(19) Евразийское патентное ведомство

(11) 040560 (13) B1

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
2022.06.23

(21) Номер заявки
201992041

(22) Дата подачи заявки
2019.09.27

(51) Int. Cl. ***G06K 9/00*** (2006.01)
G06F 40/194 (2006.01)

(54) СПОСОБ И СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ ДОКУМЕНТА

(31) **2019129251**
(32) **2019.09.17**
(33) **RU**
(43) **2021.06.30**

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

(72) Изобретатель:
Тарасов Кирилл Геннадьевич, Колесов Антон Юрьевич (RU)

(74) Представитель:
Герасин Б.В. (RU)

(56) US-A1-20130021344
US-A1-20110134494
US-A1-20140237342
RU-C2-2463660
RU-C1-2666277

(57) Представленное техническое решение относится в общем к области анализа изображений, а в частности к способам и системам интеллектуальной обработки электронного комплекта документов, например отсканированных документов клиентов банка. Техническим результатом является повышение эффективности и обеспечение высокой точности в выявлении ошибок при проведении автоматизированной интеллектуальной обработки документов. Указанный технический результат достигается благодаря осуществлению способа интеллектуальной обработки документов, выполняемого по меньшей мере одним вычислительным устройством и содержащего этапы, на которых получают по меньшей мере одно изображение документа; распознают символы на изображении документа и преобразуют их в текстовую информацию; на основе текстовой информации определяют тип документа; извлекают из текстовой информации набор сущностей с учетом типа документа; сравнивают набор сущностей с эталонным набором сущностей для данного документа; на основе результатов сравнения упомянутых наборов сущностей формируют результаты обработки документа.

040560 B1

040560 B1

Область техники

Представленное техническое решение относится в общем к области анализа изображений, а в частности к способам и системам интеллектуальной обработки электронного комплекта документов, например отсканированных документов клиентов банка.

Уровень техники

В настоящее время существует проблема оперативной и качественной обработки данных электронного комплекта отсканированных документов с целью проверки наличия обязательных заполненных полей документа из структурированных и неструктурированных документов, а также атрибутов подписанта, таких как подпись. Из уровня техники известны различные решения, выполненные с возможностью обработки документов, например клиента Банка, реализованные на базе ПО ABBYY FlexiCapture и пр. Также известно решение для проведения проверки комплекта документов, раскрытое в заявке US 2011134494 A1, опубл. 09.06.2011, в котором осуществляют чтение документа, имеющего множество страниц; проверку данных изображения каждой страницы документа, имеющего множество страниц, при этом проверяются определенные области изображения документа на наличие в них информации и ее отсутствие. Данное решение является наиболее близким аналогом.

Существенным недостатком известных решений является низкая эффективность в выявлении ошибок при проверке документов на корректность их заполнения, поскольку в очень большом количестве случаев известные решения выдают результат "ошибка есть", хотя ее на самом деле нет, все поля заполнены верно, но известное решение попросту не смогло найти их в тексте из-за того, что текст слабо структурирован. Также в известных решениях отсутствует механизм автоматизированного принятия решений по итогу упомянутой проверки.

Раскрытие изобретения

Технической проблемой или задачей, поставленной в данном техническом решении, является создание нового эффективного, простого и надежного метода автоматизированной интеллектуальной обработки любых видов документов на корректность их заполнения.

Техническим результатом является повышение эффективности и обеспечение высокой точности в выявлении ошибок при проведении автоматизированной интеллектуальной обработки документов.

Указанный технический результат достигается благодаря осуществлению способа интеллектуальной обработки документов, выполняемого по меньшей мере одним вычислительным устройством и содержащего этапы, на которых

получают по меньшей мере одно изображение документа;

распознают символы на изображении документа и преобразуют их в текстовую информацию;

на основе текстовой информации определяют тип документа;

извлекают из текстовой информации набор сущностей с учетом типа документа;

сравнивают набор сущностей с эталонным набором сущностей для данного документа;

на основе результатов сравнения упомянутых наборов сущностей формируют результаты обработки документа.

В одном из частных примеров осуществления способа документ представляет собой договор об индивидуальных условиях кредитования (ИУК) или договор поручительства (ДП).

В другом частном примере осуществления способа дополнительно выполняют этапы, на которых осуществляют детектирование на поступившем изображении документа атрибута подписанта; определяют расположение по меньшей мере одного атрибута подписанта на странице документа; при этом результаты обработки документа формируют с учетом информации о расположении по меньшей мере одного атрибута подписанта на странице документа.

В другом частном примере осуществления способа дополнительно выполняют этап, на котором определяют статус лица, которому принадлежит детектированный атрибут подписанта.

В другом частном примере осуществления способа дополнительно выполняют этапы, на которых получают идентификатор процесса; определяют набор моделей классификации текста на основе идентификатора процесса; преобразуют полученную текстовую информацию в набор векторов; обрабатывают набор векторов с помощью определенного ранее набора моделей классификации текста для определения типа документа.

В другом частном примере осуществления способа дополнительно выполняют этапы, на которых делят набор сущностей на простые сущности, состоящие из 1-3 слов, и сложные сущности, состоящие по меньшей мере из четырех слов; причем если в результате сравнения упомянутых наборов сущностей пороговые значения совпадающих слов по простым и сложным сущностям достигнуты, то формируют результаты сверки, в которые включается информация о успешном прохождении сверки данных; если упомянутые пороговые значения совпадающих слов по простым и сложным сущностям не достигнуты, то формируют результаты сверки, в которые включается информация о сущностях в наборе сущностей, не прошедших сверку; при этом результаты обработки документа формируют с учетом результатов сверки.

В другом частном примере осуществления способа дополнительно выполняют этап, на котором определяют качество сканирования документа; причем результаты обработки документа формируют с учетом качества сканирования документа.

В другом предпочтительном варианте осуществления заявленного решения представлена система интеллектуальной обработки документов, содержащая по меньшей мере одно вычислительное устройство, и по меньшей мере одно устройство памяти, содержащее машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют вышеуказанный способ.

Краткое описание чертежей

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

на фиг. 1 представлена общая схема взаимодействия элементов системы интеллектуальной обработки документов;

на фиг. 2 представлен пример отсканированного документа;

на фиг. 3 представлен пример общего вида системы интеллектуальной обработки документов.

Осуществление изобретения

Ниже будут описаны понятия и термины, необходимые для понимания данного технического решения.

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, вычислительное устройство, либо интегральная схема (микропроцессор), исполняющая машинные инструкции (программы).

Устройство обработки команд считывает и выполняет машинные инструкции (программы) с одного или более устройств хранения данных. В роли устройства хранения данных могут выступать, но не ограничиваясь, жесткие диски (HDD), флеш-память, ПЗУ (постоянное запоминающее устройство), твердотельные накопители (SSD), оптические приводы.

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

База данных (БД) - совокупность данных, организованных в соответствии с концептуальной структурой, описывающей характеристики этих данных и взаимоотношения между ними, причем такое собрание данных, которое поддерживает одну или более областей применения (ISO/IEC 2382:2015, 2121423 "database").

В соответствии со схемой, приведенной на фиг. 1, система 10 интеллектуальной обработки документов содержит соединенные между собой

модуль 11 преобразования данных; модуль 12 детекции подписей, модуль 13 извлечения данных, модуль 17 классификации документов пакета и модуль 18 бизнес-правил, состоящий из модуля 14 сверки данных, модуля 15 анализа свойств документа, модуля 16 принятия решения и модуля 19 анализа юридической валидности.

Указанные модули могут быть реализованы на базе программно-аппаратных средств системы 10 интеллектуальной обработки документов, например на базе по меньшей мере одного вычислительного устройства, в частности микропроцессора, и по меньшей мере одного устройства памяти, содержащего машиночитаемые инструкции, написанные на языке программирования Python, для осуществления выполняемых модулями функций. Например, модуль 11 преобразования данных может быть реализован на базе инструмента для оптического распознавания символов (англ. optical character recognition, OCR). Модуль 12 детекции подписей может быть реализован на базе нейронной сети архитектуры YOLOv3, заранее обученной на типовом наборе подписей и печатей. Модуль 17 классификации документов пакета может быть реализован на базе программно-аппаратных средств системы 10, сконфигурированных для представления текста в виде векторов (например, TFIDF), и включать набор моделей классификации текста, например SVM или Random Fields. Модуль 13 извлечения данных может быть реализован на базе программно-аппаратных средств системы 10 и включать набор моделей для анализа семантики естественных языков word2vec, заранее обученную математическую модель - условные случайные поля (Conditional Random Fields) и вычислительные средства для обработки естественного языка (Natural Language Processing, NLP). Модуль 18 бизнес-правил, состоящий из модуля 14 сверки данных, модуля 15 анализа свойств документа, модуля 16 принятия решения и модуля 19 анализа юридической валидности, может быть реализован на базе программно-аппаратных средств системы 10, сконфигурированных в программно-аппаратной части таким образом, чтобы выполнять приписанные им ниже функции.

На первом этапе работы системы 10 на модуль 11 преобразования данных и модуль 12 детекции подписей поступает по меньшей мере одно изображение документа, в частности отсканированного документа, например файл в формате многостраничного PDF, JPEG, TIFF или любого другого известного формата, который может использоваться для хранения в нем изображения отсканированного документа. Изображение документа может поступать от источника данных изображений 1, в частности непосредственно от устройства сканирования документов, например сканера, либо могут быть извлечены из соот-

ветствующей базы данных изображений, в которую данные изображения документов заранее сохранены.

Также в соответствии с заранее заданным программно-аппаратным алгоритмом в модуль 17 классификации документов пакета и в модуль 18 бизнес-правил поступают данные об идентификаторе процесса от автоматизированной системы (АС) 2 Банка. Идентификатор процесса от АС 2 Банка может подаваться в упомянутые модули широко известными из уровня техники методами, например перед подачей документа на сканер или перед извлечением изображения документа из БД, согласно процессу, в рамках которого осуществляется проверка документа. На основе данных об идентификаторе процесса в дальнейшем определяется набор возможных типов документов, которые могут быть на изображении документа, поступившем в модуль 11 преобразования данных; набор сущностей, которые следует извлекать модулем 13, и данные о расположении подписей в документах. Например, данные о идентификаторе процесса могут указывать на то, что на вход модулю 17 классификации документов может приходить 2 типа документов: договор об индивидуальных условиях кредитования (ИУК) или договор поручительства (ДП), поэтому срабатывает соответствующий классификатор.

Документом, изображение которого поступает на модуль 11 преобразования данных, может быть любой документ, состоящий по меньшей мере из одной страницы, которая может содержать атрибуты подписанта, и заполненный в соответствии с известным шаблоном. Документом может быть, например, документ/договор ИУК, подписанные клиентом банка, или договор поручительства (ДП). Документ может содержать поля, в которых указана информация о подписанте, например ФИО подписанта, адрес подписанта, номер карты подписанта, данные паспорта и пр., а также информацию о условиях договора, например условиях кредитования. В частности, согласно схеме, представленной на фиг. 2, в области 101 документа 100 может содержаться поле с информацией о номере упомянутого заявления, в области 102 - поле с названием города, в области 103 - поле с датой заявления, в области 104 - поля с информацией о подписанте и условиях кредитования, в области 105 или 106 документа - изображения атрибутов подписанта, например изображение подписи.

Модуль 11 преобразования данных осуществляет распознавание символов на изображении документа и преобразует их в текстовую информацию. Вместе с этим модуль 12 детекции подписей осуществляет детектирование на поступившем изображении документа атрибута подписанта, определяя его расположения на странице документа. Атрибут подписанта может отсутствовать на странице, эта информация также передается далее по схеме, приведенной на фиг. 1. Например, модуль 12 может определить, что изображение атрибута подписанта представляет собой изображение подписи в области 105 или 106 документа (см. фиг. 2), автоматически указав координаты найденных боксов 105 и 106. Соответственно данные о расположении атрибутов подписанта на странице документа или об их отсутствии модуль 12 направляет в модуль 19 анализа юридической валидности.

Для детектирования изображений атрибутов подписанта используются известные алгоритмы работы нейронной сети архитектуры YOLOv3, обученной на отобранном наборе данных подписей и печатей, раскрытые, например, в статье, опубликованной в Интернет по адресу

https://pireddie.com/media/files/papers/YOLOv3.pdf.

Если изображение документа содержит атрибуты более одного подписанта, например изображение подписи клиента Банка и изображение подписи сотрудника Банка, то модуль 19 анализа юридической валидности может быть выполнен с возможностью определения статуса лица, которому принадлежит детектированные атрибуты подписанта. Для этого в памяти модуля 19 пользователем системы 10 может быть заранее задан перечень статусов лиц и информация о местоположении их атрибутов подписанта на изображении документа исходя из идентификатора процесса, данные о котором поступили от АС 2 Банка в модуль 18, причем информация о статусе лиц может указывать на то, какому лицу принадлежит атрибут подписанта, в частности, например, клиенту Банка или сотруднику Банка. Например, для статуса лица "клиент Банка" данные о местоположении могут указывать на то, что его атрибуты подписанта должны располагаться в области 105 документа, а для статуса лица "сотрудник Банка" данные о местоположении могут указывать на то, что его атрибуты подписанта располагаются в области 106 документа.

Соответственно модуль 19 анализа юридической валидности сравнивает данные о расположении изображения атрибута подписанта на странице документа, полученные от модуля 12, с упомянутыми сохраненными в памяти данными, в частности данными о местоположении атрибутов подписанта согласно типу процесса, определенного модулем 19 на основе полученных ранее данных о идентификаторе процесса и на основе результата сравнения определяет статус лица, которому принадлежит детектированный атрибут подписанта, т.е. на основе информации о расположении атрибута подписанта на изображении страницы документа. Данные о статусе лица и данные о расположении изображений атрибутов подписантов на странице документа модуль 19 направляет в модуль 16 принятия решений. Если в модуль 19 поступила информация об отсутствии атрибутов подписанта на изображении, то эту информацию модуль 19 перенаправляет в модуль 16.

Что касается текстовой информации, то ее модуль 11 преобразования данных направляет в модуль 13 извлечения данных и в модуль 17 классификации документов пакета. Модуль 17 на основе данных об идентификаторе процесса, поступивших от АС 2, определяет набор моделей классификации текста, который могут быть заранее заданы в упомянутом модуле 17 для каждого типа процесса пользователем

системы 10, после чего полученную текстовую информацию модуль 17 преобразует в набор векторов, который обрабатывается определенным ранее набор моделей классификации текста для определения типа документа. Данные о типе документа модуль 17 передает в модуль 13, который извлекает из полученной текстовой информации от модуля 11 набор сущностей в соответствии с типом документа. Набор сущностей может включать ФИО, адрес, номер карты, дату документа, номер карты, данные паспорта, условия кредитования и т.д. Для извлечения из полученной текстовой информации набора сущностей модуль 13 выполняет токенизацию текстовой информации и подает токенизированную текстовую информацию на вход набору моделей word2vec, на выходе которого модуль 13 получает последовательность векторов.

Далее внутри модуля 13 определяется обученная модель машинного обучения CRF (Conditional Random Fields) на основе данных о типе документа и последовательность векторов обрабатывается упомянутой обученной моделью, которая определяет набор сущностей. Обученные модели машинного обучения CRF для каждого типа документа могут быть заранее заданы в упомянутом модуле 13 пользователем системы 10. Модели машинного обучения, обученные методом CRF, широко используются в различных областях ИИ, в частности в задачах распознавания речи и образов, обработки текстовой информации, а также и в других предметных областях: биоинформатике, компьютерной графике и пр.

В альтернативном варианте реализации заявленного решения сущности могут быть извлечены при помощи технологии обработки естественного языка (Natural Language Processing, NLP). Данная технология широко известна из уровня техники (см., например, статью "NLP. Основы. Техники. Саморазвитие. Часть 2: NER", опубликованную в Интернет по адресу

https://habr.com/ru/company/abbyy/blog/449514/),

и дополнительно более подробно не будет раскрываться в настоящем документе. Алгоритм обработки последовательности векторов также может выбираться в зависимости от типа документа.

Полученный набор сущностей модуль 13 извлечения данных направляет в модуль 14 сверки данных. Также в модуль 14 подается эталонный набор сущностей модулем 18 бизнес-правил. Эталонный набор сущностей модулем 18 определяется на основе поступивших ранее данных об идентификаторе процесса из АС 2 Банка. Эталонный набор сущностей для каждого типа процесса может быть заранее задан в упомянутом модуле 18 пользователем системы 10. Полученные данные наборов сущностей модуль 14 делит на простые сущности, состоящие из 1-3 слов, и сложные сущности, состоящие из по меньше четырех слов. Например, если на вход системе 10 поступил документ ИУК, то простыми сущностями будут являться, например, ФИО, сумма кредитования, дата начала договора, номер паспорта, дата выдачи паспорта и пр., а сложными сущностями будут являться, например, адрес, место выдачи паспорта и пр. Далее модуль 14 сверки данных переходит к этапу сравнения набора сущностей, полученного от модуля 13, с эталонным набором сущностей. Данные простых сущностей модуль 14 сверки данных приводит к одному формату, после чего сравнивает их. В данных сложных сущностей перед их сравнением расшифровываются общепризнанные сокращения, исключаются слова, не содержащие названия. Если установленные пользователем системы 10 пороговые значения совпадающих слов по простым и сложным сущностям достигнуты, то набор сущностей, полученный от модуля 13, проходит сверку данных. Если пороговые значения совпадающих слов по простым и/или сложным сущностям не достигнуты, то набор сущностей не проходит проверку. По итогу сравнения наборов сущностей модуль 14 сверки данных формирует результаты сверки, в которые включается информация о успешном прохождении сверки, либо в случае, если набор сущностей не прошел сверку, информация о сущностях в наборе сущностей, не прошедших сверку. Информация о наборе сущностей, полученная от модуля 13, вместе с текстовой информации и результатами сверки модулем 14 сверки данных направляются в модуль 15 анализа свойств документа.

Вся собранная модулем 15 в ходе работы всех предыдущих модулей информация, в частности текстовая информация и результаты сверки от модуля 14 и изображения документа от источника 1, модулем 15 проверяется на то, что все необходимые пункты документа (или поля документа) содержатся в тексте документа. Для этого модуль 15 осуществляет обработку полученной текстовой информации методами NLP (нечеткое вхождение ключевых слов для каждого абзаца), по результатам которой модуль 15 определяет целостность документа. Алгоритм обработки NLP также может быть выбран на основе данных об идентификаторе процесса, которые ранее поступили в модуль 18 от АС 2 Банка. Для обработки полученной текстовой информации методами NLP был проанализирован набор типовых документов на распределение слов в абзацах документа и были найдены характерные слова и/или фразы для каждого абзаца документа, причем из разных его частей (начало, середина, конец). Таким образом, стали известны для каждого значимого (который должен присутствовать в документе для проверки целостности) абзаца документа его характерные слова. Далее было создано правило, согласно которому: если определенная доля слов или фраз встречается (fuzzy search) в абзаце документа, то данный значимый абзац найден. Если все необходимые абзацы (пункты) документа найдены в тексте, то целостность проверена успешно. В альтернативном варианте реализации заявленного решения целостность документа может быть проверена с помощью средств и методов, раскрытых в заявке US 2011134494 A1. На основе данных о целостности документа и данных сверки модуль 15 определяет качества сканирования изображения документа.

Например, если сверка данных прошла успешно и данные о целостности документа указывают на то, что документ содержит все пункты, то модуль 15 присваивает изображению документа высокий показатель качества сканирования. Если результаты сверки указывают на то, что пороговые значения совпадающих слов по простым и/или сложным сущностям не достигнуты, причем данные о целостности документа указывают на то, что документ содержит не все пункты, то модуль 15 присваивает изображению документа низкий показатель качества сканирования. Информация о показателе качества сканирования модуль 15 передает в модуль 16 принятия решения.

Также модуль 15 анализа свойств документа выполнен с возможностью проверки не приложен ли документ от другого лица. Упомянутая проверка выполняется на основе данных о целостности документа и данных о уникальных сущностях набора сущностей, которые у различных клиентов отличаются или которые могут совпасть у различных клиентов с очень маленькой вероятностью (например, сущности, идентифицирующие подписанта). Анализ только лишь уникальных сущностей позволяет исключить те сущности, которые у разных клиентов могут повторяться, например, валюта кредита, которая чаще всего бывает в рублях и прочие сущности в зависимости от типа документа. Например, для документа ИУК или ПД уникальной сущностью является ФИО заемщика. Также уникальными сущностями могут быть ИНН, СНИЛС, серийный номер паспорта и т.д.

Если уникальные сущности не совпадают (например, в отношении документа ИУК - ФИО заемщика), при этом данные о целостности документа указывают на то, что все пункты в документе присутствуют, то модуль 15 определяет, что документ, изображение которого поступило в систему 10, принадлежит другому лицу. Если модулем 15 было определено, что целостность документа неполная, при этом уникальные сущности набора сущностей, например, идентифицирующие подписанта, указывают на то, что документ, изображение которого поступило в систему 10, является документом данного лица, то модуль 15 формирует список сущностей, которые не прошли сверку. Соответственно, если уникальные сущности набора сущностей, идентифицирующие подписанта, совпадают с эталонным набором сущностей и данные о целостности документа указывают на то, что все пункты в документе присутствуют, то модуль 15 определяет, что упомянутый документ является документом данного лица. Алгоритмы модуля 15 анализа свойств параметризованы идентификатором процесса.

Вся собранная в ходе работы всех предыдущих модулей документа информация, за исключением изображений документа, направляется в модуль 16 принятия решения. Если результаты сверки, полученные от модуля 14, являются положительными и данные, полученные от модуля 19, указывают на то, что все необходимые атрибуты подписантов присутствуют на изображении документа в соответствующих его областях (т.е. правило расположения всех подписей выполнено; в данном случае определяется количеству найденных подписей, по взаимному расположению их, исключая места где заведомо не может быть подписи), то модуль 16 записывает в хранилище результатов веб-сервиса 20 обработки документов информацию об успешном прохождении проверки документа. Например, если в пакете документов был только документ ИУК и ДП не требовался, то модуль 16 записывает в упомянутое хранилище веб-сервиса 20 информацию об успешном прохождении проверки документа, а также информацию о решении, в частности, о том, что можно выдавать кредит. Дополнительно в генерируемые и записываемые в хранилище результаты обработки документов модулем 16 заносится информация о наборе сущностей и результаты сверки. Если данные, полученные от модуля 19, указывают на то, что атрибут подписанта отсутствует на изображении документа в соответствующей области, то модуль 16 принятия решения генерирует информацию о том, что что документ следует проверить человеком, в которую также включается информация о результатах сверки.

Соответствующие области (допустимый диапазон координат для атрибутов подписанта) могут быть определены модулем 18 на основе типа документа, который определяется на основе данных о идентификаторе процесса, поступивших от АС 2 Банка, и в дальнейшем поступают в модуль 16. Если результаты сверки являются отрицательными, то модуль 16 принятия решений извлекает из полученных данных информацию о всех сущностях из набора сущностей, которые не прошли сверку данных, и определяет типы этих сущностей. Если тип сущности указывает на то, что сущность является простой сущностью, а информация о качестве сканирования, полученная от модуля 15, указывает на то, что изображению документа назначен высокий показатель качества сканирования, то модуль 16 принятия решения генерирует информацию о том, что документ не прошел проверку, в которую также включается информация о результатах сверки, и что в выдаче кредита следует отказать. В то же время если информация о качестве сканирования указывает на то, что изображению документа назначен низкий показатель качества сканирования, то модуль 16 генерирует и записывает в хранилище результатов обработки документов веб-сервиса 20 информацию о том, что документ следует проверить человеком, в которую также включается информация о результатах сверки.

Если сущность, не прошедшая сверку данных, является сложной сущностью, то модуль 16 принятия решения, независимо от показателя качества сканирования документа, генерирует и записывает в хранилище результатов обработки документов веб-сервиса 20 информацию о том, что документ следует проверить человеком, в которую также включается информация о результатах сверки. В сгенерированные результаты обработки документов при отрицательных результатах сверки также включается инфор-

мация о наличии или отсутствии атрибутов подписанта.

Сгенерированные модулем 16 принятия решений результаты обработки документов могут быть получены через интерфейс веб-сервиса 20 или его API. Веб-сервис 20 формирует ответ в виде json с результатами обработки документа. Данные результаты обработки документов могут быть выведены на устройство отображения данных, например дисплей вычислительного устройства, такого как портативный или стационарный компьютер, терминал связи, мобильный телефон или смартфон, планшет и пр. Например, если документом являлся документ ИУК, то на устройство отображения данных дополнительно может быть выведено решение о выдаче кредита, в отказе в выдаче или о необходимости проверить документ вручную.

Таким образом, за счет того, что результаты обработки документа формируют на основе результатов сравнения набора сущностей, извлеченного из текстовой информации с учетом типа документа, с эталонным набором сущностей для данного документа, обеспечивается высокая точность в выявлении ошибок при проведении автоматизированной интеллектуальной обработки документов, а также ее эффективность, т.е. обеспечивается достижение указанного технического результата. Также за счет использования алгоритмов машинного обучения и NLP-методов, раскрытых в настоящей заявке, и типизации данных дополнительно повышается эффективность и точность в выявлении ошибок при проведении автоматизированной интеллектуальной обработки документов. Кроме того, представленное техническое решение обладает расширенными функциональными возможностями по сравнению с известными решениями, в частности: обеспечивает возможность автоматизированного принятия решения о выдаче кредита, выявления причины отказа либо обоснования передачи документа на проверку человеку; обеспечивает механизм проверки юридической валидности и комплектности документов. В общем виде (см. фиг. 3) система (200) интеллектуальной обработки документов содержит объединенные общей шиной информационного обмена один или несколько процессоров (201), средства памяти, такие как ОЗУ (202) и ПЗУ (203), интерфейсы ввода/вывода (204), устройства ввода/вывода (205) и устройство для сетевого взаимодействия (206).

Процессор (201) (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 (200) также необходимо учитывать графический процессор, например GPU NVIDIA с программной моделью, совместимой с CUDA, или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа, а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (202) представляет собой оперативную память и предназначено для хранения исполняемых процессором (201) машиночитаемых инструкций для выполнения необходимых операций по логической обработке данных. ОЗУ (202), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом в качестве ОЗУ (202) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (203) 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 (200) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (204). 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 (200) применяются различные средства (205) В/В информации, например, клавиатура, дисплей (монитор), сенсорный дисплей, тач-пад, 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 (206) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (206) 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.

Дополнительно могут применяться также средства спутниковой навигации в составе системы (200), например GPS, ГЛОНАСС, BeiDou, Galileo. Конкретный выбор элементов устройства (200) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала.

Модификации и улучшения вышеописанных вариантов осуществления настоящего технического

решения будут ясны специалистам в данной области техники. Предшествующее описание представлено только в качестве примера и не несет никаких ограничений. Таким образом, объем настоящего технического решения ограничен только объемом прилагаемой формулы изобретения.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ интеллектуальной обработки документов для проверки на корректность их заполнения, выполняемый по меньшей мере одним вычислительным устройством, содержащий этапы, на которых

получают по меньшей мере одно изображение документа;

распознают символы на изображении документа и преобразуют их в текстовую информацию;

на основе текстовой информации, полученной на предыдущем этапе, определяют тип документа;

извлекают из текстовой информации набор сущностей с учетом типа документа;

сравнивают набор сущностей с эталонным набором сущностей для данного документа;

на основе результатов сравнения упомянутых наборов сущностей формируют результаты обработки документа, содержащие информацию об успешном прохождении проверки документа или информацию о том, что документ не прошел проверку.

2. Способ по п.1, характеризующийся тем, что документ представляет собой договор об индивидуальных условиях кредитования (ИУК) или договор поручительства (ДП).

3. Способ по п.1, характеризующийся тем, что дополнительно содержит этапы, на которых

осуществляют детектирование на поступившем изображении документа атрибута подписанта;

определяют расположение по меньшей мере одного атрибута подписанта на странице документа;

при этом результаты обработки документа формируют с учетом информации о расположении по меньшей мере одного атрибута подписанта на странице документа.

4. Способ по п.3, характеризующийся тем, что дополнительно содержит этап, на котором определяют статус лица, которому принадлежит детектированный атрибут подписанта.

5. Способ по п.1, характеризующийся тем, что этап, на котором определяют тип документа на основе текстовой информации, содержит этапы, на которых

получают идентификатор процесса;

определяют набор моделей классификации текста на основе идентификатора процесса;

преобразуют полученную текстовую информацию в набор векторов;

обрабатывают набор векторов с помощью определенного ранее набора моделей классификации текста для определения типа документа.

6. Способ по п.1, характеризующийся тем, что этап, на котором сравнивают набор сущностей с эталонным набором сущностей, содержит этапы, на которых

делят набор сущностей на простые сущности, состоящие из 1-3 слов, и сложные сущности, состоящие по меньшей мере из четырех слов;

причем если в результате сравнения упомянутых наборов сущностей пороговые значения совпадающих слов по простым и сложным сущностям достигнуты, то формируют результаты сверки, в которые включается информация об успешном прохождении сверки данных;

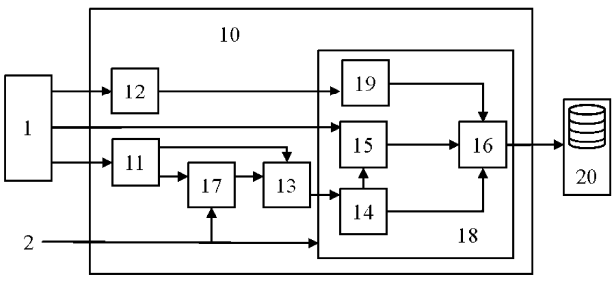
если упомянутые пороговые значения совпадающих слов по простым и сложным сущностям не достигнуты, то формируют результаты сверки, в которые включается информация о сущностях в наборе сущностей, не прошедших сверку;

при этом результаты обработки документа формируют с учетом результатов сверки.

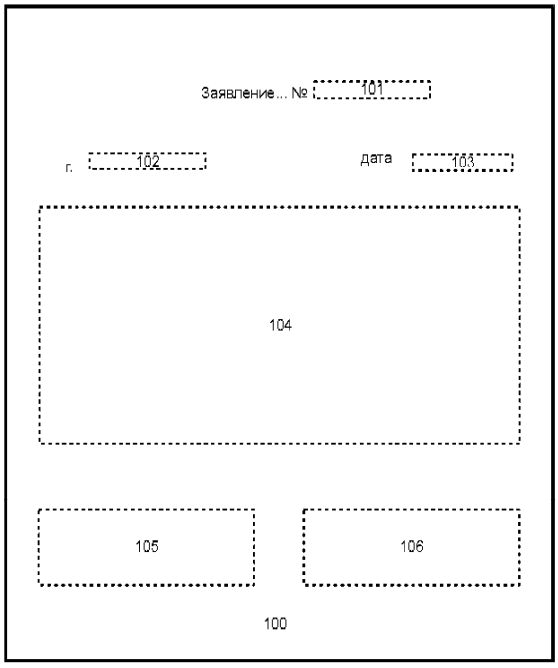
7. Способ по п.1, характеризующийся тем, что дополнительно содержит этап, на котором определяют качество сканирования документа;

причем результаты обработки документа формируют с учетом качества сканирования документа.

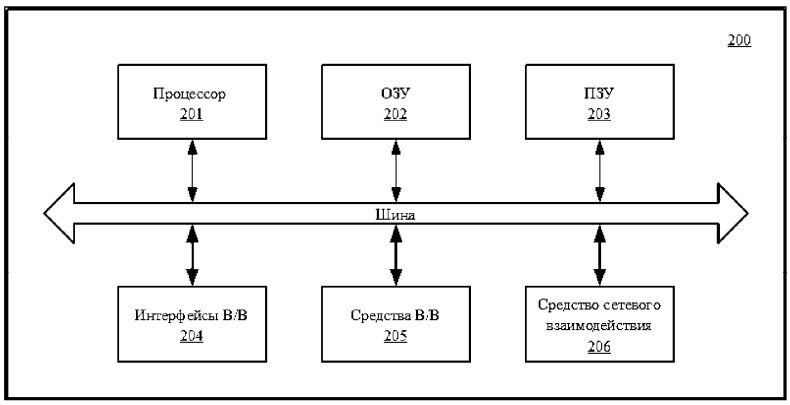
8. Система интеллектуальной обработки документов, содержащая по меньшей мере одно вычислительное устройство и по меньшей мере одно устройство памяти, содержащее машиночитаемые инструкции, которые при их исполнении по меньшей мере одним вычислительным устройством выполняют способ по любому из пп.1-7.



Фиг. 1



Фиг. 2



Фиг. 3

Евразийская патентная организация, ЕАПВ
Россия, 109012, Москва, Малый Черкасский пер., 2