(12) МЕЖДУНАРОДНАЯ ЗАЯВКА, ОПУБЛИКОВАННАЯ В СООТВЕТСТВИИ С ДОГОВОРОМ О ПАТЕНТНОЙ КООПЕРАЦИИ (PCT)

(19) Всемирная Организация Интеллектуальной Собственности
Международное бюро

(43) Дата международной публикации
07 мая 2020 (07.05.2020)

WIPO | PCT

(10) Номер международной публикации
WO 2020/091618 A1

(51) Международная патентная классификация:
G06F 17/27 (2006.01)

(21) Номер международной заявки: PCT/RU2018/000712

(22) Дата международной подачи:
30 октября 2018 (30.10.2018)

(25) Язык подачи: Русский

(26) Язык публикации: Русский

(71) Заявитель: **ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ "МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ ИНСТИТУТ (ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)" (FEDERALNOE GOSUDARSTVENNOE AVTONOMNOE OBRAZOVATELNOE UCHREZHDENIE VYSSHEGO OBRAZOVANIYA "MOSKOVSKY FIZIKO-TEKHNICHESKY INSTITUT (GOSUDARSTVENNY UNIVERSITET)")** [RU/RU]; ул. Керченская, 1А, корп.1 Москва, 119049, Moscow (RU).

(72) Изобретатель: **СУВОРОВ, Владимир Александрович (SUVOROV, Vladimir Alexandrovich)**; ул. Турку, 12, корп. 2, кв. 233 Санкт-Петербург, 192238, St.Petersburg (RU).

(74) Агент: **КОТЛОВ, Дмитрий Владимирович (KOTLOV, Dmitry Vladimirovich)**; ООО "ЦИС "Сколково", Территория инновационного центра "Сколково", д. 4, офис 402.1 Москва, 143026, Moscow (RU).

(81) Указанные государства *(если не указано иначе, для каждого вида национальной охраны)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Указанные государства *(если не указано иначе, для каждого вида региональной охраны)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), евразийский (AM, AZ, BY, KG, KZ, RU, TJ, TM), европейский патент (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Опубликована:

— *с отчётом о международном поиске (статья 21.3)*

WO 2020/091618 A1

(54) **Title:** SYSTEM FOR IDENTIFYING NAMED ENTITIES WITH DYNAMIC PARAMETERS

(54) **Название изобретения:** СИСТЕМА ОПРЕДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ДИНАМИЧЕСКИМИ ПАРАМЕТРАМИ

(57) **Abstract:** The invention relates to the use of computing systems to extract information from texts in a natural language. The technical result consists in increasing the accuracy of automatic identification of named entities from a text. The claimed system for identifying named entities with dynamic parameters comprises a module for processing original texts in a natural language, a module for determining subject areas of named entities, a decision model module and a module with a set of models for extracting named entities from the text, wherein a text in a natural language is received, said text comprising unlabelled sets of sentences or labelled sets of sentences containing defined types of entities, the obtained text is analysed so as to find parts of the text that contain named entities, the named entities are classified, it is determined which particular model for extracting named entities from among the set of models will identify named entities from the text, and the model selected from the set is launched.

(57) **Реферат:** Изобретение относится к извлечению информации из текстов на естественном языке с помощью вычислительных систем. Технический результат заключается в повышении точности автоматического определения именованных сущностей из текста. Заявленная система определения именованных сущностей с динамическими параметрами содержит модуль процессинга исходных текстов на естественном языке, модуль определения тематик именованных сущностей, модуль решающей модели и модуль набора моделей извлечения именованных сущностей из текста, при этом принимают текст на естественном языке, содержащий неразмеченные наборы предложений или размеченные наборы предложений, содержащие сущности определенного типа, анализируют полученный текст для нахождения участков текста, содержащих именованные сущности, классифицируют именованные сущности, определяют какая именно модель извлечения именованных сущностей из набора моделей будет определять из текста именованные сущности, запускают выбранную из набора модель.

СИСТЕМА ОПРЕДЕЛЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ С ДИНАМИЧЕСКИМИ ПАРАМЕТРАМИ

ОБЛАСТЬ ИЗОБРЕТЕНИЯ

[0001] Настоящее изобретение относится к извлечению информации из текстов на естественном языке с помощью вычислительных систем, в частности, к системам определения именованных сущностей с динамическими параметрами.

УРОВЕНЬ ТЕХНИКИ

[0002] Извлечение информации является одной из важнейших операций автоматизированной обработки текстов на естественном языке.

[0003] В настоящий момент задача выделения сущностей в документе требуется в ряде отраслей.

[0004] Например, в юридических документах требуется извлекать номер договора, фамилию и имя судьи, заявителя, ответчика, статью кодекса по которой ведется дело, основание для возбуждения дела, даты – все это может быть использовано, например, для поиска документа или автоматического составления шаблона заявления о опротестовании решения в высшей инстанции суда.

[0004] В медицинских документах выделение сущностей таких как возраст пациента, болезнь, лекарства которыми его лечили, даты лечения, прогноз заболевания может быть использовано для автоматического построения карты вероятностей успехов лечения

[0005] В финансовой сфере анализируя разговор клиента с финансовым советником после выделения сущностей таких как зарплата клиента, количество накоплений, количество членов семьи, желаемый срок инвестиций, отношение к определенным ситуациям в жизни можно составить риск профиль и сделать персонализированное предложение клиенту

[0006] Термин «именованная сущность» впервые был введен на шестой Конференции по Пониманию Сообщений (Message Understanding Conference, MUC-6) в 1996 году. MUC-6 и предшествующие ей Конференции по Пониманию Сообщений были посвящены задачам по извлечению информации: получение структурированной информации о компаниях и военных операциях из неструктурированных текстов, как, например, газетных статей и военных сводок.

[0007] При постановке задачи по извлечению информации было замечено, что необходимо различать такие информационные единицы, как персона, организация, локация и числовые выражения, включающие в себя время, даты, деньги и проценты. Идентификация в тексте данных сущностей была признана одной из важнейших подзадач извлечения информации и была названа извлечение именованных сущностей.

[0008] В настоящее время используется ряд подходов, каждый из которых обладает своими преимуществами и недостатками.

- Использование словарей сущностей. Этот подход удобен, когда вариантов сущностей малое количество и все их возможно описать – например требуется найти все упоминания закрытого списка членов совета директоров компании. Работает с отличным качеством, зачастую не требует контекста. Однако не справляется если сущность не попала в список. Также необходимо вести и актуализировать списки сущностей.

- Онтологический/семантический подход. Используется граф знаний и разметка частей речи.

- Нейросетевой подход. Используя большое количество размеченных данных возможно обучить систему выделять размеченные сущности, упоминаемые в том же контексте. Требуется большое количество данных, однако система позволяет выделять сущности произвольного размера.

[0009] Из заявки US20090144609A1 (МПК G06F17/00, G06F17/30, опубл. 04.06.2009) известны способ и система обработки естественного языка, в части распознавания сущностей (ERDS). В заявке описывается автоматическое определение того, к каким объектам относится текст, используя методы обработки естественного языка и анализ информации, полученной из контекстуальных данных в тексте.

[0010] Из патента US9619457B1 (МПК G06F17/27, G06K9/00, опубл. 11.04.2017) известен способ обработки текста на естественном языке и, более конкретно, к методам автоматической идентификации значимых объектов в документах. Этот способ описывает идентификацию набора упоминаний сущностей в каждом абстрактном и каждом соответствующем документе на основе их соответствующих частей речи (POS) и анализе зависимостей.

[0011] Из заявки US20140195884A1 (МПК G06F17/21, G06F17/30, опубл. 10.07.2014) известен способ автоматического извлечения и организации информации устройством обработки из множества источников данных. Для источников данных применяется

конвейер для извлечения информации об обработке естественного языка, который включает автоматическое обнаружение объектов. Информация об обнаруженных объектах идентифицируется путем анализа продуктов конвейера обработки естественного языка. Идентифицированная информация сгруппирована в классы эквивалентности, содержащие эквивалентную информацию. Создается хотя бы одно отображаемое представление классов эквивалентности. Вычисляется порядок, в котором отображается хотя бы одно отображаемое представление. Производится комбинированное представление классов эквивалентности, которое соответствует порядку отображения отображаемого представления.

[0012] Таким образом, известные из уровня техники решения, предназначенные для обработки естественного языка, имеют ограниченную функциональность в части выбора определенной архитектуры для распознавания сущностей, а также основным недостатком является то, что в известных решениях обучение осуществляется с использованием метода частичного обучения является чрезмерной генерацией шаблонов, которая для точных и полных результатов требует валидации экспертом.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

[0013] Технической проблемой, на решение которой направлено заявленное техническое решение, является создание системы определения именованных сущностей с динамическими параметрами. Также существенной решаемой проблемой подобного рода технических решений является осуществление возможности из потока текста узнавать требуемые сущности совершенно разной природы, имея примеры предложений содержащие определенные сущности.

[0014] Технический результат заключается в повышении точности автоматического определения именованных сущностей из текста.

[0015] Именованные сущности (NER - Named-Entity Recognizing)— это объекты определенного типа, чаще всего составные, например, названия организаций, имена людей, даты, места, денежные единицы и т.д. Эти категории могут быть представлены концептами заранее определенной или динамически выстраиваемой онтологии.

Примеры именованных сущностей:

- Имена персоналий: И. Сечин, Ben White;

- Географические названия: р. Ока, гор. Москва;

- Названия компаний/организаций: РЖД, ОАО «Уют»;

- Даты и временные отрезки: 02.03.1913, 2 p.m;

- Номера телефонов: +7(123)456-78-90;

- Адреса: 3-ая улица Строителей д. 25, кв.12;

- Марки товаров: Nokia, Apple, Land Rover;

- Обозначения денежных единиц: руб., $, GBP;

- Ссылки на литературу: [2], [Иванов, 1995];

- Гены, белки, хим. вещества: H2N–CH(R)–COOH.

[0016] Более сложной сущностью являются такие сущности как: приветствие, прощание, разговор о погоде. Сущность может содержать любое количество слов.

[0017] Каждая именованная сущность – это n-грамма в тексте, для которой определен класс. Некоторые примеры классов именованных сущностей:

Новостная тематика: имена персон, названия организаций и геолокаций.

Биологическая тематика: названия протеинов, клеток.

[0018] В предпочтительном варианте реализации заявлена система определения именованных сущностей с динамическими параметрами, содержащая: модуль процессинга исходных текстов на естественном языке, модуль определения тематик именованных сущностей, модуль решающей модели и модуль набора моделей извлечения именованных сущностей из текста, при этом

модуль процессинга исходных текстов на естественном языке, выполнен с возможностью:

- приема текста на естественном языке, содержащего неразмеченные наборы предложений или размеченные наборы предложений, содержащие сущности определенного типа;

- анализа полученного текста, для нахождения участков текста, содержащих именованные сущности, а именно: разбивают полученный текст на предложения, осуществляют семантический анализ предложений, после чего выполняют лемматизацию и стемминг слов, при этом образуя массив слов из которого удаляются стоп-слова и пунктуации, группируют слова со схожим смысловым значением, присваивают вес каждому слову в зависимости от частоты встречаемости его в тексте и определяют именованные сущности из сгруппированных слов;

модуль определения тематик именованных сущностей, выполнен с возможностью: классифицировать именованные сущности, которые были определены в тексте на разные классы тематик, обнаруживать и классифицировать предопределённых связи между сущностями, которые

распознаны в тексте и с помощью онтологического графа связывают классифицированные именованные сущности с соответствующими сущностями из открытой базы знаний;

модуль решающей модели, выполнен с возможностью принимать на вход классифицированные тематики именованных сущностей, слабосвязанные тематики и слова с наибольшим весом и с помощью, обученной нейросетевой модели определять какая именно модель извлечения: именованных сущностей из набора моделей будет определять из текста именованные сущности;

и модуль набора моделей извлечения именованных сущностей из текста запускает выбранную из набора модель, которая определяет именованные сущности, которые описывают определенные объекты.

[0019] В частном варианте семантическая структура из множества семантических структур представлена графом, включающим множество узлов, соответствующих множеству семантических классов, и множество дуг, соответствующих множеству семантических отношений.

[0020] В другом частном варианте стоп-слова выполнены в виде предлогов и вспомогательных частей речи, а также с использованием базы стоп-слов.

[0021] В другом частном варианте отношения между предложениями могут быть сильно связанные, это если тематики могут пересекаться и слабо связанные.

[0022] В другом частном варианте для классификации тематик используется библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов (BigARTM).

[0023] В другом частном варианте открытой базой знаний, может быть: DBpedia, Freebase, OpenCyc, Wikidata, YAGO, Wordnet.

ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0024] Реализация изобретения будет описана в дальнейшем в соответствии с прилагаемыми чертежами, которые представлены для пояснения сути изобретения и

никоим образом не ограничивают область изобретения. К заявке прилагаются следующие чертежи:

Фиг. 1 иллюстрирует общий вид заявленной системы;

Фиг. 2 иллюстрирует общую схему вычислительного устройства.

ДЕТАЛЬНОЕ ОПИСАНИЕ ИЗОБРЕТЕНИЯ

[0025] В приведенном ниже подробном описании реализации изобретения приведены многочисленные детали реализации, призванные обеспечить отчетливое понимание настоящего изобретения. Однако, квалифицированному в предметной области специалисту, будет очевидно каким образом можно использовать настоящее изобретение, как с данными деталями реализации, так и без них. В других случаях хорошо известные методы, процедуры и компоненты не были описаны подробно, чтобы не затруднять излишне понимание особенностей настоящего изобретения.

[0026] Кроме того, из приведенного изложения будет ясно, что изобретение не ограничивается приведенной реализацией. Многочисленные возможные модификации, изменения, вариации и замены, сохраняющие суть и форму настоящего изобретения, будут очевидными для квалифицированных в предметной области специалистов.

[0027] Настоящее изобретение направлено на создание системы определения именованных сущностей с динамическими параметрами.

[0028] Как представлено на Фиг. 1, заявленная система (100) включает связанные между модуль процессинга исходных текстов на естественном языке (110), модуль определения тематик именованных сущностей (120), модуль решающей модели (130) и модуль набора моделей извлечения именованных сущностей из текста (140).

[0029] Модуль процессинга исходных текстов на естественном языке (110) выполнен с возможностью приема текста на естественном языке, содержащего неразмеченные наборы предложений или размеченные наборы предложений, содержащие сущности определенного типа.

[0030] Ниже приведено два примера, как выглядит размеченный текст:

Первый пример:

«[PERSON Bill Gates] is one of the founders of [ORGANIZATION Microsoft], a company well-known in [LOCATION USA] and all over the world.»

В первом примере в тексте распознаны именованные сущности (которыми являются слова и словосочетания) и проклассифицированы по предопределенным категориям, таким, как:

- личность – PERSON;

- организации – ORGANIZATION;

- географические объекты – LOCATION.

Второй пример:

[[Андрей] PER]] купил [10 [CNT]] деталей [Лего [COM]]

Во втором примере в тексте распознаны именованные сущности (которыми являются слова) и проклассифицированы по предопределенным категориям, таким, как:

- личность – PERSON;

- количество - CNT;

- компания - COM.

[0030] Далее происходит анализ полученного текста, для того, чтобы найти те участков текста, в которых содержатся именованные сущности.

[0031] Для этого полученный текст разбивают на предложения, после чего осуществляют семантический анализ предложений, то есть определяет смысл этих предложений.

[0032] Семантическая структура из множества семантических структур представлена графом, включающим множество узлов, соответствующих множеству семантических классов, и множество дуг, соответствующих множеству семантических отношений. Семантический анализ может быть построен на базе элементов библиотеки iPavlov или StanfordNLP.

[0033] После чего выполняют лемматизацию и стемминг слов, при этом образуя массив слов из которого удаляются стоп-слова и пунктуации, группируют слова со схожим смысловым значением, присваивают вес каждому слову в зависимости от частоты встречаемости его в тексте и определяют именованные сущности из сгруппированных слов.

[0034] При выполнении этапа лемматизации, осуществляют приведение слов в тексте в их первоначальную словарную форму, например, бежала – бежать. Данный этап осуществляется с помощью словарей или с помощью интеграции правил языка в код – например, определения части речи, окончания и преобразовании окончания в нужное.

[0035] При этом стоп-слова выполнены в виде предлогов и вспомогательных частей речи, а также с использованием базы стоп-слов

[0036] В процессе стэмминга происходит приведение слов к основе, а именно, отрезание окончания и формообразующего суффикса.

[0037] Группировка слов со схожим смысловым значением осуществляется с использованием метода word2vec. Обученный на поисковых запросах Word2Vec осуществляет кластеризацию слов по принципу их семантической близости, именно, выявляет семантическую близость слов, например, к чему семантически ближе слово Интернет — к компьютерным сетям или космическим пространства). После чего слова со схожим значением располагаются ближе друг к другу.

[0038] Для присвоения веса каждому слову используется функция tf-idf или BM25.

Функция TF-IDF (TF — частота слова, IDF — обратная частота документа) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

Функция BM25 используется для упорядочивания слов в тексте по их релевантности, а именно по весу.

[0039] Модуль определения тематик именованных сущностей (120) выполнен с возможностью классифицировать именованные сущности, которые были определены в тексте на разные классы тематик, обнаруживать и классифицировать предопределённых связи между сущностями, которые распознаны в тексте и с помощью онтологического графа связывают классифицированные именованные сущности с соответствующими сущностями из открытой базы знаний.

[0040] Для классификации тематик используется библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов (BigARTM), можно также использовать метод k-средних (k-means) на векторах.

[0041] Далее с помощью онтологического графа связывают классифицированные именованные сущности с соответствующими сущностями из открытой базы знаний. Данный этап необходим для того чтобы расширять список сущностей и слов, подаваемых на вход решающей модели. Отношения могут быть разной степени связанности. Сильно связанные это когда тематики могут пересекаться – например, тематика – великие политики

и великие военачальники. Слабо связанные это, например, виды слонов и виды бабочек. Степень связанности решается по дереву онтологий. Онтологический граф может быть построен на базе WordNet. Например слово car – машина связано со словами auto, automobile, machine, motorcar (автомобиль, тачка).

[0042] Открытыми базами знаний, могут быть такими, как: DBpedia, Freebase, OpenCyc, Wikidata, YAGO, Wordnet

[0043] Модуль решающей модели (130), выполнен с возможностью принимать на вход классифицированные тематики именованных сущностей, слабосвязанные тематики и слова с наибольшим весом и с помощью, обученной нейросетевой модели например многослойной сверточной сети с функцией оптимизации softmax и 5 слоями нейронов определять какая именно модель извлечения именованных сущностей из набора моделей будет определять из текста именованные сущности.

[0044] Модуль набора моделей извлечения именованных сущностей из текста (140) запускает выбранную из набора модель, которая определяет именованные сущности, которые описывают определенные объекты.

[0045] На Фиг. 2 далее будет представлена общая схема компьютерного устройства (200), обеспечивающего обработку данных, необходимую для реализации заявленного решения.

[0046] В общем случае устройство (200) содержит такие компоненты, как: один или более процессоров (201), по меньшей мере одну память (202), средство хранения данных (203), интерфейсы ввода/вывода (204), средство В/В (205), средства сетевого взаимодействия (206).

[0047] Процессор (201) устройства выполняет основные вычислительные операции, необходимые для функционирования устройства (200) или функционала одного или более его компонентов. Процессор (201) исполняет необходимые машиночитаемые команды, содержащиеся в оперативной памяти (202).

В частности, процессор (201) может представлять собой микропроцессор с полным набором команд (CISC), микропроцессор с сокращенным набором команд (RISC), микропроцессор с командными словами сверхбольшой длины (VLIW), процессор, реализующий другой набор команд, или процессоры, реализующие комбинацию наборов команд. Процессор (201) также может представлять собой одно или несколько вычислительных устройств специального назначения, например, заказную интегральную микросхему (ASIC), программируемую пользователем вентильную матрицу (FPGA), процессор цифровых

сигналов (DSP), сетевой процессор и т.п. Процессор (201) настроен на выполнение команд для осуществления рассмотренных в настоящем документе операций и функций.

[0048] Память (202), как правило, выполнена в виде ОЗУ и содержит необходимую программную логику, обеспечивающую требуемый функционал.

[0049] Средство хранения данных (203) может выполняться в виде HDD, SSD дисков, рейд массива, сетевого хранилища, флэш-памяти, оптических накопителей информации (CD, DVD, MD, Blue-Ray дисков) и т.п. Средство (203) позволяет выполнять долгосрочное хранение различного вида информации, например, истории обработки транзакционных запросов (логов), идентификаторов пользователей и т.п.

[0050] Интерфейсы (204) представляют собой стандартные средства для подключения, например, USB, RS232, RJ45, LPT, COM, HDMI, PS/2, Lightning, FireWire и т.п.

[0051] Выбор интерфейсов (204) зависит от конкретного исполнения устройства (200), которое может представлять собой персональный компьютер, мейнфрейм, серверный кластер, тонкий клиент, смартфон, ноутбук и т.п.

[0052] В качестве средств В/В данных (205) может использоваться: клавиатура, джойстик, дисплей (сенсорный дисплей), проектор, тачпад, манипулятор мышь, трекбол, световое перо, динамики, микрофон и т.п.

[0053] Средства сетевого взаимодействия (206) выбираются из устройства, 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 (205) обеспечивается организация обмена данными по проводному или беспроводному каналу передачи данных, например, WAN, PAN, ЛВС (LAN), Интранет, Интернет, WLAN, WMAN или GSM.

[0054] Компоненты устройства (200) сопряжены посредством общей шины передачи данных (210).

[0055] В настоящих материалах заявки было представлено предпочтительное раскрытие осуществление заявленного технического решения, которое не должно использоваться как ограничивающее иные, частные воплощения его реализации, которые не выходят за рамки испрашиваемого объема правовой охраны и являются очевидными для специалистов в соответствующей области техники.

[0056] Некоторые части описания предпочтительных вариантов реализации представлены в виде алгоритмов и символического представления операций с битами данных в памяти компьютера. Такие описания и представления алгоритмов представляют собой средства, используемые специалистами в области обработки данных, чтобы наиболее эффективно передавать сущность своей работы другим специалистам в данной области. В настоящем документе и в целом алгоритмом называется самосогласованная последовательность операций, приводящих к требуемому результату. Операции требуют физических манипуляций с физическими величинами. Обычно, хотя и не обязательно, эти величины принимают форму электрических или магнитных сигналов, которые можно хранить, передавать, комбинировать, сравнивать и подвергать другим манипуляциям. Оказалось, что прежде всего для обычного использования удобно описывать эти сигналы в виде битов, значений, элементов, символов, членов, цифр и т.д.

[0057] Однако следует иметь в виду, что все эти и подобные термины должны быть связаны с соответствующими физическими величинами, и что они представляют собой просто удобные метки, применяемые к этим величинам. Если иное специально и недвусмысленно не указано в нижеследующем обсуждении, следует принимать, что везде по тексту такие термины как "определение", "вычисление", "расчет", "вычисление", "получение", "установление", "изменение" и т.п., относятся к действиям и процессам вычислительного устройства или аналогичного электронного вычислительного устройства, которое работает с данными и преобразует данные, представленные в виде физических (например, электронных) величин в регистрах и памяти вычислительного устройства, в другие данные, аналогичным образом представленные в виде физических величин в памяти или регистрах вычислительного устройства, либо других подобных устройствах хранения, передачи или отображения информации.

[0058] Следует понимать, что вышеприведенное описание носит иллюстративный, а не ограничительный характер. Различные другие варианты осуществления станут очевидны специалистам в данной области техники после прочтения и понимания приведенного выше описания. Поэтому объем раскрытия должен определяться со ссылкой на прилагаемую формулу изобретения наряду с полным объемом эквивалентов, на которые такие требования предоставляют право.

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Система определения именованных сущностей с динамическими параметрами, содержащая: модуль процессинга исходных текстов на естественном языке, модуль определения тематик именованных сущностей, модуль решающей модели и модуль набора моделей извлечения именованных сущностей из текста, при этом

модуль процессинга исходных текстов на естественном языке, выполнен с возможностью:

- приема текста на естественном языке, содержащего неразмеченные наборы предложений или размеченные наборы предложений, содержащие сущности определенного типа;

- анализа полученного текста, для нахождения участков текста, содержащих именованные сущности, а именно: разбивают полученный текст на предложения, осуществляют семантический анализ предложений, после чего выполняют лемматизацию и стемминг слов, при этом образуя массив слов из которого удаляются стоп-слова и пунктуации, группируют слова со схожим смысловым значением, присваивают вес каждому слову в зависимости от частоты встречаемости его в тексте и определяют именованные сущности из сгруппированных слов;

модуль определения тематик именованных сущностей, выполнен с возможностью: классифицировать именованные сущности, которые были определены в тексте на разные классы тематик, обнаруживать и классифицировать предопределённых связи между сущностями, которые

распознаны в тексте и с помощью онтологического графа связывают классифицированные именованные сущности с соответствующими сущностями из открытой базы знаний;

модуль решающей модели, выполнен с возможностью принимать на вход классифицированные тематики именованных сущностей, слабосвязанные тематики и слова с наибольшим весом и с помощью, обученной нейросетевой модели определять какая именно модель извлечения именованных сущностей из набора моделей будет определять из текста именованные сущности;

и модуль набора моделей извлечения именованных сущностей из текста запускает выбранную из набора модель, которая определяет именованные сущности, которые описывают определенные объекты.

2. Система по п.1, характеризующаяся тем, что семантическая структура из множества семантических структур представлена графом, включающим множество узлов,

соответствующих множеству семантических классов, и множество дуг, соответствующих множеству семантических отношений.

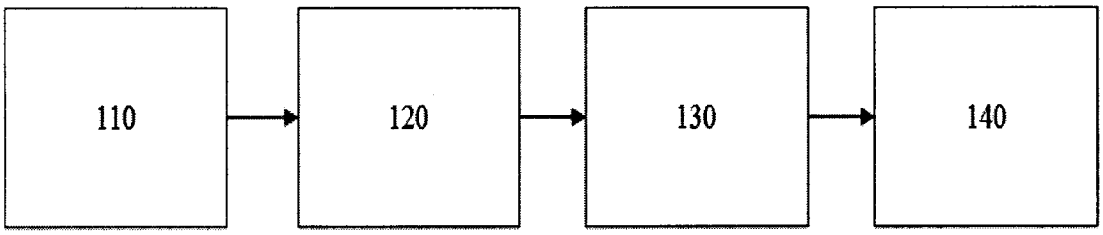
3. Система по п.1, характеризующаяся тем, что стоп-слова выполнены в виде предлогов и вспомогательных частей речи, а также с использованием базы стоп-слов.

4. Система по п.1, характеризующаяся тем, что для присвоения веса каждому слову используется функция tf-idf или BM25.

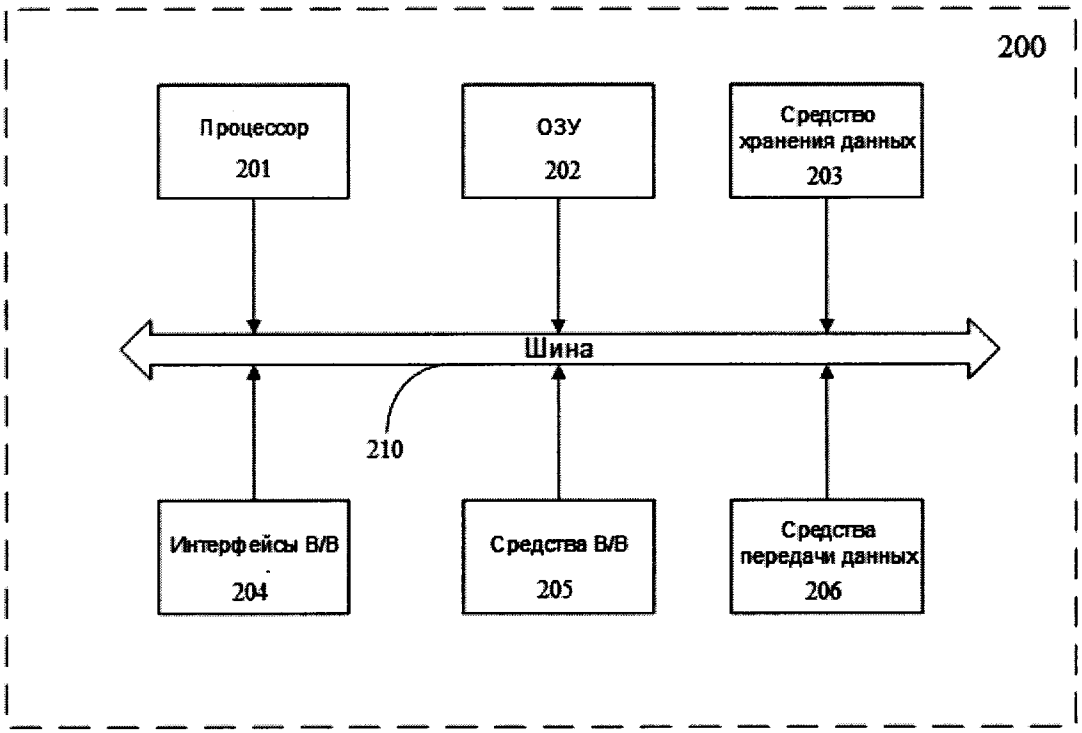
5. Система по п.1, характеризующаяся тем, что отношения между предложениями могут быть сильно связанные, это если тематики могут пересекаться и слабо связанные.

6. Система по п.1, характеризующаяся тем, что для классификации тематик используется библиотека с открытым кодом для тематического моделирования больших коллекций текстовых документов (BigARTM).

7. Система по п.1, характеризующаяся тем, что открытой базой знаний, может быть: DBpedia, Freebase, OpenCyc, Wikidata, YAGO, Wordnet.


110
120
130
140
100

Фиг. 1


200
Процессор
201
ОЗУ
202
Средство хранения данных
203
Шина
210
Интерфейсы В/В
204
Средства В/В
205
Средства передачи данных
206

Фиг. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/RU 2018/000712

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/27 (2006.01)

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatSearch (RUPTO internal), USPTO, PAJ, K-PION, Esp@cenet, Information Retrieval System of FIPS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	RU 2665239 C2 (OBSCHESTVO S OGRANICHENNOI OTVETSTVENNOSTJU "ABI PRODAKSHN") 20.08.2018	1-7
A	RU 2619193 C1 (OBSCHESTVO S OGRANICHENNOI OTVETSTVENNOSTJU "ABI INFOPOISK") 12.05.2017	1-7
A	RU 2636098 C1 (OBSCHESTVO S OGRANICHENNOI OTVETSTVENNOSTJU "ABI PRODAKSHN") 20.11.2017	1-7
A	US 9135238 B2 (GOOGLE INC) 15.09.2015	1-7

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 July 2019 (15.07.2019)	18 July 2019 (18.07.2019)
Name and mailing address of the ISA/ RU	Authorized officer
Facsimile No.	Telephone No.

Form PCT/ISA/210 (second sheet) (July 1998)

ОТЧЕТ О МЕЖДУНАРОДНОМ ПОИСКЕ

Номер международной заявки

PCT/RU 2018/000712

A. КЛАССИФИКАЦИЯ ПРЕДМЕТА ИЗОБРЕТЕНИЯ

G06F 17/27 *(2006.01)*

Согласно Международной патентной классификации МПК

B. ОБЛАСТЬ ПОИСКА

Проверенный минимум документации (система классификации с индексами классификации)

G06F

Другая проверенная документация в той мере, в какой она включена в поисковые подборки

Электронная база данных, использовавшаяся при поиске (название базы и, если, возможно, используемые поисковые термины)

PatSearch (RUPTO internal), USPTO, PAJ, K-PION, Esp@cenet, Information Retrieval System of FIPS

C. ДОКУМЕНТЫ, СЧИТАЮЩИЕСЯ РЕЛЕВАНТНЫМИ:

Категория*	Цитируемые документы с указанием, где это возможно, релевантных частей	Относится к пункту №
A	RU 2665239 C2 (ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АБИ ПРОДАКШН") 20.08.2018	1-7
A	RU 2619193 C1 (ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АБИ ИНФОПОИСК") 12.05.2017	1-7
A	RU 2636098 C1 (ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АБИ ПРОДАКШН") 20.11.2017	1-7
A	US 9135238 B2 (GOOGLE INC) 15.09.2015	1-7

☐ последующие документы указаны в продолжении графы C.

☐ данные о патентах-аналогах указаны в приложении

* Особые категории ссылочных документов:

"A" документ, определяющий общий уровень техники и не считающийся особо релевантным

"E" более ранняя заявка или патент, но опубликованная на дату международной подачи или после нее

"L" документ, подвергающий сомнению притязание(я) на приоритет, или который приводится с целью установления даты публикации другого ссылочного документа, а также в других целях (как указано)

"O" документ, относящийся к устному раскрытию, использованию, экспонированию и т.д.

"P" документ, опубликованный до даты международной подачи, но после даты испрашиваемого приоритета

"T" более поздний документ, опубликованный после даты международной подачи или приоритета, но приведенный для понимания принципа или теории, на которых основывается изобретение

"X" документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает новизной или изобретательским уровнем, в сравнении с документом, взятым в отдельности

"Y" документ, имеющий наиболее близкое отношение к предмету поиска; заявленное изобретение не обладает изобретательским уровнем, когда документ взят в сочетании с одним или несколькими документами той же категории, такая комбинация документов очевидна для специалиста

"&" документ, являющийся патентом-аналогом

Дата действительного завершения международного поиска	Дата отправки настоящего отчета о международном поиске
15 июля 2019 (15.07.2019)	18 июля 2019 (18.07.2019)
Наименование и адрес ISA/RU: Федеральный институт промышленной собственности, Бережковская наб., 30-1, Москва, Г-59, ГСП-3, Россия, 125993 Факс: (8-495) 531-63-18, (8-499) 243-33-37	Уполномоченное лицо: Дацко Ю. Телефон № (495) 531-65-15