(19) Евразийское патентное ведомство

(11) 044779 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2023.09.29

(21) Номер заявки
202293441

(22) Дата подачи заявки
2022.12.23

(51) Int. Cl. ***G06F 21/62*** (2013.01)
G06F 16/30 (2019.01)

(54) СПОСОБ И СИСТЕМА ОБЕЗЛИЧИВАНИЯ КОНФИДЕНЦИАЛЬНЫХ ДАННЫХ

(31) **2022132305**
(32) **2022.12.09**
(33) **RU**
(43) **2023.09.27**

(71)(73) Заявитель и патентовладелец:
ПУБЛИЧНОЕ АКЦИОНЕРНОЕ ОБЩЕСТВО "СБЕРБАНК РОССИИ" (ПАО СБЕРБАНК) (RU)

(72) Изобретатель:
Бабак Никита Григорьевич,
Белорыбкин Леонид Юрьевич,
Теренин Алексей Алексеевич,
Шаброва Анастасия Игоревна (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) WO-A1-2019018060
US-A1-20140019586
US-B1-9489354
US-A1-20120159637
US-A1-20180260734
US-A1-20210303725
US-B1-10963590
US20200334381
EA-B1-39466

(57) Изобретение, в общем, относится к области защиты данных, а более конкретно к обезличиванию конфиденциальных данных с сохранением структуры данных в текстовых документах. Решаемой технической проблемой при реализации решения является обеспечение возможности сохранения стилистической, семантической, лексической и морфологической структуры данных в текстовых документах при их обезличивании. В предпочтительном варианте реализации заявлен способ обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных, содержащий этапы, на которых получают документ, содержащий текстовые данные; сегментируют текстовые данные; токенизируют текстовые данные и присваивают порядковые позиции в тексте каждому токену; выполняют векторизацию токенов; осуществляют обработку векторных представлений токенов с помощью модели машинного обучения на базе нейронной сети, обученной на наборах конфиденциальных данных, в ходе которой осуществляется определение принадлежности каждого токена к категории конфиденциальных данных; обезличивают данные, относящиеся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных; формируют список замен, включающий указания порядковых позиций токенов с конфиденциальными данными в тексте, порядковых позиций изменения границ форматирования частей текста, обезличенные данные, соответствующие конфиденциальным данным; заменяют исходные конфиденциальные данные в текстовом документе на обезличенные данные по списку замен, причем в процессе замены обезличенные данные форматируют в соответствии с позициями изменения форматирования частей текста.

044779 B1

044779 B1

Область техники

Настоящее техническое решение, в общем, относится к области защиты данных, а более конкретно к обезличиванию конфиденциальных данных с сохранением структуры данных в текстовых документах.

Уровень техники

В настоящее время в любой организации существуют ограничения на хранение и обработку конфиденциальных данных. Так, организации обязаны хранить и обрабатывать конфиденциальные данные в определенных базах данных и обеспечивать полную сохранность таких данных, например, в соответствии с законодательными актами, такими как Федеральный закон "О персональных данных", Федеральный закон "О банках и банковской деятельности" и т.д. Конфиденциальные данные содержат персональные данные клиентов и сотрудников, сведения, составляющие банковскую и коммерческую тайну, сведения о медицинском страховании, медицинские записи и т.д., что объясняет такие высокие требования к их обработке и распространению, и, соответственно, не позволяет использовать такие данные во внешних структурах. При этом, у организаций сохраняется потребность и необходимость в применении таких данных, например, при разработке и тестировании программного обеспечения, для передачи третьим лицам, таким как агентства по переводу документов, консалтинговые, аудиторские компании, для использования в разработке моделей искусственного интеллекта и т.д. Одним из способов применения конфиденциальных данных во внешних источниках, известный из уровня техники, является обезличивание конфиденциальных данных. Так, конфиденциальные данные подвергаются необратимой модификации, исключающей возможность отнести эти данные к конкретному субъекту прямым или косвенным образом.

Однако такой подход неприменим в сферах, где требуется сохранение целостности структуры данных, таких как высокая релевантность, синтаксические и морфологические особенности данных, например, при создании AI-моделей, переводе документов подрядчиками, прототипировании аналитического приложения или витрины данных, ввиду невозможности корректного перевода и/или точного обучения модели.

Так, из уровня техники, также известно решение, раскрытое в заявке на патент США № US 2012131075 A1 (MAWDSLEY GARY [GB], et al.), опубл. 24.05.2012. Указанное решение, в частности, раскрывает способ обратимой анонимизации персональных данных, хранящихся в базе данных, в котором для подмножества элементов данных определяют отклонение каждого из указанных элементов данных в подмножестве относительно справочных элементов данных и присваивают идентификаторы отклонения каждому из указанных определенных отклонений в указанных данных для анонимизации элементов данных в указанном подмножестве данных; создают таблицу преобразования, отображающую указанные элементы данных в указанном подмножестве на указанные идентификаторы отклонения; сохраняют упомянутую таблицу перевода; и хранят упомянутые идентификаторы отклонения, определяющих упомянутые анонимизированные элементы данных.

Недостатками такого решения является низкая точность и эффективность обезличивания данных в сферах, где требуется сохранение целостности данных, из-за невозможности применения способа на неструктурированных данных, осуществления распознавания конфиденциальной информации на основе правил, что может привести к пропуску данных, которые не учтены в правилах, а также невозможности сохранения форматирования данных и их семантических и морфологических особенностей. Общими недостатками существующих решений является отсутствие точного и эффективного способа обезличивания данных в текстовых документах, обеспечивающего возможность сохранения формата стиля и структуры данных, а также позволяющего обнаруживать данные, которые раньше не встречалась. Кроме того, такого рода решение должно обеспечивать семантическую структуру данных и осуществлять проверку контрольных разрядов некоторых типов данных. Также, такое решение должно обеспечивать возможность выполнения обратимого обезличивания.

Раскрытие изобретения

Данное техническое решение направлено на устранение недостатков, присущих существующим решениям, известным из уровня техники.

Решаемой технической проблемой в данном техническом решении является создание нового и эффективного способа обезличивания конфиденциальных данных в текстовых документах.

Основным техническим результатом, проявляющимся при решении вышеуказанной проблемы, является обеспечение возможности сохранения стилистической, семантической, лексической и морфологической структуры данных в текстовых документах при их обезличивании.

Дополнительным техническим результатом, является повышение точности обезличивания данных в текстовых документах, за счет определения конфиденциальных данных в текстовых документах с помощью модели машинного обучения.

Указанные технические результаты достигаются благодаря осуществлению способа обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных, содержащего этапы, на которых:

a) получают документ, содержащий текстовые данные;

b) сегментируют текстовые данные, полученные на этапе a), причем в ходе сегментации данные разбиваются по границам изменения форматирования текста на части текста, и присваивают указанным частям текста порядковую позицию в тексте;

c) токенизируют текстовые данные, полученные на этапе a), и присваивают порядковые позиции в тексте каждому токену;

d) выполняют векторизацию токенов, полученных на этапе c);

e) осуществляют обработку векторных представлений токенов, полученных на этапе d), с помощью модели машинного обучения на базе нейронной сети, обученной на наборах конфиденциальных данных, в ходе которой осуществляется определение принадлежности каждого токена к категории конфиденциальных данных;

f) обезличивают данные, относящиеся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных;

g) формируют список замен, включающий указания порядковых позиций токенов с конфиденциальными данными в тексте, порядковых позиций изменения границ форматирования частей текста, обезличенные данные, соответствующие конфиденциальным данным;

h) заменяют исходные конфиденциальные данные в текстовом документе, на обезличенные данные по списку замен, причем в процессе замены обезличенные данные форматируют в соответствии с позициями изменения форматирования частей текста.

В одном из частных примеров осуществления способа текстовые данные получают в виде неструктурированного текстового документа.

В другом частном примере осуществления способа категории конфиденциальных данных представляют собой, по меньшей мере, одно из:

числовые сущности;

именованные сущности.

В другом частном примере осуществления способа числовые сущности представляют, по меньшей мере, одно из: основной номер держателя карты, номер ИНН, номер телефона, номер счета, номер страхового свидетельства, IP - адрес, MAC - адрес.

В другом частном примере осуществления способа именованные сущности представляют собой, по меньшей мере, одно из: персоны, локации, организации.

В другом частном примере осуществления способа структурой данных является соответствие формата, вида и смыслового содержания конфиденциальных и обезличенных данных.

В другом частном примере осуществления способа обезличенные данные для числовых сущностей генерируются с сохранением идентифицирующих признаков.

В другом частном примере осуществления способа обезличенные данные для именованных сущностей приводятся к соответствующей конфиденциальным данным морфологической форме.

В другом частном примере осуществления способа повторяющиеся конфиденциальные данные обезличиваются одинаковыми данными.

Кроме того, заявленные технические результаты достигаются за счет системы обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных, содержащих:

по меньшей мере один процессор;

по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных.

Краткое описание чертежей

Признаки и преимущества настоящего технического решения станут очевидными из приводимого ниже подробного описания и прилагаемых чертежей, на которых:

фиг. 1 иллюстрирует блок-схему выполнения заявленного способа;

фиг. 2 иллюстрирует блок-схему выполнения способа обратимого обезличивания данных;

фиг. 3 иллюстрирует блок-схему выполнения способа деобезличивания данных;

фиг. 4 иллюстрирует пример системы автоматического обезличивания конфиденциальных данных;

фиг. 5 иллюстрирует пример системы автоматического обратимого обезличивания конфиденциальных данных;

фиг. 6 иллюстрирует пример вычислительного устройства для реализации заявленных систем.

Осуществление изобретения

Заявленное техническое решение предлагает новых подход, обеспечивающий обезличивание конфиденциальных данных в текстовых документах. Основной особенностью заявленного решения является обеспечение возможности обезличивания конфиденциальных данных в текстовых документах с сохранением структуры обезличенных данных, за счет сохранения стилистической структуры конфиденциальных данных и обезличивания данных с сохранением семантической, лексической и морфологической формы данных. Кроме того, реализация настоящего технического решения, повышает точность обезличивания данных, за счет реализации модели машинного обучения на базе нейронной сети, обеспечивающей поиск конфиденциальных данных, которые раньше не встречались, в том числе в неструктурированных данных. Также, еще одним дополнительным преимуществом, достигаемым при использовании заявленного решения, является возможность выполнения как обратимого обезличивания данных, так и необратимого обезличивания данных.

Заявленное техническое решение может выполняться, например, системой, машиночитаемым носителем, сервером и т.д. В данном техническом решении под системой подразумевается, в том числе компьютерная система, ЭВМ (электронно-вычислительная машина), ЧПУ (числовое программное управление), ПЛК (программируемый логический контроллер), компьютеризированные системы управления и любые другие устройства, способные выполнять заданную, четко определенную последовательность операций (действий, инструкций).

Под устройством обработки команд подразумевается электронный блок либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы). Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройства хранения данных, например, таких устройств, как оперативно запоминающие устройства (ОЗУ) и/или постоянные запоминающие устройства (ПЗУ). В качестве ПЗУ могут выступать, но, не ограничиваясь, жесткие диски (HDD), флеш-память, твердотельные накопители (SSD), оптические носители данных (CD, DVD, BD, MD и т.п.) и др.

Программа - последовательность инструкций, предназначенных для исполнения устройством управления вычислительной машины или устройством обработки команд.

Термин "инструкции", используемый в этой заявке, может относиться, в общем, к программным инструкциям или программным командам, которые написаны на заданном языке программирования для осуществления конкретной функции, такой как, например, получение и обработка данных, формирование профиля пользователя, прием и передача сигналов, анализ принятых данных, идентификация пользователя и т.п. Инструкции могут быть осуществлены множеством способов, включающих в себя, например, объектно-ориентированные методы. Например, инструкции могут быть реализованы, посредством языка программирования C++, Java, Python, различных библиотек (например, "MFC"; Microsoft Foundation Classes) и т.д. Инструкции, осуществляющие процессы, описанные в этом решении, могут передаваться как по проводным, так и по беспроводным каналам передачи данных, например, Wi-Fi, Bluetooth, USB, WLAN, LAN и т.п.

На фиг. 1 представлена блок схема способа 100 обезличивания конфиденциальных данных в текстовых документах, который раскрыт поэтапно более подробно ниже. Указанный способ 100 заключается в выполнении этапов, направленных на обработку различных цифровых данных. Обработка, как правило, выполняется с помощью системы, которая может представлять, например, сервер, компьютер, мобильное устройство, вычислительное устройство и т.д. Более подробно элементы системы раскрыты на фиг. 6. В одном частном варианте осуществления, также может выполняться системой 400, которая более подробно раскрыта ниже.

Под термином конфиденциальные данные в данном решении стоит понимать данные, доступ к которым ограничен в соответствии с политиками безопасности организаций и/или законодательными актами. Так, конфиденциальные данные могут представлять персональные данные, сведения, составляющие банковскую и коммерческую тайну, ФИО сотрудников и клиентов, партнеров и поставщиков, адреса, номера телефонов, адреса электронных почтовый ящиков, номера социального страхования, информация о банковских картах, номер ИНН, регистрационный номер машины, номер БИК, IP-адрес, данные геолокации, номер документа о браке, номер документа об образовании, дата, URL-адрес, MAC-адрес, номер трудовой книжки, номер военного билета, код ОКПО и т.д., не ограничиваясь.

Обезличивание данных - действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность конфиденциальных данных конкретному субъекту конфиденциальных данных.

На этапе 110 происходит получение документа, содержащего текстовые данные.

На указанном этапе 110 документ, содержащий текстовые данные, поступает в систему, например, систему 400. В одном частном варианте осуществления, документ может быть загружен в систему 400 посредством сети связи, такой как Интернет, ЛВС и т.д. В еще одном частном варианте осуществления, документ может быть импортирован непосредственно из флэш-накопителя и/или встроенной памяти системы 400. Под документом, содержащим текстовые данные в данном решении следует понимать любой файл данных, содержащий текстовые данные. Так, документ может представлять собой любой неструктурированный документ, например, файл формата Word, PDF, текстовый документ, фотографию, оцифрованный документ, файл электронной почты и т.д.

Далее способ 100 переходит к этапу 120.

На этапе 120 сегментируют текстовые данные, полученные на этапе 110, причем в ходе сегментации данные разбиваются по границам изменения форматирования текста на части текста, и присваивают указанным частям текста порядковую позицию в тексте.

Как уже отмечалось выше, стиль форматирования является существенным в определенных задачах, связанных с обработкой данных, например, при разработке моделей искусственного интеллекта (ИИ), переводе текста и т.д. При этом, известные системы обезличивания данных не предполагают и не содержат такой возможности в связи с тем, что стили форматирования текста не привязаны непосредственно к тексту, а содержатся в сопутствующих данных документа, например, в метаданных. Кроме того, некоторые модели машинного обучения также учитывают форматирование данных, что соответственно повышает точность самого процесса обучения.

Соответственно, для реализации указанной возможности в заявленном способе предложен следующий подход. Указанный подход, в одном частном варианте осуществления, решает, как проблему предоставления сторонним разработчикам обезличенных данных, так и обеспечивает формирование точной и максимально приближенной, в том числе и по стилю форматирования, выборки данных для обучения модели.

После получения текстового документа системой 400, указанная система 400, извлекает текст из документа частями так, чтобы сохранить исходные стили форматирования документа. Под извлечением текста из документа в данном решении следует понимать выделение из текстового документа определенных частей текста и сохранение указанных частей в памяти системы 400, например, в виде файла данных, т.е. на указанном этапе 120 исходный текст преобразуется в несколько частей, разбитых по границам изменения форматирования. Под форматированным текстом в данном решении понимается результат процесса оформления страницы, абзаца, строки, символа документа. Так, стилем форматирования текста может являться изменение цвета, подчеркивание, выделение курсивом, начертание, изменение отступов, регистров и т.д., определенных слов, символов, абзацев в текстовом документе. Стиль форматирования текста может определяться посредством разметки XML ("расширяемый язык разметки"), которая может хранится, например, в файле текстового документа, являться машиночитаемым приложением к текстовому документу и т.д. Стоит отметить, что особенностью данного этапа является то, что текст разбивается не на целые слова и/или токены окруженные разметкой XML, а разбивается именно по границам изменения форматирования разметки. Такой подход обеспечивает избегание нарушения стилей и дальнейшее неправильное применение стилей к обезличенным частям.

Рассмотрим пример извлечения текста в соответствии с указанным этапом 120. Так, следующий текст: "Иванов Иван получил карту 4485 2911 2679 2812 в ООО "Бетта". Теперь Ивану доступны онлайн платежи, с чем его поздравил Андрей." будет преобразован на несколько частей, разбитых по границам изменения форматирования (табл. 1).

Таблица 1

Части текста
Иванов Иван получил карту
4485
2911
2679
2812
в ООО «
Б
етта»
.
Теперь
Ивану
доступны онлайн платежи
, с чем его поздравил Андрей.

Как видно из табл. 1, символ Б является частью слова Бетта, однако имеет другой стиль форматирования, и, соответственно, вынесен в отдельную часть текста. Для специалиста в данной области техники будет очевидно, что, хотя пример приведен на основе регистра и подчеркивания символа, данный подход применим для любого изменения форматирования текста.

После получения частей текста, указанным частям присваиваются индексы начала и конца в рамках общего текста. Указанные индексы в дальнейшем необходимы для преобразования обезличенных данных в исходный стиль форматирования.

В качестве инструмента определения начала и конца части текста (длина текста), могут использоваться как внутренние средства документа, обеспечивающие подсчет символов в тексте, например, текстовые редакторы, так и инструменты, например, программные средства для анализа текстовых документов, в том числе и встроенные в систему 400.

Итак, продолжая предыдущий пример, указанным частям текста будут присвоены следующие индексы (табл. 2).

Таблица 2

Части текста	**Индексы**
Иванов Иван получил карту	[0, 26]
4485	[26, 30]
	[30, 31]
2911	[31, 35]
	[35, 36]
2679	[36, 40]
	[40, 41]
2812	[41, 45]
	[45, 46]
в ООО «	[46, 53]
Б	[53, 54]
етта»	[54, 59]
.	[59, 60]
Теперь	[60, 68]
Ивану	[68, 73]
доступны онлайн платежи	[73, 97]
, с чем его поздравил Андрей.	[97, 126]

Таким образом, на указанном этапе 120 выполняют сегментацию текстовых данных и присваивают указанным частям текста порядковую позицию в тексте.

Далее способ 100 переходит к этапу 130.

На этапе 130 выполняют токенизацию текстовых данных, полученных на этапе 110, и присваивают порядковые позиции в тексте каждому токену.

На этапе 130 выполняется обработка полученных текстовых данных. Входной текст токенизируется, т.е. сегментируется на части, например, на предложения, слова или символы. Под токеном в данном решении следует понимать последовательность символов в документе, которая имеет значение для анализа. Так, токенами могут являться, например, отдельные слова, части слов и т.д. Токенизация на предложения может проводиться при помощи, например, лексических анализаторов, таких как razdel, rusenttokenize, NLTK и т.д. Лексический анализ - это процесс аналитического разбора входной последовательности символов на распознанные группы (лексемы) с целью получения на выходе идентифицированных последовательностей, называемых "токенами". Кроме того, токенизация входного текста может быть осуществлена на основе регулярных выражений. Для специалиста в данной области техники очевидно, что может быть применен любой лексический анализатор, известный из уровня техники, и данное решение не ограничивается приведенными выше примерами. После процесса токенизации, по аналогии с присвоением индексов начала и конца на этапе 120, токенам текста присваиваются начальный и конечный индекс. Как указывалось выше, определение позиции слова в текстовом документе может быть выполнено, например, при помощи анализаторов текста.

При этом, стоит отметить, что в процессе токенизации текста стили форматирования текста будут стерты, т.к. указанный процесс не может учитывать стиль текста в документе.

Продолжая приведенный выше пример, представленный текст будет разбит на следующие токены, которым будут присвоены следующие индексы (табл. 3).

Таблица 3

Токен	Начальный индекс	Конечный индекс
Иванов	0	6
Иван	7	11
получил	12	19
карту	20	25
4485	26	30
2911	31	35
2679	36	40
2812	41	45
в	46	47
ООО	48	51
«	52	53
Бетта	53	58
»	58	59
.	59	60
Теперь	61	67
Ивану	68	73
доступны	74	82
онлайн	83	89
платежи	90	97
,	97	98
с	99	100
чем	101	104
его	105	108
поздравил	109	118
Андрей	119	125
.	125	126

На этапе 140 выполняют векторизацию токенов, полученных на этапе 130.

На указанном этапе 140 выполняется векторизация каждого токена, полученного в процессе токенизации, например, с помощью эмбеддингов (embeddings) или прямого кодирования (one hot encoding). Так, например, при токенизации, каждый токен, представлен в словаре своим индексом, отображающий позицию в указанном словаре. Таким образом, каждый токен представляет индекс в словаре, и, соответственно процесс векторизации осуществляется путем замены каждого токена на его индекс в словаре. Затем индексы группируются с учётом разряженности словаря и семантической близости токенов. Для специалиста в данной области техники будет очевидно, что для векторизации токенов могут применять и другие алгоритмы векторизации, например, с помощью алгоритмов TransformersBertEmbedder, Word2vec, fastText и т.д., не ограничиваясь. Указанный процесс векторизации является подготовительным этапом к обработке данных моделью машинного обучения, выполняющей распознавание конфиденциальных данных (этап 150).

На этапе 150 осуществляют обработку векторных представлений токенов, с помощью модели машинного обучения на базе нейронной сети, обученной на наборах конфиденциальных данных, в ходе которой осуществляется определение принадлежности каждого токена к категории конфиденциальных данных.

Для обработки неструктурированных данных и определения в неструктурированном тексте конфиденциальных данных была разработана и применена модель машинного обучения. Указанная модель предназначена для решения задачи классификации текста (задача NER). Распознавание именованных сущностей (Named-entity recognition, NER) - это подзадача извлечения информации, которая направлена на поиск и классификацию упоминаний именованных сущностей в неструктурированном тексте по заранее определенным категориям, таким как имена лиц, организации, местоположения, денежные значения, проценты и т.д.

Обучение модели МО производилось на заранее размеченных данных. На момент создания модели был использован датасет (набор данных) из размеченных конфиденциальных данных, состоящий из более чем 2 миллионов токенов. Набор данных сформирован по оригинальным текстовым документам, содержащим персональные данные, сведения, составляющие банковскую и коммерческую тайну, и т.д. Обучение модели распознаванию конфиденциальных данных в текстовых документах заключалось в определении классов именованных сущностей в тексте (ФИО, номер банковской карты, номер социального страхования и т.д.). В качестве схемы для разметки данных может применятся, например, BIO/IOB-схема, BILUO-схема и т.д., не ограничиваясь. Датасет может быть представлен, например, в виде упорядоченного списка токенов, отделённых разделителем от класса именованной сущности, например, символом пробела. Для специалиста в данной области техники очевидно, что может быть применена любая схема разметки и представления датасета известная из уровня техники и данное решение не ограничивается приведенными выше примерами. Для обучения модели МО использовалось 80% токенов из датасета, а для расчёта метрик качества - 20%.

Основная метрика качества модели МО, взвешенная по классам F1 мера, составляет около 95,3%, что превышает результаты автоматического машинного обучения (AutoML) более чем на 15%. При этом

разброс между полнотой (recall) и точностью (precision) не превышает 0,8%, что говорит о высоком уровне качества модели как с точки зрения безопасности - низкая вероятность пропустить конфиденциальную информацию, так и с точки зрения конечного пользователя - малое количество ложных обнаружений конфиденциальной информации. Степень падения метрики качества при аугментации текстов на уровне символов, показывающая стабильность модели, не превышает 5%.

Указанная модель машинного обучения на базе нейронной сети была успешно внедрена и протестирована в организации в рабочих процессах подразделений.

В одном частном варианте осуществления, в качестве нейронной сети может быть использована, например, нейронная сеть архитектуры трансформер (Transformer), рекуррентная нейронная сеть (RNN) и т.д., не ограничиваясь. Особенностью указанной модели машинного обучения является возможность эффективного распознавания конфиденциальных данных в текстовых документах, в том числе благодаря анализу семантической близости токенов.

Таким образом, на этапе 150 выполняют, с помощью представленной обученной модели машинного обучения, обработку данных, осуществляя распознавание конфиденциальных данных. Продолжая пример, результат обработки данных будет иметь следующий вид (табл. 4).

Таблица 4

Токен	**Тег**	**Начальный индекс**	**Конечный индекс**
Иванов	LASTNAME	0	6
Иван	NAME	7	11
получил	O	12	19
карту	O	20	25
4485	CARD	26	30
2911	CARD	31	35
2679	CARD	36	40
2812	CARD	41	45
в	O	46	47
ООО	OPF	48	51
«	O	52	53
Бетта	UL	53	58
»	O	58	59
.	O	59	60
Теперь	O	61	67
Ивану	NAME	68	73
доступны	O	74	82
онлайн	O	83	89
платежи	O	90	97
,	O	97	98
с	O	99	100
чем	O	101	104
его	O	105	108
поздравил	O	109	118
Андрей	NAME	119	125
.	O	125	126

Так, из указанных текстовых данных извлекаются конфиденциальные данные с присвоенной категорией, к которой указанные данные относятся. Например, как видно из таблицы 4, все данные, относящиеся к конфиденциальным, помечены тегами - короткими строками, которые взаимно однозначно соответствуют видам конфиденциальных данных. Тэги соответствуют категории, присвоенной указанным данным, например, CARD - номер карты, NAME - имя и т.д. Тэги пишутся на латинице, для того, чтобы они имели общий вид на всех кодировках.

В одном частном варианте осуществления, категории конфиденциальных данных, которые присваиваются текстовым данным представляют собой, по меньшей мере категорию данных числовых сущностей и категорию данных именованных сущностей. Указанный этап может выполняться моделью машинного обучения и является грубой классификацией, когда данные разделяют на числовые и именованные сущности, для последующей идентификации типа конфиденциальных данных, например, с помощью мультиклассовой классификации (Multiclass classification), распознавания именованных сущностей (NER) и т.д. Указанное разделение необходимо для выбора корректного алгоритма обезличивания, соответствующего типу конфиденциальных данных, и для установления более жёстких требований к сохранению морфологических особенностей именованных сущностей. Так, после отнесения данных к числовым, может быть определен тип числовых данных. Числовые сущности могут представлять, следующие данные: Основной номер держателя карты (PAN), Номер ИНН ФЛ, Номер ИНН ЮЛ, Номер дома, Номер ДУЛ, Номер счёта ФЛ, Номер счёта ЮЛ, Телефон, Номер ОГРНИП, Номер ОГРН, Номер СНИЛС, Номер полиса ОМС, Паспорт транспортного средства, Номер БИК, IP-адрес, Номер социального страхования SNN, Идентификационный номер работодателя EIN, Индивидуальный номер налогоплательщика ITIN, Данные геолокации, Номер документа о браке, Номер документа об образовании, Дата, Номер трудовой книжки, Номер военного билета, Код ОКПО, IMEI и т.д. Для повышения точности распознавания

конфиденциальных данных, в настоящем решении, также применяется процедура проверки контрольных разрядов к числовым сущностям. Алгоритм проверки контрольных разрядов проверяет данные на соответствие контрольным разрядам, которые обычно вычисляются с помощью алгоритма Луна или других алгоритмов. Алгоритм Луна - алгоритм вычисления контрольной цифры некоторых видов данных. Не является криптографическим средством, а предназначен в первую очередь для выявления ошибок, вызванных непреднамеренным искажением данных. Контрольный разряд используется в различных номерах, таких как: номера банковских карт, СНИЛС, ОКПО, ОГРН, ИНН и т.д. не ограничиваясь. Контрольный разряд необходим, для того, чтобы исключить вероятность некорректного распознавания типа конфиденциальных данных.

Соответственно, именованные сущности также могут быть поделены по типу сущностей. Так, именованные сущности могут представлять: персоны, локации, организации, например, Фамилия, Имя, Отчество, Должность, Полное и краткое наименование юридического лица, логин, улица, город и т.д.

Указанные извлеченные сущности могут сохраняться, например, в защищенной памяти системы 400 для дальнейшего их обезличивания. В одном частном варианте осуществления, токены с одинаковыми тегами конфиденциальных данных объединяются в единый токен (например, токены, характеризующие номер банковской карты). Так, токены, принадлежащие тегу CARD могут быть объединены в единый токен. Кроме того, в еще одном частном варианте осуществления, одинаковые токены с одинаковым классом также могут быть идентифицированы и помечены для дальнейшей заменены на одинаковые обезличенные значения, например, одно и тоже встречающееся имя будет обезличено одним и тем же значением.

Таким образом, на указанном этапе 150 из текстового документа извлекаются все сущности, содержащие конфиденциальную информацию.

Далее, на этапе 160 выполняется обезличивание данных, относящихся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных.

Так, после определения типа конфиденциальных данных, способ 100 переходит к этапу 160. На этапе 160 выполняется обезличивание определенных конфиденциальных данных с сохранением структуры данных. Основной особенностью заявленного технического решения является возможность обезличивания данных с сохранением морфологических особенностей, структуры текста и с заменой одинаковых сущностей на одинаковые обезличенные значения. Указанная возможность позволяет использовать обезличивание для тех задач, где важны морфологические особенности и смысл текста, например, обучение моделей искусственного интеллекта, перевод текста и т.д. Под сохранением структуры текста понимается псевдонимизация конфиденциальных данных, при которой конфиденциальные или личные данные заменяются данными того же типа, пола и языка/региона. Например, женское имя заменяется другим женским именем, распространенным в местной культуре, улица заменяется на улицу со схожей морфологической и фонетической структурой, например, Екатерина заменяется на Елизавету, и т.д. Для указанного процесса на основе типа конфиденциальных данных в словаре, например, базе данных системы 400, содержащий набор данных для обезличивания, определяются схожие атрибуты с фрагментом текстовых данных, подлежащему замене. Например, при сопоставлении атрибутов Иванова Ивана, атрибут Елена Смит будет обладать низкой степенью схожести, т.к. указанные данные обладают разными классами (пол. муж. и жен.) и национальностями (Смит этнически принадлежит к немецкой национальности). При этом, атрибут Петров Пётр будет иметь высокую степень схожести, т.к. пол и национальность совпадает. Как указывалось выше, определение типа данных в категории может быть выполнено на этапе 150 с помощью модели машинного обучения. Кроме того, в одном частном варианте осуществления, часть текстовых данных может быть заменена данными с равнозначным количеством букв, например, слово "два" может быть обезличено как "три". Указанная особенность может быть достигнута посредством заранее размеченной базы данных, содержащей указания на вышеперечисленные структурные особенности текста (например, база данных может хранить разделы, относящиеся к персональным данным по национальностям, половому признаку и т.д.), т.е. база данных может представлять собой таблицу с идентифицированными именем столбца и типом данных.

Кроме того, в еще одном частном варианте осуществления, конфиденциальные данные, представленные датами, могут сдвигаться на фиксированное смещение, например, на 1, 10 дней и т.д. Указанная особенность позволяет получить репрезентативные данные, например, для обучающего набора данных, где дата совершения события может являться критичной, например, в задачах поиска аномальных событий.

Стоит отметить, что для указанных задач анонимизации в системе 400 дополнительно могут содержаться классификаторы типов данных, например, контекстные классификаторы, классификаторы морфологического анализа, комбинация указанных классификаторов и т.д. Указанные классификаторы выполнены с возможностью определения конкретного типа данных (подкласса данных), к которым относятся данные (национальность, географическая принадлежность и т.д.). Более подробно пример классификатора раскрыт в статье, найденной в Интернет, см. https://arxiv.org/abs/1708.07903. Так, например, классификаторы морфологического анализа, такие как pymorphy2 и др., могут включать в себя класси-

фикатор типов на основе префиксов и суффиксов или классификатор типов на основе подслов и составных слов. Классификация по национальности и этнической принадлежности может осуществляться, например, с помощью NamePrism и других классификаторов, использующих закономерности гомофилии, но не ограничивающихся ими.

Так, продолжая этап 160, обезличенные значения с сохранением косвенно идентифицирующих признаков генерируются для токенов, относящихся к типу и категории конфиденциальных данных. В одном частном варианте осуществления, для числовых сущностей значения могут генерироваться на основе заданных шаблонов. Так, например, для номеров телефонов, шаблон будет содержать код страны и количество цифр номера. Например, для России код страны является +7, соответственно номера телефонов, содержащиеся в конфиденциальных данных и начинающиеся на +7, будут заменены с сохранением указанного кода. Для специалиста в данной области техники очевидно, что указанные особенности могут быть сохранены и для других числовых сущностей (ИНН, социальное страхование и т.д.).

В еще одном частном варианте осуществления, как указывалось выше, именованные сущности обезличиваются в соответствии с нужной морфологической формой для токенов, относящихся к конфиденциальным данным с помощью хранилища обезличенных данных (например, базы данных). Так, указанная база данных или словарь изначально наполнены уникальными данными в нормальной форме и сгруппированы по видам конфиденциальной информации: Имя, Фамилия, Должность и т.д.

Так, данные из текстового документа, приведенные в примере выше, будут заменены на следующие данные (табл. 5).

Таблица 5

Токен	**Тег**	**Начальный индекс**	**Конечный индекс**	**Обезличенное значение**
Иванов	LASTNAME	0	6	Петров
Иван	NAME	7	11	Петр
получил	O	12	19	
карту	O	20	25	
4485 2911 2679 2812	CARD	26	45	4485 2952 7892 2812
в	O	46	47	
ООО	OPF	48	51	
«	O	52	53	
Бетта	UL	53	58	Гамма
»	O	58	59	
.	O	59	60	
Теперь	O	61	67	
Ивану	NAME	68	73	Петру
доступны	O	74	82	
онлайн	O	83	89	
платежи	O	90	97	
,	O	97	98	
с	O	99	100	
чем	O	101	104	
его	O	105	108	
поздравил	O	109	118	
Андрей	NAME	119	125	Потап
.	O	125	126	

Как видно из таблицы, тип данных CARD, был обезличен с сохранением структуры, а именно были сохранены первые 6 цифр, обозначающие банковский идентификационный номер, и последние 4 цифры, служащие для идентификации карты и проверки её подлинности с помощью контрольного разряда. Кроме того, имена, фамилии, названия юридического лица также были обезличены с сохранением структуры данных.

Стоит отметить, что в одном частном варианте реализации, при обратимом процессе обезличивания данных, таблица замен сохраняется в защищенное хранилище данных, памяти 400.

Таким образом, на указанном этапе 160 выполняют обезличивание данных, относящихся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных (пола, национальной принадлежности и т.д.).

Далее способ переходит к этапу 170.

На этапе 170 формируют список замен, включающий указания порядковых позиций токенов с конфиденциальными данными в тексте, порядковых позиций изменения границ форматирования частей текста, обезличенные данные, соответствующие конфиденциальным данным.

На этапе 170, на основе сформированной таблицы замен, таблицы форматирования и порядковых позиций токенов в тексте (индексы начала и конца), составляется список замен на обезличенные значения для частей текста. Список замен содержит указание на порядковый номер части текста, граничные индексы замены в рамках части текста, вид найденных конфиденциальных данных и обезличенное значение. Если конфиденциальный токен представлен несколькими частями текста, то обезличенное значение разбивается на соответствующее количество частей.

Указанный список также может быть сформирован в системе 400.

Список замен для приведенного примера будет выглядеть следующим образом (табл. 6).

Таблица 6

ID части текста	Тег	Начальный индекс	Конечный индекс	Обезличенное значение
0	LASTNAME	0	6	Петров
0	NAME	7	11	Петр
1	CARD	0	4	4485
2	CARD	0	1	
3	CARD	0	4	2952
4	CARD	0	1	
5	CARD	0	4	7892
6	CARD	0	1	
7	CARD	0	4	2812
10	UL	0	1	Г
11	UL	0	4	амма
14	NAME	0	5	Петру
16	NAME	22	28	Потап

На этапе 180 заменяют исходные конфиденциальные данные в текстовом документе, на обезличенные данные по списку замен, причем в процессе замены обезличенные данные форматируют в соответствии с позициями изменения форматирования частей текста.

Так, в исходном документе части текста заменяются на обезличенные значения по списку замен, который был сформирован на этапе 170. Таким образом сохраняется исходное форматирование документа.

Например, текст из примера после обезличивания будет выглядеть следующим образом: "Петров Петр получил карту 4485 2952 7892 2810 в ООО "Гамма". Теперь Петру доступны онлайн платежи, с чем его поздравил Потап."

В одном частном варианте осуществления, если выбран обратимый режим обезличивания, то, как указывалось выше, в системе 400 сохраняется таблица подстановок, например, в защищенной области памяти. В этом случае система 400 выполнена с возможностью восстановления исходных сущностей из обезличенных значений.

После замены частей текста, содержащих конфиденциальные данные, на обезличенные данные, система 400 выполнена с возможностью формирования нового текстового документа и предоставления указанного документа пользователю системы и/или выгрузки документа в сеть Интернет.

Рассмотрим более подробно вариант реализации заявленного изобретения, относящийся к обратимому обезличиванию.

Более подробно способ 200 обратимого обезличивания конфиденциальных данных в текстовых документах с сохранением структуры показан на фиг. 2.

На этапе 210 система, такая так система 500 получает текстовый документ. Указанный этап аналогичен этапу 110, и, соответственно, для получения документа могут быть использованы приведенные выше средства.

На этапе 220 выполняется сегментация текстовых данных, полученные на этапе 210, причем в ходе сегментации данные разбиваются по границам изменения форматирования текста на части текста, и присваивают указанным частям текста порядковую позицию в тексте.

На этапе 230 выполняют токенизацию текстовых данных, полученные на этапе 210, и присваивают порядковые позиции в тексте каждому токену.

На этапе 240 выполняют векторизацию токенов, полученных на этапе 230.

На этапе 250 осуществляют обработку векторных представлений токенов, полученных на этапе 240, с помощью модели машинного обучения на базе нейронной сети, обученной на наборах конфиденциальных данных, в ходе которой осуществляется определение принадлежности каждого токена к категории конфиденциальных данных.

На этапе 260 обезличивают данные, относящиеся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных.

На этапе 270 формируют таблицу подстановок, характеризующую соответствие конфиденциальных данных и обезличенных данных, и сохраняют указанную таблицу в память. В частности, указанную таблицу подстановок сохраняют в защищенную память системы 500, например, в хранилище данных 506. Указанная особенность позволяет в дальнейшем деобезличить обезличенные данные, т.е. восстановить исходные данные в документе.

На этапе 280 формируют список замен, включающий указания порядковых позиций конфиденциальных данных в тексте, порядковых позиций изменения границ форматирования частей текста, обезличенные данные, соответствующие конфиденциальным данным. Указанный список также может быть сохранен в хранилище 406.

На этапе 290 заменяют исходные конфиденциальные данные, на обезличенные данные по списку замен, причем в процессе замены обезличенные данные форматируют в соответствии с позициями изме-

нения форматирования частей текста. После замены конфиденциальных данных на обезличенные, система 500 может сформировать новый текстовый документ.

Как было указано выше, основной особенностью обратимого обезличивания является возможность восстановления в текстовом документе исходных конфиденциальных данных. Такая потребность может возникать, например, при переводе текста за пределами организации. Так, если требуется перевести договор, содержащий конфиденциальные данные, наличие таких данных будет влиять на контекст перевода, и, соответственно, нерелевантные обезличенные данные могут привести к некорректному переводу. Кроме того, после предоставления переведенного документа, требуется восстановить исходные данные. Соответственно, в еще одном частном варианте осуществления заявленное решение раскрывает способ 300 деобезличивания данных, раскрытый более подробно на фиг. 3.

Так, указанный способ 300 может являться обратным процессом к этапам способа 200.

На этапе 310 в систему, например, систему 500 поступает текстовый документ в виде файла данных, обезличенный в соответствии с этапами способа 200. В одном частном варианте, документ может быть загружен в систему 500.

На этапе 320 система 500, определяет в текстовом документе обезличенные данные, которые необходимо деобезличить, с помощью таблицы подстановок.

Указанный этап 320 основан на таблице подстановок, сохраненной в процессе обезличивания документа. Так, в текстовом документе определяются все сущности, содержащиеся в таблице подстановок, например, ФИО, название организации и т.д. и для последующей замены на конфиденциальные данные, которые содержались в исходном документе. Стоит отметить, что процесс определения сущностей, подлежащих деобезличиванию, может быть выполнен с помощью поиска порядковых индексов по указанной таблице подстановок.

Далее, на этапе 330 осуществляют процесс замены обезличенных данных на конфиденциальные данные в соответствии с таблицей подстановок.

На указанном этапе 330 все обезличенные значения заменяют на исходные конфиденциальные данные. В ходе процесса деобезличивания стиль документа не затрагивается, что обеспечивает получение исходного документа с сохранением стиля форматирования.

Так, например, после преобразования обезличенного документа третьими лицами, указанный документ загружают в систему 500. Т.е. получают текстовый документ, обезличенный в соответствии с этапами способа 200. Далее, на основе списка замен, хранящегося в системе, определяют в текстовом документе обезличенные данные, которые необходимо деобезличить, с помощью таблицы подстановок. После этого, осуществляют процесс замены обезличенных данных на конфиденциальные данные в соответствии с таблицей подстановок.

Таким образом, в вышеприведенных материалах был описан способ обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных.

Теперь рассмотрим сценарий применения некоторых вариантов заявленного решения.

Так, одним из сценариев применения может являться перевод подрядчиками договоров, соглашений и других документов организации. Как было отмечено ранее, для осуществления перевода, переводчику требуется контекст и смысловое содержание данных, кроме того, при переводе также учитываются лексические и морфологические особенности текста.

Таким образом, для реализации заявленного изобретения, документ, подлежащий переводу, загружается посредством интерфейса пользователя (например, интерфейса 501) в систему, такую как система 500. Далее, система 500 сегментирует текстовые данные, например, в соответствии с этапом 120 или этапом 220, токенизирует и векторизует исходные данные (этапы 130 и 140 или этапы 230 и 240) и выполняет обработку векторизованных токенов посредством модели машинного обучения, обученной на распознавание конфиденциальных данных (этап 150 или этап 250). Причем конфиденциальные данные распознаются по типам и категориям данных (этап 160 или этап 260). После этого, выполняется обезличивание конфиденциальных данных и формирование списка замен (этапы 170 и 180). Кроме того, если выбрано обратимое обезличивание, то в таком случае таблица подстановок дополнительно сохраняется в защищенную память системы, например, системы 500. (этапы 270 и 280). Последним этапом работы системы 500 является осуществление замены конфиденциальных данных в соответствии со списком замен в текстовом документе (этап 190 или этап 290). Как видно из представленного описания примера, указанный способ может быть выполнен и системой 400, в случае необратимого обезличивания. При этом, обезличенные конфиденциальные данные сохраняют структуру данных, а именно, стиль форматирования, морфологические, лексические и семантические особенности. Сформированный новый документ с обезличенными данными далее может быть отображен в пользовательском интерфейсе и/или выгружен в виде файла данных. Указанный файл далее может быть направлен переводчикам.

Соответственно, после выполнения перевода, измененный файл данных может быть загружен в систему для замены обезличенных данных на исходные конфиденциальные данные.

Кроме того, стоит отметить, что сохранение стилей форматирования является критичным для некоторых сфер. Так, например, некоторые автоматизированные системы хранят документы во внутреннем представлении, причем стиль документа учитывается указанными системами, поэтому если стили доку-

ментов будут отличаться, то система не сможет соотнести исходный и обезличенный документ. Также, стили в документах могут использоваться как часть стандартизированных форм или для выделения ключевой информации, например, без сохранения стилей переводчик не сможет применять их в своей работе, а значит не сможет выдать корректный перевод. Также, стиль форматирования может учитываться в области разработки искусственного интеллекта. Так, кроме самого этапа обучения существует этап парсинга документа и подготовки данных. Если не сохранять стили, то они и не будут учитываться при парсинге, что в дальнейшем приведёт к некорректной подготовке данных и неправильному обучению моделей. Т.е. одним из преимуществ раскрытого изобретения является возможность встраивания нового инструмента в уже существующие процессы, где происходит взаимодействие с документами, имеющими определенные стили. Удаление этих стилей может привести к нарушению работы автоматизированных систем, снижению качества обработки информации, нарушению стандартов.

На фиг. 6 представлена система 600, реализующая этапы заявленного способа (100).

В общем случае система 600 содержит такие компоненты, как: один или более процессоров 601, по меньшей мере одну память 602, средство хранения данных 603, интерфейсы ввода/вывода 604, средство В/В 605, средство сетевого взаимодействия 606, которые объединяются посредством универсальной шины.

Процессор 601 выполняет основные вычислительные операции, необходимые для обработки данных при выполнении способа 100. Процессор 601 исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти 602. Память 602, как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

Средство хранения данных 603 может выполняться в виде HDD, SSD дисков, рейд массива, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средства 603 позволяют выполнять долгосрочное хранение различного вида информации, например, истории таблицы подстановок, таблицы форматирования и т.п.

Для организации работы компонентов системы 600 и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В 604. Выбор соответствующих интерфейсов зависит от конкретного исполнения вычислительного устройства, которые могут представлять собой, не ограничиваясь: PCI, AGP, PS/2, IrDa, FireWire, LPT, COM, SATA, IDE, Lightning, USB (2.0, 3.0, 3.1, micro, mini, type C), TRS/Audio jack (2.5, 3.5, 6.35), HDMI, DVI, VGA, Display Port, RJ45, RS232 и т.п.

Выбор интерфейсов 604 зависит от конкретного исполнения системы 600, которая может быть реализована на базе широко класса устройств, например, персональный компьютер, мейнфрейм, ноутбук, серверный кластер, тонкий клиент, смартфон, сервер и т.п.

В качестве средств В/В данных 605 может использоваться: клавиатура, джойстик, дисплей (сенсорный дисплей), монитор, сенсорный дисплей, тач-пад, манипулятор мышь, световое перо, стилус, сенсорная панель, трекбол, динамики, микрофон, средства дополненной реальности, оптические сенсоры, планшет, световые индикаторы, проектор, камера, средства биометрической идентификации (сканер сетчатки глаза, сканер отпечатков пальцев, модуль распознавания голоса) и т.п.

Средства сетевого взаимодействия 606 выбираются из устройств, обеспечивающий сетевой прием и передачу данных, например, Ethernet карту, WLAN/Wi-Fi модуль, Bluetooth модуль, BLE модуль, NFC модуль, IrDa, RFID модуль, GSM модем и т.п. С помощью средств 606 обеспечивается организация обмена данными между, например, системой 600, представленной в виде сервера и вычислительным устройством пользователя, на котором могут отображаться полученные данные (обезличенный текстовый документ) по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

На фиг. 4 показан частный случай реализации системы 600.

Указанная система 400 состоит из графического пользовательского интерфейса 401, модуля управления 402, модуля обезличивания 403, модуля распознавания конфиденциальной информации 404.

Модули 402-404 могут являться программно-аппаратными средствами и могут представлять собой, по меньшей мере, сервер, компьютер и т.д., выполняющий предписанную ему функцию. Кроме того, для специалиста очевидно, что указанные модули могут быть реализованы с помощью по меньшей мере одного процессора и могут являться логическими модулями. Так, модуль распознавания конфиденциальной информации (КИ) 404 может содержать модель машинного обучения и предназначен для определения категорий и типов конфиденциальной информации (этап 150). Модуль 403 соответственно выполнен с возможностью замены определенных конфиденциальных данных на обезличенные данные. В одном частном варианте осуществления, указанный модуль 403 также может содержать средства классификации типов конфиденциальной информации и базу данных с наборами данных для обезличивания.

На фиг. 5 показан еще один частный случай реализации системы 600.

Указанная система 500 состоит из графического пользовательского интерфейса 501, модуля управления 502, модуля обезличивания 503, модуля распознавания конфиденциальной информации 504, модуля управления таблицей подстановок 505, хранилища данных 506.

Модули 502-506 могут являться программно-аппаратными средствами и могут представлять собой, по меньшей мере, сервер, компьютер и т.д., выполняющий предписанную ему функцию. Кроме того, для

специалиста очевидно, что указанные модули могут быть реализованы с помощью по меньшей мере одного процессора и могут являться логическими модулями. Так, модуль распознавания конфиденциальной информации (КИ) 504 может содержать модель машинного обучения и предназначен для определения категорий и типов конфиденциальной информации (этап 250). Модуль 503 соответственно выполнен с возможностью замены определенных конфиденциальных данных на обезличенные данные. В одном частном варианте осуществления, указанный модуль 503 также может содержать средства классификации типов конфиденциальной информации и базу данных с наборами данных для обезличивания. Хранилище 506 может представлять удаленный сервер, встроенную защищенную память и/или защищенную область в памяти и т.д.

Представленные материалы заявки раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных, содержащий этапы, на которых:

a) получают документ, содержащий текстовые данные;

b) сегментируют текстовые данные, полученные на этапе a), причем в ходе сегментации данные разбиваются по границам изменения форматирования текста на части текста, и присваивают указанным частям текста порядковую позицию в тексте;

c) токенизируют текстовые данные, полученные на этапе a), и присваивают порядковые позиции в тексте каждому токену;

d) выполняют векторизацию токенов, полученных на этапе c);

e) осуществляют обработку векторных представлений токенов, полученных на этапе d), с помощью модели машинного обучения на базе нейронной сети, обученной на наборах конфиденциальных данных, в ходе которой осуществляется определение принадлежности каждого токена к категории конфиденциальных данных;

f) обезличивают данные, относящиеся к токенам с конфиденциальными данными, причем в ходе обезличивания данные заменяют на данные из той же категории с сохранением структуры данных;

g) формируют список замен, включающий указания порядковых позиций токенов с конфиденциальными данными в тексте, порядковых позиций изменения границ форматирования частей текста, обезличенные данные, соответствующие конфиденциальным данным;

h) заменяют исходные конфиденциальные данные в текстовом документе на обезличенные данные по списку замен, причем в процессе замены обезличенные данные форматируют в соответствии с позициями изменения форматирования частей текста.

2. Способ по п.1, характеризующийся тем, что текстовые данные получают в виде неструктурированного текстового документа.

3. Способ по п.1, характеризующийся тем, что категории конфиденциальных данных представляют собой по меньшей мере одно из:

числовые сущности;

именованные сущности.

4. Способ по п.3, характеризующийся тем, что числовые сущности представляют по меньшей мере одно из: основной номер держателя карты, номер ИНН, номер телефона, номер счета, номер страхового свидетельства, IP-адрес.

5. Способ по п.3, характеризующийся тем, что именованные сущности представляют собой по меньшей мере одно из: персоны, локации, организации.

6. Способ по п.1, характеризующийся тем, что структурой данных является соответствие формата, вида и смыслового содержания конфиденциальных и обезличенных данных.

7. Способ по п.3, характеризующийся тем, что обезличенные данные для числовых сущностей генерируются с сохранением идентифицирующих признаков.

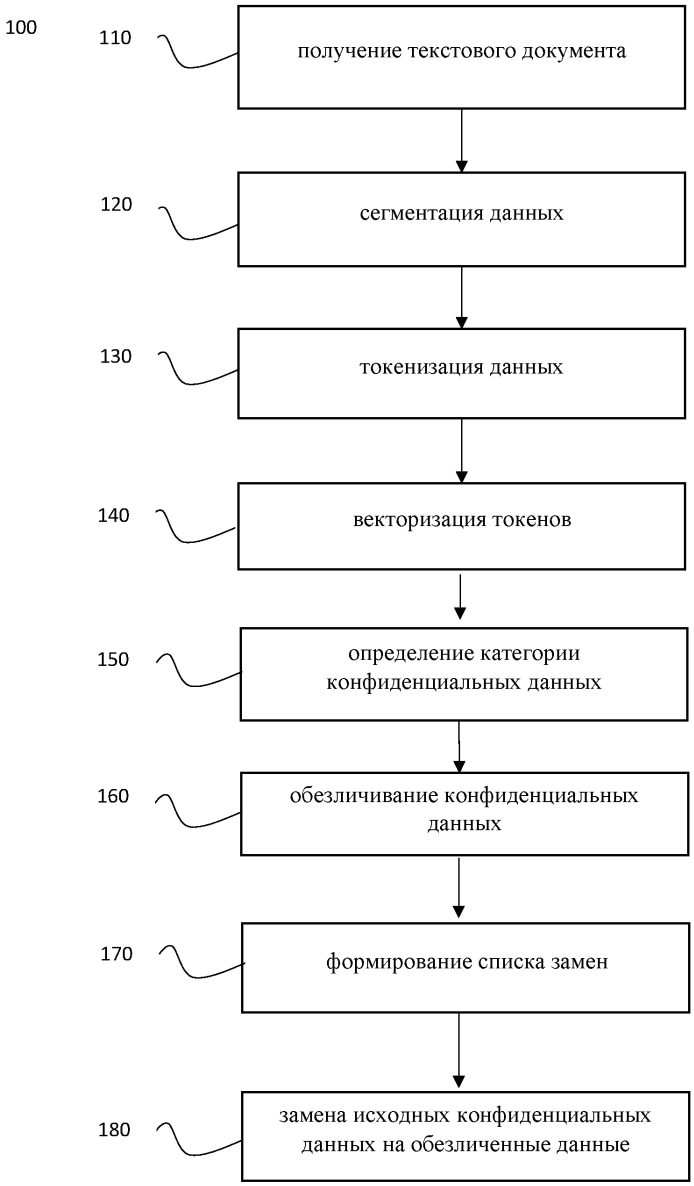
8. Способ по п.3, характеризующийся тем, что обезличенные данные для именованных сущностей приводятся к соответствующей конфиденциальным данным морфологической форме.

9. Способ по п.1, характеризующийся тем, что повторяющиеся конфиденциальные данные обезличиваются одинаковыми данными.

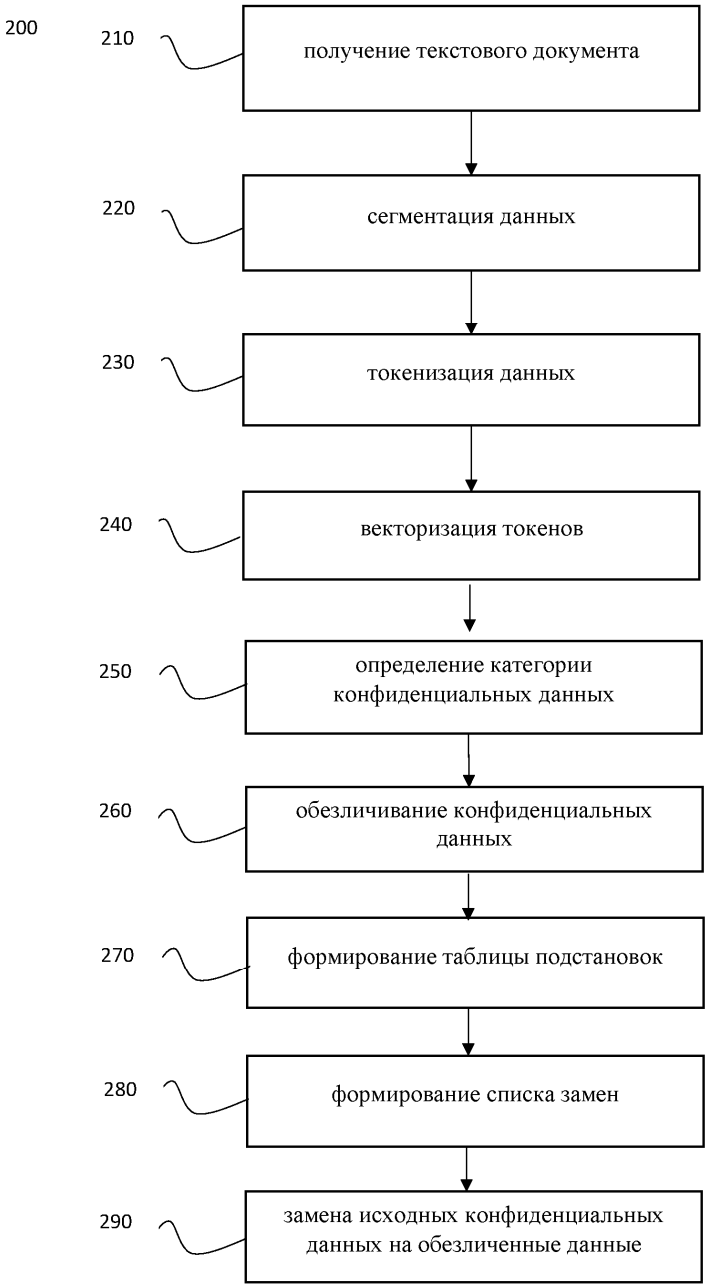
10. Система обезличивания конфиденциальных данных в текстовых документах с сохранением структуры данных, содержащая

по меньшей мере один процессор;

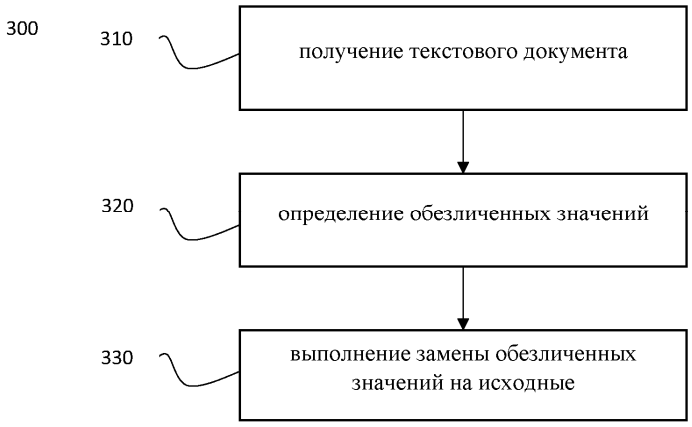
по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из пп.1-9.


100
110
получение текстового документа
120
сегментация данных
130
токенизация данных
140
векторизация токенов
150
определение категории конфиденциальных данных
160
обезличивание конфиденциальных данных
170
формирование списка замен
180
замена исходных конфиденциальных данных на обезличенные данные

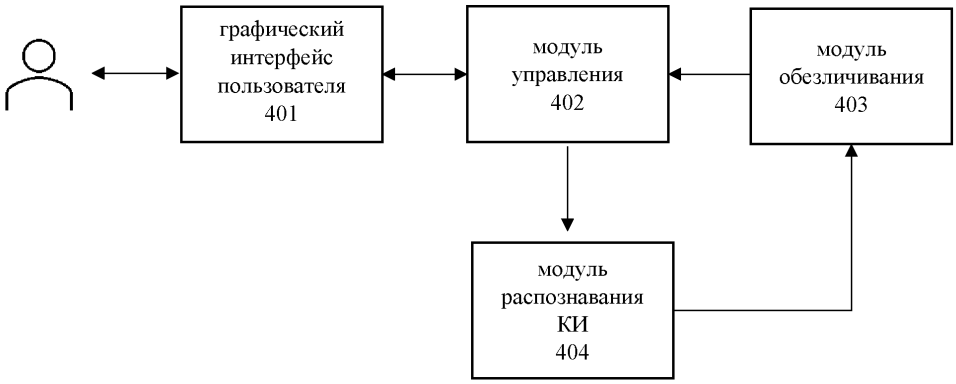
Фиг. 1


200
210
получение текстового документа
220
сегментация данных
230
токенизация данных
240
векторизация токенов
250
определение категории конфиденциальных данных
260
обезличивание конфиденциальных данных
270
формирование таблицы подстановок
280
формирование списка замен
290
замена исходных конфиденциальных данных на обезличенные данные

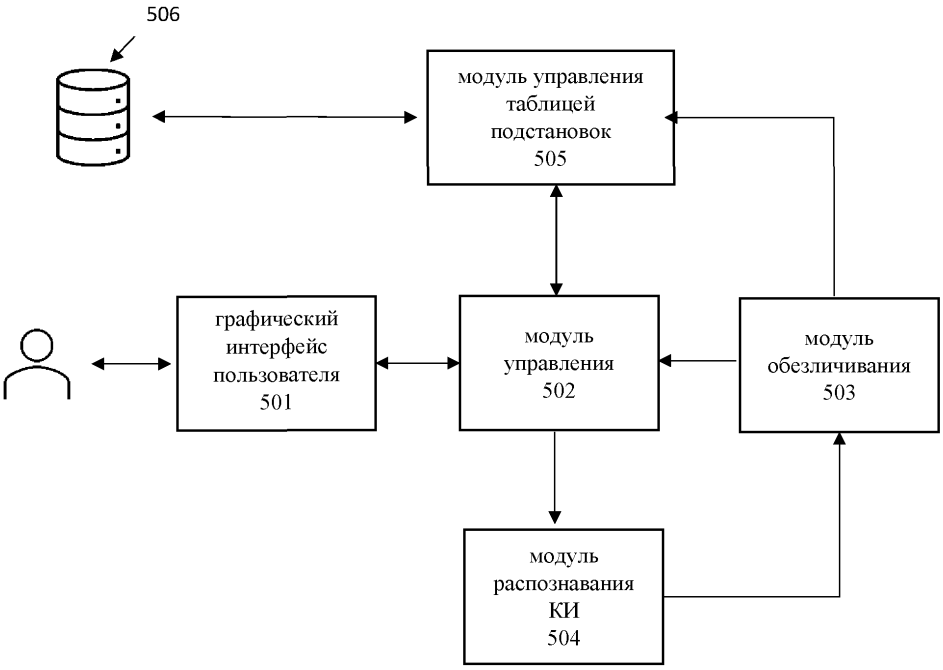
Фиг. 2


300
310
получение текстового документа
320
определение обезличенных значений
330
выполнение замены обезличенных значений на исходные

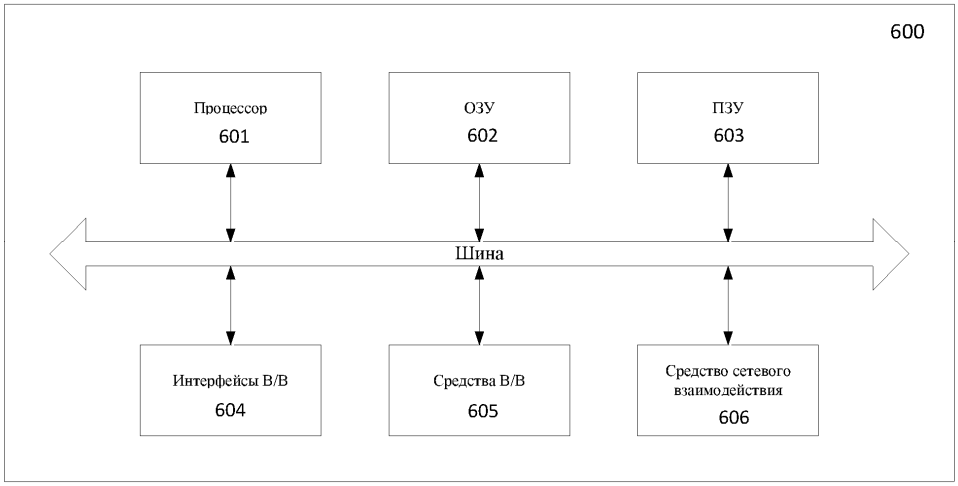
Фиг. 3



Фиг. 4



Фиг. 5



Фиг. 6