ПРОГРАММИРОВАНИЕ, 2023, № 5, с. 19–34

АНАЛИЗ ДАННЫХ

УДК 004.622

ПРИМЕНЕНИЕ ИМИТАЦИОННОГО КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ К ЗАДАЧЕ ОБЕЗЛИЧИВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ. МОДЕЛЬ И АЛГОРИТМ ОБЕЗЛИЧИВАНИЯ МЕТОДОМ СИНТЕЗА

© 2023 г. А. В. Борисов[a,*] (ORCID: 0000-0002-3124-2147), А. В. Босов[a,] (ORCID: 0000-0001-7163-341X), А. В. Иванов[a,***] (ORCID: 0000-0001-7811-7645)**

[a]*Федеральный исследовательский центр "Информатика и управление" РАН, 119333, Москва, ул. Вавилова, д. 44, кор. 2, Россия*

**e-mail: aborisov@ipiran.ru*

***e-mail: avbosov@ipiran.ru*

****e-mail: aivanov@ipiran.ru*

Поступила в редакцию 14.02.2023 г.
После доработки 12.03.2023 г.
Принята к публикации 14.05.2023 г.

Представлена вторая часть исследования, посвященного тематике автоматизированного обезличивания персональных данных. Обзор и анализ перспектив для исследований, выполненный ранее, здесь дополнен практическим результатом. Предложена модель процесса обезличивания, сводящая задачу обеспечения анонимности персональных данных к манипулированию выборками разнотипных случайных элементов. Соответственно, ключевой идеей преобразования данных для обеспечения их анонимности при условии сохранения полезности является применение метода синтеза, т.е. полной замены всех необезличенных данных синтетическими значениями. В предлагаемой модели выделен набор типов элементов, для которых предложены шаблоны синтеза. Совокупность шаблонов составляет алгоритм обезличивания методом синтеза. Методически каждый шаблон опирается на типовой статистический инструмент – частотные оценки вероятностей, ядерные оценки плотностей Розенблатта–Парзена, статистические средние и ковариации. Применение алгоритма иллюстрируется простым примером из области гражданских авиаперевозок.

DOI: 10.31857/S0132347423050023, **EDN:** ZXUVBM

1. ВВЕДЕНИЕ

Проблематика персональных данных в контексте их автоматизированной обработки в последние годы привлекала внимание исследователей, обладающих самыми разными компетенциями. Помимо очевидной содержательности для специалистов в областях права и информационной безопасности обезличивание персональных данных (ПД) и их последующая безопасная обработка стали источником постановок задач для специалистов в области анализа данных, распределенных вычислений, разработчиков баз данных, программистов и математиков. Причем разнообразие исследуемых вопросов оказалось очень широким, поэтому их содержательному обзору была посвящена отдельная работа [1]. Настоящая статья представляет вторую часть исследования, начатого в [1], и посвящена практическим результатам.

Анализ методов и алгоритмов обезличивания привел к двум принципиальным выводам. Во-первых, методов, обеспечивающих гарантированную анонимность обезличенным ПД, нет. Исключением может быть только метод синтеза, подразумевающий, что вместо исходного набора необезличенных ПД пользователю предоставляется набор синтетических данных заданного объема, которые связаны с исходным набором только совпадением некоторого набора характеристик, например, средних величин или размахов выборок.

Так, метод синтеза хорошо показал себя в задаче обезличивания строковых данных [2], особенности которых требуют для обезличивания больших усилий. Второй вывод – это реалистичность поддержки высокого уровня анонимности при обезличивании только числовых данных. Тексты (строки), потоковые, медийные и бинарные данные более-менее универсальных решений предложить не позволяют. Но несмотря на

очевидную привлекательность метода синтеза, в исследованиях ему уделяется довольно мало внимания. Вместе с тем даже самые простые статистические методы позволяют предлагать довольно универсальные решения для большинства возможных числовых атрибутов ПД. Такому алгоритму посвящена данная статья.

Ориентируясь на типовые статистические инструменты, семантику решаемой задаче и применяемым терминам дает, видимо самая распространенная концепция обезличивания *K*-анонимность [3, 4]. Но в отличие от многих известных реализаций этой концепции [5–13] предлагаемый подход ставит целью формирование результирующего обезличенного набора только из синтетических данных. Кроме того, сочетание статистики и *K*-анонимности дает возможность определить и вычислить все сопровождающие процесс обезличивания показатели, такие как уровни анонимности и полезности [14–18], а также используемые пользователями-аналитиками корреляции атрибутов ПД.

Статья организована следующим образом. В следующем разделе приведено описание модели процесса обезличивания – преобразования набора необезличенных ПД в набор обезличенных данных заданного объема. В третьем разделе приведен алгоритм обезличивания. Реализованный авторами прототип автоматизированной системы, модельные данные и результаты экспериментов обсуждаются в разделе 4. В заключение подведен итог исследования и кратко сформулированы возможные перспективы.

2. МОДЕЛЬ ОБЕЗЛИЧИВАНИЯ

2.1. Основные понятия

Предлагаемая для описания ПД модель использует несколько дополнительных понятий и терминов, которые объясняются в данном разделе. При моделировании процесса обезличивания следует исходить из того, что любые данные, которые отнесены законодательством к персональным, критически чувствительны к анонимности, поскольку позволяют идентифицировать субъекта ПД. Набор ПД может содержать множество различных свойств субъекта – данных определенного типа, называемых *атрибутами*. Предполагается, что эта совокупность атрибутов идентифицирует субъекта, так что их восстановление полностью или частично является деобезличиванием. При этом вклад разных атрибутов в идентификацию различен. Часть атрибутов может нести явный идентификационный характер, если атрибуты сами по себе являются прямыми указателями на конкретного субъекта ПД (типичный пример – паспортные данные). Такие атрибуты называются *персональными идентификаторами* и даже с самыми мягкими требованиями к анонимности при обезличивании должны исключаться. Аналогично следует поступать и с любыми другими чувствительными данными, которые не являются персональными идентификаторами, не могут быть использованы для прямой идентификации субъекта, но носят очевидный критический характер в связи с угрозой нарушения анонимности. Такие данные называют *чувствительными* (типичные примеры – пароли, коды, ключевые слова и т.п.) и они также должны исключаться при обезличивании.

После исключения из исходного необезличенного набора ПД чувствительной информации в нем остаются атрибуты, по отдельности или в совокупности представляющие угрозу нарушения анонимности. Предполагая, что степень угрозы от этих данных несопоставимо меньше, чем от персональных идентификаторов, все равно будем учитывать такую угрозу, называя такие атрибуты *квазиидентификаторами*. Можно считать, что квазиидентификаторы по отдельности являются гораздо менее чувствительными по отношению к анонимности, но влияние на анонимность больших совокупностей квазиидентификаторов преуменьшать нельзя [14, 16]. При этом именно атрибуты-квазиидентификаторы будут представлять значительный интерес для конечного потребителя обезличенных данных, т.е. будут подвергаться дальнейшей обработке, которая может привести к деобезличиванию. Соответственно, принципиально важными являются вопросы чувствительности квазиидентификаторов к анонимности, выбора применяемого алгоритма обезличивания и оценки защищенности от возможного нарушения анонимности обезличенных данных. Именно, на этих аспектах сосредоточено внимание в описанной далее модели и алгоритме обезличивания.

Определенный вызов представляет вопрос, а есть ли в необезличенном наборе данных те атрибуты, которые не являются квазиидентификаторами. Формальный ответ на этот вопрос – нет, потому что при наличии неограниченного ресурса для проведения идентификации субъекта ПД может быть использован любой атрибут. Да, знание малозначительного факта о субъекте не позволит его идентифицировать, но если этих фактов окажется очень много или к факту будут добавлены другие факты из дополнительных источников, то и малозначительный факт сыграет свою роль. Однако, чтобы эти рассуждения не привели к переоценке потенциальной угрозы, предлагается все-таки выделять в необезличенных ПД *нечувствительную* информацию, т.е. те данные, которые сколь-либо существенной угрозы анонимности не несут и могут быть использованы для идентификации субъекта лишь при наличии неограниченного ресурса.

В [1] упомянута небольшая часть резонансных фактов реализации угрозы нарушения анонимности обезличенных ПД. И эти факты, и многие другие не включали традиционного “бытового” понимания идентификации – точного установления личности. Исследователи справедливо считают, что возможность интерпретации выявленных ПД для установления личности доказана, а механизм этой интерпретации непринципиален. Например, определение атрибутов адреса и возраста субъекта, не означает его идентификации, нет фамильно-именной группы, тем более, номера паспорта или кредитной карты. Но субъекта по этим данным установить несложно, а нужные для этого методы и их законность в рамках научного исследования не обсуждаются. Важным для анализа защищенности от угрозы нарушения анонимности обезличенных данных является именно возможность выборки из множества обезличенных данных атрибутов одного или группы субъектов ПД. Таким образом, формулируя определение и формальное описание понятия деобезличивания, следует использовать именно такую интерпретацию идентификации. При этом для подлежащего обезличиванию обезличенного набора данных должна быть определена процедура оценки показателя защищенности ПД, который называется *уровнем анонимности*. Соответственно, следующий шаг – это применение к необезличенному набору такой процедуры обезличивания, чтобы уровень анонимности был велик настолько, насколько этого требует законодательство или/и владелец информации.

Понимая, что под *обезличиванием ПД понимаются действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность ПД конкретному субъекту ПД,* для предлагаемой математической модели требуется дать формальное определение “обратному” процессу деобезличивания, применимому для формирования объективной (числовой) оценки уровня анонимности обезличенных данных.

Пусть есть набор данных НД, содержащий необезличенные ПД вида

$$\text{НД} = \left\{\langle \text{ID-субъекта},\ \mathrm{A}_1,\ldots,\mathrm{A}_n\rangle_i\right\}_{i=1}^{N},$$

т.е. множество N однотипных кортежей, каждый из которых содержит данные по одному субъекту ПД, и эти данные представляют собой набор разнотипных атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_n$. Пусть далее в результате обезличивания сформирован набор данных

$$\text{ОД} = \left\{\langle \widehat{\text{ID}}\text{-субъекта},\ \hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_n\rangle_i\right\}_{i=1}^{N}.$$

Деобезличиванием данных ОД называется выборка по условиям, применяемым к атрибутам $\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_n$, из набора ОД любого подмножества $\left\{\langle \widehat{\text{ID}}\text{-субъекта}\rangle_j\right\}_{j=1}^{K}$ с целью точного или приближенного установления отвечающего этим идентификаторам набора атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_n$. Отметим, что в определении не фигурирует идентификатор ID-субъекта необезличенного набора, что отвечает сделанному замечанию о конкретном механизме интерпретации с целью идентификации субъекта ПД.

Следует отметить, что любая обработка обезличенных данных, сохраняющая идентификаторы субъектов, является деобезличиванием. Не приводит к деобезличиванию данных в смысле данного определения только такая обработка набора ОД, в результатах которой не сохраняются связи обработанных данных и исходных идентификаторов, т.е. обработка, включающая только операции агрегирования. Поскольку владелец НД и оператор, выполняющий обезличивание, не могут гарантированно контролировать содержание выполняемых над обезличенными данными операций, то формировать объективную оценку уровня защищенности обезличенных данных следует, исходя из потенциальной возможности любых выборок. Отсюда получаем следующее определение.

Путь при любой выборке, т.е. применении любых условий α к атрибутам данных ОД

$$\alpha = \left\{\hat{\mathrm{A}}_i \in \alpha_1,\ldots,\hat{\mathrm{A}}_n \in \alpha_n\right\},$$

где α_i – множество $\hat{\mathrm{A}}_i$ возможных значений $\hat{\mathrm{A}}_i$, из ОД формируется результирующий набор $\{\langle \widehat{\text{ID}}\text{-субъекта}\rangle_j\}_{j=1}^{K_\alpha}$ и существует число K такое, что $K \le K_\alpha$ для всех α. Такое число K называется *гарантированной оценкой уровня K-анонимности*.

Данное понятие следует концепции K-анонимности, но не означает, что предлагаемый далее алгоритм обезличивания направлен на обеспечение заданного показателя K-анонимности. Однако такая величина является хорошей объективной оценкой уровня анонимности. Величина K определяет минимальное число идентификаторов $\widehat{\text{ID}}$-субъекта, которое можно получить, максимально уточнив значения атрибутов $\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_n$. То есть какие бы действия ни выполнялись по уточнению атрибутов набора ОД, результирующий набор всегда будет содержать не менее K идентификаторов, так что, если целью обработки ОД является идентификация субъекта по признакам $\mathrm{A}_1,\ldots,\mathrm{A}_n$, то вероятность утери анонимности составит менее, чем $\frac{1}{K}$. И это верхняя гарантированная оценка, на реальную потерю анонимности влияет в сторону понижения и искажения

$\hat{A}_1,\ldots,\hat{A}_n$, и ограниченность в формировании условий $\alpha_1,\ldots,\alpha_n$, а главное – остающиеся "за скобками" действия по установлению связи между выявленными идентификаторами и верными атрибутами $A_1,\ldots,A_n$ и реальным субъектом ПД.

Данное понятие гарантированной оценки уровня анонимности K не зависит от размера N набора данных ОД. Так что в дополнение к нему имеет смысл рассматривать относительную характеристику уровня анонимности набора ОД. Именно, *относительным уровнем K-анонимности* называется величина $k = \frac{K}{N}100\%$. Получается, что 100%-обезличенные данные содержат все одинаковые атрибуты $\hat{A}_1,\ldots,\hat{A}_n$, что возможно только в том случае, если в ОД попадают только агрегаты по всему набору НД (это "идеальная" в смысле анонимности ситуация). При $k = \frac{100\%}{N}$ получается, что существует хотя бы один набор атрибутов $\hat{A}_1,\ldots,\hat{A}_n$, которому отвечает ровно один $\widehat{ID}$-субъекта (это наихудшая ситуация в смысле K-анонимности).

Величины K и k дают возможность формально определить следующие понятия. Набор данных ОД называется:

– *не допускающим деобезличивание*, если $k = 100\%$,

– *частично допускающим деобезличивание*, если $K > 1$,

– *допускающим идентификацию*, если $K = 1$.

Если при обезличивании формируется не допускающий деобезличивание набор ОД, то это означает фактическое отсутствие в ОД поля $\widehat{ID}$-субъекта, что возможно только в том случае, если ОД содержит только агрегированные данные (средняя температура, минимальная зарплата, максимальный тариф и т.п.). В таком случае выборки данных по заданным для деобезличивания требованиям достаточно для получения защищенного (с высоким уровнем анонимности) ОД и не требуются дополнительные алгоритмы обезличивания.

Любой набор ОД с полями $\widehat{ID}$-субъекта будет частично допускающим деобезличивание (иначе у всех записей должны совпадать значения всех атрибутов, что делает такой набор бесполезным). Если уровень анонимности, оцениваемый числами K и k для исходного набора данных НД, отобранного из имеющихся ПД по заданным требованиям к атрибутам, окажется неудовлетворительным, то требуется применение алгоритма обезличивания такого, чтобы уровень анонимности стал удовлетворительным. Если при этом результирующий ОД потеряет свойства применимости, то от предоставления ОД придется отказаться по причине невозможности обеспечить анонимность.

ОД, допускающий идентификацию – это такой набор данных, из которого можно выбрать ровно один $\widehat{ID}$-субъекта, задав некоторый набор условий $\alpha_1,\ldots,\alpha_n$. Такая выборка называется *идентификацией* и интерпретируется как реализация угрозы нарушения анонимности.

Величины K и k довольно грубо, но интуитивно понятно характеризуют набор данных с точки зрения угрозы нарушения анонимности. Грубость оценки как раз и обеспечивает ее гарантированный характер, т.е. ориентацию на наихудший случай из возможных. При этом нетрудно представить ситуацию, когда $K = 1$, но реальная угроза нарушения анонимности отсутствует. Эта ситуация может быть связана с любым атрибутом, имеющим очень большое множество возможных значений, вплоть до того, что все значения атрибута уникальны. Таким образом, указание любого из имеющихся значений $\langle\hat{A}_j\rangle$ такого атрибута $\hat{A}_j$ в качестве условия α_j будет идентификацией в смысле данного выше определения. Реальная же угроза нарушения анонимности может и не иметь места, т.к. на идентифицируемость знание $\langle\hat{A}_j\rangle$ может никак не влиять. Можно формализовать понятие такого невлияния. Для этого предлагается использовать следующее расширение понятия K-анонимности.

Если возможно для всех значений $\langle\hat{A}_j\rangle$ атрибута $\hat{A}_j$ задать ε-окрестность, т.е. интервал $\Delta_j = (\langle\hat{A}_j\rangle - \varepsilon_j, \langle\hat{A}_j\rangle + \varepsilon_j)$ с помощью малого числа ε_j такого, что семантические содержания значения $\langle\hat{A}_j\rangle$ и интервала Δ_j, т.е. равенств $\hat{A}_j = \langle\hat{A}_j\rangle$ и выражения $\hat{A}_j \in \Delta_j$, можно считать неотличимыми, то для определения оценки уровня анонимности согласно данному выше определению следует использовать условие

$$\alpha = \left\{\hat{A}_1 \in \alpha_1,\ldots,\hat{A}_j \in \Delta_j,\ldots,\hat{A}_n \in \alpha_n\right\},$$

для всех атрибутов $\hat{A}_j$, размер области значения которых существенно превосходит объем N имеющихся данных. Соответствующее число K_ε называется *гарантированной оценкой уровня ε-анонимности*. Основным претендентом на такую интерпретацию являются данные, область значений которых имеет непрерывную структуру. Сложность реализации этого понятия уровня анонимности будет состоять в том, что определение малости числа ε_j всегда будет носить некоторый субъективный характер.

Данные формальные определения трех оценок уровня анонимности (K, k, K_ε) потенциально да-

ют эксперту инструмент для качественной оценки уровня анонимности для конкретных наборов данных. Заметим, что нет формальных оснований для ограничения применимости этих характеристик только к обезличенным наборам, поэтому на практике оценки (K, k, K_ε) могут рассчитываться для любых наборов ПД как обезличенных, так и необезличенных (исходных).

При практическом применении предложенных характеристик в распоряжении эксперта окажется набор данных объемом N кортежей, перечень n атрибутов $A_1,\ldots,A_n$ и их типов. Для оценки уровня анонимности эксперт должен:

• исключить из атрибутов $A_1,\ldots,A_n$ персональные идентификаторы и чувствительную информацию;

• исключить из атрибутов $A_1,\ldots,A_n$ несущественные и нечувствительные атрибуты, не влияющие на оценку уровня анонимности;

• сформировать из оставшихся атрибутов перечень $A_1,\ldots,A_m$, отнеся к нему атрибуты, принимающие только числовые непрерывные значения;

• задав подходящую ε-окрестность для каждого атрибута, вычислить оценку K_ε (конкретные реализации расчета могут использовать разные варианты определения подходящего интервала Δ_j, например можно положить $\varepsilon_j = 0.05|\langle \hat{A}_j \rangle|$ или $\varepsilon_j = 0.05(x_{MAX} - x_{MIN})$, где x_{MAX}, x_{MIN} – максимальное и минимальное значение среди всех имеющихся значений $\langle \hat{A}_j \rangle$, что соответствует экспертной оценке 10% длины интервала Δ_j;

• сформировать из оставшихся атрибутов перечень $A_{m+1},\ldots,A_l$, отнеся к нему атрибуты, принимающие только числовые дискретные значения (сюда же следует относить и словарные типы) и вычислить оценку K;

• сравнить K_ε и K и, если обе величины принимают сопоставимые значения, вычислить оценку относительного уровня анонимности как $k = \dfrac{\min(K_\varepsilon, K)}{N} 100\%$.

2.2. Структура наборов необезличенных и обезличенных данных

Семантическое и структурное многообразие необезличенных ПД существенно затрудняют унификацию понятий и функций обработки данных, в т.ч. преобразований с целью обезличивания. Здесь представлен вариант унифицированного представления наборов данных НД и ОД, удобный для формального определения служебных структур и понятий, используемых далее в алгоритме обезличивания. Фактические представления необезличенных данных могут существенно отличаться, что предполагает наличие промежуточного функционального элемента, содержащего средства автоматизации, от которых потребуется выполнение преобразований:

$$\text{НД} \to \text{НД}(y) \to \text{ОД}(y) \to \text{ОД},$$

где НД – исходный набор ПД, ОД – целевой набор обезличенных ПД, НД(y), ОД(y) – исходный и целевой наборы данных в унифицированном представлении. Заметим, что через $\text{НД}(y) \to \text{ОД}(y)$ обозначено обезличивание данных, алгоритм которого является целью работы. Следуя введенной модели $\text{НД} = \{\langle \text{ID-персоны}, A_1,\ldots,A_n\rangle_i\}_{i=1}^N$, уточним ее. Будем предполагать, что структурное представление НД отвечает реляционной модели данных, т.е. обезличиваемые данные представлены в виде набора таблиц, таблицы обладают первичными ключами и используют вторичные ключи для установления связей, т.е. исходные данные реализованы типовой реляционной базой данных.

На рис. 1 показан простой, но довольно типичный пример структуры набора НД:

• таблица “Персональные идентификаторы” содержит все чувствительные атрибуты,

• набор “простых” атрибутов, представленных непосредственно своим значением, размещен в таблице “Атрибуты-значения” и связан ключом “Паспорт”,

• набор словарных атрибутов размещен в таблице “Атрибуты-словарные” и связан ключом “E-mail”,

• словарные значения определены своими ключами и данными в таблицах словарей “Словарь-3”, “Словарь-4”.

Элемент (строку, кортеж) НД(y), обозначенный выше $\langle$ID-субъекта, $A_1,\ldots,A_n\rangle$ составляют:

• ключ ID-субъекта, заменяющий всю совокупность атрибутов из “Персональные идентификаторы”,

• $A_1,\ldots,A_m$ – перечень “простых” атрибутов-значений,

• $A_{m+1},\ldots,A_n$ – перечень значений “словарных” атрибутов.

Если НД допускает атрибуты с множественными значениями, то такой НД корректно (полностью) может преобразовываться в НД(y), если

• m – это число атрибутов-значений, имеющихся в НД у субъекта, имеющего максимальное число атрибутов-значений, и допускается наличие в НД(y) пустых значений атрибутов $A_1,\ldots,A_m$,

• $(n-m)$ – это число словарных атрибутов, имеющихся в НД у субъекта, имеющего максимальное число словарных атрибутов, и допуска-

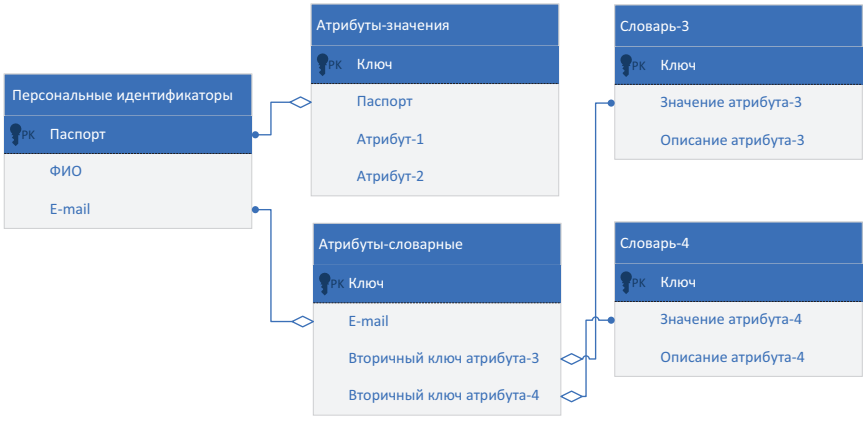

Рис. 1. Пример типового необезличенного набора.

ется наличие в НД(y) пустых значений атрибутов $A_{m+1},\ldots,A_n$.

Тогда набор данных вида $\{\langle$ID-субъекта, $A_1,\ldots,A_n\rangle_i\}_{i=1}^N$ называется *унифицированным набором необезличенных данных* НД(y).

Заметим, что такое определение является семантически зависимым, т.к. число атрибутов, определяющее структуру НД(y), зависит непосредственно от данных, содержащихся в НД. Это полезное свойство, т.к. дает возможность, во-первых, лучше понимать и контролировать потенциальные действия по деобезличиванию, во-вторых, удобнее оценивать уровень угрозы нарушения анонимности в ОД(y). Очевидным недостатком такого подхода к унификации НД являются потенциально значительные вычислительные трудности (требования к применяемым компьютерным ресурсам, необходимым при обезличивании больших объемов НД). НД(y), соответствующий примеру на рис. 1, может иметь вид

$$\left\{\begin{array}{c}\langle\text{ID-субъекта}\,(\text{GUID}),\ \text{Атрибут-1},\\ \text{Атрибут-2},\text{Значение атрибута-3},\\ \text{Значение атрибута-4}\rangle_i\end{array}\right\}_{i=1}^{N}.$$

Для ПД в унифицированном виде НД(y) уже начат процесс обезличивания, т.к. набор НД лишен персональных идентификаторов и дополнен уникальным идентификатором субъекта (все чувствительные атрибуты заменены не имеющим содержания ключом GUID – Globally Unique Identifier, гарантирующим при каждом новом формировании НД(y) новое значение, идентифицирующее запись НД). Причем для целей обезличивания подходящей будет любая из существующих реализаций GUID. Завершить первый этап обезличивания может служебная процедура, позволяющая в последующем выполнить деобезличивание. Для этого требуется набор данных

$$\text{Контракт} = \{\langle\text{ID-субъекта, ID-субъекта}\,(\text{GUID})\rangle_i\}_{i=1}^N.$$

Требуется или нет формировать набор Контракт, хранить его и иметь возможность деобезличивания в дальнейшем – зависит как от цели обезличивания, так и от статуса заказчика. Можно увидеть диаметрально разные ситуации, когда данные обезличиваются в интересах аналитики торговой сети (тогда ни о каком деобезличивании речи быть не может) и когда данные обезличиваются в интересах силового ведомства, выполняющего, к примеру, анализ нелегальных ресурсов в сети. Наконец, необходимо понимать, что сам факт наличия набора Контракт является источником угрозы нарушения анонимности.

Следующий шаг обезличивания НД также является формальным, но более содержательным. С НД(y) должны быть выполнены операции:

- при наличии заявленной цели обезличивания определены не отвечающие ей атрибуты ПД и соответствующие поля устранены как несущественные данные,
- определены квазиидентификационные данные.

Таким образом, в НД(y) останутся поля, которые определены в качестве квазиидентификаторов, и поля, которые признаны нечувствительными данными.

И наконец из структуры НД(y), к которой далее применяется алгоритм обезличивания, удаляются нечувствительные атрибуты. Если такие атрибуты есть, то они возвращаются неизменными уже в набор ОД(y).

2.3. Типы атрибутов

Исходя из того, что обезличивается набор данных вида $\{\langle$ID-субъекта, $A_1,\ldots,A_n\rangle_i\}_{i=1}^N$, содержа-

щий только квазиидентификаторы, остается определить только допустимые типы для атрибутов.

2.3.1. Числовое дискретное значение. Атрибут A_i набора НД(y) принимает числовое дискретное значение, если множество его потенциально возможных значений (т.е. не только в данном НД(y), но и в других возможных наборах) конечно или счетно, т.е. множество возможных значений можно перечислить и перенумеровать: $\langle A_i \rangle \in \{x_1, \ldots, x_l, \ldots\}$. Рассматривая величины x_l, надо понимать такие их свойства, как фактические значения, разницу между соседними величинами, равномерность или точки сгущения и т.п. Но кроме этих свойств принципиально важным является фактическое распределение значений $x_1, \ldots, x_l$, ... в данном конкретном наборе НД(y), т.е. то, как часто/редко те или иные значения x_l появляются в наборе НД(y).

2.3.2. Числовое непрерывное значение. Атрибут A_i набора НД(y) принимает числовое непрерывное значение, если множество его потенциально возможных значений (т.е. не только в данном НД(y), но и в других возможных наборах) задается диапазоном (интервалом) числовой оси: $\langle A_i \rangle \in \left(\left[x_{l_i}, x_{r_i}\right]\right)$. Здесь круглые скобки (,) или квадратные [,] используются при необходимости исключить или, наоборот включить, левую границу x_{l_i} или правую границу x_{r_i} в область возможных значений, при этом x_{l_i} может принимать значение $-\infty$, а x_{r_i} может принимать значение $+\infty$. Для дальнейших преобразований, выполняемых для обезличивания, принципиально важным является фактическое распределение значений такого атрибута A_i в данном конкретном наборе НД(y).

2.3.3. Словарное значение. По своему содержанию атрибут со словарным значением близок атрибуту с числовым дискретным значением, потому что для такого атрибута возможные значения можно пересчитать и пронумеровать. Более того, форма номера в словаре такого атрибута неважна, поэтому всегда можно считать, что значения такого атрибута $\langle A_i \rangle \in \{1, \ldots, n_i\}$, т.е. выбираются из натурального ряда, нумерующего словарь. Однако здесь есть важное отличие. Для числовых дискретных атрибутов принципиально важно конкретное числовое значение атрибута, т.е. всех величин x_l, для словарного это не имеет значения, поэтому они и нумеруются с помощью натурального ряда. Соответственно, выполнять обезличивание таких атрибутов придется способом, имеющим некоторые отличия от числовых дискретных атрибутов.

2.3.4. Дата/время. Темпоральные значения могут как нести самостоятельное содержание, например, дату рождения, время покупки, так и дополнять фактографию, представленную другими атрибутами, например, время изменения географического положения. Можно считать, что эти значения имеют вид день-месяц-год-час-минута-секунда, таким образом, тип дата/время тождественен вещественному числовому типу (типичное машинное представление для этого типа). Ясно, однако, что применять к нему те же методы, что для числового непрерывного значения неверно. Преобразования обезличивания здесь должны учитывать его интервальный характер (дней, месяцев, лет, часов, минут, секунд), а не распределение величин – значений соответствующего атрибута в конкретном НД(y).

2.3.5. Потоковые данные. Собственно значения этого типа представляют собой упорядоченные группы числовых величин (или структур, составленных из числовых величин), например, последовательности географических положений. Но особенностью этого типа является то, что его данные поступают постоянно, последовательно одно за другим в течение некоторого времени. Такого рода информация характерна для трекинговых систем, чатов, соцсетей и т.д. Обезличивание таких данных возможно в рамках предложенной модели, если допускается использовать вертикальное разбиение (обезличивание путем ограничения данных по каждому субъекту) или горизонтальное разбиение (обезличивание путем ограничения объема данных для субъектов).

2.3.6. Текстовые и бинарные данные. Текстовые данные – наиболее сложные с точки зрения обезличивания, унифицированных решений для них реализовать практически невозможно, любое решение будет исходить из конкретной семантики конкретного атрибута. Фиксированный формат строки решает этот вопрос, но, только в том случае, если текстовое значение атрибута гарантированно удовлетворяет некоторому формату. Примерами таких атрибутов могут быть адрес электронной почты, телефон, т.п. Также сюда можно отнести разного рода адресную информацию, например, адрес регистрации, проживания. Но даже в этих случаях обезличивание может выполняться только ограничением объема предоставляемой информации: для электронной почты – только домен, для телефона – только код и т.п. Если текст имеет произвольный формат, то он, по-видимому, не может быть гарантировано обезличен. Аналогична ситуация с бинарными данными. Если это квазиидентификаторы (а в модели НД(y) остались только такие атрибуты), то эти атрибуты должны исключаться.

Гарантированный способ обезличить как текстовые, так и бинарные данные – это замена их на синтетические. Успешные примеры такого рода хорошо известны – это примеры синтеза текста, синтеза изображений лиц и сцен. Сложность здесь состоит в том, что каждая задача такого рода

является существенным исследовательским вызовом и возможно даже представляет самостоятельный научный интерес. Соответственно, универсальной рекомендации по обезличиванию методом синтеза таких данных нет.

3. АЛГОРИТМ ОБЕЗЛИЧИВАНИЯ

3.1. Подготовка к обезличиванию

Описание основных преобразований предлагаемого алгоритма дается в предположении, что выполнены перечисленные в предыдущем разделе статьи подготовительные шаги, так что в обрабатываемом наборе НД(y) остались только атрибуты-квазидентификаторы перечисленных типов (2.3.1–2.3.5), в т.ч. при необходимости выполнена декомпозиция данных, и создан набор Контракт.

Сам алгоритм обезличивания состоит в последовательном применении *шаблонов обезличивания*, определенных далее для каждого типа атрибутов.

Упомянем здесь еще и финальный этап обезличивания – переход от унифицированного представления наборов данных, введенного для удобства описания алгоритма обезличивания, т.е. преобразование ОД(y) $\to$ ОД. Формальный смысл этого преобразования – вернуть исходную структуру набора, в частности, восстановить вторичные ключи, выделив и заполнив словари и другие сущности, характерные для исходной модели данных НД. Собственно, к обезличиванию этот переход прямого отношения не имеет, поэтому далее к нему возвращаться не будем.

3.2. Шаблоны обезличиввания

3.2.1. Числовое дискретное значение. Пусть из имеющегося набора атрибутов $A_1,\dots,A_n$ атрибуты $A_1,\dots,A_m$, т.е. для простоты первые m штук, принимают числовые дискретные значения, каждый из атрибутов в своей области допустимых значений. Будем считать, что каждый элемент набора НД(y), составленный из атрибутов $A_1,\dots,A_m$, является реализацией случайного вектора $\xi = \mathrm{col}(\xi_1,\dots,\xi_m)$, ξ_j – дискретная случайная величина со своей областью значений $\{x_j^1,\dots,x_j^{N(j)}\}$, $j = 1,\dots, m$. Выберем из НД(y) набор $\{\langle A_1,\dots,A_m\rangle_l\}_{l=1}^{L(m)}$ неповторяющихся реализаций ξ, упорядоченный последовательно по значениям $\langle A_1\rangle,\dots,\langle A_m\rangle$. Дополним полученный набор порядковым номером l, принимающим значения $0,\dots,L(m)-1$. Величина $l(m)$ зависит от набора данных НД(y) и от числа m участвующих атрибутов. Заметим, что эта величина не является произведением $N(j)$, $j = 1,\dots,m$, потому что не все комбинации возможных значений $x_1^{l_1},\dots,x_m^{l_m}, l_1 = 1,..,N(1),\dots, l_m = 1,\dots, N(m)$ могут встретиться в НД(y) Далее будем обозначать эту величину просто L. Полученный набор $\{\langle l, A_1,\dots,A_m\rangle\}_{l=0}^{L-1}$ можно интерпретировать как множество значений $\{x_l\}_{l=0}^{L-1}$ дискретной случайной величины ξ, дополненный номерами значений. Для расчета обезличенного значения вместо ξ применяется *шаблон синтеза числовых дискретных значений*. Он состоит в восстановлении (оценке) по имеющимся данным распределения исходного набора значений атрибутов $A_1,\dots,A_m$ в наборе НД(y) и формировании в ОД(y) полностью синтетических значений. Для их моделирования набор $\{\langle A_1,\dots,A_m\rangle_l\}_{l=0}^{L-1}$ из НД(y) дополняется сначала частотами $\{\langle p_l, A_1,\dots,A_m\rangle_l\}_{l=0}^{L-1}$, где p_l – частота появления значения x_l атрибутов $A_1,\dots,A_m$ в наборе НД(y), т.е. $p_l = \dfrac{Num(x_l)}{N}$, где $Num(x_l)$ – число появлений реализаций x_l в наборе НД(y). Эти величины могут быть вычислены вместе с формированием множества $\{A_1,\dots,A_{ml}\}_{l=0}^{L-1}$ неповторяющихся реализаций ξ. Записи для набора ОД(y) формируются путем моделирования псевдослучайных величин $\xi = \mathrm{col}(\hat{A}_1,\dots,\hat{A}_m)$ в соответствии с дискретным распределением $\{(p_l, x_l)\}_{l=0}^{L-1}$ для каждой записи $\{\langle \text{ID-субъекта}(\mathrm{GUID}), A_1,\dots,A_m\rangle_i\}_{i=1}^{N}$ в наборе НД(y).

3.2.2. Числовое непрерывное значение. Пусть из имеющегося набора атрибутов $A_1,\dots,A_n$ атрибуты $A_1,\dots,A_m$, т.е. для простоты первые m штук, принимают числовые непрерывные значения, каждый из атрибутов в своей области допустимых значений. Будем считать, что каждый элемент набора НД(y), составленный из атрибутов $A_1,\dots,A_m$, является реализацией случайного вектора $\xi = \mathrm{col}(\xi_1,\dots,\xi_m)$, ξ_j – непрерывная случайная величина со своей областью значений $\xi_j \in [x_{l_j}, x_{r_j}]$, где x_{l_j} – минимальное значение атрибута A_j в наборе НД(y), x_{r_j} – максимальное. Для каждой ξ_j, $j = 1,\dots,m$, из значений $\{\langle A_j\rangle_i\}_{i=1}^{N}$ набора НД(y) вычислим оценку среднего и дисперсии:

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N}\langle A_j\rangle_i, \qquad \sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\langle A_j\rangle_i\right)^2 - \mu_j^2.$$

Заметим, что в отличие от числовых дискретных значений, в данном случае не нужны операции по выявлению неповторяющихся реализаций ξ, не нужно упорядочивать и нумеровать значения. Для расчета обезличенного значения вместо ξ применяется *шаблон синтеза числового непрерывного*

значения. Он состоит в восстановлении (оценке) по имеющимся данным распределения исходного набора значений атрибутов $A_1,\ldots,A_m$ в наборе НД(y) и формировании в ОД(y) полностью синтетических значений. Для их моделирования распределение вектора ξ (набора атрибутов $A_1,\ldots,A_m$) аппроксимируется ядерной оценкой Розенблатта–Парзена [19, 20] с дополнением эмпирического правила Сильвермана [21]. Это значит, что плотность вероятности $f_\xi(x_1,\ldots,x_m)$ вектора ξ приближенно представляется в виде

$$f_\xi(x_1,\ldots,x_m) \approx \frac{1}{N\sqrt{\det(H^T H)}} \times$$

$$\times \sum_{i=1}^{N} \Phi(x_1,\ldots,x_m; \mathrm{col}(\langle A_1,\ldots,A_m\rangle_i), H^2),$$

$$H^2 = \mathrm{diag}(h_1^2,\ldots,h_m^2), \quad h_j = \sqrt[m+4]{\left(\frac{4}{m+2}\right)} \frac{1}{\sqrt[m+4]{N}} \sigma_j,$$

где $\Phi(x_1,\ldots,x_m; M, \Sigma)$ – m-мерная гауссовская плотность вероятности с вектором средних значений M и ковариационной матрицей Σ. Таким образом, исходное распределение вектора ξ (набора атрибутов $A_1,\ldots,A_m$) из НД(y) представляется гауссовской смесью из N слагаемых (ровно столько, сколько элементов в НД(y)), слагаемые имеют математические ожидания, задаваемые истинными значениями $\langle \hat{A}_1,\ldots,\hat{A}_m\rangle_i$ атрибутов, ковариации для каждого слагаемого – диагональные матрицы (т.е. элементы соответствующих m-мерных векторов независимы), а дисперсии задаются эмпирически обоснованным "сужением" оцененной по данным набора НД(y) дисперсии вектора ξ.

Для моделирования синтетических значений атрибутов $A_1,\ldots,A_m$ для набора ОД(y): в i-м элементе набора $\langle$ID-субъекта (GUID), $A_1,\ldots,A_m\rangle_i$ для j-го атрибута A_j моделируется случайная величина τ, имеющая равновероятное распределение на множестве числе $\{1,\ldots,N\}$, моделируется стандартная гауссовская величина ε, вычисляется значение $\langle \hat{A}_j\rangle_i = \langle A_j\rangle_\tau + h_j\,\varepsilon$. Отметим, что такая схема крайне проста в практической реализации.

3.2.3. Словарное значение. Характер данных этого типа аналогичен типу числовое дискретное значение. Действительно, не ограничивая общности, можно считать, что если атрибут A_1 словарный, то его значения можно описать реализациями ξ_1 – дискретной случайной величины с областью значений $\{1,\ldots,n\}$. Отличие же состоит в том, что для данных этого типа не измеряется "расстояние" между реализациями в традиционном арифметическом смысле. Разница между реализациями определяется семантикой словаря, так что унифицированного числового представления не существует. Поскольку есть характеристика атрибута A_1 в качестве квазиидентификатора, то ключевым является его влияние на оценку уровня анонимности, формируемого ОД(y), поэтому здесь применяется *шаблон устранения аномалий*. Его действие таково. Задается уровень T аномальных значений. Это величина в процентах, характеризующая приемлемое число повторений значений словарных атрибутов, не представляющее угрозы нарушения анонимности. Интуитивно приемлемое начальное значение $T=10\%$ означает, что для рассматриваемого словаря, содержащего n уникальных значений, приемлемым считается наличие не менее $\frac{10}{n}\%$ элементов для каждого из имеющихся n значений. Заметим, что шаблон применяется к НД(y), поэтому все n возможных словарных значений в обезличиваемом наборе есть, в отличие от НД, где словарь мог быть наполнен и неиспользуемыми значениями. Для выбора величины T можно использовать расчет уровня анонимности по данному атрибуту, описанный далее.

При нарушении условия для выбранного уровня T словарные значения, для которых условие нарушено, исключаются из словаря. Вместо всех исключенных словарных значений вводится одно нейтральное значение – "не определено", "не задано", "неизвестно". Область значений величины ξ_1 изменяется на $\{1,\ldots,n-m\}$, где m – число исключенных словарных значений. Далее атрибут A_1 включается в состав атрибутов типа числовое дискретное значение и вместе с ними подвергается обработке шаблоном синтеза числового дискретного значения.

3.2.4. Дата/время. Для атрибута A_1 этого типа применяется *шаблон разбиения*. Этот шаблон решает две задачи: во-первых, маскируются данные, потенциально угрожающие нарушением анонимности, во-вторых, тип дата/время заменяется простым словарным типом. Для применения шаблона разбиения к атрибуту A_1 выполняется следующее. Определяется диапазон дат – имеющихся в наборе НД(y) значений атрибута A_1. На основании значений диапазона выбирается одно из интервальных разбиений (выполняется семплирование диапазона). Сформировать коллекцию разбиений (семплов) – задача применяемого автоматизированного средства. Начальный набор, например, может включать:

- вариант для дат рождения – {⟨1 год, 1–3 года, 3–7 лет, 7–12 лет, 12–17 лет, ..., 95–100 лет⟩, 100 лет},
- другой вариант для дат рождения – {1913 г.р., 1914 г.р., ..., 2021 г.р.},
- вариант для данных трекинговых систем – {01.01.2023 0:00–01.01.2023 0:10, 01.01.2023 10:00–01.01.2023 0:20, ..., 31.12.2023 23:50–01.01.2023 0:00},

• вариант для средств коммерции – {01.01.2023, 02.01.2023, ..., 31.12.2023}.

Далее в наборе НД(y) атрибут A_1, имевший тип дата/время, становится атрибутом, имеющим тип словарное значение.

3.2.5. Потоковые данные. Свойство повторяемости этого типа неизбежно приведет к наличию в наборе НД(y) множества однотипных атрибутов A_j, $j = 1,\ldots,m$, с собственно содержательными данными и атрибутов A_{m+l}, $l = 1,\ldots,m$, содержащих "привязки" данных A_j (временные и/или географические). При этом свойство быть квазиидентификатором для всей этой совокупности атрибутов в большей степени учтено при подготовке к обезличиванию исключением избыточных данных. Именно, для потоковых данных в наборе НД(y) выбран из потенциально неограниченного множества атрибутов A_j, $j = 1,\ldots,M$, с потоковой информацией небольшой набор A_j, $j = 1,\ldots,m$, $m \ll M$, из соображений уменьшения угрозы нарушения анонимности, в частности, такого, чтобы данные в атрибутах A_j были во всех элементах (или в большинстве) набора НД(y). Дальнейшая обработка потоковых данных сводится, таким образом, к определению типа и характеристики атрибутов A_j, $j = 1,\ldots,m$, с содержательной информацией, т.е. применением к ним одного из перечисленных выше шаблонов, а для атрибутов A_{m+l}, $l = 1,\ldots,m$, типа дата/время применяется шаблон разбиения. Если эти атрибуты содержат данные геопривязки, то к ним естественным будет применение шаблона синтеза числового непрерывного значения.

3.3. Оценка результатов обезличивания

Применение ко всем атрибутам набора НД(y) перечисленных шаблонов алгоритма обезличивания выполняется в обратном порядке – от шаблонов потоковых данных до универсальных шаблонов синтеза числовых величин. Полученный в результате набор ОД(y) содержит данные, которые следует интерпретировать как обезличенные. Будет ли этот набор итоговым, должна подтверждать оценка результата с точки зрения анонимности и полезности полученного результата. Для формирования такой оценки алгоритм обезличивания дополняется возможностями расчета некоторых числовых характеристик.

3.3.1. Объединение атрибутов. Корреляции. Важным вопросом, обязательным к рассмотрению при обезличивании данных, должно стать взаимное влияние значений разных атрибутов. Основные шаблоны алгоритма обезличивания, применяемые к группам атрибутов типов числовое дискретное значение и числовое непрерывное значение смогут сохранить характер зависимостей между значениями атрибутов в исходном наборе НД(y). Именно для этого формировались преобразования не отдельного атрибута A_1, а групп атрибутов $A_1,\ldots,A_m$. При этом надо отметить, что применение шаблонов алгоритма к группе в сравнении с последовательным применением этих же шаблонов к каждому отдельному атрибуту – более вычислительно затратный процесс. Эффективнее всего применять шаблоны к небольшим группам атрибутов, т.е. разбивать группу $A_1,\ldots,A_m$ на несколько небольших подгрупп, к каждой из которых уже применять соответствующий шаблон.

Для выявления групп зависимых атрибутов, а также для качественной оценки результатов обезличивания с точки зрения сохранения важных зависимостей, рекомендуется применение простого статистического метода оценки корреляций. Коэффициент корреляции корр_{12} значений двух атрибутов A_1 и A_2 определяется как

$$\text{корр}_{12}(A_1, A_2) = \frac{\sum\limits_{i=1}^{N}\langle A_{1i}\rangle\langle A_{2i}\rangle - \left(\sum\limits_{i=1}^{N}\langle A_{1i}\rangle\right)\left(\sum\limits_{i=1}^{N}\langle A_{2i}\rangle\right)}{\sqrt{\left(\sum\limits_{i=1}^{N}(\langle A_{1i}\rangle)^2 - \left(\sum\limits_{i=1}^{N}\langle A_{1i}\rangle\right)^2\right)\left(\sum\limits_{i=1}^{N}(\langle A_{2i}\rangle)^2 - \left(\sum\limits_{i=1}^{N}\langle A_{2i}\rangle\right)^2\right)}}.$$

Эта величина, принимающая значения от 0 до 1, характеризует степень линейной зависимости значений, принимаемых атрибутами A_1 и A_2 в наборе НД(y). Соответственно, при подготовке обезличивания можно вычислять коэффициент корреляции для любых пар атрибутов и принимать решение об отсутствии зависимостей (близости коэффициента к нулю) и относить атрибуты к разным подгруппам числовых типов, либо о ее наличии.

Аналогично вычисляются корреляции применительно к набору ОД(y) и, таким образом, есть возможность сравнения $\text{корр}_{12}(A_1, A_2)$ и $\text{корр}_{12}(\hat{A}_1, \hat{A}_2)$. Заметим, что ни один из методов обезличивания гарантированного сохранения корреляций не обеспечивает, поэтому анализ результирующего

набора ОД(y) таким инструментом вполне целесообразен.

3.3.2. Оценка уровня анонимности. Для анализа уровня анонимности в соответствии с данным в разделе 2.1 определением в наборе ОД(y) требуется:

• выбрать атрибуты $\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m$, участвующие в расчете,

• определить величины ε_j приемлемого малого отклонения значений атрибута $\hat{\mathrm{A}}_j$ для каждого j-го атрибута, принимающего числовые непрерывные значения,

• вычислить число K – гарантированную оценку уровня K-анонимности,

• вычислить число k – относительный уровень K-анонимности,

• вычислить число K_ε – гарантированную оценку уровня ε-анонимности.

Все эти вычисления имеет смысл также проводить на наборе НД(y) для соответствующих атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_m$.

Для величин ε_j для тех атрибутов $\hat{\mathrm{A}}_j$, что принимают числовые непрерывные значения, предлагается следующий метод. Задается уровень нечувствительности T в процентах, например $T = 10\%$ Среди значений $\langle\hat{\mathrm{A}}_j\rangle_i, i = 1,\ldots,N$, находится минимальное x_{min_j} и максимальное x_{max_j}, величина ε_j рассчитывается как

$$\varepsilon_j = \left(x_{max_j} - x_{min_j}\right)\frac{T}{200}.$$

Смысл этой величины таков. Если взять размах выборки $(x_{max_j} - x_{min_j})$ и рассмотреть случай, когда значения $\langle\hat{\mathrm{A}}_j\rangle_i$ атрибута $\hat{\mathrm{A}}_j$, представленные в наборе данных ОД(y), распределены максимально равномерно, то получится равномерная сетка с шагом $(x_{max_j} - x_{min_j})\frac{1}{N}$ (подразумевается, что поскольку тип значений атрибута числовое непрерывное значение, то в выборке нет повторяющихся значений, поэтому разных уникальных значений ровно N). Выбранное так значение ε_j направлено на то, чтобы в ε_j-окрестность каждого значения $\langle\hat{\mathrm{A}}_j\rangle_i$ попадало примерно $T\%$ имеющихся значений $\langle\hat{\mathrm{A}}_j\rangle_i$ (исключение составят крайние точки, близкие к x_{max_j} и x_{min_j}, при приближении к которым число точек в окрестности будет уменьшаться до $T/2$). Это "идеальное" распределение, в реальных данных будут точки, разбросанные как дальше, так и ближе друга от друга, чем в равномерной сетке. Соответственно, судить об уровне K_ε-анонимности можно, сравнивая разницу между идеальным значением T и фактической величиной K_ε.

Наконец надо отметить, что формального смысла анализировать уровень анонимности для атрибутов, обработанных шаблонами синтеза, вообще говоря, нет, т.к. данные – значения таких атрибутов – являются синтетическими, т.е. полностью анонимны и не могут в принципе рассматриваться в качестве угрозы нарушения анонимности, т.к. по своей природе не являются де-факто ПД.

3.3.3. Анализ полезности. Второй объект анализа результативности выполненного обезличивания – оценка полезности (применимости) данных результирующего набора ОД(y). Полезность результирующего набора ОД(y) предлагается определять разницей между наборами значений $\{\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ и $\{\langle\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m\rangle_i\}_{i=1}^N$ в каждой из групп, выделенных в результате анализа корреляций. Поскольку эти наборы интерпретировались как реализации некоторых случайных векторов, то следует и различие между ними оценивать как различие между случайными векторами, т.е. соответствующими выборкам $\{\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ и $\{\langle\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m\rangle_i\}_{i=1}^N$ вероятностными распределениями. Для определения количественной оценки разницы между распределениями значений атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_m$ в НД(y) и значений атрибутов $\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m$ в ОД(y) предлагается использовать типовое решение – расстояние Кульбака–Лейблера [22]. Его традиционно используют как числовую оценку меры сходства/расхождения между распределениями вероятностей двух случайных векторов (величин) ξ и $\hat{\xi}$.

Для определения расстояния Кульбака–Лейблера в случае атрибутов типа числовое дискретное значение сформируем из всех имеющихся данных $\{\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ – значений атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_m$ в наборе НД(y) область значений $\{\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l\}_{l=0}^{L-1}$ – все неповторяющиеся значения $\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i$, дополним их частотами p_l появления значения $\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l$ в наборе $\{\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ и частотами $\hat{p}_l$ появления значения $\langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l$ в наборе $\{\langle\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m\rangle_i\}_{i=1}^N$, обозначим для простоты через $x_l = \langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l$. Будем считать, что таким образом заданы величины ξ и $\hat{\xi}$ с одинаковой областью значений $\{x_l\}_{l=0}^{L-1}$ и рядами распределений $\{p_l\}_{l=0}^{L-1}$ и $\{\hat{p}_l\}_{l=0}^{L-1}$, соответственно, т.е.

$$p_l = \mathbf{P}\left(\xi = x_l = \langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l\right),$$

$$\hat{p}_l = \mathbf{P}(\hat{\xi} = x_l = \langle\mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_l),$$

где $\mathbf{P}(\cdot)$ – вероятность события.

По предложенным обозначениям заметим следующее. Во-первых, строго говоря частоты

p_l, $\hat{p}_l$ нужно рассматривать как оценки неизвестных вероятностей, а не точные значения указанных вероятностей. В предположении, что выборка данных в НД(y) достаточно велика и статистически представительна (не содержит заведомых статистических искажений исходного распределения значений атрибутов), можно пренебречь погрешностью между частотной оценкой и истинной вероятностью. Во-вторых, будем учитывать, что величины $p_l > 0$ для всех $l = 0,\ldots, L-1$, по построению, потому что в область значений включены только те значения, что действительно имеются в наборе $\{\langle \mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$. При этом среди величин $\hat{p}_l$ могут оказаться нулевые в силу случайного характера примененного шаблона обезличивания.

Мерой полезности распределения $\{(x_l, \hat{p}_l)\}_{l=0}^{L-1}$ по отношению к распределению $\{(x_l, p_l)\}_{l=0}^{L-1}$ называется расстояние Кульбака–Лейблера $D(\hat{\xi},\xi)$:

$$D(\hat{\xi},\xi) = \sum_{l=0}^{L-1} \hat{p}_l \ln\left(\frac{\hat{p}_l}{p_l}\right).$$

Заметим, что величина $D(\hat{\xi},\xi)$ является несимметричной ($\hat{\xi},\xi$ нельзя поменять местами), в рассматриваемой постановке существует (может быть вычислена) из-за определения величин p_l (их положительность гарантирует существование логарифма). Известно, что значение $D(\hat{\xi},\xi) \geq 0$, причем равенство $D(\hat{\xi},\xi) = D(\xi,\hat{\xi})$ означает, что $\hat{p}_l = p_l$ для всех $l = 0,\ldots, L-1$. Масштабной характеристики (большой-малый) величина $D(\hat{\xi},\xi)$ не имеет, так что степень близости к 0 определяется экспериментально.

В случае числового непрерывного значения имеющиеся данные двух наборов – значения $\{\langle \mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ атрибутов $\mathrm{A}_1,\ldots,\mathrm{A}_m$ в наборе НД(y) и значения $\{\langle \hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m\rangle_i\}_{i=1}^N$ атрибутов $\hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m$ в наборе ОД(y) интерпретируются как выборки из распределений случайных векторов $\xi = \mathrm{col}(\xi_1,\ldots,\xi_m)$ и $\hat{\xi} = \mathrm{col}(\hat{\xi}_1,\ldots,\hat{\xi}_m)$, каждый элемент которых предполагается непрерывной случайной величиной. Таким образом, законы распределения ξ и $\hat{\xi}$ определяются плотностями вероятности – функциями $f_\xi(x_1,\ldots,x_m)$ и $f_{\hat{\xi}}(x_1,\ldots,x_m)$ соответственно:

$$\mathbf{P}(\xi \in X) = \int_X f_\xi(x_1, \ldots, x_m)dx_1 \ldots dx_m,$$
$$\mathbf{P}(\hat{\xi} \in X) = \int_X f_{\hat{\xi}}(x_1, \ldots, x_m)dx_1 \ldots dx_m.$$

Поскольку истинные распределения f_ξ и $f_{\hat{\xi}}$, вообще говоря, неизвестны, то в качестве плотностей будут использоваться, как и в случае числового дискретного значения, их ядерные оценки.

Именно, вычислим величины $\mu = \mathrm{col}(\mu_1, \ldots, \mu_n)$, $\sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_m)$, $\hat{\mu} = \mathrm{col}(\hat{\mu}_1,\ldots,\hat{\mu}_m)$, $\hat{\sigma} = \mathrm{diag}(\hat{\sigma}_1,\ldots,\hat{\sigma}_m)$, т.е. выборочные средние значения и среднеквадратические отклонения имеющихся выборок $\{\langle \mathrm{A}_1,\ldots,\mathrm{A}_m\rangle_i\}_{i=1}^N$ и $\{\langle \hat{\mathrm{A}}_1,\ldots,\hat{\mathrm{A}}_m\rangle_i\}_{i=1}^N$. В дополнение к определенной выше шаблоном синтеза числового непрерывного значения плотности $f_\xi(x_1, \ldots, x_m)$, положим

$$\hat{f}_\xi(x_1, \ldots, x_m) = \frac{1}{N\sqrt{\det(\hat{H}^T\hat{H})}} \times$$
$$\times \sum_{i=1}^N \Phi(x_1, \ldots, x_m; \mathrm{col}(\langle \hat{\mathrm{A}}_1, \ldots, \hat{\mathrm{A}}_m\rangle_i), \hat{H}^2),$$
$$\hat{H}^2 = \mathrm{diag}(\hat{h}_1^2,\ldots,\hat{h}_m^2), \quad \hat{h}_j = \sqrt[m+4]{\left(\frac{4}{m+2}\right)}\frac{1}{\sqrt[m+4]{N}}\hat{\sigma}_j.$$

Мерой полезности распределения $f_\xi(x_1,\ldots,x_m)$ по отношению к распределению $f_\xi(x_1,\ldots,x_m)$ называется расстояние Кульбака–Лейблера $D(\hat{f}_\xi, \hat{f}_{\hat{\xi}})$ распределений $\hat{f}_\xi(x_1,\ldots,x_m)$ и $\hat{f}_{\hat{\xi}}(x_1,\ldots,x_m)$:

$$D(\hat{f}_\xi, f_{\hat{\xi}}) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \hat{f}_\xi(x_1,\ldots,x_m) \times$$
$$\times \ln\left(\frac{\hat{f}_\xi(x_1,\ldots,x_m)}{f_\xi(x_1,\ldots,x_m)}\right)dx_1 \ldots dx_m.$$

Величина $D(\hat{f}_\xi, f_{\hat{\xi}})$ имеет те же свойства, что и величина $D(\hat{\xi},\xi)$ полезности для типа числовое дискретное значение.

4. ПРАКТИЧЕСКИЙ ПРИМЕР ОБЕЗЛИЧИВАНИЯ

4.1. Входные и выходные данные

Для выполнения модельных расчетов с предложенным алгоритмом требуется прежде всего решить вопрос с источником исходных необезличенных данных. Поскольку реальные данные в исследовательских целях применяться не могут, нужно получить какие-либо макетные данные, имитирующие реальные ПД. С этой целью была создана простая база данных по теме гражданских авиационных перевозок. Идея для создания набора ОД, имеющего практический смысл, состояла в использовании открытых данных по авиаперевозкам, выполненным в РФ в 2019 году. Во-первых, мы использовали действующее на тот момент расписание, которое позволило сформиро-

вать служебные конструкции – списки аэропортов и авиакомпаний. Кроме того, мы получили типы воздушных судов для большей части рейсов, что дало возможность имитировать число перевезенных пассажиров, исходя из средней загруженности рейсов 75–85%. Для пассажиров имитировался перелет туда и обратно по имеющемуся расписанию (по этой причине достаточно только информации об аэропорте отправления), и еще случайным образом задавался возраст (для этого использовалось дискретное распределение, полученное из гауссовского). В итоге в базу данных включены следующие сведения:

1) синтетический идентификатор пассажира,

2) возраст (лет),

3) дата-время отправления,

4) дата-время прибытия,

5) аэропорт отправления,

6) авиакомпания,

7) атрибут отправления,

8) атрибут прибытия.

Последние два атрибута включены для имитации атрибутов ПД типа числовое непрерывное значение и получены следующим образом. “Атрибут отправления” является производным значения “дата-время отправления”: тип дата/время преобразован в тип вещественный и зашумлен случайным значением, равномерно распределенным на отрезке [–1, 1] (это очень маленький шум, единственная задача которого убрать прямую зависимость этих двух атрибутов, имеющую место из-за того, что фактическая дата-время отправления однозначно определяет дату-время прибытия). Аналогично “атрибут прибытия” получается из “дата-время прибытия”. Таким образом, в НД есть два атрибута непрерывного типа, и они зависимы друг от друга.

Далее, к атрибуту “возраст (лет)” применялся описанный выше шаблон разбиения. Результирующий атрибут “возраст (группа)” вместе с атрибутами “аэропорт отправления”, “авиакомпания” дал три атрибута дискретного типа, и они независимы друг от друга. Независимость очевидна для возраста, поскольку его значения моделировались независимо. Для аэропортов и авиакомпаний независимость (точнее отсутствие выраженной линейной зависимости) объясняется тем, что в выборку всегда входит пара прямой и обратный рейс. Это объяснение подтверждается расчетами.

Итого в ОД вошли: синтетический идентификатор пассажира, возраст (группа), аэропорт отправления, авиакомпания, атрибут отправления, атрибут прибытия.

4.2. Результаты расчетов

Для 10000 записей о перелетах в наборе НД в результате применения алгоритма обезличивания формировался набор ОД такого же размера. Для анализа результатов вначале рассмотрим корреляционные показатели для двух групп атрибутов – дискретных и непрерывных. Для группы дискретных атрибутов корреляции $\text{корр}(A_i, A_j)$ в наборах:

$$\text{НД: корр (возраст, аэропорт)} = 0.021,$$

$$\text{корр(возраст, авиакомпания)} = 0.077,$$

$$\text{корр(аэропорт, авиакомпания)} = 0.03,$$

$$\text{ОД: корр(возраст, аэропорт)} = 0.059,$$

$$\text{корр(возраст, авиакомпания)} = 0.06,$$

$$\text{корр(аэропорт, авиакомпания)} = 0.029.$$

Для группы непрерывных атрибутов корреляции в наборах:

$$\text{НД: корр (атрибут отправления, атрибут прибытия)} = 0.966,$$

$$\text{ОД: корр (атрибут отправления, атрибут прибытия)} = 0.931.$$

Отметим, что эти результаты нужны не только для того, чтобы убедиться в работоспособности частотных и ядерных оценок при моделировании выборок случайных значений, что и так очевидно, но и для анализа влияния на эти показатели шаблона разбиения, который применялся к атрибуту возраст.

Другой показатель дает введенная в предыдущем разделе мера полезности (полезн(A_i, ..., A_j)). Поскольку масштаб этой меры заранее определить нельзя, то для сравнения эксперимент по обезличиванию был повторен с заменой метода синтеза на метод зашумления. Вот некоторые результаты:

для метода синтеза:

$$\text{полезн(атрибут отправления)} = 0.396,$$

$$\text{полезн (атрибут отправления, атрибут прибытия)} = 0.315,$$

$$\text{полезн (возраст, авиакомпания, аэропорт, атрибут отправления, атрибут прибытия)} = 0.129,$$

для метода зашумелния

$$\text{полезн (атрибут прибытия)} = 0.537,$$

$$\text{полезн (атрибут отправления, атрибут прибытия)} = 0.427,$$

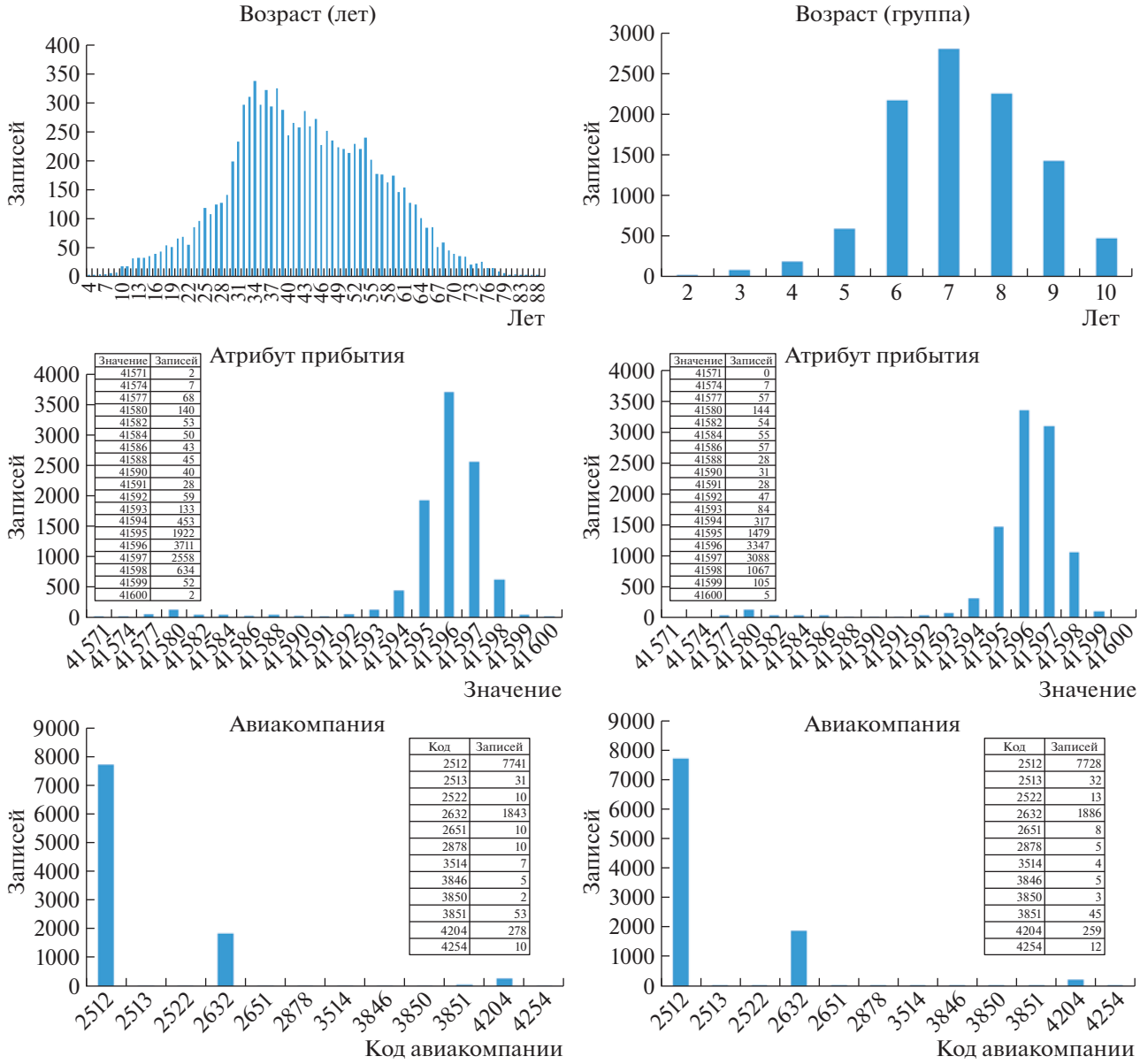

Рис. 2. Пример гистограмм необезличенных и обезличенных атрибутов.

полезн(возраст, авиакомпания, аэропорт,
атрибут отправления,
атрибут прибытия) = 1.725.

Для визуальной оценки этих результатов можно использовать гистограммы распределений. Примеры для трех атрибутов представлены парами (необезличенный-обезличенный) на рис. 2.

Наконец, нужно сделать замечание об уровне анонимности. Непрерывные атрибуты, использованные для данного примера, были выбраны так, чтобы принципиально усложнить реализацию концепции K-анонимности. Оставив атрибуты “дата-время отправления” и “дата-время прибытия” неизменными, об уровне K-анонимности можно было бы говорить вполне предметно. Изменение этих атрибутов с помощью малого непрерывного шума привело к тому, что уровень K-анонимности в обоих наборах данных стал равным 1. Более того, малость шума приводит к тому, что и уровень K_ε-анонимности столь же мал. Именно, в НД он также равен 1, в ОД значение вырастает до 3, что конечно же существенно ситуацию не меняет. Не столь ожидаемая, но такая же неудачная для K-анонимности ситуация оказалась и для дискретных атрибутов. Получилось, что при использованном числе записей и распределении пассажиров по возрасту есть возрастные

группы младенцев, в которых на некоторые авиакомпании попадает по одному рейсу для пассажира такого возраста. И ситуация воспроизводится при синтезе. Так что и здесь уровень *K*-анонимности оказывается равным 1. Эта неприятность могла бы не возникнуть, если бы был применен шаблон устранения аномалий, как и предусмотрено алгоритмом обезличивания. Не сделав этого, мы еще больше подчеркнули преимущество концепции синтеза.

Для целей данной работы это хороший пример, потому что для полученного результата, для данных, обезличенных предложенным методом синтеза, расчет уровня *K*-анонимности, строго говоря, ничего не означает. Все синтезированные записи набора ОД анонимны на 100%, потому что в принципе нет субъекта ПД для этих записей из-за их полностью синтетического характера.

5. ЗАКЛЮЧЕНИЕ

В этой и предыдущей статьях был охвачен довольно широкий спектр вопросов по теме автоматизации обезличивания персональных данных. Внимание к тематике ПД в нашей стране столь же велико, как и во всех государствах, обладающих значительной информационной инфраструктурой. Но это внимание ограничено нетехническими сообществами. Лучше всего проработаны правовые вопросы, на хорошем уровне находится нормативная база, предметно работают уполномоченные госорганы, ведется социальная дискуссия. Но при этом значимое внимание со стороны научного сообщества, исследовательская активность отсутствуют. В двух статьях была сделана попытка показать широту и разнообразие именно технических проблем и задач в данной области и включиться в исследовательскую активность со своими идеями.

6. БЛАГОДАРНОСТИ

Работа выполнялась с использованием инфраструктуры Центра коллективного пользования “Высокопроизводительные вычисления и большие данные” (ЦКП “Информатика” ФИЦ ИУ РАН, Москва).

СПИСОК ЛИТЕРАТУРЫ

1. *Борисов А.В., Босов А.В., Иванов А.В.* Применение имитационного компьютерного моделирования к задаче обезличивания персональных данных. Оценка состояния и основные положения // Программирование, 2023. № 4, с. 58–74.
2. *Aggarwal C.C., Yu P.S.* On Privacy-Preservation of Text and Sparse Binary Data with Sketches // SIAM Conference on Data Mining, 2007.
3. *Sweeney L. K*-anonymity: a model for protecting privacy // International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002. V. 10. № 5. P. 557–570.
4. *Samarati P., Sweeney L.* Generalizing Data to Provide Anonymity when Disclosing Information (Abstract) // Proc. of ACM Symposium on Principles of Database Systems, 1998. P. 188.
5. *Samarati P.* Protecting Respondents’ Identities in Microdata Release // IEEE Trans. Knowl. Data Eng., 2001. V. 13. № 6. P. 1010–1027.
6. *Bayardo R.J., Agrawal R.* Data Privacy through Optimal k-Anonymization // Proceedings of the ICDE Conference, 2005. P. 217–228.
7. *Fung B., Wang K., Yu P.* Top-Down Specialization for Information and Privacy Preservation // ICDE Conference, 2005.
8. *Wang K., Yu P., Chakraborty S.* Bottom-Up Generalization: A Data Mining Solution to Privacy Protection // ICDM Conference, 2004.
9. *Domingo-Ferrer J., Mateo-Sanz J.* Practical data-oriented micro-aggregation for statistical disclosure control // IEEE TKDE, 2002. V. 14. № 1.
10. *Winkler W.* Using simulated annealing for k-anonymity // Technical Report 7, US Census Bureau, Washington D.C. 20233, 2002.
11. *Iyengar V.S.* Transforming Data to Satisfy Privacy Constraints // KDD Conference, 2002.
12. *Lakshmanan L., Ng R., Ramesh G.* To Do or Not To Do: The Dilemma of Disclosing Anonymized Data // ACM SIGMOD Conference, 2005.
13. *Aggarwal C.C., Yu P.S.* On Variable Constraints in Privacy-Preserving Data Mining // SIAM Conference, 2005.
14. *Aggarwal C.C.* On k-anonymity and the curse of dimensionality // VLDB Conference, 2005.
15. *Iyengar V.S.* Transforming Data to Satisfy Privacy Constraints // KDD Conference, 2002.
16. *Machanavajjhala A., Gehrke J., Kifer D., Venkitasubramaniam M. L*-Diversity: Privacy Beyond k-Anonymity // ICDE Conference, 2006.
17. *Fung B., Wang K., Yu P.* Top-Down Specialization for Information and Privacy Preservation // ICDE Conference, 2005.
18. *Wang K., Yu P., Chakraborty S.* Bottom-Up Generalization: A Data Mining Solution to Privacy Protection // ICDM Conference, 2004.
19. *Rosenblatt M.* Remarks on Some Nonparametric Estimates of a Density Function // Ann. Math. Statist., 1956. V. 27. № 3. P. 832–837.
20. *Parzen E.* On Estimation of a Probability Density Function and Mode // Ann. Math. Statist., 1962. V. 33. № 3. P. 1065–1076.
21. *Silverman B.W.* Density Estimation for Statistics and Data Analysis. London: Chapman & Hall/CRC, 1986.
22. *Kullback S., Leibler R.A.* On information and sufficiency // Ann. Math. Statist., 1951. V. 22. № 1. P. 79–86.

APPLICATION OF SIMULATED COMPUTER SIMULATION TO THE TASK OF PERSONAL DEPERSONALIZATION DATA. MODEL AND ALGORITHM FOR DECONTAMINATION BY SYNTHESIS

© 2023 г. S. A. Borisov[a,#], A. A. Bosov[a,##], and D. E. Ivanov[a,###]

[a] *Federal Research Center "Informatics and Management" RAS, Moscow, Russia*

[#] *e-mail: aborisov@ipiran.ru*

[##] *e-mail: avbosov@ipiran.ru*

[###] *e-mail: aivanov@ipiran.ru*

The second part of the study on the topic of automated depersonalization of personal data is presented. The review and analysis of the prospects for research, performed earlier, is supplemented here by a practical result. A model of the depersonalization process is proposed, reducing task of ensuring anonymity of personal data to manipulation of samples of different types of random elements. Accordingly, the key idea of transforming data to ensure their anonymity, provided that utility is maintained, is to apply the synthesis method, i.e. complete replacement of all unpublished data with synthetic values. The proposed model identifies a set of element types for which synthesis patterns are proposed. The set of patterns compiles the depersonalization algorithm by the synthesis method. Methodically, each template is based on a typical statistical tool – frequency probability estimates, nuclear Rosenblatt-Parsen density estimates, statistical averages and covariances. The application of the algorithm is illustrated by a simple example from the field of civil air transportation.