Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

На правах рукописи

УДК 004.056.5

ГУСЕВ
Илья Николаевич

РАЗРАБОТКА И АНАЛИЗ ЭФФЕКТИВНЫХ АЛГОРИТМОВ ОБЕЗЛИЧИВАНИЯ И ДЕ-ОБЕЗЛИЧИВАНИЯ ПЕРСОНАЛЬНЫХ ДАННЫХ

АВТОРЕФЕРАТ
диссертации на соискание степени
магистра технических наук

по специальности 1-38 80 04 – Технология приборостроения

Минск 2017

Работа выполнена на кафедре проектирования информационно-компьютерных систем учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: **ГОНОВ Александр Николаевич,**
кандидат технических наук, доцент, доцент кафедры проектирования информационно-компьютерных систем учреждения образования «Белорусский государственный университет информатики и радиоэлектроники»

Рецензент: **НОВИКОВ Сергей Олегович,**
кандидат технических наук, доцент кафедры информационных систем и технологий «Международный институт дистанционного образования Белорусский национальный технический университет»

Защита диссертации состоится «22» июня 2017 г. года в $9^{\underline{00}}$ часов на заседании Государственной комиссии по защите магистерских диссертаций в учреждении образования «Белорусский государственный университет информатики и радиоэлектроники» по адресу: 220013, Минск, ул. П.Бровки, 6, копр. 1, ауд. 415, тел. 293-20-80, e-mail: kafpiks@bsuir.by

С диссертацией можно ознакомиться в библиотеке учреждения образования «Белорусский государственный университет информатики и радиоэлектроники».

ВВЕДЕНИЕ

В настоящее время, в связи с участившимися преступлениями в сфере информационных разработок, связанных со взломом баз данных и хищения информации, необходимо обратить особое внимание на вопросы защиты персональных данных (ПД).

Выполнение всех требований предъявляемых к информационным системам персональных данных (ИСПД), как правило связано с существенными финансовыми и материальными затратами, вызванными необходимостью создания системы защиты, обеспечением высокой квалификации персонала, получением разрешительных документов, что не всегда возможно для большого числа пользователей и операторов. В связи с этим представляют интерес исследования направленные на разработку и анализ методов обработки ПД, позволяющих снизить затраты и увеличить эффективность ИСПД.

В настоящее время основной проблемой построения систем защиты персональных данных является отсутствие, или недостаточная проработанность законодательных актов и иных нормативных документов, что приводит к возникновению сложностей при проектировании и разработке систем защиты ПД.

На сегодняшний день существует большое число работ в области защиты информации в информационных системах персональных данных. Наиболее значимые результаты были получены российскими и белорусскими учеными, которые проводили исследования в области информационного права, и правового обеспечения информационной безопасности и защиты информации (Т.В. Борботько, Л.М. Лыньков, Б.И. Беляев, И.Л. Бачило, В.А. Копылов, и др.); проблемы функционирования системы с точки зрения технических наук (Е.В. Касперский, В.А. Минаев и др.). Среди зарубежных авторов особый интерес вызывают работы А. Конхейма, К. Девиса, Б. Крейбса, Г. Олмана, в работах которых представлено описание некоторых механизмов информационной безопасности, и в частности защиты информации в информационных системах персональных данных и проектирования систем баз данных.

Разработка научных и технических основ защиты персональных данных в информационных системах, построение центров обработки данных и, как следствие, повышение качества оказания услуг в сфере информационной безопасности доказывает актуальность темы.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

С развитием процессов информатизации общества, созданием новых и интеграцией существующих информационных систем, ориентированных на обслужи-

вание населения, все большее внимание уделяется организации обработки персональных данных. Персональные данные составляют важную часть информационного пространства, содержащую сведения о физических лицах – субъектах персонализации. Выделение таких данных в отдельное подмножество обусловлено особыми требованиями к организации их обработки (действиями с ПД), связанными с возможностью нанесения вреда субъектам ПД. Поэтому как в зарубежных странах, так и в Республики Беларусь развивается и совершенствуется законодательная база, регламентирующая правила обработки ПД и реализацию прав граждан на конфиденциальность касающейся их информации.

Степень разработанности проблемы

Исследование в области безопасности ИСПД осуществлялось на основе построения теоретических моделей с использованием работ российских и белорусских ученых: Т.В. Борботько, Л.М. Лыньков, Б.И. Беляев, И.Л. Бачило, А.А. Фатьянова, А.А. Шиверскова, М.А. Вуса, В.А. Герасименко, Е.В. Касперского, В.А. Минаева и др. Одним из недостатков исследований, представленных в современной технической литературе, является неполное рассмотрение особенностей проектирования информационных систем и сложность предложенных алгоритмов защиты ПД.

Предложенное исследование направлено на устранение этого недостатка на основе модификации алгоритма защиты персональных данных методом обезличивания

Цель и задачи исследования

Целью диссертации является разработка и анализ эффективных алгоритмов обезличивания и де-обезличивания персональных данных.

Поставленная цель работы определяет следующие основные задачи:

1. Провести обзор и анализ законодательных актов, регулирующих порядок работы с персональными данными, определить группы и категории информационных систем персональных данных.

2. Исследовать и проанализировать известные подходы к обезличиванию ПД, а также разработать алгоритм обезличивания и де-обезличивания персональных данных.

3. Описать организацию безопасной работы с персональными данными в дата-центрах внешних операторов.

Область исследования

Содержание диссертации соответствует образовательному стандарту высшего образования второй ступени (магистратуры) специальности 1-38 80 04 «Технология приборостроения».

Теоретическая и методологическая основа исследования

В основу диссертации легли работы белорусских и зарубежных ученых в области создания систем защиты информации, а также анализ технических нормативных правовых актов по рассматриваемой тематике.

Информационная база исследования сформирована на основе литературы, открытой информации, технических нормативно-правовых актов, сведений из электронных ресурсов, а также материалов научных конференций и семинаров.

Научная новизна

Научная новизна и значимость полученных результатов работы заключается в модификации алгоритма защиты персональных данных основанного на использовании комбинации метода изменения состава или семантики и метода перестановки, позволяющая проводить обезличивание больших массивов персональных данных.

Теоретическая значимость работы заключается в анализе существующих методов обезличивания персональных данных.

Практическая значимость диссертации состоит в разработанном алгоритме обезличивания персональных данных, который позволит проводить обезличивание большого массива информации.

Основные положения, выносимые на защиту

1. Анализ нормативно-технических актов в области защиты информации, позволяющий сделать выводы о состоянии законодательной базы, а так же о возможностях её совершенствования.

2. Алгоритм обезличивания и де-обезличивания персональных данных, основанный на перемешивании данных и уменьшении их состава, позволяющий проводить обезличивание больших массивов данных при минимальной сложности.

3. Требования к организации обработки персональных данных в дата-центрах внешних операторов, основанные на законодательных актах, позволяющие обеспечить конфиденциальность данных.

Апробация диссертации и информация об использовании ее результатов

Результаты исследований, вошедшие в диссертацию, представлены в публикациях на 53-ой научно-технической конференции аспирантов, магистрантов и студентов БГУИР (Минск, Беларусь, 2017 г.).

Публикации

Изложенные в диссертации основные положения и выводы опубликованы в 4 печатных работах. В их числе 4 статьи в сборниках материалов научных конференций.

Общий объем публикаций по теме диссертационной работы составляет 0,5 авторских листа.

Структура и объем работы

Диссертация состоит из введения, общей характеристики работы, трех глав с краткими выводами по каждой главе, заключения, библиографического списка и приложений.

В первой главе приведен обзор современного состояния проблемы защиты персональных данных в информационных системах, а также рассмотрены законодательные акты и иные нормативные документы касающиеся данной тематики. **Во второй главе** представлены разработанные алгоритмы обезличивания и де-обезличивания персональных данных в информационных системах. **В третьей главе** представлено описание системы обработки персональных данных, а также приведены правила работы с персональными данными для Операторов. **В приложении** представлены публикации автора и акт внедрения.

Общий объем диссертационной работы составляет 79 страниц. Из них 50 страницы основного текста, 9 иллюстраций на 6 страницах, 10 таблиц на 3 страницах, библиографический список из 64 наименований на 5 страницах, список собственных публикаций соискателя из 4 наименований на 1 странице, 3 приложения на 13 страницах.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы повышения защищенности ИСПД, указаны основные направления исследований, проводимых по данной тематике, а также описано обоснование актуальности темы.

В **общей характеристике работы** показана актуальность проводимых исследований, степень разработанности проблемы, сформулированы цель и задачи диссертации, обозначена область исследований, научная (теоретическая и практическая) значимость исследований, а также апробация работы.

В первой главе приведен обзор современного состояния проблемы защиты персональных данных в информационных системах, а также рассмотрены законодательные акты и иные нормативные документы, касающиеся данной тематики.

Из анализа следует, что проблема защиты персональных данных заключается в отсутствии точных алгоритмов и методик и в несовершенстве законодательной базы Республики Беларусь. Также в данной главе представлена классификация ИСПД и способов ее защиты, позволившая выявить, что разработка системы защиты ИСПД в каждом конкретном случае носит субъективных характер, поскольку дать объективную оценку вреда, нанесенного субъектам при доступе к их персональным данным практически невозможно. В большинстве случаев, предлагаемые решения для разработки систем защиты ПД могут привести к необоснованным затратам, так как в большинстве случаев они являются избыточными.

Во второй главе сделан анализ всех возможных методов обезличивания персональных данных, дана их подробная характеристика, и проведено их сравнение. Анализ позволил выделить общие слабые места методов и обосновать целесообразность проведения разработки новых методов обезличивания ПД, основанных на процедурах перемешивания данных, относящихся к различным субъектам. На основе этого анализа был выбран комбинированных метод обезличивания ПД, основанный на методе декомпозиции и методе перемешивания.

Сущность предложенного алгоритма заключается в том что первоначально массив персональных данных (таблица 1) разбивается на несколько подмножеств, каждая из которых содержит заданный набор атрибутов всех субъектов (таблицы 2 и 3).

Таблица 1 – Исходное множество персональных данных

Атрибут t_1	Атрибут t_2	Атрибут t_3	Атрибут t_4
a_1	b_1	c_1	d_1
a_2	b_2	c_2	d_2
a_3	b_3	c_3	d_3
a_4	b_4	c_4	d_4
a_5	b_5	c_5	d_5

Таблица 2 – Подмножество U_0

Атрибут t_1	Атрибут t_2	№
a_1	b_1	1
a_2	b_2	2
a_3	b_3	3
a_4	b_4	4
a_5	b_5	5

Таблица 3 – Подмножество U_1

Атрибут t_3	Атрибут t_4	№
c_1	d_1	1
c_2	d_2	2
c_3	d_3	3
c_4	d_4	4
c_5	d_5	5

Затем данные из таблицы 3 делятся на подмножества $U_2(5)$, соответствующее атрибуту t_3 и $U_3(5)$, соответствующее атрибуту t_4. Эти множества считаем

упорядоченными, так как они имеют 5 занумерованных в порядке возрастания номеров мест.

Проведем циклическую перестановку каждого из полученных подмножеств $U_2(5)$и $U_3(5)$ с $r_0 = 2$ и $r_1 = 3$. При циклической перестановке каждый элемент исходного множества переставляется на новое место, номер которого отличится от номера элемента на постоянную величину. Назовем эту величину параметром цикла и обозначим её r. На множестве, содержащем n элементов, параметр цикла может принимать значения от 1 до $n(1 \le |r| < n)$, при этом мы получим различные результаты перестановок.

В начале, следует определить направление циклического сдвига, так как возможны варианты, когда $r > 0$ (сдвиг в право) или $r < 0$ (сдвиг влево). Новый порядковый номер элемента k, имеющего до перестановки номер j вычисляется по формулам:

$$k = (j + r) mod\ n, \text{если } j + r \neq n \text{ и } k = n, \text{если } j + r = n\ (r > 0) \quad (1)$$
$$k = (j - r + n) mod\ n, \text{если } j - r \neq 0 \text{ и } k = n, \text{если } j - r = 0\ (r < 0)$$

В результате проведенной циклической перестановки получим обезличенные методом перестановки значения подмножеств $U_2(5)$ и $U_3(5)$, приведенные в таблице 4

Таблица 4 – Обезличенные данные

Атрибут t_3	Атрибут t_4	№
c_3	d_4	1
c_4	d_5	2
c_5	d_1	3
c_1	d_2	4
c_2	d_3	5

Далее необходимо построить таблицу связи подмножеств U_1 и U_2, данные связи приведены в таблице 5.

Де-обезличивание осуществляется по заданному набору связей (таблица 5) между раздельно хранимыми частями.

Правила разделения исходного массива данных определяются таким образом, чтобы каждая из частей не содержала сведений, позволяющих однозначно идентифицировать субъекта.

Таблица 5 – Связи между подмножествами

№ строки таблицы 1	№ строки таблицы 2
1	1
2	2
3	3
4	4
5	5

При де-обезличивании массива данных, необходимо провести обратный порядок действий. Во-первых происходит де-обезличивание подмножества U_1, для определения правильного порядка хранимых данных, а во-вторых происходит де-обезличивание исходного множества с помощью таблицы связей.

В третьей главе представлено описание системы обработки персональных данных. Для организации автоматизированной системы обработки персональных данных и разработке ИСПД необходимо учитывать интересы пользователей системы, субъектов и операторов, требования законодательных актов и иных нормативных документов. В связи с этим практический интерес представляют типовые решения, которые можно применять при разработке и организации системы обработки данных.

В любой информационной системе, занимающийся обработкой персональных данных, существуют следующие участники системы:

– владельцы персональных данных – субъекты персональных данных – физические лица, предоставляющие свои персональные данные в систему;

– пользователи системы – юридические или физические лица, заказывающие обработку персональных данных;

– операторы – юридические лица, осуществляющие автоматизированную обработку персональных данных с использованием ИСПД.

Взаимодействия между участниками и процедуры обработки персональных данных должны обеспечивать конфиденциальность данных на всех этапах их обработки, представления, хранения и передачи. При этом возможность идентификации персональных данных возможна только на этапе их регистрации в системе субъектом и предоставления результатов пользователю, либо субъекту. Обработка должна проводиться с данными, которые нельзя идентифицировать. Такой подход к организации обработки приводит к уменьшению затрат оператора на обеспечение конфиденциальности персональных данных.

Так же существует необходимость создания внешних организаций в виде специализированных центров обработки персональных данных (ЦОД), которые будут обслуживать группу организаций, обеспечиваю хранение и защиту персональных данных, а так же предоставление этих данных пользователям.

ЦОД используются по всему миру уже долгое время, и состоят из стандартного набора оборудования:

– информационная инфраструктура – серверное оборудование, обеспечивающее основные функции дата-центра (обработка и хранение информации);

– телекоммуникационная инфраструктура – обеспечивает взаимосвязь элементов дата-центра, а также передачу данных между центром и пользователем;

– инженерная инфраструктура – обеспечивает нормальное функционирование основных систем.

Пример структуры такого центра представлен на рисунке 1.

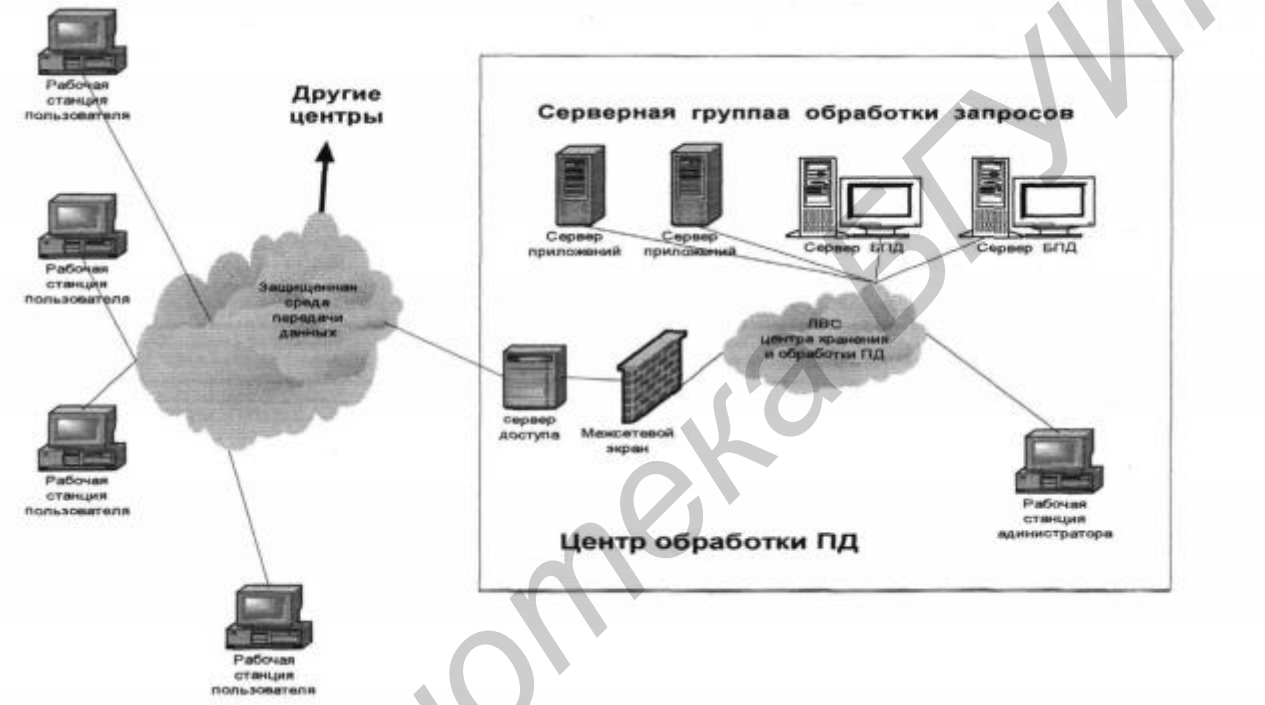

Рисунок 1 – Структура центра обработки персональных данных

Важными вопросами являются организация работы с данными и взаимодействие между центром, субъектами персональных данных и пользователями.

Здесь сложность заключается в том, что сотрудники организации - пользователя должны иметь доступ к персональным данным с одновременной идентификацией субъекта, как на этапе сбора данных, так и на этапах их обработки, а сотрудники центра не должны иметь возможности идентификации данных.

При этом:

– центр получает и хранит только обезличенные данные, для сохранения конфиденциальности, при регистрации субъектов у пользователя каждому субъекту присваивается идентификатор, который используется при поиске данных и их обработке, как в центре, так и у пользователей;

– процедуры назначения идентификаторов субъектов, обезличивания и де-обезличивания персональных данных реализуются на рабочих станциях пользователей или субъектов (входят в состав клиентского программного обеспечения);

– представление персональных данных в сочетании с идентификационными данными субъекта возможно только для конкретного сотрудника пользователя при наличии у него специального разрешения.

Выполнение процедур обезличивания и де-обезличивания можно проводить для блоков данных, даже при запросе ПД только по одному субъекту, это позволит сократить число обращений к центру, повысит эффективность обработки запросов. Однако от кэширования данных, в целях безопасности, целесообразно отказаться.

При такой организации решается проблема персистентных (долговременных) данных, которые могут оставаться в центре после завершения обработки, так как в центре имеются только обезличенные данные, а де-обезличивание средствами центра невозможно.

ЗАКЛЮЧЕНИЕ

Основные научные результаты диссертации

1. Выполнен обзор и анализ законодательных актов, регулирующих порядок работы с персональными данными, определены группы и категории информационных систем персональных данных [1, 2].

2. Исследованы и проанализированы известные подходы к обезличиванию ПД, а также разработан алгоритм обезличивания и де-обезличивания персональных данных, позволяющий проводить обезличивание больших массивов данных, при минимальной сложности.

3. Описана организация дата-центров, а также даны рекомендации по организации безопасной работы с ПД в них [3, 4].

Рекомендации по практическому использованию результатов

Полученные результаты внедрены в учебный процесс на кафедре проектирвоания информационно-компьютерных систем учреждения образования “Белорусский государственный университет информатики и радиоэлектроники в учебный курс “Программные средства защиты информации и защита информации в автоматизированных офисных и банковских системах”.

СПИСОК ПУБЛИКАЦИЙ СОИСКАТЕЛЯ

1. Гусев И.Н. Адаптивные системы защиты информации / И.Н. Гусев, Е.В. Осакович, С.А. Мигалевич // материалы 52−ой науч. конф. аспирантов, магистрантов и студентов «Проектирование информационно−компьютерных систем», Минск, Респ. Беларусь, 25–30 апреля 2016 г. / УО «БГУИР». – Минск, 2016. – С.41–42.

2. Гусев, И.Н. Обеспечение защиты персональных данных в Республике Беларусь / И.Н. Гусев // материалы 53−ой науч. конф. аспирантов, магистрантов и студентов «Проектирование информационно−компьютерных систем», Минск, Респ. Беларусь, 02–06 мая 2017 г. / УО «БГУИР». – Минск, 2016. – В печати.

3. Гусев, И.Н. Защита информации в сети интернет / И.Н. Гусев // материалы 53−ой науч. конф. аспирантов, магистрантов и студентов «Проектирование информационно−компьютерных систем», Минск, Респ. Беларусь, 02–06 мая 2017 г. / УО «БГУИР». – Минск, 2016. – В печати.

4. Гусев, И.Н. Проблемы защиты персональных данных / И.Н. Гусев // материалы 53−ой науч. конф. аспирантов, магистрантов и студентов «Проектирование информационно−компьютерных систем», Минск, Респ. Беларусь, 02–06 мая 2017 г. / УО «БГУИР». – Минск, 2016. – В печати.

РЭЗЮМЭ

Гусеў Ілля Мікалаевіч
Распрацоўка і аналіз эфектыўных алгарытмаў абезлічвання і дэ-абезлічвання персанальных данных

Ключавыя словы: персанальныя данныя, абезлічванне

Мэта працы: распрацоўка і аналіз эфектыўных алгарытмаў абезлічвання і дэ-абезлічвання персанальных данных.

Атрыманыя вынікі і іх навізна: выкананы агляд і аналіз дакументаў па інфармацыйнай бяспекі. Выяўлена, што ў заканадаўчых і нарматыўных актах, звязаных з абаротам персанальных дадзеных, вызначаецца толькі парадак і прававыя асновы дзеянняў пэўных дзяржаўных органаў, якія вядуць рэгістр з персанальнымі дадзенымі. Пры такой пастаноўцы пытання закона не рэгулюе і не абараняе персанальныя дадзеныя фізічных асоба пры іх атрыманні, апрацоўка, захоўванне, перадача і знішчэнне іншых суб'ектаў; распрацаваны алгарытм абезлічвання і дэ-абезлічвання персанальных данных, які дазваляе праводзіць обезличивание вялікіх масіваў дадзеных, пры мінімальнай складанасці.

Ступень выкарыстання: вынікі ўкаранёны ў навучальны працэс на кафедре праектавання інфармацыйна-камп'ютэрных сістэм ўстановы образования «Беларускі дзяржаўны універсітэт інфарматыкі і радиоэлектроники ў навучальныя курс" Метады і тэхнічныя сродкі забеспячэння бяспекі "і" Асновы абароны інфармацыі".

Вобласць ужывання: сістэмы абароны інфармацыі.

РЕЗЮМЕ

Гусев Илья Николаевич

Разработка и анализ эффективных алгоритмов обезличивания и де-обезличивания персональных данных

***Ключевые слова*:** персональные данные, обезличивание.

***Цель работы*:** разработка и анализ эффективных алгоритмов обезличивания и де-обезличивания персональных данных.

***Полученные результаты и их новизна*:** выполнен обзор и анализ документов по информационной безопасности. Выявлено, что в законодательных и нормативных актах, связанных с оборотом персональных данных, определяется лишь порядок и правовые основы действий определенных государственных органов, ведущих регистр с персональными данными. При такой постановке вопроса закон не регулирует и не защищает персональные данные физических лиц при их получении, обработке, хранении, передаче и уничтожении иными субъектами; разработан алгоритм обезличивания и де-обезличивания персональных данных, позволяющий проводить обезличивание больших массивов данных, при минимальной сложности.

***Степень использования*:** результаты внедрены в учебный процесс на кафедре проектирования информационно−компьютерных систем учреждения образования “Белорусский государственный университет информатики и радиоэлектроники в учебные курс “Методы и технические средства обеспечения безопасности ” и “Основы защиты информации”.

***Область применения*:** системы защиты информации.

SUMMARY

Ilya N. Gusev

Development and analysis of efficient algorithms depersonalization and de-anonymization of personal data

***Keywords*:** personal data, depersonalization.

***The object of study*:** development and analysis of efficient algorithms depersonalization and de-anonymization of personal data.

***The results and novelty*:** a review and analysis of information security documents. It was revealed that in the laws and regulations related to the turnover of personal data, is determined only by the procedure and the legal basis of actions of certain public bodies, leading register of personal data. Put that the law does not regulate or protect the personal data of individuals in their manufacture, processing, storage, transfer and destruction of other subjects; The algorithm developed by depersonalization and de-anonymization of personal data, allowing to carry out the depersonalization of large data sets, with E-minimality complexity.

***Degree of use*:** results are introduced in the educational process at the department of information and computer design systems and institutions to analyze "the Belarusian State University of Informatics and Radio Electronics" in the training course "Methods and means of security" and "Fundamentals of information security".

Sphere of application: information security systems