УДК 004.056.5:336.71(076)

ОЦЕНКА СТЕПЕНИ РАЗРУШЕНИЯ СТАТИСТИКИ ЕСТЕСТВЕННОГО ЯЗЫКА ПРИ ОБЕЗЛИЧИВАНИИ ПЕРСОНАЛЬНЫХ ДАННЫХ НА ОСНОВЕ МЕТОДА ИЗМЕНЕНИЯ СОСТАВА ИЛИ СЕМАНТИКИ

Я. А. Клиндухов

Национальный детский технопарк,
Минск, Беларусь, klinduhov.science@gmail.by

В данной работе предложена процедура обезличивания персональных данных, основанная на методе изменения состава или семантики с использованием доверительных вычислительных баз (ДВБ). Разработана структурная схема процедуры обезличивания, включающая разбиение данных на блоки и их последовательную обработку с применением различных ДВБ. Для оценки эффективности метода проведён анализ индекса совпадения Фридмана до и после обезличивания. Результаты показали снижение индекса совпадения более чем в 2,5 раза, что свидетельствует о значительном разрушении статистики естественного языка и подтверждает высокий уровень информационной безопасности. Полученные значения приближаются к характеристикам случайного текста, что подчеркивает практическую применимость метода при сохранении высокого уровня информационной безопасности персональных данных.

Ключевые слова: персональные данные; обезличивание персональных данных; метод изменения состава или семантики; индекс совпадения Фридмана.

ASSESSING THE DEGREE OF NATURAL LANGUAGE STATISTICS DESTRUCTION DURING PERSONAL DATA ANONYMIZATION BASED ON THE METHOD OF COMPOSITION OR SEMANTICS ALTERATION

Y. A. Klindukhov

National Children's Technopark,
Minsk, Belarus, klinduhov.science@gmail.by

This paper proposes a procedure for personal data anonymization based on the method of altering composition or semantics using trusted computing bases (TCBs). A structural scheme of the anonymization procedure has been developed, involving partitioning data into blocks and their sequential processing using various TCBs. To evaluate the method's effectiveness, an analysis of the Friedman coincidence index before and after anonymization was conducted. The results showed a reduction in the coincidence index by more than 2.5 times, indicating a significant disruption of natural language statistics and confirming a high level of information security. The obtained values approach the characteristics of random text, highlighting the method's practical applicability while maintaining a high level of personal data information security.

Keywords: personal data; personal data anonymization; method of altering composition or semantics; Friedman coincidence index.

1. Введение

В условиях цифровой трансформации и стремительного роста объемов обрабатываемых персональных данных важную роль играет процесс их обезличивания [1–3]. Под обезличиванием персональных данных будем понимать действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных.

Одним из таких методов является метод изменения состава или семантики, сущность которого заключается в том, что выполняют обобщение, изменение или удаление части сведений, позволяющих идентифицировать субъекта персональных данных. При этом полученные обезличенные персональные данные и правила их изменения необходимо хранить раздельно.

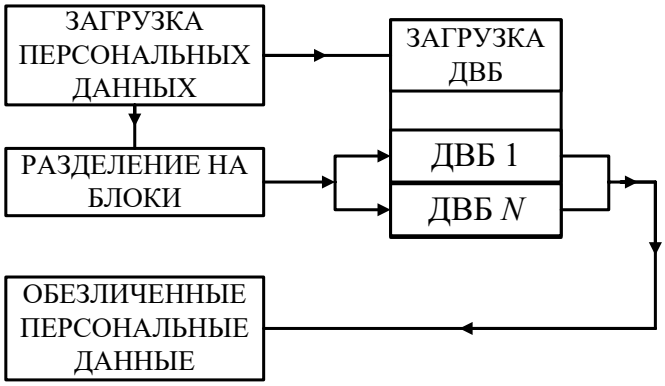
Обезличивание персональных данных указанным выше методом не требует наличие высоких вычислительных мощностей, однако характеризуется следующими недостатками. Во-первых, удаление части сведений, позволяющих идентифицировать субъекта персональных данных, без их сохранения в отдельной базе данных приводит к утрате такого свойства обезличенных персональных данных, как полнота. Во-вторых, простые замены исходных символов персональных данных обезличенными сохраняет вероятности появления соответствующих символов обезличенных персональных данных. В этом случае нарушитель информационной безопасности имеет возможность, получив доступ к обезличенным персональным данным, рассчитать вероятности появления отдельных символов и различных их сочетаний, что позволит ему раскрыть содержимое доверительной вычислительной базы. В связи с этим целью данной работы являлась оценка степени разрушения статистически естественного языка обезличенных персональных данных при обезличивании на основе метода изменения состава или семантики с использованием набора доверительных вычислительных баз.

В качестве объекта исследования использован метод изменения состава или семантики. Данный метод выбран в качестве объекта исследования, поскольку он является одним из методов обезличивания персональных данных в соответствии с требованиями Оперативно-аналитического центра при Президенте Республики Беларусь [4].

Предметом исследования являлись количественные показатели разрушения статистики естественного языка при применении предложенной структурной схемы обезличивания персональных данных на основе метода изменения состава или семантики.

2. Обезличивание персональных данных с использованием блочной замены и набора различных ДВБ

На рисунке показана структурная схема, реализующая метод изменения состава или семантики.



Общая структурная схема функционирования компьютерной программы, реализующей схему обезличивания персональных данных с применением различных ДВБ

Сущность функционирования данной схемы заключается в следующем. Персональные данные, подлежащие обезличиванию, загружают в доверительную вычислительную базу (ДВБ), которая содержит набор таблиц подстановки ДВБ 1, ДВБ 2, …, ДВБ N. Персональные данные разбивают на блоки, каждый из которых обезличивают с помощью ДВБ 1, ДВБ 2, …, ДВБ N, ДВБ 1, ДВБ 2, …, ДВБ N и т.д.

Важно отметить, что содержимое ДВБ необходимо соблюдать сохранять в секрете и обновлять в соответствии с требованиями в месте эксплуатации информационной системы.

3. Оценка степени разрушения статистики естественного языка при обезличивании персональных данных на основе метода изменения состава или семантики

Для оценки уровня информационной безопасности обезличенных персональных данных определена вероятность появления каждого символа исходных персональных данных и персональных данных, обезличенных с использованием различных ДВБ. На основании полученных данных был вычислен индекс совпадения Фридмана. Данный метод криптоанализа был разработан американским криптографом и криптоаналитиком Уильямом Фридманом и опубликован в 1920 году. Выбор именно этого метода обусловлен его высокой чувствительностью к статистическим свойствам текста: индекс совпадения Фридмана позволяет количественно

оценить степень отклонения распределения символов от характеристик естественного языка и приближения к случайному тексту, что является ключевым критерием эффективности процедуры обезличивания персональных данных. Результаты математических вычислений представлены в табл. 1 и 2. Для данных вычислений использовалась следующая расчетная формула:

$$IC = \sum_{i=1}^{n} (p_i)^2, \quad p_i = \frac{f_i}{N}, \tag{1}$$

где IC – индекс совпадения Фридмана; p_i – вероятность появления символа в тексте; f_i – частота появления символа в тексте; N – количество символов в тексте.

Таблица 1

Индекс совпадений для исходных персональных данных

№	Символ	Количество	Вероятность появления	Вклад в индекс совпадений
1	␣	120002	0,1015	0,0103
2	(	20	0,0169	0,0003
3	)	20	0,0169	0,0003
4		30	0,0254	0,0006
5	-	60	0,0507	0,0026
...	...	...	...	...
50	я	6259	0,0053	0,0000
				0,0398

Примечание. Индекс совпадения исходных персональных данных составил 0,0398.

Таблица 2

Индекс совпадений персональных данных после процедуры обезличивания

№	Символ	Количество	Вероятность появления	Вклад в индекс совпадений
1	$	1961	0,0017	0,0000
2	+	5968	0,0050	0,0000
3	/	1189	0,0101	0,0001
4	5	949	0,0008	0,0000

Окончание табл. 2

№	Символ	Количество	Вероятность появления	Вклад в индекс совпадений
5	7	20746	0,0175	0,0003
...	...	...	...	...
120	è	1	0,0000008	0,0000
				0,0147

Примечание. Индекс совпадения персональных данных, обезличенных с применением различных ДВБ, составил 0,0147.

Полученные данные представлены в виде табл. 3.

Таблица 3

Сравнение полученных результатов

Тип данных	Индекс совпадений
Исходные	0,0398
Сложная замена	0,0147
Естественный язык (русский)	0,053
Случайный текст ($1/n$, где $n = 120$)	0,0083

Примечание. Полученные результаты показали снижение индекса совпадения более чем в 2,5 раза по сравнению с исходными персональными данными.

4. Заключение

Разработанная структурная схема процедуры обезличивания персональных данных, основанная на использовании различных ДВБ, показала свою эффективность и практическую применимость.

Результатом исследования стало количественное подтверждение эффективности метода через оценку степени разрушения статистики естественного языка персональных данных. Применение индекса совпадения Фридмана в качестве критерия оценки оказалось оправданным, поскольку данный показатель обладает высокой чувствительностью к статистическим свойствам текста и позволяет объективно измерять степень приближения обезличенных персональных данных к характеристикам случайного текста.

Полученные экспериментальные результаты демонстрируют снижение индекса совпадения Фридмана с 0,0398 для исходных данных до

0,0147 после применения процедуры обезличивания, что составляет уменьшение более чем в 2,5 раза. Это свидетельствует о существенном разрушении статистики естественного языка. Показательным является сравнение с эталонными значениями: полученный показатель (0,0147) значительно отличается от характерного для русского языка значения (0,053) и приближается к теоретическому значению для случайного текста с равномерным распределением (0,0083).

Библиографические ссылки

1. *Ворона В. А.* Биометрическая идентификация личности. М. : Горячая линия-Телеком, 2023.

2. *Коллинз М.* Защита сетей. Подход на основе анализа данных. М. : ДМК Пресс, 2020.

3. *Остапенко Г. А.* Информационные операции и атаки в социотехнических системах: организационно-правовые аспекты противодействия. М. : Горячая линия-Телеком, 2020.

4. Об изменении приказов Оперативно-аналитического центра при Президенте Республики Беларусь от 28 марта 2014 г. № 26 и от 20 февраля 2020 г. № 66 : приказ Оперативно-аналитического центра при Президенте Республики Беларусь от 10 дек. 2024 г. № 259. URL: https://www.oac.gov.by/public/content/files/files/law/prikaz-oac/2024-259.pdf (дата обращения: 24.02.2025).