(19) Евразийское патентное ведомство

(11) 038241 (13) B1

# (12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ЕВРАЗИЙСКОМУ ПАТЕНТУ

(45) Дата публикации и выдачи патента
**2021.07.29**

(21) Номер заявки
**201990538**

(22) Дата подачи заявки
**2019.03.19**

(51) Int. Cl. ***G06F 16/22*** (2006.01)
***G06F 17/27*** (2006.01)

## (54) СПОСОБ И СИСТЕМА ПОИСКА РЕЛЕВАНТНЫХ НОВОСТЕЙ

(31) **2019107328**
(32) **2019.03.14**
(33) **RU**
(43) **2020.09.30**

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

(72) Изобретатель:
**Федоров Федор Борисович, Липачева Александра Евгеньевна, Кузнецов Владимир Алексеевич, Черкасов Роман Владиславович (RU)**

(74) Представитель:
**Герасин Б.В. (RU)**

(56) RU-C2-2382401
RU-C2-2608884
RU-C2-2629449
US-A1-20110309139
US-B1-9384211

(57) Настоящее решение относится к области информационных технологий, в частности к поисковым механизмам, предназначенным для выявления релевантной информации из разнородных источников данных. Техническим результатом является обеспечение формирования связанного набора информации из новостных источников с группировкой по компаниям, являющимся объектами новостей, и заданными типами событий. В первом предпочтительном варианте осуществления заявленного решения представлен компьютерно-реализуемый способ поиска релевантных новостей, в котором получают на управляющем сервере набор новостей по меньшей мере от одного сервера новостного агрегатора; осуществляют на управляющем сервере анализ полученного набора новостей, который включает в себя лемматизацию текстов каждой новости из упомянутого набора новостей; обработку полученных лемм текстов новостей с помощью модели машинного обучения, которая содержит установленный набор данных компаний и список событий, причем для каждого события в модели машинного обучения установлен заданный набор лемм; определение новостей, содержащих леммы, идентифицирующие заданные события, и формирование связи выявленных событий по меньшей мере с одной компанией; после чего формируют список релевантных новостей на основании выполненного анализа.

038241 B1

038241 B1

## Область техники

Настоящее техническое решение в общем относится к области информационных технологий, а в частности к поисковым механизмам, предназначенным для выявления релевантной информации из разнородных источников данных.

## Уровень техники

В настоящее время сбор данных (англ. "Data Mining") является важной составляющей для различных сфер бизнеса, в особенности в сферах аналитики и прогнозирования. Зачастую источником данных об информации по интересующим темам являются общедоступные ресурсы в сети Интернет, например новостные ресурсы (веб-сайты, каналы в мессенджерах и т.п.).

При анализе данных основной проблемой является агрегирование массива новостных источников, в частности привязка действительных событий к компаниям для целей последующего поиска. Как правило, на сегодняшний день нет эффективных средств для фильтрации собираемого новостного контента для создания агрегированных массивов информации с привязкой по объектам новостей, например компаниям. Из существующего уровня техники известны различные алгоритмы для сбора данных, например решение, описанное в заявке WO 1999005614 (автор: Louis Gay et al., опубликовано 04.02.1999), которое позволяет агрегировать данные из множества источников и отслеживать ретроспективную актуальность собираемых данных. Из патента RU 2382401 (патентообладатель "Майкрософт корпорейшн", опубликовано 20.02.2010) известен подход для анализа и сравнения совокупностей документов, в соответствии с чем документы могут быть предположительно организованы в группы по своему содержимому или источнику и проанализированы на предмет межгрупповых и внутригрупповых различий и общностей. Например, сопоставление двух групп документов, посвященных одной теме, но полученных из двух различных источников, к примеру, информационного обзора происшествия в различных частях мира, может показать интересные различия мнений и общих истолкований ситуаций. За счет перемещения содержимого из статичных совокупностей в наборы статей, генерируемых во времени, может быть рассмотрено его развитие. Например, поток новостных статей по общему описанию может быть рассмотрен во времени с целью выделения действительно информативных свежих новостей и фильтрования множества статей, которые в значительной степени передают "практически то же самое". Общим недостатком существующих подходов является отсутствие способа выявления релевантных новостей относительно привязки к объекту новости, например компании и соответствующему событию, связанному с ней, что не позволяет эффективно осуществить сбор релевантной информации из множества источников данных.

## Раскрытие изобретения

Решаемой технической проблемой или технической задачей с помощью заявленного подхода является обеспечение процесса поиска и формирования набора новостей с привязкой к заданному набору наименования компаний как объектов новостей и событий, о которых появляется информация в открытых источниках данных. Техническим результатом, достигаемым при решении вышеуказанной технической задачи, является обеспечение формирования связанного набора информации из новостных источников с группировкой по компаниям, являющимся объектами новостей, и заданными типами событий.

Дополнительным техническим результатом является повышение точности выявления информации о компаниях для заданного типа событий в общедоступных источниках информации.

Указанный технический результат достигается благодаря осуществлению компьютерно-реализуемого способа поиска релевантных новостей, в котором

получают на управляющем сервере набор новостей по меньшей мере от одного сервера новостного агрегатора;

осуществляют на управляющем сервере анализ полученного набора новостей, который включает в себя

лемматизацию текстов каждой новости из упомянутого набора новостей;

обработку полученных лемм текстов новостей с помощью модели машинного обучения, которая содержит установленный набор данных компаний и список событий, причем для каждого события в модели машинного обучения установлен заданный набор лемм;

определение новостей, содержащих леммы, идентифицирующие заданные события и

формирование связи выявленных событий по меньшей мере с одной компанией;

формируют список релевантных новостей на основании выполненного анализа.

В одном из частных примеров осуществления способа при получении набора новостей осуществляется фильтрация дублирующих новостей.

В другом частном примере осуществления способа фильтрация осуществляется с помощью вычисления меры Жаккарда между сигнатурами новостей.

В другом частном примере осуществления способа события, присвоенные новости о компании, сохраняют в базе данных.

В другом частном примере осуществления способа новостной агрегатор обновляет список новостей с помощью информационных каналов.

В другом частном примере осуществления способа информационные каналы представляют собой веб-сайты в сети Интернет и/или мессенджер-каналы.

В другом частном примере осуществления способа в ходе анализа новостей выполняется определение принадлежности события основной или дочерней компании.

В другом частном примере осуществления способа принадлежность компании, упоминаемой в новости, определяется с помощью алгоритма решающих деревьев.

В другом частном примере осуществления способа в ходе лемматизации текстов новостей осуществляется их очистка от знаков пунктуации, стоп-слов и именованных сущностей.

В другом частном примере осуществления способа в ходе лемматизации для каждой леммы текста новости рассчитывается статистическая мера.

В другом частном примере осуществления способа алгоритм машинного обучения представляет собой логическую регрессию, классифицирующий принадлежность новости событию на основании анализа статистической меры лемм.

В другом частном примере осуществления способа для каждого текста новости выполняется определение частотных словосочетаний длиной от 2 до 10 лемм.

В другом частном примере осуществления способа алгоритм машинного обучения представляет собой градиентный бустинг, обученный для классификации события на основе количества предложений, содержащих леммы, идентифицирующие событие из поискового запроса.

В другом частном примере осуществления способа после присвоения события из новости компании выполняется выделение лемм и/или предложений, содержащих леммы, идентифицирующее упомянутое событие.

В другом предпочтительном варианте осуществления заявленного решения представлена система поиска релевантных новостей, содержащая по меньшей мере один процессор и по меньшей мере одну память, которая содержит машиночитаемые инструкции, которые при их исполнении по меньшей мере одним процессором выполняют вышеуказанный способ.

**Краткое описание чертежей**

Признаки и преимущества настоящего технического решения станут очевидными из приводимого ниже подробного описания и прилагаемых чертежей.

Фиг. 1 иллюстрирует взаимодействие элементов, входящих в заявленное решение.

Фиг. 2 иллюстрирует общий процесс выполнения способа.

Фиг. 3 иллюстрирует процесс обработки текстовых данных.

Фиг. 4 - представлен пример графического интерфейса пользователя при взаимодействии с сервисом по подбору релевантных новостей.

Фиг. 5 иллюстрирует общий вид вычислительного устройства.

**Осуществление изобретения**

На фиг. 1 представлена общая вычислительная архитектура (100) представленного решения. Основной функционал по сбору и обработке информации выполняется на управляющем сервере (110), который посредством канала передачи данных получает информацию сервера (120) новостного агрегатора, который связан посредством сети Интернет (150) со множеством новостных ресурсов (130). Сервер (110) обеспечивает взаимодействие с пользователями (10) для отображения данных по собранной новостной информации, а также дополнительный функционал, который будет раскрыт далее в материалах заявки.

В качестве канала передачи данных между управляющем сервером (110) и сервером новостного агрегатора (120) может выступать Интернет или Интранет. При этом сервер новостного агрегатора (120) может представлять собой несколько устройств, входящих в состав различного сетевого окружения, например совокупность серверов, маршрутизаторов, кластеров и т.п. Канал передачи данных может быть организован с помощью различного вида известных протоколов передачи данных, как проводных, так и беспроводных, например TCP/IP, 802.11, Ethernet, FTP и др., обеспечивая формирование различного сетевого взаимодействия, в частности LAN, WAN, PAN, WLAN и т.п. Управляющий сервер (110) выполняет основную обработку информации, получаемой от сервера новостного агрегатора (120), хранит и формирует данные для отображения пользователям (10). Отображение информации может формироваться с помощью специализированного графического интерфейса пользователя. Пользователи (10) могут взаимодействовать с управляющим сервером (110) с помощью веб-портала или иного типа программного приложения, обеспечивающего доступ к агрегированной новостной информации. Доступ может предоставляться, например, посредством API. Взаимодействие пользователей (10) может осуществляться с помощью различных электронных устройств, в качестве которых могут выступать, например, компьютер, ноутбук, смартфон, планшет, игровая приставка, умное носимое электронное устройство, тонкий клиент, а также устройства дополненной, смешанной или виртуальной реальности и др.

Сервер новостного агрегатора (120) связан посредством сети Интернет (150) с различными информационными ресурсами (130) или информационными каналами, предоставляющими новостную информацию. Такими ресурсами (130) могут выступать, например, веб-сайты, каналы мессенджеров (Telegram™, WatsApp™, Viber™ и др.), социальные сети (Facebook™, Вконтакте™ и т.п.). Сохранение полученной информации на сервере (110) может осуществляться в формате JSON в хранилище данных, например базе данных. При этом может учитываться источник получения новостной информации и дата ее размещения на соответствующем ресурсе (130).

На фиг. 2 представлен общий процесс выполнения заявленного способа поиска релевантной новостной информации (200). Информация из новостных источников, собранная и хранимая на сервере новостного агрегатора (120), передается (201) на управляющий сервер (110). Информация от сервера новостного агрегатора (120) может передаваться в режиме онлайн или офлайн. В онлайн режиме данные из сети Интернет (150) передаются по факту их появления на веб-ресурсе, к которому имеется подключение у сервера новостного агрегатора (120). В режиме офлайн новости сохраняются на сервере новостного агрегатора (120), например в базе данных, и в установленное время (например, каждый час, раз в день и т.п.) или по запросу от управляющего сервера (110) передаются на него.

Данные от сервера новостного агрегатора (120) могут передаваться в различных форматах, например xml, html, txt и т.п. Формат данных для передачи также может изменяться в зависимости от режима передачи информации на управляющий сервер (110). Помимо самого текста новости, данные содержат информацию о компаниях, упомянутых в тексте.

На управляющем сервере (110) находится сформированный список, содержащий наименования компаний (2021) и событий (2022), на предмет которых осуществляется анализ входящей новостной информации от сервера новостного агрегатора (120). Указанные данные хранятся в базе данных управляющего сервера (110). В качестве событий могут выступать, например, арест/заморозка счетов компании, банкротство компании, наличие исков к компании, обвал/рост акций и т.п. Список событий (2022) и компаний (2021) может обновляться или изменяться в течение времени. Поиск релевантной информации по данным, полученным от сервера новостного агрегатора (120), осуществляется с помощью обработки (202) полученного массива данных с помощью модели машинного обучения, которая обучена осуществлять поиск по наименованиям компаний (2021) и соответствующих событий (2022) в массиве текстовой информации и выдавать суждение о релевантности соответствующей информации. Обработка данных на сервере (110) выполняется по факту получения нового массива данных от сервера новостного агрегатора (120) либо по заранее установленному сценарию. В качестве сценария может настраиваться автоматический скрипт, который в установленное время осуществляет активацию модели машинного обучения для обработки данных (202).

При выполнении этапа обработки (202) выполняется обращение к хранилищу информации управляющего сервера (110), которое содержит полученные от сервера новостного агрегатора (120) данные из новостных источников (130). При доступе к сохраненной на управляющем сервере (110) информации осуществляется ее обработка (202) для выявления релевантных данных и привязки данных (203) из новостей к соответствующим типам событий в ходе обработки информации с помощью модели машинного обучения.

На фиг. 3 представлен процесс (300) осуществления обработки новостных данных, полученных от сервера новостного агрегатора (120), которая осуществляется в процессе выполнения этапов (202)-(203). На первом шаге (301) новостные текстовые данные, полученные от сервера новостного агрегатора (120), проходят лемматизацию, в ходе которой выполняется разделение на леммы корпуса текста каждой новости. Из полученных данных извлекается текст новости и метаданные из файлов. В ходе выполнения процесса лемматизации текстов (301) тело новости разделяется на слова по всем пунктуационным разделителям, после чего приводится к нормальной форме, например с помощью библиотеки pymorphy2. Затем осуществляется преобразование текста, в частности выполняется очистка текста от знаков пунктуации, стоп-слов (предлоги, союзы, местоимения) и именных сущностей. Именной сущностью в данном случае считается любое слово, начинающееся с большой буквы и не являющееся при этом первым словом в предложении. Также может выполняться процесс N-грамминга (https://ru.wikipedia.org/wiki/N-грамма), при котором в тексте выделяются наиболее частотные словосочетания длины от 2 до 10 лемм. Список наиболее частотных словосочетаний получен путем автоматического анализа большого корпуса текста и содержит более 9 млн объектов.

Также входящие новости проходят процедуру дедупликации, в ходе которой отфильтровываются повторяющиеся новости. В ходе выполнения процедуры дедупликации для каждой новости считается сигнатура MinHash (см. https://en.wikipedia.org/wiki/MinHash), после чего для каждой пары новостей вычисляется схожесть сигнатур по мере Жаккара (иногда коэффициент Жаккара). Если схожесть пары новостей превышает заданный порог, например, 0.7, то более короткая новость из пары корпусов текстов считается дублирующей и не подвергается дальнейшей обработке. На следующем шаге (302) после лемматизации текстов новостей выполняется обработка нормализованного текста. В тексте новости осуществляется поиск наименования компаний, не имеющих омонимов (например, "Сбербанк™"). Находятся все словосочетания с большой буквы и в кавычках, после чего проводится поиск лемм найденных словосочетаний в списке компаний, хранимого в базе данных сервера (110). Найденные наименования компаний классифицируются по признаку "основная" или "дополнительная" компания (т.е. которая является косвенно упоминаемой в тексте новости). Компания считается "основной", если она является предметом новости, и "дополнительной", если наименование компании просто упоминается в теленовости. Классификация осуществляется с помощью модели машинного обучения, в частности алгоритма принятия решений, например с помощью решающих деревьев. Список признаков решающего дерева выглядит следующим образом.

1) Номер предложения первого упоминания компании (0, если это заголовок).

2) Номер предложения первого упоминания компании, нормированный на число предложений.

3) Длина текста в символах; значение порога классификации - 0.5.

Для определения релевантности того или иного события для компаний, указываемых в теленовостях, осуществляется обработка полученных лемм из тела новости на шаге (303) с помощью моделей машинного обучения.

В качестве одного примера модели машинного обучения может применяться логическая регрессия с помощью расчета статистической меры TF-IDF для лемм текста (см. https://ru.wikipedia.org/wiki/TF-IDF). Для каждой леммы в тексте считается статистическая мера, после чего на полученных признаках делается суждение заранее обученной логистической регрессии. Помимо обработки с помощью модели машинного обучения, составляется список заданных лемм, например, список может содержать 30-40 лемм, имеющих наибольший вес в логистической регрессии. Список строится для каждого события после процесса обучения логистической регрессии. На выходе модели определяется вес каждой леммы, по которым осуществляется отбор лемм для списка на основании значений их весов.

Если полученное значение вероятности суждения модели (303) выше заранее установленного порога и в тексте новости встречается хотя бы одна лемма из упомянутого списка, то по меньшей мере одно событие присваивается новости (304) для выявленного в тексте наименования компании.

Дополнительно для каждого события может задаваться набор лемм, например 10-15 лемм, наиболее соответствующих событию, которые выделяются из ранее определенного списка лемм, и если событие было присвоено новости на этапе (304), то все найденные в тексте леммы из упомянутого набора выделяются в тексте. Вторым примером применения модели машинного обучения является классифицирующий алгоритм в виде градиентного бустинга, например LightGBM (https://lightgbm.readthedocs.io). Для каждого текста новости считается количество предложений, содержащих пары характерных для события лемм. Пары характерных лемм подбираются для каждого события в ходе обучения классификатора. Характерные леммы (и их количество) подбираются автоматически в ходе обучения.

На полученных таким образом признаках (парах лемм) делается суждение с помощью упомянутой модели машинного обучения (303). Если полученное значение вероятности выше заранее подобранного порога, то событие присваивается новости (304) для одной или нескольких компаний, указанных в новости. Дополнительно в каждом тексте могут выделяться предложения, содержащие пары характерных для события лемм. Если в ходе обработки новостных данных не осуществляется выявление релевантных событий для указанных наименований компаний, то такая информация не учитывается (305).

На фиг. 4 представлен пример графического интерфейса пользователя (400) для взаимодействия с сервисом по подбору релевантной новостной информации. Интерфейс (400) предоставляет функционал по отображению и управлению содержанием предоставляемых данных. Формирование поискового запроса выполняется с помощью панели ввода информации о наименовании компании (401). В основном поле (404) для отображения текущей или найденной информации представлен перечень компаний, для которых осуществляется обработка выявления релевантной информации из базы данных сервера (110).

Компании в поле (404) могут отображаться в различном иерархическом порядке, например в алфавитном, по количеству новостей и т.п. Информация может отфильтровываться по временному диапазону, который устанавливается в поле ввода дат (402).

Также, интерфейс (400) содержит панель управления для настройки параметров поисковых запросов (403). С помощью панели управления (403) можно осуществлять настройку выявления тех или иных типов событий, осуществлять привязку компаний, конфигурировать параметры сервиса и т.п. В поле (405) отображается список выявленных новостных источников в соответствии с заданными событиями для компаний.

Пользователи (10) также могут устанавливать функцию оповещения для выбранных наименований компаний. Оповещения о поступлении новых новостей могут передаваться посредством сообщений электронной почты, PUSH уведомлений, SMS уведомлений и т.п. При настройке функции оповещения пользователь (10) может настраивать требуемые параметры, например наименование компании, тип событий, связанных с компаниями.

Сформированная информация по обработанным новостям также может отображаться с применением фильтра, настроенным относительно роли пользователя (10), взаимодействующего с интерфейсом (400). С учетом параметров учетной записи пользователя (10) ему могут отображаться только те новости, которые содержат связанный с его ролью тип событий.

На фиг. 5 представлен пример общего вида устройства (500), которое обеспечивает реализацию представленного решения. На базе устройства (500) может реализовываться различный спектр вычислительных устройств, например управляющий сервер (110), сервер новостного агрегатора (120), устройства пользователей (10) и т.д. В общем виде устройство (500) содержит объединенные общей шиной информационного обмена один или несколько процессоров (501), средства памяти, такие как ОЗУ (502) и ПЗУ (503), интерфейсы ввода/вывода (504), устройства ввода/вывода (505) и устройство для сетевого взаимодействия (506).

Процессор (501) (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 устройстве (500) также необходимо учитывать графический процессор, например GPU NVIDIA или Graphcore, тип которых также является пригодным для полного или частичного выполнения способа (200), а также может применяться для обучения и применения моделей машинного обучения в различных информационных системах.

ОЗУ (502) представляет собой оперативную память и предназначено для хранения исполняемых процессором (501) машиночитаемых инструкций для выполнения необходимых операций по логической обработке данных. ОЗУ (502), как правило, содержит исполняемые инструкции операционной системы и соответствующих программных компонент (приложения, программные модули и т.п.). При этом в качестве ОЗУ (502) может выступать доступный объем памяти графической карты или графического процессора.

ПЗУ (503) представляет собой одно или более средств для 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 устройства (500) и организации работы внешних подключаемых устройств применяются различные виды интерфейсов В/В (504). 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 (500) применяются различные средства (505) 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 (506) обеспечивает передачу данных посредством внутренней или внешней вычислительной сети, например Интранет, Интернет, ЛВС и т.п. В качестве одного или более средств (506) 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 устройства (500), например GPS, ГЛОНАСС, BeiDou, Galileo.

Конкретный выбор элементов устройств (500) для реализации различных программно-аппаратных архитектурных решений может варьироваться с сохранением обеспечиваемого требуемого функционала от того или иного типа устройства. Представленные материалы изобретения раскрывают предпочтительные примеры реализации технического решения и не должны трактоваться как ограничивающие иные, частные примеры его воплощения, не выходящие за пределы испрашиваемой правовой охраны, которые являются очевидными для специалистов соответствующей области техники.

## ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Компьютерно-реализуемый способ поиска релевантных новостей, содержащий этапы, на которых

получают на управляющем сервере набор новостей по меньшей мере от одного сервера новостного агрегатора;

осуществляют на управляющем сервере анализ полученного набора новостей, который включает в себя

лемматизацию текстов каждой новости из упомянутого набора новостей;

обработку полученных лемм текстов новостей с помощью модели машинного обучения, которая содержит установленный набор данных компаний и список событий, причем для каждого события в модели машинного обучения установлен заданный набор лемм;

определение новостей, содержащих леммы, идентифицирующие заданные события, и формирование связи выявленных событий по меньшей мере с одной компанией;

формируют список релевантных новостей на основании выполненного анализа набора новостей.

2. Способ по п.1, характеризующийся тем, что при получении набора новостей осуществляется фильтрация дублирующих новостей.

3. Способ по п.2, характеризующийся тем, что фильтрация осуществляется с помощью вычисления меры Жаккара между сигнатурами новостей.

4. Способ по п.1, характеризующийся тем, что события, присвоенные новости о компании, сохраняют в базе данных.

5. Способ по п.1, характеризующийся тем, что новостной агрегатор обновляет список новостей с помощью информационных каналов.

6. Способ по п.5, характеризующийся тем, что информационные каналы представляют собой вебсайты в сети Интернет и/или каналы мессенджеров.

7. Способ по п.1, характеризующийся тем, что в ходе анализа новостей выполняется определение принадлежности события основной или дочерней компании.

8. Способ по п.7, характеризующийся тем, что принадлежность компании, упоминаемой в новости, определяется с помощью алгоритма решающих деревьев.

9. Способ по п.1, характеризующийся тем, что в ходе лемматизации текстов новостей осуществляется их очистка от знаков пунктуации, стоп-слов и именных сущностей.

10. Способ по п.1, характеризующийся тем, что в ходе лемматизации для каждой леммы текста новости рассчитывается статистическая мера.

11. Способ по п.10, характеризующийся тем, что алгоритм машинного обучения представляет собой логическую регрессию, классифицирующий принадлежность новости событию на основании анализа статистической меры лемм.

12. Способ по п.1, характеризующийся тем, что для каждого текста новости выполняется определение частотных словосочетаний длиной от 2 до 10 лемм.

13. Способ по п.1, характеризующийся тем, что алгоритм машинного обучения представляет собой градиентный бустинг, обученный для классификации события на основе количества предложений, содержащих леммы, идентифицирующие событие из поискового запроса.

14. Способ по п.1, характеризующийся тем, что после присвоения события из новости компании выполняется выделение лемм и/или предложений, содержащих леммы, идентифицирующее упомянутое событие.

15. Устройство для поиска релевантных новостей, содержащее по меньшей мере один процессор и по меньшей мере одну память, содержащую машиночитаемые инструкции, которые при их исполнении по меньшей мере одним процессором выполняют способ по любому из пп.1-14.

16. Система поиска релевантных новостей, содержащая

по меньшей мере один управляющий сервер;

по меньшей мере один сервер новостного агрегатора, выполненный с возможностью получения новостных данных от по меньшей мере одного новостного источника, причем

управляющий сервер выполнен с возможностью

получения новостных данных от по меньшей мере одного сервера новостного агрегатора;

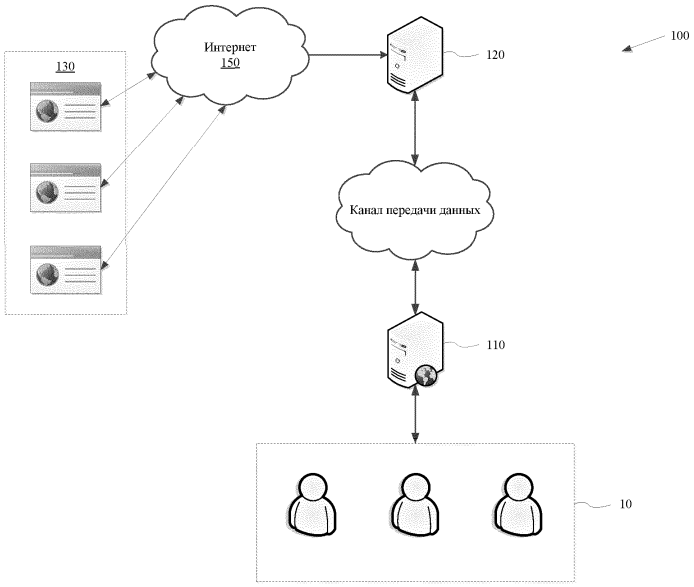
анализа полученных данных, в ходе которого выполняется

лемматизация текстов каждой новости из упомянутого набора новостей;

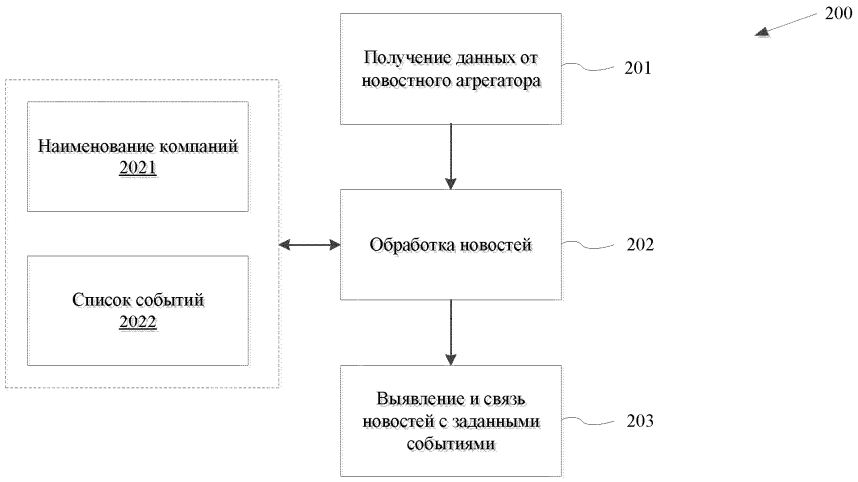
обработка полученных лемм текстов новостей с помощью модели машинного обучения, которая содержит установленный набор данных компаний и список событий, причем для каждого события в модели машинного обучения установлен заданный набор лемм;

определение новостей, содержащих леммы, идентифицирующие заданные события, и формирование связи выявленных событий по меньшей мере с одной компанией;

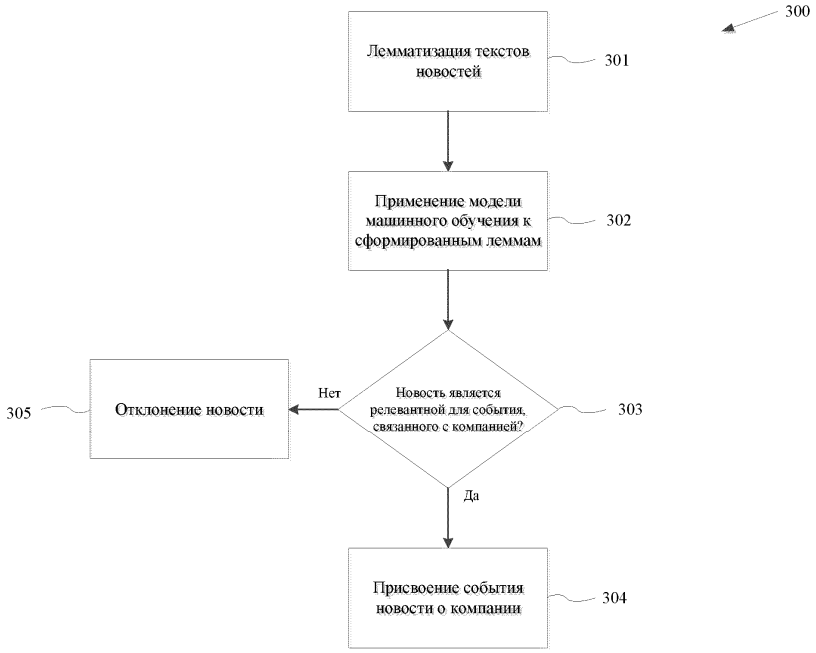
формирование списка релевантных новостей на основании выполненного анализа набора новостей.



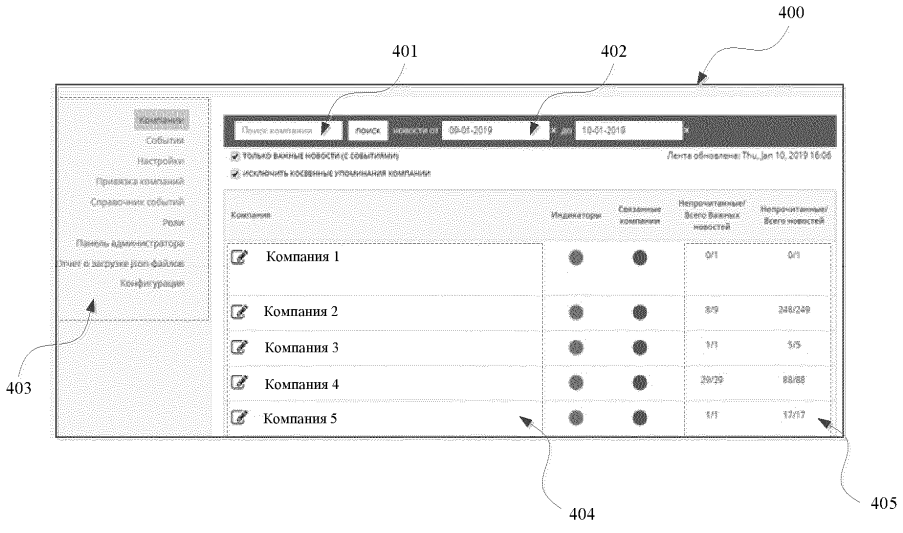

Фиг. 1



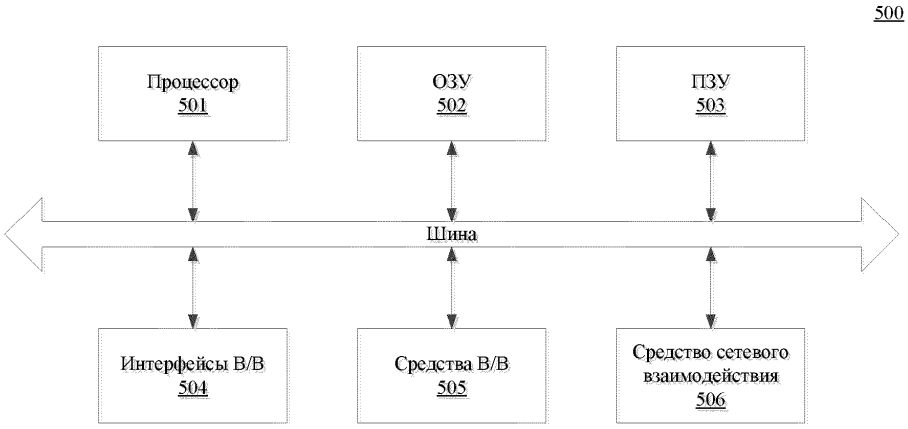

Фиг. 2




Фиг. 3




Фиг. 4


500
Процессор
501
ОЗУ
502
ПЗУ
503
Шина
Интерфейсы В/В
504
Средства В/В
505
Средство сетевого взаимодействия
506

Фиг. 5