(N-1)!}{(k-1)!(N-k)!}. \qquad (10)$$

Общее число возможных разбиений N элементов атрибута с учетом (10) равно

$$U_0(N) = \sum_{k=1}^{N-1} \frac{(N-1)!}{(k-1)!(N-k)!} = \sum_{k=1}^{N-1} \left[\prod_{i=1}^{k-1} \left(\frac{N}{(k-i)} - 1 \right) \right]. \qquad (11)$$

Нижняя оценка величины $U_0(N)$ составляет

$$U_0(N) = \sum_{k=1}^{N-1} \left[\prod_{i=1}^{k-1} \left(\frac{N}{(k-i)} - 1 \right) \right] > \sum_{k=1}^{N-1} \left(\frac{N}{k} \right)^{k-1}.$$

Далее на этапе каждое разбиение из k блоков подвергается перестановке, и возможное число перестановок для этого разбиения равно $k!$.

Следовательно, общее число возможных вариантов преобразования атрибута, содержащего N элементов, на одном этапе, с учетом (11) равно

$$U(N) = \sum_{k=1}^{N-1} (U_0(N)k!) = \sum_{k=1}^{N-1} \frac{(N-1)!\,k!}{(k-1)!(N-k)!} = \sum_{k=1}^{N-1} \frac{(N-1)!\,k}{(N-k)!} = \sum_{k=1}^{N-1} \left[k \prod_{i=1}^{k-1} (N-i) \right]. \qquad (12)$$

Если полное преобразование одного атрибута номер j проводится за R_j этапов, то число вариантов полного преобразования этого атрибута вычисляется с использованием выражения (12) по формуле

$$U_j(N) = \prod_{m=1}^{R_j} U(N) = \prod_{m=1}^{R_j} \left[\sum_{k=1}^{N-1} \left(k \prod_{i=1}^{k-1} (N-i) \right) \right]. \qquad (13)$$

Для исходной таблицы $\mathbf{T}_0(N, M)$, содержащей M атрибутов, число вариантов преобразований вычисляется с учетом соотношения (13) по формуле

$$T(N, M) = \prod_{j=1}^{M} U_j(N). \qquad (14)$$

Формулы (10)—(14) показывают, что число возможных вариантов преобразования таблицы $\mathbf{T}_0(N, M)$ очень велико (особенно с учетом того, что реальные значения N лежат в диапазоне от 10^3 до 10^8). Например, при $N = 100$, $M = 10$, $R_j = 1$, $Z_{j1} = 10$ ($j = 1,2, ..., 10$) получим $U_0(10,100) = 10^5 \cdot 17310308 \geqslant 10^{12}$. При фиксированном $k = 10$ $U(100) \geqslant 10^{13}$, откуда $T(100,10) \geqslant 10^{130}$.

Для увеличения числа вариантов при небольших значениях N можно увеличить число M за счет разбиения элементов одного атрибута на несколько частей, каждая из которых будет входить в состав своего атрибута (создание новых атрибутов). Например, можно телефон, название улицы в адресе, фамилию субъекта

разделить на несколько частей и рассматривать их как отдельные структурные единицы, а их множества рассматривать как атрибуты.

Пример 2.

В табл. 2 первом столбце представлены уникальные номера элементов для каждого атрибута.

Ключи преобразований атрибутов табл. 2:

- **Атрибут 1 (Фамилия)**

$$\mathbf{K}_1 = (3, (\mathbf{K}_{11} = \{3, (4,6,4), \mathbf{X}_{11} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}\}),$$

$$(\mathbf{K}_{12} = \{3, (6,5.3), \mathbf{X}_{12} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}\}),$$

$$(\mathbf{K}_{13} = \{2, (5,9),\} \mathbf{X}_{13} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\}));$$

- **Атрибут 2 (Имя)**

$$\mathbf{K}_2 = (2, (\mathbf{K}_{21} = \{4, (3,5,2,4), \mathbf{X}_{21} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{pmatrix}\}),$$

$$(\mathbf{K}_{22} = \{3, (7,3,4), \mathbf{X}_{22} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}\}));$$

- **Атрибут 3 (Отчество)**

$$\mathbf{K}_3 = (3, (\mathbf{K}_{31} = \{2, (6,8), \mathbf{X}_{31} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\}),$$

$$(\mathbf{K}_{32} = \{4, (5,3,2,4), \mathbf{X}_{32} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 1 & 4 & 2 \end{pmatrix}\}),$$

$$(\mathbf{K}_{33} = \{4, (2,3,4,5), \mathbf{X}_{33} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 3 & 1 & 2 \end{pmatrix}\}));$$

- **Атрибут 4 (Место рождения)**

$$\mathbf{K}_4 = (3, (\mathbf{K}_{41} = \{4, (4,3,2,5), \mathbf{X}_{41} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 1 & 3 \end{pmatrix}\}),$$

$$(\mathbf{K}_{42} = \{3, (3,6,5), \mathbf{X}_{42} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}\}),$$

$$(\mathbf{K}_{43} = \{3, (6,1,7), \mathbf{X}_{43} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}\}));$$

- **Атрибут 5 (Год рождения)**

$$\mathbf{K}_5 = (2, (\mathbf{K}_{51} = \{2,(5,9), \mathbf{X}_{51} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\}),$$

$$(\mathbf{K}_{52} = \{3, (3,5,6), \mathbf{X}_{52} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}\})).$$

В табл. 3 в первом столбце представлены порядковые номера элементов атрибутов после преобразования, уникальные номера нужно вычислять по исходной таблице (см. табл. 2).

Таблица 2

Исходная таблица T_0 (14,5)

№	Фамилия	Имя	Отчество	Место рождения	Год рождения
1	Иванов	Петр	Сергеевич	Москва	1940
2	Петров	Иван	Петрович	Рязань	1954
3	Иванов	Сергей	Андреевич	Москва	1947
4	Сидоров	Петр	Иванович	Тверь	2000
5	Николаев	Сергей	Васильевич	Киев	2002
6	Сергеев	Евгений	Петрович	Москва	1987
7	Митин	Алексей	Дмитриевич	Калуга	1991
8	Петров	Игорь	Николаевич	Ржев	1958
9	Власов	Анатолий	Алексеевич	Мурманск	1954
10	Куракин	Сергей	Александрович	Казань	1936
11	Лебедев	Иван	Николаевич	Уфа	2005
12	Крылов	Иван	Андреевич	Минск	2003
13	Рублев	Антон	Семенович	Москва	1950
14	Пушкин	Александр	Сергеевич	Курск	1928

Таблица 3

Таблица обезличенных данных T^* (14,5), полученная после преобразования таблицы T_0 (14,5)

№	Фамилия	Имя	Отчество	Место рождения	Год рождения
1	Рублев	Сергей	Васильевич	Москва	1954
2	Пушкин	Евгений	Петрович	Калуга	1936
3	Иванов	Алексей	Андреевич	Рязань	2005
4	Николаев	Игорь	Семенович	Москва	2003
5	Сергеев	Петр	Сергеевич	Тверь	1950
6	Митин	Иван	Александрович	Ржев	1987
7	Петров	Сергей	Николаевич	Мурманск	1991
8	Власов	Иван	Андреевич	Казань	1958
9	Куракин	Иван	Иванович	Уфа	1928
10	Петров	Антон	Сергеевич	Минск	1940
11	Иванов	Александр	Петрович	Москва	1954
12	Сидоров	Анатолий	Дмитривич	Курск	1947
13	Лебедев	Сергей	Николаевич	Москва	2000
14	Крылов	Петр	Алексеевич	Киев	2002

Заключение

Предлагаемый способ обладает следующими достоинствами:

- исходная таблица ПД может быть получена из множества различных хранилищ ПД, представленных в структурированной форме;
- число атрибутов исходной таблицы может превышать число структурных единиц в начальной форме представления ПД, что позволит увеличить защищенность деобезличенных данных;
- для осуществления преобразования исходной таблицы используются уникальные параметры — ключи преобразования, которые хранятся в защищенной форме отдельно от места хранения персональных данных;
- при обезличивании не затрагиваются сами ПД, что обеспечивает сохранение их целостности;
- имеется возможность проводить поиск и обработку ПД по таблице обезличенных данных без предварительного деобезличивания всей таблицы;
- при проведении процедуры деобезличивания достаточно задавать значения одного или нескольких атрибутов, по которым восстанавливаются записи исходной таблицы, имеющие заданные значения атрибутов.
- высокая защищенность от атак, направленных на деобезличивание, при неизвестных ключах преобразования, обеспечивается очень большим числом возможных вариантов преобразований исходной таблицы;
- имеется возможность динамического изменения ключей преобразования на каждом этапе и для каждого атрибута;
- существует возможность хранения атрибутов таблицы обезличенных данных в различных базах данных и на различных компьютерах (серверах);
- есть возможность добавления новых атрибутов в исходную таблицу, и их преобразования без деобезличивания таблицы обезличенных данных;
- преобразование состоит из повторяющихся простых однотипных процедур, что существенно упрощает его программную реализацию.

Для практической реализации предложенного способа целесообразно использовать следующие рекомендации:

- проводить формирование исходной таблицы ПД на отдельном сервере с передачей таблицы на сервер обезличивания по его запросу. Это позволит совместить во времени различные операции;
- при формировании исходной таблицы ПД провести нумерацию строк для формирования уникальных номеров элементов;
- для занесения новых ПД использовать либо предварительно подготовленную избыточную исходную таблицу ПД, содержащую пустые строки, либо создавать несколько исходных таблиц меньшей размерности добавляя в них новые данные.

Большая размерность пространства ключей может быть значительно сокращена, если задавать функциональные связи между ключами. Например, при заданном числе блоков Z_{jr} при разбиении можно задать число элементов в блоке формулой $k_{jr}(i) = f_{jr}(i, Z_{jr})$. Для матриц перестановок также могут быть заданы аналитические зависимости (формальные правила) для их составления.

Предложенный способ обобщает известные способы, основанные на перемешивании данных, увеличивая при этом рассеивание элементов таблицы и защищенность обезличенных данных.

Список литературы

1. **Федеральный** закон от 27 июля 2006 г. N 152-ФЗ "О персональных данных".

2. **Столбов А. П.** Обезличивание персональных данных в здравоохранении // Врач и информационные технологии. 2017. № 3. С. 76—91.

3. **Рябко С. Д.** Об обезличивании персональных данных // Информационная безопасность. 2009. № 5. URL: www.itsec.ru/articles2/bypub/insec-5-2009.

4. **Sweeney L.** k-anonymity: a model for protecting privacy // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems. 2002. Vol. 10, N. 5. P. 557—570.

5. **Трифонова Ю. В., Жаринов Р. Ф.** Возможности обезличивания персональных данных в системах, использующих реляционные базы данных // ДокладыТУСУР. 2014. № 2 (32). С. 188—194.

6. **Бондаренко К. О., Козлов В. А.** Универсальный быстродействующий алгоритм процедур обезличивания данных // Изв. ЮФУ. Технические науки. 2015. № 11 (172). С. 130—142.

7. **Приказ** Роскомнадзора от 5 сентября 2013 года N 996 "Об утверждении требований и методов по обезличиванию персональных данных".

8. **Методические** рекомендации по применению приказа Роскомнадзора от 5 сентября 2013 г. № 996 "Об утверждении требований и методов по обезличиванию персональных данных". Утверждены руководителем Роскомнадзора 13.12.2013.

9. **Саксонов Е. А., Шередин Р. В.** Процедура обезличивания персональных данных // Наука и образование: элек-

тронное научно-техническое издание. 2011. № 3. Эл № ФС 77 — 30569. Государственная регистрация № 0421100025.

10. **Куракин А. С.** Алгоритм деперсонализации персональных данных // Научно-технический вестник информационных технологий, механики и оптики. СПб: СПбГУ ИТМО, 2012. № 6. С. 130—135.

11. **Карпова И. П.** О реализации метода обезличивания персональных данных // Вестник компьютерных и информационных технологий. 2013. № 6. С. 56—60.

12. **Бондарец Е. А.** Модификация алгоритма перемешивания персональных данных // Ученые заметки ТОГУ: электронное научное издание. 2015. Т. 6, № 2. С. 282—288.

E. A. Saksonov, Chief Specialist, Professor, e-mail: saksmiem@mail.ru,
Moscow Technical University of Communications and Informatics, Moscow, Russian Federation

Method of Depersonalization of Personal Data

A method of depersonalization of large amounts of personal data is proposed. The method preserves the structure and semantics of data, allows you to increase the security of depersonalized data and process personal data without prior depersonalization. A mathematical model of the method is developed. Estimates of security depersonalized data are obtained.

Keywords: *personal data, depersonalization of personal data, transformation of personal data tables, security of depersonalized personal data*

DOI: 10.17587/it.27.314-321

References

1. **Federal** Law of July 27, 2006 No. 152-FZ "About personal data".

2. **Stolbov A. P.** Depersonalization of personal data in healthcare, *Vrach I informacionnie tehnologii,* 2017, no. 3, pp. 76—91 (in Russiun).

3. **Ryabko S. D.** On depersonalization of personal data, *Informacionnaya bezopasnost*, 2009, no. 5, available at: www.itsec.ru/articles2/bypub/insec-5-2009 (in Russiun).

4. **Sweeney L.** k-anonymity: a model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 2002, vol.10, no. 5, pp. 557—570.

5. **Trifonova Yu. V., Zarinov R. F.** Opportunities for depersonalization of personal data in systems using relational databases, *Doklady TUSUR*, 2014, no. 2 (32), pp. 188—194 (in Russiun).

6. **Bondarenko K. O., Kozlov V. A.** Universal fast-acting algorithm for the depersonalization of personal data, *Izv. YuFU. Tehnicheskie nauki*, 2015, no. 11 (172), pp. 130—142 (in Russiun).

7. **Order** of Roskomnadzor dated September 5, 2013 N 996 "On approval of requirements and methods for anonymization of personal data."

8. **Recommendations** for the application of the Roskomnadzor order from 5.09.2013 № 996 "On approval of requirements and methods for depersonalization of personal data". Utverzdeny rukovoditelem Roskomnadzora 13.12.2013.

9. **Saksonov E. A., Sheredin R. V.** Procedure for depersonalization of personal data, *Nauka I obrasovanie: electronnoe nauchno-tehnicheskoe izdanie*, 2011, no. 3, EL № FS 77-30569 (in Russiun).

10. **Kurakin A. S.** The algorithm of the anonymisation of personal data*, Nauchno-tehnicheskii vestnik informationnych tehnologij, mehaniki I optiki*, 2012, no. 6, pp. 130—135 (in Russiun).

11. **Karpova I. P.** About the implementation of the method of depersonalization of personal data, *Vestnik Komputernych i Informacionnych Tehnologij*, 2013, no. 6, pp. 56—60 (in Russiun).

12. **Bondarec E. A.** Modification of the algorithm for mixing personal data, *Uchenie zametki TOGU: electronnoe nauchnoe izdanie*, 2015, vol. 6, no. 2, pp. 282—288 (in Russiun).