(19) Евразийское патентное ведомство

(11) 038259 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2021.07.30

(21) Номер заявки
201991625

(22) Дата подачи заявки
2019.07.31

(51) Int. Cl. ***G06N 3/08*** (2006.01)
G06N 20/20 (2019.01)

(54) СПОСОБ И СИСТЕМА КЛАССИФИКАЦИИ ДАННЫХ ДЛЯ ВЫЯВЛЕНИЯ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ

(31) **2019121020**
(32) **2019.07.05**
(33) **RU**
(43) **2021.01.31**

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

(72) Изобретатель:
Теренин Алексей Алексеевич, Смирнов Дмитрий Владимирович, Струков Дмитрий Константинович, Коряковский Денис Александрович (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20170116519
US-A1-20190122136
Ladislav Lenc et al., "Ensemble of Neural Networks for Multi-Label Document Classification", 2017, CEUR Workshop Proceedings, vol. 1885, ISSN 1613-0073, p. 186-192, размещено в Интернет: http://ceur-ws.org/Vol-1885/186.pdf

(57) Изобретение в общем относится к области вычислительной обработки данных, а в частности к методам классификации данных для выявления конфиденциальной информации. Компьютерно-реализуемый способ классификации данных для выявления конфиденциальной информации, выполняемый с помощью по меньшей мере одного процессора и содержащий этапы, на которых получают данные, представленные в табличном формате; осуществляют обработку полученных данных с помощью ансамбля нейронных сетей, в ходе которой данным в каждой ячейке таблицы присваивается тег, соответствующий заданному типу конфиденциальной информации, причем для каждой нейронной сети сформирована матрица классификации, на основании которой вычисляется F-мера для каждого типа данных; осуществляют обработку полученных данных с помощью алгоритмов определения контрольных разрядов на предмет выявления в ячейках таблицы данных, обладающих контрольным разрядом; на основе полученных от каждой нейронной сети таблиц с проставленными тегами и соответствующей нейронным сетям матрицы F-мер формируют итоговую таблицу с проставленными тегами с учетом данных, обладающих контрольным разрядом; выполняют классификацию данных итоговой таблицы по классам конфиденциальности на основе сравнения проставленных тегов итоговой таблицы с заданными тегами конфиденциальной информации.

038259 B1

038259 B1

Область техники

Изобретение в общем относится к области вычислительной обработки данных, а в частности к методам классификации данных для выявления конфиденциальной информации.

Уровень техники

В настоящее время выявление конфиденциальной информации из большого массива данных и последующая ее классификация является приоритетной задачей для многих отраслей. Наиболее широкое применение данных технологий наблюдается в финансовом секторе, где среди больших объемов различных данных необходимо отдельно выявлять и классифицировать конфиденциальную информацию. Для этого используются различные инструменты и технологии, позволяющие так или иначе выявлять конфиденциальную информацию из больших объемов общих данных. Ключевой особенностью в работе таких инструментов является преобразование данных в табличный формат и последующий их анализ с помощью алгоритмов машинного обучения. Данные хранятся и обрабатываются в различных автоматизированных системах и файловых ресурсах, имеющих различные уровни конфиденциальности, способы доступа, атрибутивный состав. Проверка на наличие чувствительных данных осуществляется различными инструментами. В связи с этим появилась необходимость создать единое техническое решение, позволяющее с помощью нейронных сетей автоматически обрабатывать большое количество данных и выявлять конфиденциальную информацию. Значительный объем данных обычно структурирован и хранится в базах данных в табличном формате, поэтому данное техническое решение направлено на выявление конфиденциальной информации из массива табличных данных.

На сегодняшний момент из уровня техники известны решения, направленные на хранение и классификацию данных по заданным пользователем критериям. Известны сервисы защиты конфиденциальной информации Amazon Macie и Google Cloud DLP. В их основе используются машинные алгоритмы обучения для обнаружения, классификации и защиты конфиденциальной информации. В данных сервисах для классификации информации используются регулярные выражения. Недостатки использования регулярных выражений заключаются в том, что для каждого вида конфиденциальной информации необходимо прописывать несколько регулярных выражений, которые не учитывают редкие особенности данных или могут быть более общими, например, содержать в себе лишние данные.

Сущность изобретения

Заявленное изобретение предлагает новый подход в области выявления и классификации конфиденциальной информации с помощью создания моделей машинного обучения для обработки большого объема данных.

Решаемой технической проблемой или технической задачей является создание нового способа классификации данных, обладающего высокой степенью точности и высокой скоростью распознавания конфиденциальной информации.

Основным техническим результатом, достигающимся при решении вышеуказанной технической проблемы, является повышение точности классификации конфиденциальной информации.

Дополнительным техническим результатом, достигающимся при решении вышеуказанной технической проблемы, является повышение скорости классификации конфиденциальной информации.

Заявленные результаты достигаются за счет компьютерно-реализуемого способа классификации данных для выявления конфиденциальной информации, выполняемого с помощью по меньшей мере одного процессора и содержащего этапы, на которых

получают данные, представленные в табличном формате;

осуществляют обработку полученных данных с помощью ансамбля нейронных сетей, в ходе которой данным в каждой ячейке таблицы присваивается тег, соответствующий заданному типу конфиденциальной информации, причем для каждой нейронной сети сформирована матрица классификации, на основании которой вычисляется F-мера для каждого типа данных;

осуществляют обработку полученных данных с помощью алгоритмов определения контрольных разрядов на предмет выявления в ячейках таблицы данных, обладающих контрольным разрядом;

выполняют классификацию каждой ячейки в таблице на основе полученных от каждой нейронной сети таблиц с проставленными тегами и соответствующей нейронным сетям матрицы F-мер и формируют итоговую таблицу с проставленными тегами с учетом данных, обладающих контрольным разрядом;

выполняют классификацию данных итоговой таблицы по классам конфиденциальности на основе сравнения проставленных тегов итоговой таблицы с заданными тегами конфиденциальной информации.

В одном из частных вариантов осуществления способа для каждой нейронной сети вычисляются показатели F-меры для каждого типа данных.

В другом частном варианте осуществления способа конфиденциальная информация представлена, по меньшей мере, в виде текстовых данных и/или числовых данных.

Также указанные технические результаты достигаются за счет осуществления системы классификации данных для выявления конфиденциальной информации, которая содержит 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 вышеуказанного способа.

Описание чертежей

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

фиг. 1 иллюстрирует блок-схему выполнения заявленного способа;

фиг. 2 иллюстрирует пример данных распознаваемых нейронными сетями;

фиг. 3 иллюстрирует пример архитектуры нейронной сети;

фиг. 4 иллюстрирует результат тестирования моделей;

фиг. 5 иллюстрирует сравнение обучающих моделей;

фиг. 6 иллюстрирует метрику качества распознавания данных первой моделью;

фиг. 7 иллюстрирует метрику качества распознавания данных второй моделью;

фиг. 8 иллюстрирует общий вид заявленной системы.

Осуществление изобретения

В данном изобретении могут использоваться для ясности понимания работы такие термины, как "оператор", "клиент", "сотрудник банка", которые в общем виде следует понимать, как "пользователь" системы.

Заявленный способ (100) классификации данных для выявления конфиденциальной информации, как представлено на фиг. 1, заключается в выполнении ряда последовательных этапов, осуществляемых процессором вычислительного устройства.

Начальным шагом (101) является получение массива данных в табличном формате. Таблицы с данными поделены на столбцы и ячейки, каждая из которых содержит информацию. Информация может представлять собой номера банковских карт, СНИЛС, ОКПО, ОГРН, ИНН, дату, номер паспорта, номер телефона, фамилию, имя, отчество, электронную почту, адрес, должность, адрес сайта, и др., не ограничиваясь. Следующим шагом (102) осуществляют обработку полученных данных с помощью ансамбля нейронных сетей, в ходе которой данным в каждой ячейке таблицы присваивается тег, соответствующий заданному типу конфиденциальной информации, причем для каждой нейронной сети сформирована матрица классификации, на основании которой вычисляется F-мера для каждого типа данных.

Обучение нейронных сетей происходит на заранее размеченных данных. Проверка результата обучения производится на тестовых данных, не пересекающихся с обучающими данными. Способ обучения нейронных сетей будет раскрыт далее в настоящих материалах патента.

В проверенных таблицах данные помечаются тэгами - короткими строками, которые взаимно однозначно соответствуют видам конфиденциальной информации. Тэги подбираются таким образом, чтобы пользователь мог интуитивно понять, что этот тэг обозначает, например, CARD - номер карты, NAME - имя и т.д. Тэги пишутся на латинице, для того, чтобы они имели общий вид на всех кодировках. Виды конфиденциальной информации входят в одну из категорий законодательно регулируемых данных, например персональные данные, банковская тайна, коммерческая тайна и т.д.

Матрица классификации - стандартный инструмент для оценки статистических моделей, в ней отображены вероятности распознавания действительного значения как прогнозируемого, для каждого заданного прогнозируемого варианта.

На основе классификации тестовых данных вычисляются F-меры. F-мера или (F1-score) представляет собой совместную оценку точности и полноты. Данная метрика вычисляется по следующей формуле: F-мера = 2 * Точность * Полнота/(Точность + Полнота). F-мера вычисляется в каждом алгоритме для каждого вида данных.

Далее на шаге (103) осуществляют обработку полученных данных с помощью алгоритмов определения контрольных разрядов на предмет выявления в ячейках таблицы данных, обладающих контрольным разрядом.

Алгоритм проверки контрольных разрядов проверяет данные на соответствие контрольным разрядам, которые обычно вычисляются с помощью алгоритма Луна. 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, номер паспорта, номер телефона и т.д., не ограничиваясь. Контрольный разряд необходим, для того, чтобы исключить вероятность неумышленной ошибки при вводе информации.

Следующим шагом (104) выполняют классификацию каждой ячейки в таблице на основе полученных от каждой нейронной сети таблиц с проставленными тегами и соответствующей нейронным сетям матрицы F-мер и формируют итоговую таблицу с проставленными тегами с учетом данных, обладающих контрольным разрядом.

Табличные данные классифицируются по одному столбцу за раз. Каждый фрагмент данных классифицируется несколькими нейронными сетями. Результаты записываются в датафреймы с тэгами классификации. На основе классификации нейронными сетями и F-мер выбирается вид данных для классификации.

На шаге (105) выполняют классификацию данных итоговой таблицы по классам конфиденциально-

сти на основе сравнения поставленных тегов итоговой таблицы с заданными тегами конфиденциальной информации.

Для построения модели обучения был создан алгоритм, имеющий в своей основе нейронную сеть, по архитектуре аналогичный алгоритму NER (Named-entity recognition - алгоритм распознавания именованных сущностей). Данный алгоритм предназначен для поиска данных в текстах и учитывает синтаксические особенности, что позволяет качественнее классифицировать ячейки, в которых больше одного слова.

Модель нейронной сети может быть сверточной, рекуррентной и т.д. На фиг. 2 представлены виды данных, распознаваемые нейронной сетью. Виды распознаваемых данных содержат один из основных и распространенных видов персональных данных.

Модели, обученные классифицировать данные, указанные выше, демонстрируют разницу в распознавании числовых и тестовых типов данных.

При обучении использовалось две модели. Первая модель учитывает синтаксические особенности - последовательность слов (последовательность символов, разделяемых пробелом) - и расценивает каждый экземпляр данных как упорядоченный массив. Вторая модель не учитывает синтаксические особенности и расценивает каждый экземпляр данных как единый неделимый элемент. Сравнение моделей производилось на процедурно генерируемой таблице, содержащей все используемые в модели виды данных и состоящей из 1000 экземпляров каждого вида данных.

На фиг. 3 представлен пример архитектуры нейронной сети (200), применяемой для реализации заявленного способа (100). Нейронная сеть выполняется из совокупности взаимосвязанных модулей, обеспечивающих ее работу для целей обработки данных на предмет выявления и классификации конфиденциальной информации.

Модуль проверки файлов и процесса обучения нейронных сетей (210) обеспечивает загрузку и исполнение всех нейронных сетей. Нейронные сети для осуществления той или иной классификации подгружаются из библиотеки (220) с помощью модуля обучения нейронных сетей и проверки с помощью нейронных сетей (211). Модуль (211) позволяет обучать определенную нейронную сеть и проверять с ее помощью объект класса pandas DataFrame (табличный файл в библиотеке pandas на языке Python, позволяет преобразовывать в таблицу данные из файлов формата xls, xlsx, csv, json).

Модуль проверки на регулярные выражения (212) позволяет проверять pandas DataFrame с помощью регулярных выражений. Для проверки использует список регулярных выражений (221).

Модуль проверки на контрольные разряды (213) осуществляет классификацию данных в pandas DataFrame с помощью проверки контрольных разрядов.

Модуль классификации типов конфиденциальной информации (214) классифицирует проверенные файлы по типам конфиденциальной информации, загружая их из списка типов конфиденциальной информации (222).

Модуль формирования обучающих выборок и тестовых файлов (215) производит тестирование и проверку моделей нейронных сетей, используя информацию из списка типов конфиденциальной информации (222) и из базы обучающих данных (223).

Модуль формирования статистики (216) формирует статистику проверки файлов.

Далее будет представлен принцип обучения нейронных сетей для целей осуществления заявленного способа.

На первом этапе обучения производят выбор параметров нейронной сети. Далее осуществляется создание тренирующих выборок. Из файлов в формате .txt или .csv, содержащихся в модуле списка типов конфиденциальной информации (222) и представляющие из себя столбец с данными строго определенного вида конфиденциальной информации, создаются тренирующие выборки в формате .xlsx. Далее из файлов, содержащихся в модуле списка типов конфиденциальной информации (222), создается тестовый файл. На следующем этапе осуществляется обучение модели на полученных обучающих выборках. Далее производится создание матрицы классификации, которая показывает, как классифицируется каждый вид данных. И на заключительном шаге результат выводится пользователю.

На фиг. 4 показан результат тестирования моделей. На диаграмме отображены вероятности классификации различных видов конфиденциальной информации. По ней можно определить, какие данные распознаются каждой моделью лучше, чем другие. Чем дальше точка, соответствующая своему типу данных, расположена от центра, тем точнее распознаются данные этого вида.

На фиг. 5 отображено сравнение обучающих моделей. В таблице показаны вероятности верной классификации конфиденциальной информации различными моделями. По таблице можно определить, какая модель распознает лучше и на сколько тот или иной вид конфиденциальной информации. Чем больше вероятность - тем лучше модель распознает данные. Для того чтобы определить на сколько одна модель распознает лучше или хуже определенные данные, необходимо вычислить разницу между значениями для первой и второй модели.

На фиг. 6 и 7 представлены метрики качества первой и второй модели. На матрицах показаны вероятности распознавания реальных экземпляров конфиденциальной информации как вид конфиденциальной информации. Матрицы позволяют вычислить точность и полноту классификации каждого вида кон-

фиденциальной информации. Точность системы в пределах класса - это доля объектов, действительно принадлежащих данному классу относительно всех объектов, которые система отнесла к этому классу (отношение значения на диагонали к сумме всех значений столбца). Полнота системы - это доля найденных классификатором объектов, принадлежащих классу относительно всех объектов этого класса (отношение значения на диагонали к сумме всех значений строки). На фиг. 8 представлен пример общего вида вычислительной системы (300), которая обеспечивает реализацию заявленного способа (100) или является частью компьютерной системы, например, сервером, персональным компьютером, частью вычислительного кластера, обрабатывающим необходимые данные для осуществления заявленного технического решения.

В общем случае, система (300) содержит объединенные общей шиной информационного обмена один или несколько процессоров (301), средства памяти, такие как ОЗУ (302) и ПЗУ (303), интерфейсы ввода/вывода (304), устройства ввода/вывода (1105), и устройство для сетевого взаимодействия (306).

Процессор (301) (или несколько процессоров, многоядерный процессор и т.п.) может выбираться из ассортимента устройств, широко применяемых в настоящее время, например, таких производителей, как Intel™, AMD™, Apple™, Samsung Exynos™, MediaTEK™, Qualcomm Snapdragon™ и т.п. Под процессором или одним из используемых процессоров в системе (300) также необходимо учитывать графический процессор, например, GPU NVIDIA или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа (100), а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (302) представляет собой оперативную память и предназначено для хранения исполняемых процессором (301) машиночитаемых инструкций для выполнения необходимых операций по логической обработке данных. ОЗУ (302), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом в качестве ОЗУ (302) может выступать доступный объем памяти графической карты или графического процессора.

ЗУ (303) представляет собой одно или более устройств постоянного хранения данных, например жесткий диск (HDD), твердотельный накопитель данных (SSD), флэш-память (EEPROM, NAND и т.п.), оптические носители информации (CD-R/RW, DVD-R/RW, BlueRay Disc, MD) и др.

Для организации работы компонентов системы (300) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (304). 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. Для обеспечения взаимодействия пользователя с вычислительной системой (300) применяются различные средства (305) В/В информации, например клавиатура, дисплей (монитор), сенсорный дисплей, тачпад, джойстик, 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.

Средство сетевого взаимодействия (306) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (306) может использоваться, но не ограничиваться: Ethernet карта, GSM модем, GPRS модем, LTE модем, 5G модем, модуль спутниковой связи, NFC модуль, Bluetooth и/или BLE модуль, Wi-Fi модуль и др.

Представленные материалы патента раскрывают предпочтительные примеры реализации изобретения 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 техники.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Компьютерно-реализуемый способ классификации данных для выявления конфиденциальной информации, выполняемый с помощью по меньшей мере одного процессора и содержащий этапы, на которых

получают данные, представленные в табличном формате;

осуществляют обработку полученных данных с помощью ансамбля нейронных сетей, в ходе которой данным в каждой ячейке таблицы присваивается тег, соответствующий заданному типу конфиденциальной информации, причем для каждой нейронной сети сформирована матрица классификации, на основании которой вычисляется F-мера для каждого типа данных;

осуществляют обработку полученных данных с помощью алгоритмов определения контрольных разрядов на предмет выявления в ячейках таблицы данных, обладающих контрольным разрядом;

выполняют классификацию каждой ячейки в таблице на основе полученных от каждой нейронной

сети таблиц с проставленными тегами и соответствующей нейронным сетям матрицы F-мер и формируют итоговую таблицу с проставленными тегами с учетом данных, обладающих контрольным разрядом;

выполняют классификацию данных итоговой таблицы по классам конфиденциальности на основе сравнения проставленных тегов итоговой таблицы с заданными тегами конфиденциальной информации.

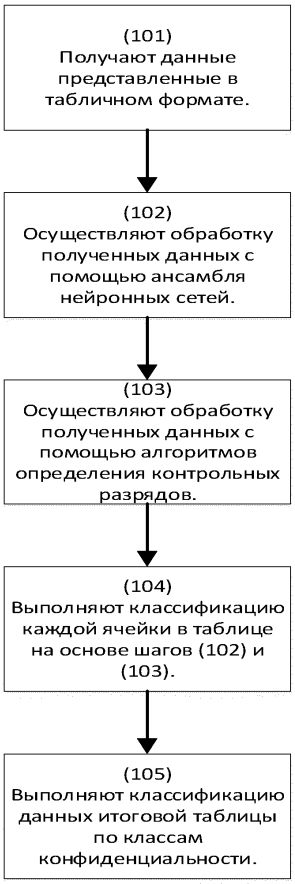
2. Способ по п.1, характеризующийся тем, что для каждой нейронной сети вычисляются показатели F-меры для каждого типа данных.

3. Способ по п.1, характеризующийся тем, что конфиденциальная информация представлена, по меньшей мере, в виде текстовых данных и/или числовых данных.

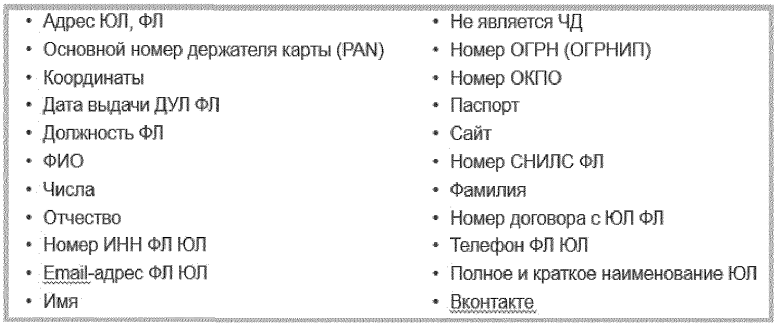
4. Система классификации данных для выявления конфиденциальной информации, содержащая

по меньшей мере один процессор;

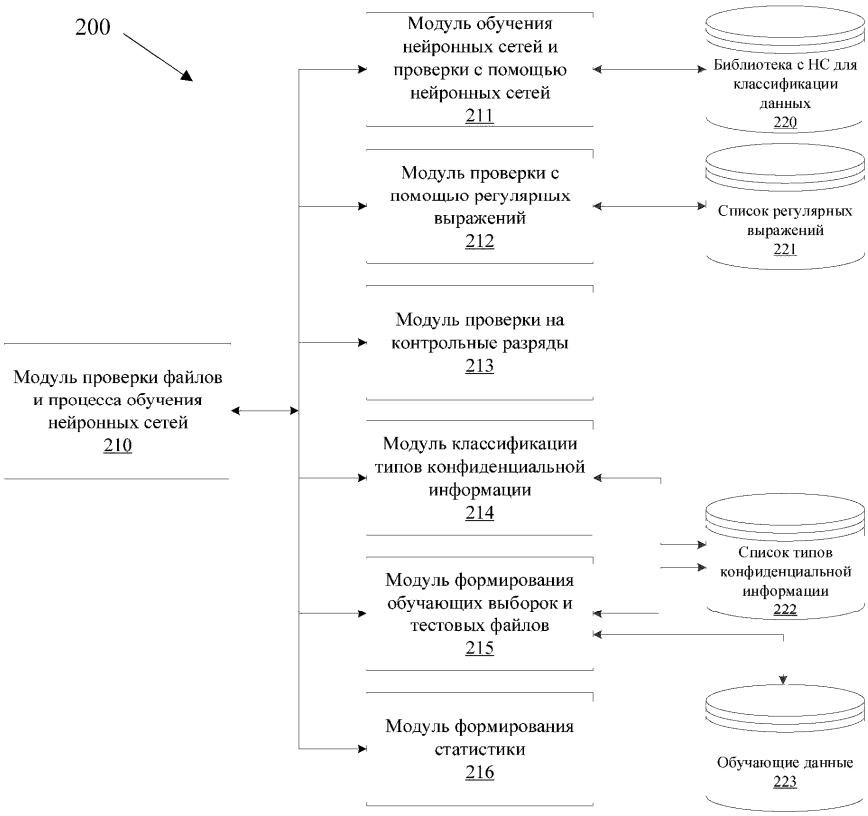
по меньшей мере одну память, соединенную с процессором, которая содержит машиночитаемые инструкции, которые при их выполнении по меньшей мере одним процессором обеспечивают выполнение способа по любому из пп.1-3.



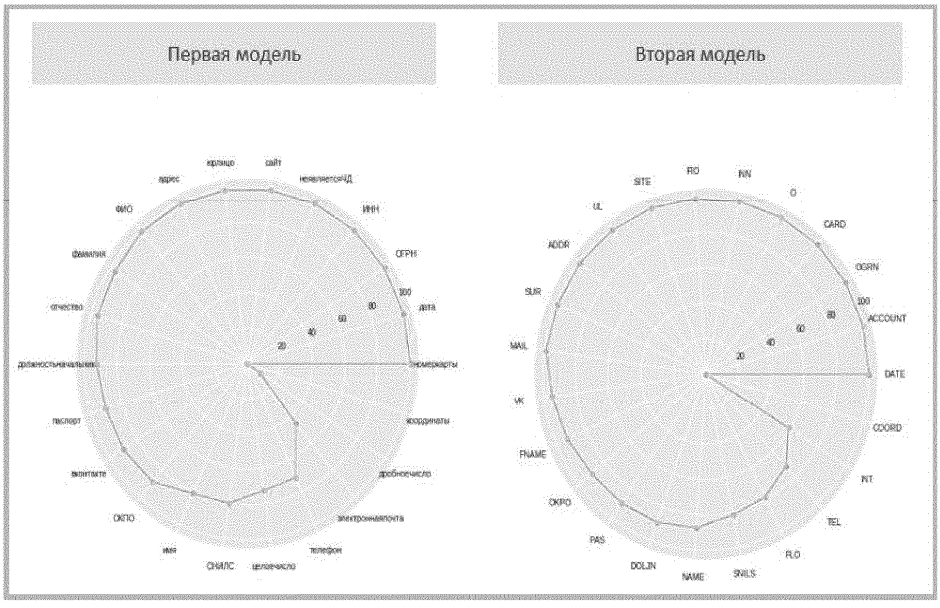
Фиг. 1



Фиг. 2


200
Модуль обучения нейронных сетей и проверки с помощью нейронных сетей
211
Библиотека с НС для классификации данных
220
Модуль проверки с помощью регулярных выражений
212
Список регулярных выражений
221
Модуль проверки на контрольные разряды
213
Модуль проверки файлов и процесса обучения нейронных сетей
210
Модуль классификации типов конфиденциальной информации
214
Список типов конфиденциальной информации
222
Модуль формирования обучающих выборок и тестовых файлов
215
Модуль формирования статистики
216
Обучающие данные
223

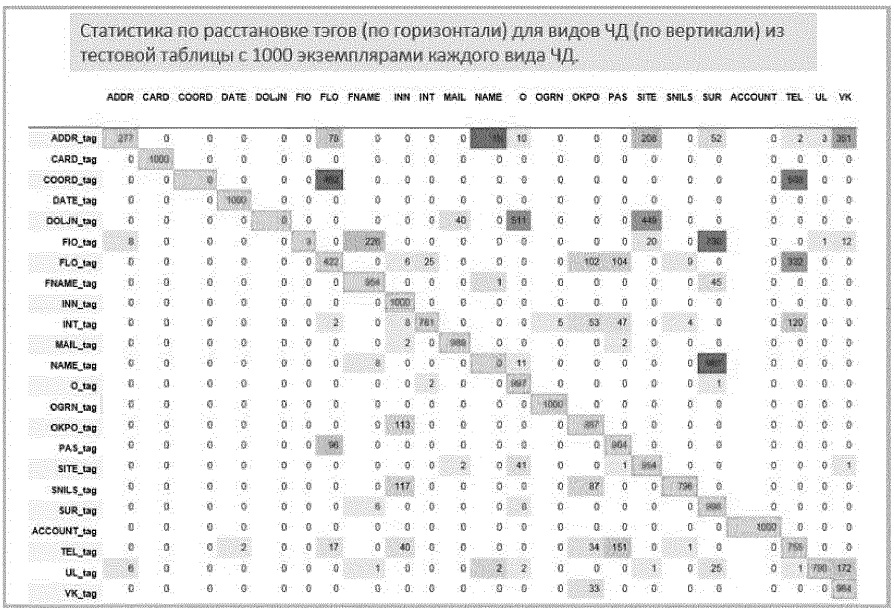
Фиг. 3


Первая модель
Вторая модель
ссылка
сайт
адрес
неявляетсяИД
ФИО
ИНН
ОГРН
фамилия
дата
отчество
номеркарты
должностьначальника
координаты
паспорт
дробноечисло
вконтакте
электроннаяпочта
ОКПО
имя
телефон
СНИЛС
целоечисло
FIO
INN
SITE
O
UL
CARD
ADDR
OGRN
SUR
ACCOUNT
MAIL
DATE
VK
COORD
FNAME
INT
OKPO
TEL
PAS
FLO
DOLIN
NAME
SNILS

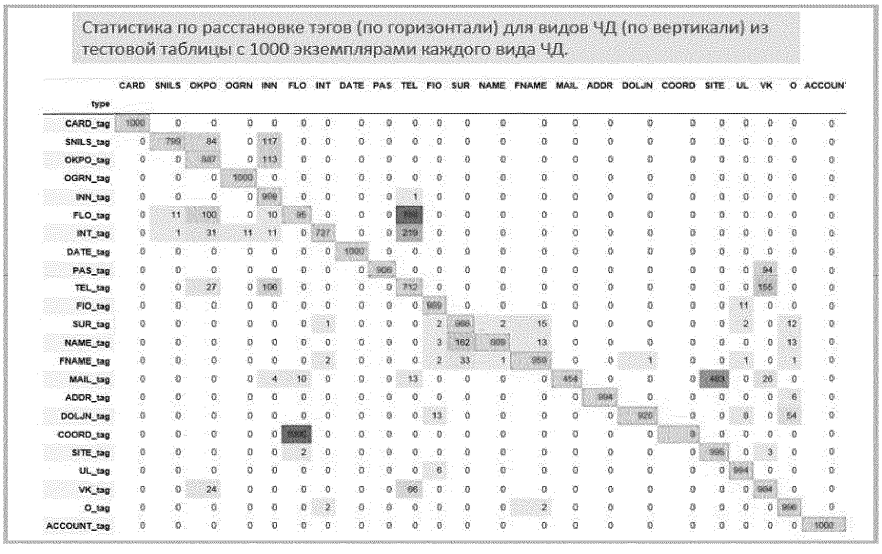
Фиг. 4

Тип ЧД	Точность определения	
	Первая модель, %	Вторая модель, %
СНИЛС	81 6	79 9
ОКПО	90 0	88 7
ОГРН	100 0	100 0
ИНН	99 9	99 9
дата	100 0	100 0
паспорт	89 6	90 6
телефон	71 8	71 2
ФИО	99 7	98 9
фамилия	99 1	96 6
имя	87 2	80 9
отчество	92 7	95 9
электронная почта	98 9	45 4
адрес	99 5	99 4
должность начальника	89 2	92 5
сайт	99 7	99 5
юрлицо	99 6	99 4
вконтакте	95 0	90 4
не является ЧД	99 9	99 6
Номер счета	100	100

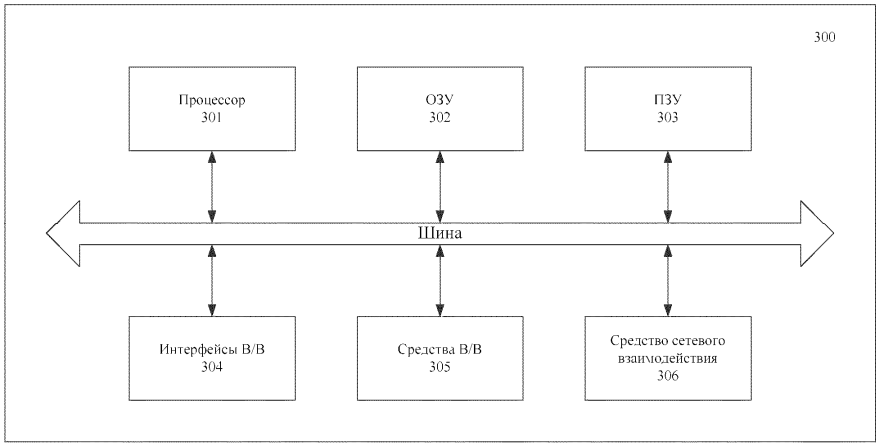
Фиг. 5



Фиг. 6



Фиг. 7


300
Процессор
301
ОЗУ
302
ПЗУ
303
Шина
Интерфейсы В/В
304
Средства В/В
305
Средство сетевого
взаимодействия
306

Фиг. 8