РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(19) **RU** (11) **2 674 331** (13) **C2**

(51) МПК
G06F 17/22 (2006.01)
G06F 17/30 (2006.01)

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 17/2211 (2006.01); G06F 17/3053 (2006.01)

(21)(22) Заявка: **2017110788, 03.09.2015**

(24) Дата начала отсчета срока действия патента:
03.09.2015

Дата регистрации:
06.12.2018

Приоритет(ы):
(30) Конвенционный приоритет:
03.09.2014 US 62/045,398

(43) Дата публикации заявки: **03.10.2018** Бюл. № **28**

(45) Опубликовано: **06.12.2018** Бюл. № **34**

(85) Дата начала рассмотрения заявки PCT на национальной фазе: **03.04.2017**

(86) Заявка PCT:
US 2015/048322 (03.09.2015)

(87) Публикация заявки PCT:
WO 2016/036940 (10.03.2016)

Адрес для переписки:
129090, Москва, ул. Б.Спасская, 25, строение 3, ООО "Юридическая фирма Городисский и Партнеры"

(72) Автор(ы):
СКРИФФИНЬЯНО Энтони Дж. (US),
СУНБХАНИКХ Йем (US),
ДЭВИС Робин Фрай (US),
МЭТТЬЮЗ Уорвик (AU)

(73) Патентообладатель(и):
ДЗЕ ДАН ЭНД БРЭДСТРИТ КОРПОРЕЙШН (US)

(56) Список документов, цитированных в отчете о поиске: **US 2010/0332424 A1, 30.12.2010. US 2008/0208820 A1, 28.08.2008. US 2005/0108630 A1, 19.05.2005. US 2005/0187923 A1, 25.08.2005. RU 2480822 C2, 27.04.2013.**

(54) СИСТЕМА И ПРОЦЕСС ДЛЯ АНАЛИЗА, КВАЛИФИЦИРОВАНИЯ И ПРОГЛАТЫВАНИЯ ИСТОЧНИКОВ НЕСТРУКТУРИРОВАННЫХ ДАННЫХ ПОСРЕДСТВОМ ЭМПИРИЧЕСКОЙ АТРИБУЦИИ

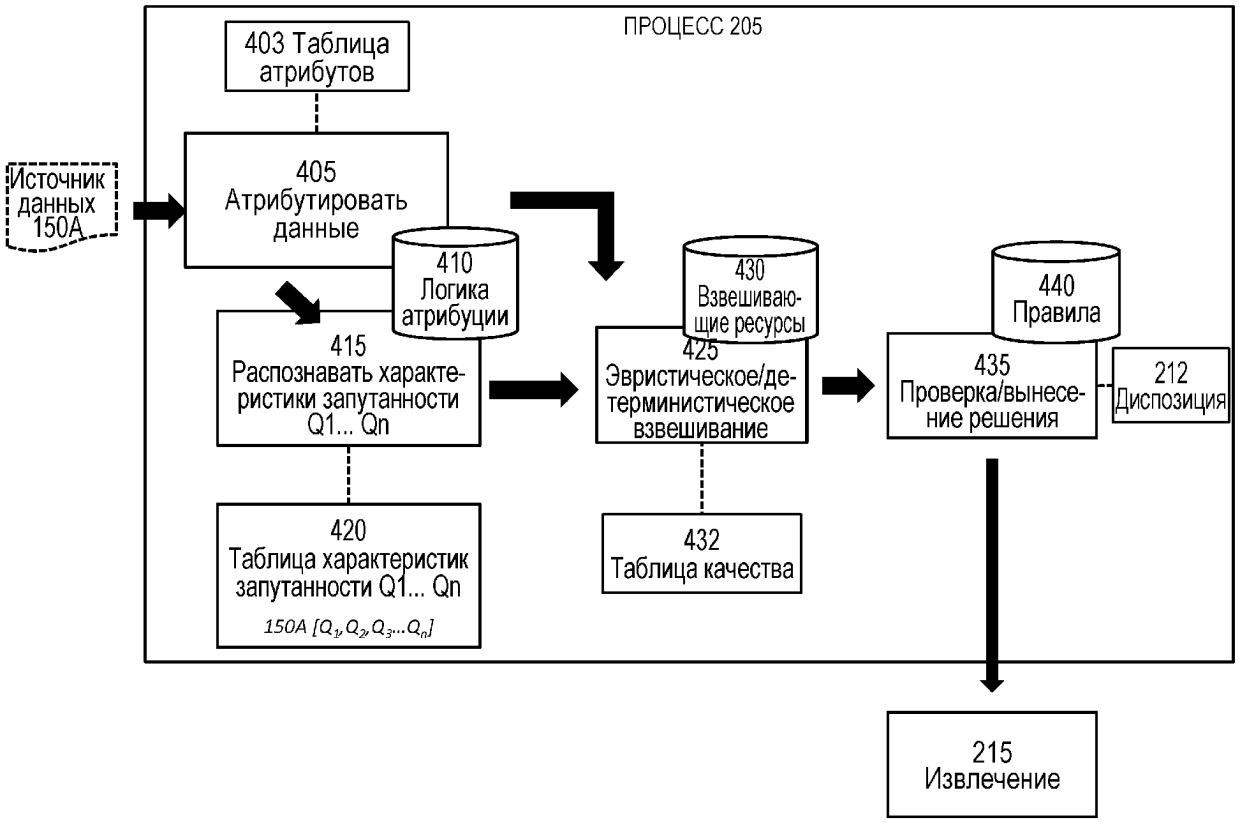
(57) Реферат:

Изобретение относится к процессам атрибуции и дифференциации для генерации описательных и контекстных атрибутов данных из плохо курированных или плохо структурированных, неструктурированных или частично структурированных источников. Техническим результатом является повышение скорости, обеспечение масштабируемости, гибкости и согласованности процессов атрибуции и дифференциации данных. В способе анализа источника данных принимают данные из источника данных. Атрибутируют источник данных в соответствии с правилами и, таким образом, выдают атрибут. Анализируют данные для идентификации характеристики запутанности в данных. Вычисляют качественную меру атрибута и, таким образом, выдают взвешенный атрибут. Вычисляют качественную меру

RU 2674331 C2

RU 2674331 C2

характеристики запутанности и, таким образом, выдают взвешенную характеристику запутанности. Анализируют взвешенный атрибут и взвешенную характеристику запутанности, для создания диспозиции. Фильтруют данные в соответствии с диспозицией и, таким образом, выдают извлеченные данные. 3 н. и 12 з.п. ф-лы, 4 ил., 10 табл.



ФИГ. 4

RU 2674331 C2

RU 2674331 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 674 331** (13) **C2**

(51) Int. Cl.
G06F 17/22 (2006.01)
G06F 17/30 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 17/2211 (2006.01); G06F 17/3053 (2006.01)

(21)(22) Application: **2017110788, 03.09.2015**

(24) Effective date for property rights:
03.09.2015

Registration date:
06.12.2018

Priority:
(30) Convention priority:
03.09.2014 US 62/045,398

(43) Application published: **03.10.2018** Bull. № **28**

(45) Date of publication: **06.12.2018** Bull. № **34**

(85) Commencement of national phase: **03.04.2017**

(86) PCT application:
US 2015/048322 (03.09.2015)

(87) PCT publication:
WO 2016/036940 (10.03.2016)

Mail address:
129090, Moskva, ul. B.Spasskaya, 25, stroenie 3, OOO "Yuridicheskaya firma Gorodisskij i Partnery"

(72) Inventor(s):
SCRIFFIGNANO, Anthony, J. (US),
SUNBHANICH, Yiem (US),
DAVIES, Robin, Fry (US),
MATTHEWS, Warwick (AU)

(73) Proprietor(s):
THE DUN & BRADSTREET CORPORATION (US)

RU 2674331 C2

(54) SYSTEM AND PROCESS FOR ANALYSIS, QUALIFICATION AND ACQUISITION OF SOURCES OF UNSTRUCTURED DATA BY MEANS OF EMPIRICAL ATTRIBUTION

(57) Abstract:

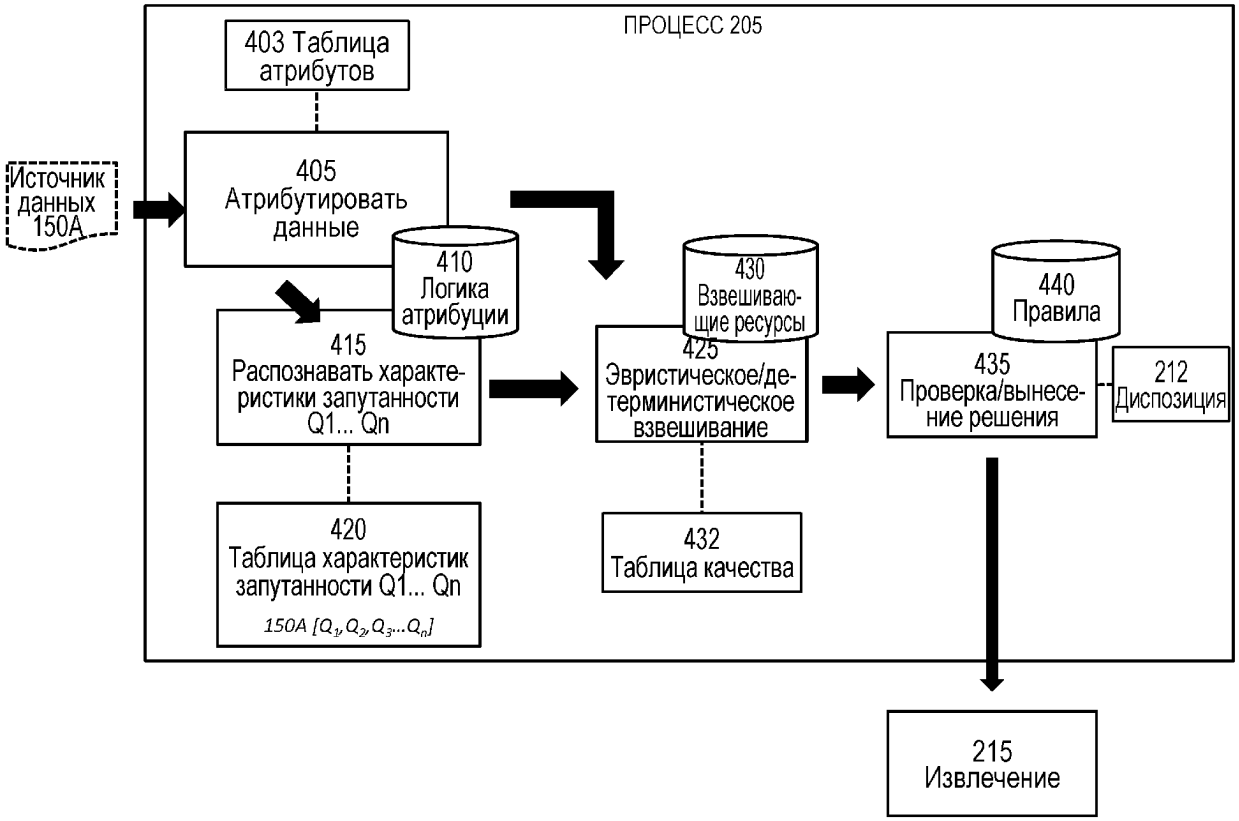
FIELD: computing; counting.

SUBSTANCE: invention relates to the processes of attribution and differentiation for the generation of descriptive and contextual attributes of data from poorly managed or poorly structured, unstructured or partially structured sources. Receive data from a data source in a data source analysis method. Attribute the data source in accordance with the rules and, thus, give the attribute. Analyze the data to identify the characteristics of entanglement in the data. Qualitative measure of the attribute is calculated and, thus, a weighted attribute is determined. Qualitative measure of the entanglement characteristic is calculated and, thus, a weighted characteristic of entanglement is determined. Analyze the weighted attribute and weighted characteristic of entanglement, to create a disposition. Filter data in accordance with the disposition and, thus, determine the extracted data.

EFFECT: technical result is to increase the speed, ensure scalability, flexibility and consistency of the processes of attribution and differentiation of data.

15 cl, 4 dwg, 10 tbl

RU 2674331 C2



ФИГ. 4

RU 2674331 C2

RU 2674331 C2

ПЕРЕКРЕСТНАЯ ССЫЛКА НА РОДСТВЕННЫЕ ЗАЯВКИ

[0001] По настоящей заявке испрашивается приоритет предварительной патентной заявки США №62/045,398, поданной 3 сентября 2014 г., содержание которой включено сюда посредством ссылки.

УРОВЕНЬ ТЕХНИКИ

1. Область техники, к которой относится изобретение

[0002] Настоящее изобретение относится к системе, которая использует новые, эмпирические, т.е. научные и воспроизводимые, процессы атрибуции и дифференциации, также именуемые здесь возможностями, для генерации описательных и контекстных атрибутов данных из плохо курированных или плохо структурированных, неструктурированных или частично структурированных источников и, в частности, источников социальной среды общения. Затем атрибуты используются для характеризации, проверки, дифференциации и окончательного принятия решения по наиболее подходящей диспозиции или обработки данных, с использованием способов, которые выходят за пределы существующих рекурсивно-совершенных процессов и режимов. Присущая проблема, которую решает это изобретение, состоит в том, что в настоящее время невозможно согласованно проверять, рассматривать и проглатывать данные в масштабе в отсутствие достаточной онтологии или канонической формы для структурирования процесса проглатывания и курирования.

[0003] Описанная здесь возможность может использоваться при обработке данных, полученных из файлов, загруженных непосредственно из онлайновых источников или по запросу, инициированному конечным пользователем, системой, приложением, или любым другим способом, который обеспечивает данные для проглатывания, обработки и использования с той или иной целью. В этом случае, "обработанной и используемой с той или иной целью" может быть любая последующая система или функция, которая использует данные и будет пользоваться возможностью, то есть, делать заключение, помогать в наблюдении шаблонов, осуществляться лучше, быстрее, более эффективно или таким образом, что имеет тенденцию к увеличению значения этих данных в контексте этой системы или функции.

[0004] Эта возможность может работать на уровне контекста, уровне файла источника или уровне содержания и может информироваться совокупным опытом предыдущих итераций самого процесса. Атрибуция "уровня контекста" действует на уровне обстоятельств получения и проглатывания источника данных. Атрибуция "уровня файла источника" обычно, но не исключительно, действует на уровне файла данных, выдаваемого источником или получаемого из него. Атрибуция "уровня содержания" действует на базовом уровне данных и обычно, но не исключительно, на основании анализ отдельных элементов данных и/или соотношений между ними.

[0005] Примером атрибуции "уровня контекста" является создание метаданных для описания частоты доставки данных из конкретного источника и "срока хранения" данных в этом источнике, то есть, как долго данные обычно считываются "текущими". Примером атрибуции "уровня файла источника" является проверка метаданных из самого файла, например, даты создания. Примером атрибуции "уровня содержания" является обнаружение системы письма, используемой для представления данных, например, упрощенной китайской.

[0006] Промышленные оценки указывают, что более 80% вновь создаваемых данных являются неструктурированными. Для получения достаточной полезной информации из данных, которые все больше присутствуют в неструктурированных или только слабо понятных форматах, или, напротив, во избежание разрастания данных, которые, в

конце концов, оказываются неточными, обманчивыми или вредоносными, если добавляются к существующему в настоящее время курированному корпусу данных или подаются в конкретный сценарий использования, например, функцию принятия решения в бизнесе, важно иметь возможность предварительно оценивать эти данные относительно значимых, но не обязательно заранее определенных, критериев и/или измеренных в известных аспектов. Преимущество предварительного оценивания состоит в том, что данные, которые не проходят определенные проверки или не демонстрируют достаточно высокий уровень качества, будут отклонены, что снижает опасность негативного эффекта. Дополнительное преимущество состоит в помощи или даже предписании попыток курирования, когда ограничения ресурсов или другие соображения не позволяют проглатывать все имеющиеся источники новых данных. Заметим, что слово "качество" используется здесь в смысле любой меры соответствия с конкретной целью и не обязательно предусматривает конкретное внутреннее значение.

[0007] Появились различные технологии для осуществления функций устранения неоднозначности и дифференциации на неструктурированных данных, включающие в себя:

a) извлечение сущности - выделение из текста отдельных компонентов, представляющих интерес, например, существительных, глаголов и модификаторов.

b) анализ отношения - присвоение атрибутов определенному тону и эмоциональной окраске содержания.

c) устранение семантической неоднозначности - приведение текста к более вычислимым конструкциям (например, токенизация).

d) лингвистическое преобразование - включает в себя транслитерацию, перевод и интерпретацию посредством обработки естественного языка (NLP).

[0008] Вышеупомянутые опасность и необходимость в ослаблении особенно применимы, когда сами данные являются данными социальной среды общения, которые неизменно имеют основной неструктурированный компонент "свободного текста", ограниченный размер, имеют "массовый источник", то есть происходят из неограниченного набора непроверенных участников, и которые, весьма вероятно, содержат одну или более "характеристик запутанности".

[0009] Некоторыми примерами этих характеристик запутанности являются:

a) сарказм: слова или предикаты, стоящие рядом, несут скрытый смысл, противоположный тому, который вытекает из поверхностной интерпретации.

- пример: XYZ Oil Co. is an excellent company to do business with, if you like destroying nature (XYZ Oil Co. является превосходной компанией для сотрудничества, если вам нравится разрушать все созданное).

b) неологизм: слова или выражения, недавно построенные и получившие массовое распространение в некотором общепринятом смысле.

- пример: Hashtags (хэштег)

c) вариации грамматики или неправильно фразированный текст: использование намеренно или ненамеренно искаженных слов, приводящее к неоднозначной или недиспозитивной интерпретации.

- пример: FBI is Hunting Terrorists With Explosives

d) пунктуация: использование пунктуации нестандартным или несогласованным образом или отсутствие пунктуации, приводящее к неоднозначной или противоречивой интерпретации.

- пример: "Eats shoots and leaves"(Казнить нельзя помиловать) в сравнении с "Eats, shoots, and leaves" (Казнить нельзя, помиловать)

e) многоязычные данные: вставка слов и выражений иностранного языка. Включает в себя официальные, неофициальные и неформальные заимствованные слова, заимствованные выражения и кальки.

- пример: He had a certain je ne sals quoi that made it difficult to understand his meaning completely.

f) орфография: выдуманная, неверная или заимствованная орфография, которая приводит к несогласованной, неверной или недиспозитивной интерпретации

- пример: RU There?

g) обфускация/шифрование: намеренное преобразование данных для запутывания умозаключения или интерпретации

h) контекст: повышенная зависимость от внешней непрерывности или внешне поддерживаемого контекста вследствие недостатка контекста, обеспеченного в самих данных.

- пример: "He had an awesome slice!" [Cake? Pizza? Tennis shot?]

i) разнородные информационные материалы: текстовые и другие формы информационных материалов объединяются в одном сообщении или фрагменте данных для создания смысла, который является неоднозначным или непознаваемым без общего понимания.

- пример: изображение, сопровождаемое надписью "This is what we think of XYZ Beverage Co.'s new flavor!"

2. Описание уровня техники

[0010] Подходы, описанные в этом разделе, являются подходами, которым можно следовать, но не обязательно подходами, которые были выдвинуты ранее, или которым ранее следовали. Таким образом, подходы, описанные в этом разделе, могут не быть традиционными для формулы изобретения в этой заявке и не признаваться традиционными за счет включения в этот раздел.

[0011] Существующие системы могут пытаться осуществлять вышеупомянутые функции (извлечение сущности, анализ отношения, устранение семантической неоднозначности, лингвистическое преобразование и т.д.) и, таким образом, измерять и проверять данные, но очень трудно понять, какие проверки и метрики применять, предварительно не поработав с данными из конкретного источника. Таким образом для создания достаточно эффективных и воспроизводимых уровней дифференциации и принятия решения, системы которые пытаются проглатывать неструктурированные данные, социальную среду общения и другие подобные данные, могут делать это рекурсивно, то есть система может переконфигурироваться на основании предыдущего опыта. Такие системы также могут реализовывать сценарии закрытого цикла, также известные как "обратная связь с хостом", с использованием имеющей обратную силу обратной связи по качеству для оказания влияния будущие результаты. Однако эти системы сталкиваются с ограничениями масштабируемости и автоматизации в виду неизменно ручной реализации, и даже когда применяется "машинное обучение", оно осуществляется только на самом элементарном эмпирическом уровне, т.е. на основании частотного и семантического анализа самих детализированных данных. Существуют также ограничения, обусловленные влиянием вышеописанных характеристик запутанности языка.

СУЩНОСТЬ ИЗОБРЕТЕНИЯ

[0012] Обеспечен способ, который включает в себя (a) прием данных из источника данных, (b) атрибутирование источника данных в соответствии с правилами и, таким образом, выдачу атрибута, (c) анализ данных для идентификации характеристики

запутанности в данных, (d) вычисление качественной меры атрибута и, таким образом, выдачу взвешенного атрибута, (e) вычисление качественной меры характеристики запутанности и, таким образом, выдачу взвешенной характеристики запутанности, (f) анализ взвешенного атрибута и взвешенной характеристики запутанности, для создания диспозиции, (g) фильтрацию данных в соответствии с диспозицией и, таким образом, выдачу извлеченных данных, и (h) передачу извлеченных данных последующему процессу. Предусмотрены также система, которая выполняет способ, и запоминающее устройство, которое содержит инструкции, управляющие процессором для осуществления способа.

[0013] Описанные здесь методы включают в себя возможности, недостижимые в уровне техники. В частности, описанные здесь методы обеспечивают методологию, которая использует новые аспекты атрибуции, которые, в свою очередь, обеспечивают новые автоматизированные реализации принятия решения на проглатывание данных, позволяющие строить системы, более быстрые, более масштабируемые, более гибкие и более согласованные, чем допускают подходы на основе уровня техники.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0014] фиг. 1 - блок-схема системы для проглатывания, атрибутирования, создания стратегий диспозиции и экспорта источников данных посредством эмпирической атрибуции.

[0015] фиг. 2 - функциональная блок-схема способа осуществляемого системой, показанной на фиг. 1.

[0016] фиг. 3 - графическое представление уровней атрибуции источника и их иерархического соотношения.

[0017] фиг. 4 - функциональная блок-схема процесса, который является частью способа, показанного на фиг. 2.

[0018] Компонент или признак, общий для более одного чертежа указан одной и той же ссылочной позицией в каждом из чертежей.

ОПИСАНИЕ ИЗОБРЕТЕНИЯ

[0019] Необходимо улучшать существующие процессы, которые пытаются анализировать и квалифицировать источники данных до проглатывания. Для удовлетворения этой потребности, предусмотрена система, которая осуществляет способ, который включает в себя (a) присвоение атрибутов данным, поступающим из источника, на нескольких уровнях, (b) создание правил диспозиции для извлечения квалифицирующего поднабор данных, при наличии, из источника, на основании критериев, которые измеряют присвоенные атрибуты в нескольких аспектах, таким образом, выдавая квалифицированные данные, (c) проглатывание квалифицированных данных, и (d) получение обратной связи и осуществление изменения в системе на основании обратной связи.

[0020] Таким образом, в настоящем документе раскрыты автоматизированные система и способ присвоения атрибутов данным из источника, принятия решений, помимо прочего, на основании атрибутов, проглатывания данных и получения обратной связи на основании опыта системы по проглатыванию (причем этот опыт регистрируется системой и сохраняется как новые атрибуты самого процесса). Способ осуществляется без вмешательства человека, что обеспечивает согласованность и масштабируемость и позволяет человеку сосредотачиваться на ситуациях, когда для надлежащего распоряжения данными требуется проницательность или дополнительное исследование. Термин "масштабируемость" означает, что этот подход не ограничивается конкретной технологией или техническим решением.

[0021] В нескольких следующих абзацах, даны определения ряда используемых здесь терминов.

[0022] Атрибут: при использовании в качестве глагола, этот термин означает вычисление и связывание метаданных (т.е. описательных данных) или других данных (например, эмпирических данных) с существующими в настоящее время данными. Присоединенные таким образом данные являются "атрибутами".

[0023] Корпус: существенная часть предмета, например, файла данных, в отличие от данных об этом предмете, например, даты его создания. Корпус относится, если из контекста не следует обратное, к предмету в целом.

[0024] Курирование: классификация, преобразование, хранение и управление предмета/ом, а именно, данных/ми в настоящем изобретении.

[0025] Проглатывание: прием и сохранение данных. Процесс проглатывания обычно предусматривает преобразование или перестройку в целевой/ую формат или таксономию.

[0026] Эмпирическая атрибуция: атрибуция атрибутов на основе научного метода. В случае настоящего изобретения, алгоритмических и математических процессов.

[0027] Методология:

1. Выбрать несколько источников данных на основании утвержденных критериев, подлежащих установлению, с учетом таких факторов, как:

a. Доступность данных, включающая в себя стоимость и допустимое использование;

b. Богатство содержания, способность наблюдать достаточно примеров для формирования эмпирических заключений;

c. Степень перекрытия с уже существующими источниками, ранее включенными в исследование; и

d. Известное смещение в источнике данных.

2. Построить автоматизированную или ручную/гибридную проверку A/B/C для измерения:

a. Существования;

b. Диспозитивной атрибуции; и

c. Степени наблюдения по экстраполированному миру.

3. Выполнить проверку и оценить результаты, включающие в себя:

a. Простую описательную статистику; и

b. Базовые визуализации.

4. Измерить смещение, например, оптимизм/пессимизм оценщиков.

5. Сформировать заключение, в какой степени каждая из гипотез наблюдается и влияет на общую оценку относительно остального мира, который не демонстрирует гипотетические критерии.

[0028] Оценивание результатов:

a. Оценить влияние каждой из гипотез на выбранные образцы.

b. Предполагая, что можно доказать релевантность, разработать систему оценивания для оценивания разных источников согласно гипотетическим аспектам.

[0029] Возможны дополнительные аспекты запутанности, которые возникают в течение периода наблюдения, например:

a. Включения других языков;

b. Влияние однородности групповой речи;

c. Обобществленная метафора из групповой речи (введенная либо окружением, либо обобществленным опытом);

d. Заимствованные слова одного языка в другом; и

e. Мультимодальность говорящих (например, говорящих на родном языке в отличие

от говорящих на неродном языке, цифровых местных в отличие от цифровых иммигрантов).

[0030] Исследование социальной среды общения составляет часть более широкого исследования неструктурированных данных. Работа в целом составляет часть текущего развития возможностей выявления, курирования и синтеза данных, относящихся к бизнесам и к людям в контексте бизнеса.

[0031] Настоящее изобретения, в основном, сосредоточено на возможностях, которые способствуют пониманию в целом полной опасности и/или полной возможности. Смежные потребности относятся к уважению закона, независимости и этике, и к обнаружению преступления.

[0032] На фиг. 1 показана блок-схема системы 100, для проглатывания, атрибутирования, создания стратегий диспозиции и экспорта источников данных посредством эмпирической атрибуции. Система 100 включает в себя компьютер 105, подключенный к сети 135.

[0033] Сеть 135 является сетью передачи данных. Сеть 135 может быть частной сетью или публичной сетью и может включать в себя некоторые или все из (a) персональной сети, например, обеспечивающей покрытие комнаты, (b) локальной сети, например, обеспечивающей покрытие здания, (c) сетью студенческий городок, например, обеспечивающее покрытие студенческого городка, (d) городской сетью, например, обеспечивающей покрытие города, (e) глобальной сетью, например, обеспечивающей покрытие зоны, простирающейся в городских, региональных или национальных границах, или (f) интернетом. Связь осуществляется по сети 135 посредством электронных сигналов и оптических сигналов.

[0034] Компьютер 105 включает в себя процессор 110 и память 115, подключенную к процессору 110. Хотя компьютер 105 представлен здесь как автономное устройство, это не является ограничением, и, вместо этого, он может быть подключен к другим устройствам (не показаны) в системе распределенной обработки.

[0035] Процессор 110 является электронным устройством, образованным логическими схемами, которое реагирует на инструкции и выполняет их.

[0036] Память 115 представляет собой материальный компьютерно-считываемый носитель данных, где хранится компьютерная программа. В связи с этим, в памяти 115 хранятся данные и инструкции, т.е. программный код, которые считываются и исполняются процессором 110 для управления работой процессора 110. Память 115 может быть реализована в виде оперативной памяти (RAM), жесткого диска, постоянной памяти (ROM) или их комбинации. Одним из компонентов памяти 115 является программный модуль 12 0.

[0037] Программный модуль 120 содержит инструкции, предписывающие процессору 110 выполнять описанные здесь процессы. Хотя в настоящем документе описаны операции, осуществляемые компьютером 105, или способом или процессом или подчиненными ему процессами, операции фактически осуществляются процессором 110.

[0038] Термин "модуль" используется здесь для обозначения функциональной операции, которая может быть реализована либо как автономный компонент, либо как интегрированная конфигурация множества подчиненных компонентов. Таким образом, программный модуль 120 может быть реализован как единственный модуль или как а множество модулей, которые работают в кооперации друг с другом. Кроме того, хотя программный модуль 120 описан здесь как установленный в памяти 115 и, таким образом реализованный программными средствами, его можно реализовать в

любом из оборудования (например, электронной схемы), программно-аппаратного обеспечения, программного обеспечения или их комбинации.

[0039] Хотя указано, что программный модуль 120 уже загружен в память 115, он может быть сконфигурирован на запоминающем устройстве 140 для последующей загрузки в память 115. Запоминающее устройство 140 является материальным компьютерно-считываемым носителем данных, где хранятся программный модуль 120. Примеры запоминающего устройства 140 включают в себя компакт-диск, магнитную ленту, постоянную память, оптические носители данных, жесткий диск или блок памяти, состоящий из нескольких параллельных жестких дисков и флэш-носитель с интерфейсом на основе универсальной последовательной шины (USB). Альтернативно, запоминающее устройство 140 может представлять собой оперативную память или электронное запоминающее устройство другого типа, расположенное в удаленной системе хранения (не показана) и подключенное к компьютеру 105 по сети 135.

[0040] Система 100 также включает в себя источник 150A данных и источник 15 0 B данных, которые совместно именуются здесь источниками 150 данных и коммуникативно подключенные к сети 135. На практике, источники 150 данных могут включать в себя любое количество источников данных, т.е. один или более источников данных. Источники 150 данных содержат неструктурированные данные и могут включать в себя социальную среду общения.

[0041] Система 100 также включает в себя пользовательское устройство 130, которое эксплуатируется пользователем 101 и подключено к компьютеру 105 по сети 135. Пользовательское устройство 130 включает в себя устройство ввода, например, клавиатуру или подсистему распознавания речи, позволяющее пользователю 101 передавать информацию и выборы команд на процессор 110. Пользовательское устройство 130 также включает в себя устройство вывода, например, дисплей или принтер, или синтезатор речи. Орган управления курсором, например, мышь, шаровой манипулятор или сенсорный экран, позволяет пользователю 101 манипулировать курсором на дисплее для передачи дополнительной информации и выборов команд на процессор 110.

[0042] Процессор 110 выводит на пользовательское устройство 130 результат 122 выполнения программного модуля 120. Альтернативно, процессор 110 может направлять выходной сигнал в запоминающее устройство 125, например, базу данных или память или удаленное устройство (не показано) по сети 135.

[0043] Последовательность операций, в которой применима система 100, относится к приему, выявлению и курированию источников неструктурированных данных, например, источников 150 данных. Эти прием, выявление и курирование могут составлять часть использования, обслуживающего любое количество случаев использования, включающих в себя, но без ограничения, формирование мнений о коллективном отношении в социальной среде общения, понимание сдвигов маркетинговой позиции по отношению сделанных заявлений, обнаружение нюанса, приводящего к выявлению кражи личных данных или другого преступления, умозаключение по поводу социальных сигналов, предвещающих предстоящее событие или поведение, или просто оценивание возрастающего значения проглатывания нового неструктурированного источника в уже существующий процесс.

[0044] На фиг. 2 показана функциональная блок-схема способа 200, осуществляемого системой 100, и, в частности, процессором 110 в соответствии с программным модулем 120. Способ 200 является общим процессом приема данных, атрибутирования источников данных и их данных на нескольких уровнях (то есть вышеупомянутых уровня контекста,

уровня источника и уровня содержания) и принятия решений в отношении диспозиции источников данных и данных, передачи данных, например, конкретных их поднаборов, на одну или более последующих систем, инициирования функций для обеспечения обратной связи по диспозиции и функций для инициирования выявления и потребления дополнительных источников данных. Способ 200 осуществляет доступ к данным из одно или более источников 150 и их обработку, но для простоты объяснения выполнение способа 200 далее будет описано на примере единственного источника данных, а именно, источника 150A данных. Способ 200 начинается с процесса 205.

[0045] Процесс 205 осуществляет доступ к источнику 150A данных, анализирует и атрибутирует его на нескольких уровнях, а именно, уровнях "контекста", "файла источника" и "содержания", как упомянуто выше, и принимает решение по наиболее подходящей диспозиции данных, содержащихся в источнике 150A данных для выдачи диспозиции 212.

[0046] На фиг. 3 показано графическое представление уровней атрибуции источника и их иерархического соотношения.

[0047] На любом уровне атрибуции источника и, в частности, на уровне содержания, атрибуция может включать в себя функции устранения неоднозначности и дифференциации, которые работают в вышеописанных аспектах, т.е. извлечения сущности, анализа отношения, устранения семантической неоднозначности и лингвистического преобразования. Кроме того, используя эти функции устранения неоднозначности и дифференциации, процесс 205 будет пытаться решить проблемы, связанные с атрибуцией, обусловленные, помимо прочего, вышеописанными характеристиками запутанности, т.е. сарказмом, неологизмом и т.д.

[0048] На фиг. 4 показана функциональная блок-схема процесса 205. Процесс 205 начинается с процесса 405.

[0049] Процесс 405 принимает данные из источника 150A данных, и атрибутирует источник 150A данных с использованием правил и ссылочной информации, хранящейся в логике 410 атрибуции, таким образом создавая таблицу 403 атрибутов. Правила и ссылочная информация являются, например, набор алгоритмов, которые сканируют данные для определения, являются ли данные текстом или разнородными информационными материалами. Например, процесс 405 анализирует источник 150A данных и определяет, что он является сторонним, например, приобретенным, источником данных, и что он создан 1 января 2015 г.

[0050] Таблица 1 является иллюстративным представлением таблицы 403 атрибутов и включает в себя несколько иллюстративных атрибутов и их значения.

Таблица 1

(пример таблицы 403 атрибутов)

Атрибут	Значение
Тип файла	Текст
Разделенный	Да
Источник (автор из свойств файла)	файлы данных ACME
Формат	DFC001
Дата создания:	1 января 2015 г.
ID веб-выявления:	- отсутствует -
Кодировка	UTF-8
Обнаружены скрипты	органы управления C0 и базовый латинский

[0051] "Тип файла" является атрибутом уровня источника и определяется в результате процесса, который сканирует метаданные и содержание файла данных для характеризации типа данных файла. Другие значения могут быть "изображение", "видео", "двоичный", "неизвестный" и т.д.

[0052] "Разделенный" это флаг Да/Нет, который представляет заключение, сделанное на основании сканирования файла для определения, содержится ли данные в разных строках.

[0053] "Источник", в примере, представляет поставщика файла; в этом случае считывается из метаданных "автор" (или "свойства") файла данных.

[0054] "Дата создания" также может считываться из метаданных файла.

[0055] "ID веб-выявления" представлен в качестве примера атрибута, который не найден, будучи явной меткой, вставленной в файл процессом выявления, инициированным функцией 210 (описанной ниже).

[0056] "Кодировка" также считывается из метаданных файла и указывает, как был построен файл. Другие значения могут включать в себя "ASCII", "BIG5", "SHIFT-JIS, "EBCDIC" и т.д.

[0057] "Обнаружены скрипты" обеспечен в примере для указания атрибута, который получен не из метаданных, а из сканирования корпуса самих данных, чтобы понять, какие диапазоны Unicode присутствуют в файле. Значение "органы управления CO и базовый латинский" фактически является стандартный латинским набором данных.

[0058] Типы и значения атрибутов, показанные в таблице 1, являются лишь примерами и не обязательно представляют типы или значения атрибутов, которые система 100 будет присоединять к конкретному/ым файлу или данным. Система 100 может быть выполнена с возможностью создания любых метаданных, которые считаются полезными.

[0059] Процесс 415 анализирует корпус источника 150A данных для генерации атрибутов в многочисленных аспектах, в том числе (но без ограничения):

a) извлечение сущности

b) устранение семантической неоднозначности

c) анализ отношения

d) извлечение языка

e) базовые метаданные

[0060] Процесс 415 также атрибутирует и измеряет наличие и распространенность "характеристик запутанности" в источнике 150A данных, и, таким образом, создает таблицу 420 характеристик запутанности, где перечислены характеристики Q1, Q2, Q3 … Qn запутанности. Выше упомянуто несколько примеров характеристик запутанности.

[0061] Таблица 2 является примером таблицы 420 характеристик запутанности, и включает в себя несколько примеров метрик и их значения.

Таблица 2

(пример таблицы 420 характеристик запутанности)

Метрика	Значение
распространенность неологизмов	AX2
Отклонение грамматики	0,56
Показатель пунктуации	0
Отношение	-0,5
Особенности орфографии	низкое
Показатель обфускации	0
Однородность информационных материалов	1,0
Отклонение фрагмента	0,01

[0062] В примере в таблице 2, масштаб и диапазон значений независимы. Некоторые могут быть числовыми, другие могут быть кодами, которые требуют неарифметических средств для создания вменяемых показателей.

[0063] Заметим, что перечисленные и проиллюстрированные здесь меры характеристик запутанности полностью независимы и не образуют замкнутый класс, в том смысле, что система способна добавлять новые характеристики запутанности по мере их идентификации. Например, в вышеприведенной таблице 2 не существует записи для "многоязычных данных", поскольку меры для этой характеристики запутанности и ее влияние еще предстоит идентифицировать в иллюстративной реализации системы.

[0064] "распространенность неологизмов" представляет показатель, вычисленный из сканирования экземпляра источника 150A данных и генерирования показателя, который измеряет, сколько неологизмов, т.е. новых и/или необычных слов, присутствуют в корпусе источника 150A данных. В этом примере, "AX2" может представлять наличие большого количества совершенно понятных неологизмов, "ZA9" может представлять малое количество неологизмов, но с распространенностью в этом наборе очень необычных или непонятных неологизмов.

[0065] "Отклонение грамматики" это мера однородности грамматического стиля. Алгоритмы, используемые для установления метрики, могут быть стандартными в данной области подходами, например, алгоритм Кока - Янгера - Касами или специальные алгоритмы и меры, или алгоритм, который объединяет несколько мер. Сами эти подмеры могут сохраняться как метрики в таблице 420 характеристик запутанности и затем объединяться для создания других записей в таблице 420 характеристик запутанности.

[0066] "Показатель пунктуации" это мера наличия пунктуации. В этом примере обнаружено мало или пренебрежимо мало пунктуации, поэтому значение этой метрики равно нулю.

[0067] "Отношение" показывает, передает ли "говорящий" в тексте позитивное отношение к предмету (то есть одобрение, рекомендацию, похвалу и т.д.), негативное отношение (то есть критику или неодобрение), или нейтральное отношение (ни позитивное, ни негативное, или, возможно, неопределенное). Отрицательное число указывает негативное отношение (критику), нуль указывает нейтральное отношение, и положительное число указывает позитивное отношение (одобрение). Приведенное здесь иллюстративное значение отношения -0,5 указывает то, что можно описать как "умеренно негативное отношение".

[0068] "Особенности орфографии" это мера распространенности орфографических ошибок, которые не являются распознанными неологизмами. Значение "низкое" здесь указывает низкую частотность орфографических ошибок. Заметим, что "орфографические ошибки" используется здесь просто для указания отклонения от известного лексикона; "высокий" показатель может указывать, например, высокую распространенность правильных существительных, которые не распознаются, а не истинных типографских или орфографических ошибок.

[0069] "Показатель обфускации" это мера, определяющая заметность попыток намеренно скрыть смысл, простым примером чего является шифрование текста. Приведенное значение нуль указывает, что обфускации не обнаружено.

[0070] "Однородность информационных материалов" указывает, выглядят ли данные как единственный тип данных (например, текст) или как смешанные информационные материалы (например, текст с внедренными изображениями или гиперссылками). В этом примере, показатель равен 1,0, и это указывает, что в файле присутствует только один тип информационных материалов. Эта информация может объединяться процессом 435 (описанным ниже) с атрибутами, выведенными процессом 405 и показанными в таблице 1, для заключения, что иллюстративный файл данных полностью состоит из текста, выстроенного колонками.

[0071] "Отклонение фрагмента" это показатель от 0 до 1, который описывает общую согласованность по размеру отдельных блоков файла. В таблице 2, показатель 0,01 указывает, что фрагменты очень однородны. Примером является хорошо структурированный файл данных, поэтому такое значение ожидаемо, поскольку фрагменты представляют строки в файле. Файл, образованный сообщениями из онлайновой социально-сетевой службы, которая позволяет пользователям отправлять и читать короткие, например 140-символьные, сообщения, может иметь средний показатель, поскольку фрагменты изменяются, но колеблются вокруг 128 символов. Для данных из социально-сетевой службы, которая допускает более длинные посты, фрагменты, предположительно, будут иметь очень высокий показатель, поскольку в этой разновидности данных возможна сильная изменчивость.

[0072] Метрики и значения, показанные в таблице 2, являются лишь примерами и не обязательно представляют значения, которые система 100 будет присоединять к конкретному/ым файлу или данным.

[0073] Как указано выше, процесс 415 может иметь отношение к многочисленным мерам для каждой метрики. Например, несколько алгоритмов можно применять для измерения значения метрики "отклонение грамматики". Например, одна или более мер фактически могут быть другими метриками в таблице 420 характеристик запутанности, другие могут быть значениями в таблице 403 атрибутов или выводиться из них.

[0074] В нижеследующей таблице 3 показаны три примера алгоритмических мер отношения. Эти меры могут объединяться в общий показатель отношения, приведенный выше в таблице 2.

Таблица 3

(иллюстративный перечень алгоритмических мер характеристики запутанности отношения)

Метрика
Среднее арифметическое отношения
Взвешенное среднее отношения
Среднеквадратическое отклонение отношения

[0075] По завершении процессов 405 и 415, процесс 205 переходит к процессу 425.

[0076] Процесс 425 является процессом эвристического/детерминистического взвешивания, который принимает таблицу 403 атрибутов и таблицу 420 характеристик запутанности и вычисляет качественные меры атрибутов перечисленных в таблице 403 атрибутов и таблице 420 характеристик запутанности, таким образом создавая таблицу 432 качества. Качественные меры в таблице 432 качества генерируются согласно взвешивающие ресурсы 430 и могут быть показателями, коэффициентами или весовыми функциями, которые измеряют источник 150A данных в многочисленных аспектах.

[0077] Таблица 4 является иллюстративным представлением таблицы 432 качества. В таблице 4, "весовой коэффициент" является качественной мерой и получается из взвешивающих ресурсов 430. Процесс 425 назначает весовой коэффициент метрике.

Таблица 4

(пример таблицы 432 качества)

Метрика	Значение	Весовой коэффициент
Распространенность	AX2	10

неологизмов		
Отклонение грамматики	0,56	50
Показатель пунктуации	0	1
Отношение	−0,5	77
Особенности орфографии	низкое	30
Показатель обфускации	0	70
Однородность информационных материалов	1,0	60
Отклонение фрагмента	0,01	44
Языки	1	80
Источник	S1	55
Возраст	76	44

[0078] Таблица 4 является простым примером. Фактические качественные измерения могут иметь отношение к весьма сложным комбинациям факторов.

[0079] Таблица 4A демонстрирует пример использования объединенных факторов.

Таблица 4A

Метрика	Значение	Весовой коэффициент
Источник	S1	10
Источник > возраст	S1:25	76

[0080] В примере, приведенном в таблице 4A, поиск метрики источника производился в другой таблице (не показана), где перечислены известные источники данных и весовые коэффициенты, назначенные, соответственно этим источникам данных. Весовой коэффициент этого источника, распознанного как источник "S1", и назначенный процессом 425, в этом случае равен 10. Однако процесс 425 способен вычислять весовые коэффициенты более сложной природы. Весовой коэффициент "источник>возраст" (призванный показывать, что он принадлежит семейству весовых коэффициентов "источник") показывает, что существует другой весовой коэффициент, который действует для источника S1 и что конкретный коэффициент (а именно, 25) применяется на основании возраста данных в источнике S1 (то есть, как давно создан файл или, альтернативно, явно указанной даты, если присутствует) для выдачи весового коэффициента 76.

[0081] По завершении процесса 425, процесс 205 переходит к процессу 435.

[0082] Процесс 435 является процессом проверки/вынесения решения, который принимает таблицу 432 качества, таблицу 420 характеристик запутанности и таблицу 403 атрибутов и использует правила 440 для определения подходящей диспозиции источника 150A данных и, таким образом, создания диспозиции 212. Правила 440 могут принимать форму матриц, поисковых таблиц, оценочных таблиц, недетерминистических конечных автоматов, деревьев решений или любой комбинации этой или другой логики принятия решения.

[0083] Диспозиция 212 может включать в себя инструкций или рекомендаций для:

a) установления правила, согласно которому файлы, аналогичные источнику 150A данных, проглатываются целиком.

b) разбиения файлов из источника 150A данных и проглатывания только частей, которые отвечают определенным критериям.

c) проглатывания всего файла из источника 150A данных, но помечания данных указателем уровня качества, зависящим от источника.

d) установления правила, согласно которому файлы из источника 150A данных всегда отклоняются.

e) осторожного проглатывания файлов из источника 150A данных, но с удержанием их в ожидании дополнительного подтверждения, и инициирования целенаправленного веб-выявления посредством функции 210.

[0084] Заметим также, что пример таблицы 432, показанный в таблице 4, является двухмерной ссылочной таблицей со значениями и весовыми коэффициентами, но это является только иллюстрацией. Процесс 435 может, посредством правил 440, использовать другие процессы, например, поиски по таблицам и недетерминистические конечные автоматы для достижения диспозиции 212.

[0085] Возвращаясь к фиг. 2, способ 200, по завершении процесса 205, способ 200 переходит к процессу 215.

[0086] Процесс 215 принимает данные в форме источника 150A данных и диспозиции 212 и выполняет процессы для подразделения и фильтрации принятых данных, для выдачи извлеченных данных 217. В связи с этим, процесс 215 использует данные, сгенерированные процессом 205, т.е. диспозицию 212, для:

a) квалификации источника 150A данных;

b) разделения содержания источника 150A данных на значимые поднаборы; и

c) проглатывания данных из источника 150A данных в последующий процесс 220 (не показан), являющийся потребителем/потребителями данных.

[0087] Процесс 220 принимает извлеченные данные 217 и передает извлеченные данные 217 последующему процессу (не показан).

[0088] Способ 200 также выполняет функцию 225 для генерации эмпирической, например, статистической, и качественной, например, согласованной с пользователем, обратной связи и возвращения обратной связи процессу 205, для улучшения процесса 205. Функция 225 информируется (т.е. принимает входные сигналы) от диспозиции 212, таблицы 432 качества, таблицы 420 характеристик запутанности и таблицы 403 атрибутов. Функция 225 инициируется обработкой диспозиции 212 процессом 215.

[0089] Способ 200 также выполняет функцию 210 как асинхронный и потенциально непрерывный процесс. Функция 210 исследует новые и существующие источники 150 данных, например, посредством автоматизированного веб-выявление, с использованием данных, сгенерированных в процессе 205, а именно, диспозиции 212, таблицы 432 качества, таблицы 420 характеристик запутанности и таблицы 403 атрибутов. Эти данные передаются функции 210 для инициирования, руководства или ограничения

автоматизированных процессов выявления источников данных. Этот интеллект может, например, принимать форму "идентификации промежутков" (которая идентифицирует области, где в ранее проглоченном корпусе наблюдался, что не только недостаток данных, но и их низкое качество или потеря ценности вследствие "устаревания"), или "генерации аналогов" (которая таргетирует классы источников данных на основании идентификации сходных или аналогичных классов источников данных и определения эффективности, согласованности или точности классов).

[0090] Функция 210 конфигурирует и выполняет внешние процедуры, приложения и функции выявления данных. Функция 210 выдает входные сигналы на эти процессы выявления данных, чтобы они служили для дополнения данных, ранее принятых способом 200. Примером таких входных сигналов является унифицированный указатель ресурса (URL) веб-сайта, из которого можно получить нужные данные, и перечень поисковых терминов на основании содержания источника 150A данных.

[0091] Система 100 допускает автоматизированное, конфигурируемое, повторяемое и адаптивное использование новых источников данных, в особенности, неструктурированных данных. Поскольку система 100 работает полностью автоматически, она обладает масштабируемостью, что позволяет значительно повысить эффективность, скорость и согласованность управления потреблением данных.

[0092] Для иллюстрации примера выполнения способа 200, начнем с файла EX1 источника, показанного ниже в таблице 5.

Таблица 5

(файл EX1 источника)

ID	Адрес электронной почты	Дата и время	Сообщение
funkyDave	dave.smith@gmail.com	2015-01-02 22:23:45:660	Gah! Dnt u just luv it when they leave half the toppings off ur pizza?
2PacGlue	Fnky2rtm@yahoo7.com.au	2015-02-02 22:25:05:424	Gonna try the new Coke flavor. NOT.

[0093] Таблица 6 демонстрирует таблицу 403 атрибутов для файла EX1 источника.

Таблица 6

(пример таблицы 403 атрибутов для файла EX1 источника)

Атрибут	Значение
Тип файла	Текст
Разделенный	Да
Источник	GNIP
Формат	GNIP01
Дата создания:	1 июля 2015 г.
Кодировка	UTF-8

[0094] Таблица 7 демонстрирует таблицу 420 характеристик запутанности для файла EX1 источника.

Таблица 7

(пример таблицы 420 характеристик запутанности для файла EX1 источника)

Метрика	Значение
Распространенность неологизмов	AG7
Отклонение грамматики	0,88
Показатель пунктуации	55
Диапазон сарказма/искренности	-3
отношение	-0,95
Особенности орфографии	высокое
Показатель обфускации	0
Однородность информационных материалов	1,0

[0095] При заполнении таблицы 420 характеристик запутанности для фрагмента данных "Gonna try the new Coke flavor. NOT.", процесс 415 будет осуществлять анализ, включающий в себя семантический анализ содержания в строках, представленных в таблице 8.

Таблица 8

(пример анализа, осуществляемого для заполнения таблицы 420 характеристик запутанности)

Слово	Анализ
Gonna	Скорее всего, это высказывание о намерении совершить действие в будущем. ″Gonna″ – неологизм, буквально означающий ″Going to″, но используемый для указания ″I think I will″.
try	
the	Это предмет предложения. Продукт, идентифицированный: Маловероятно, что речь идет о ″new Coke″, весьма вероятно, что это какой-то новый неизведанный вкус. Заключение: Coke имеет новый продукт.
new	
Coke	
flavor	
NOT	Классический неологизм отрицания – что подтверждает написание заглавными буквами. Весьма вероятно, для указания сарказма, а также буквального противопоставления предыдущему высказыванию.

[0096] Анализ, представленный в таблице 8, является "в прямом смысле" деконструкцией алгоритмического и статистического анализа, осуществляемого процессом 415. Этот анализ будет использоваться для заполнения распространенности неологизмов, поскольку слова "Gonna" и "NOT" являются неологизмами в том смысле, в котором они используются, но фактически не являются новыми словами как таковые. Это также показывает, почему показатель распространенности неологизмов не является просто числом. Неологизмы представляют собой как новые слова, так и старые слова в новом использовании. Показатель пунктуации также определяется использованием пунктуации в примере, т.е. законные предложения и заглавные буквы используются согласованно. Диапазон сарказма/искренности имеет здесь большое значение и существенно определяется использованием "NOT" как для отрицания предыдущего высказывания, так и для указания сарказма. Эти данные имеют, в целом, очень низкую искренность, хотя полная конструкция является "искренней", поскольку она призвана отчетливо передавать негативное намерение.

[0097] Заметим, что анализ, представленный в таблице 8, является "стенографией" созданной в целях этого примера. Процесс 415 будет использовать несколько усложненных функций для выделения выражений, осуществления семантического анализа и компенсации характеристик запутанности. Заметим также, что процесс 415 осуществляет анализ и регистрирует результаты по всему файлу или источнику данных.

[0098] Таблица 9 демонстрирует результат таблицы 432 качества с полученными "процентными показателями" для файла EX1 источника, показанными в самом правом столбце, для обеспечения упрощенного представления выполнения процесса 435 и

правил 440. На практике, будут конфигурироваться вычислительные процессы и алгоритмы, в общем случае, гораздо более сложные, чем пример в таблице 9.

Таблица 9

(пример таблицы 432 качества для файла EX1 источника)

Метрика	Значение	Весовой коэффициент	Показатель
Распространенность неологизмов	AG7	10	34
Отклонение грамматики	0,88	50	44
Показатель пунктуации	55	1	55
Диапазон сарказма/искренности	−3	77	23
Отношение	−0,95	30	33
Особенности орфографии	высокое	70	80
Показатель обфускации	0	60	0
Однородность информационных материалов	1,0	44	44
Языки	1	80	80
Источник	S1	55	23
Возраст	76	44	50

[0099] В таблице 10 показана "в прямом смысле" интерпретация диспозиции 212.

Таблица 10

(пример диспозиции 212 для файла EX1 источника)

1	Использовать как исходную запись начального значения	ложь
2	Использовать как запись подтверждения	истина
3	Сопоставлять с базой данных бизнесов	ложь
4	Сопоставлять с база данных контактов	истина
5	индекс точности по первому впечатлению	0,1
6	Использовать как начальное значение для автоматизированного выявления	ложь
7	Использовать как начальное значение для настройки правил [способа 100]	истина

[00100] Заметим, что в таблице 10, запись 6 указывает, что функция 210 не будет инициироваться этими данными (или данными из этого источника в будущем), и запись 7 указывает, что функция 225 будет инициироваться данными, сгенерированными в способе 100, поскольку он обработал файл EX1 источника.

[00101] Описанные здесь методы являются иллюстративными и не подлежат рассмотрению как налагающие какие-либо конкретные ограничения на настоящее изобретение. Следует понимать, что специалисты в данной области техники могут предложить различные альтернативы, комбинации и модификации. Например, этапы, связанные с описанными здесь процессами, могут осуществляться в любом порядке, если иное не указано или не определено самими этапами. Настоящее изобретение призвано охватывать все подобные альтернативы, модификации и вариации, отвечающие объему нижеследующей формулы изобретения.

[00102] Термины "содержит" или "содержащий" следует интерпретировать как указывающие наличие упомянутых признаков, целых чисел, этапов или компонентов, но не исключающие возможности наличия одного или более других признаков, целых чисел, этапов или компонентов или их групп. Употребление их наименований в едином числе не исключает возможности вариантов осуществления, предусматривающих их наличие в том или ином количестве.

(57) Формула изобретения

1. Способ анализа источника данных, содержащий этапы, на которых:

принимают данные из источника данных;

атрибутируют упомянутый источник данных в соответствии с правилами, таким образом, выдавая атрибут;

анализируют упомянутые данные для идентификации характеристики упомянутых данных, которые запутывают смысл упомянутых данных, таким образом, выдавая характеристику запутанности;

вычисляют качественную меру упомянутого атрибута, таким образом, выдавая взвешенный атрибут;

вычисляют качественную меру упомянутой характеристики запутанности, таким образом, выдавая взвешенную характеристику запутанности;

анализируют упомянутый взвешенный атрибут и упомянутую взвешенную характеристику запутанности, для создания диспозиции; и

фильтруют данные в соответствии с диспозицией, таким образом, выдавая извлеченные данные;

при этом упомянутую диспозицию выбирают из группы, состоящей из (a) устанавливать правило, согласно которому файлы, аналогичные упомянутому источнику данных, принимаются и сохраняются целиком, (b) разбивать файлы из упомянутого источника данных и принимать и сохранять только части, которые отвечают определенным критериям, (c) принимать и сохранять весь файл из упомянутого источника данных, но помечать данные указателем уровня качества, зависящим от источника, (d) устанавливать правило, согласно которому файлы из упомянутого источника данных всегда отклоняются, и (e) с осторожностью принимать и сохранять файлы из упомянутого источника данных, но удерживать их в ожидании дополнительного подтверждения.

2. Способ по п. 1, дополнительно содержащий этапы, на которых:

генерируют обратную связь на основании упомянутой диспозиции.

3. Способ по п. 1, дополнительно содержащий этапы, на которых:

конфигурируют и выполняют автоматизированный процесс выявления данных, на основании упомянутой диспозиции, для выявления нового источника данных; и

исследуют упомянутый новый источник данных.

4. Способ по п. 1, в котором упомянутый анализ осуществляют в аспекте, выбранном из группы, состоящей из извлечения сущности, устранения семантической неоднозначности, анализа отношения, извлечения языка, лингвистического преобразования и базовых метаданных.

5. Способ по п. 1, в котором упомянутую характеристику запутанности выбирают из группы, состоящей из сарказма, неологизма, вариации грамматики, неправильно фразированного текста, пунктуации, многоязычных данных, орфографии, обфускации, шифрования, контекста и использования комбинации информационных материалов.

6. Система для анализа источника данных, содержащая:

процессор; и

память, которая содержит инструкции, которые при считывании упомянутым процессором предписывают упомянутому процессору:

принимать данные из источника данных;

атрибутировать упомянутый источник данных в соответствии с правилами, таким образом, выдавая атрибут;

анализировать упомянутые данные для идентификации характеристики упомянутых данных, которые запутывают смысл упомянутых данных, таким образом, выдавая характеристику запутанности;

вычислять качественную меру упомянутого атрибута, таким образом, выдавая взвешенный атрибут;

вычислять качественную меру упомянутой характеристики запутанности, таким образом, выдавая взвешенную характеристику запутанности;

анализировать упомянутый взвешенный атрибут и упомянутую взвешенную характеристику запутанности, для создания диспозиции; и

фильтровать упомянутые данные в соответствии с упомянутой диспозицией, таким образом, выдавая извлеченные данные;

при этом упомянутую диспозицию выбирать из группы, состоящей из

(a) устанавливать правило, согласно которому файлы, аналогичные упомянутому источнику данных, принимаются и сохраняются целиком, (b) разбивать файлы из упомянутого источника данных и принимать и сохранять только части, которые отвечают определенным критериям, (c) принимать и сохранять весь файл из упомянутого источника данных, но помечать данные указателем уровня качества, зависящим от источника, (d) устанавливать правило, согласно которому файлы из упомянутого источника данных всегда отклоняются, и (e) с осторожностью принимать и сохранять файлы из упомянутого источника данных, но удерживать их в ожидании дополнительного подтверждения.

7. Система по п. 6, в которой упомянутые инструкции также предписывают упомянутому процессору:

генерировать обратную связь на основании упомянутой диспозиции.

8. Система по п. 6, в которой упомянутые инструкции также предписывают упомянутому процессору:

конфигурировать и выполнять автоматизированный процесс выявления данных, на основании упомянутой диспозиции, для выявления нового источника данных; и

исследовать упомянутый новый источник данных.

9. Система по п. 6, в которой упомянутые инструкции, которые предписывают упомянутому процессору анализировать упомянутые данные, предписывают упомянутому процессору анализировать упомянутые данные в аспекте, выбранном из группы, состоящей из извлечения сущности, устранения семантической неоднозначности, анализа отношения, извлечения языка, лингвистического преобразования и базовых метаданных.

10. Система по п. 6, в которой упомянутая характеристика запутанности выбрана из группы, состоящей из сарказма, неологизма, вариации грамматики, неправильно фразированного текста, пунктуации, многоязычных данных, орфографии, обфускации, шифрования, контекста и использования комбинации информационных материалов.

11. Запоминающее устройство, содержащее инструкции, которые считываются процессором и предписывают процессору осуществить способ анализа источников данных, при этом указанные инструкции предписывают процессору:

принимать данные из источника данных;

атрибутировать упомянутый источник данных в соответствии с правилами, таким образом, выдавая атрибут;

анализировать упомянутые данные для идентификации характеристики упомянутых данных, которые запутывают смысл упомянутых данных, таким образом, выдавая характеристику запутанности;

вычислять качественную меру упомянутого атрибута, таким образом, выдавая взвешенный атрибут;

вычислять качественную меру упомянутой характеристики запутанности, таким образом, выдавая взвешенную характеристику запутанности;

анализировать упомянутый взвешенный атрибут и упомянутую взвешенную характеристику запутанности, для создания диспозиции; и

фильтровать упомянутые данные в соответствии с упомянутой диспозицией, таким образом, выдавая извлеченные данные;

при этом упомянутую диспозицию выбирают из группы, состоящей из

(a) устанавливать правило, согласно которому файлы, аналогичные упомянутому источнику данных, принимаются и сохраняются целиком, (b) разбивать файлы из упомянутого источника данных и принимать и сохранять только части, которые отвечают определенным критериям, (c) принимать и сохранять весь файл из упомянутого источника данных, но помечать данные указателем уровня качества, зависящим от источника, (d) устанавливать правило, согласно которому файлы из упомянутого источника данных всегда отклоняются, и (e) с осторожностью принимать и сохранять файлы из упомянутого источника данных, но удерживать их в ожидании дополнительного подтверждения.

12. Запоминающее устройство по п. 11, в котором упомянутые инструкции также предписывают упомянутому процессору:

генерировать обратную связь на основании упомянутой диспозиции.

13. Запоминающее устройство по п. 11, в котором упомянутые инструкции также предписывают упомянутому процессору:

конфигурировать и выполнять автоматизированный процесс выявления данных, на основании упомянутой диспозиции, для выявления нового источника данных; и

исследовать упомянутый новый источник данных.

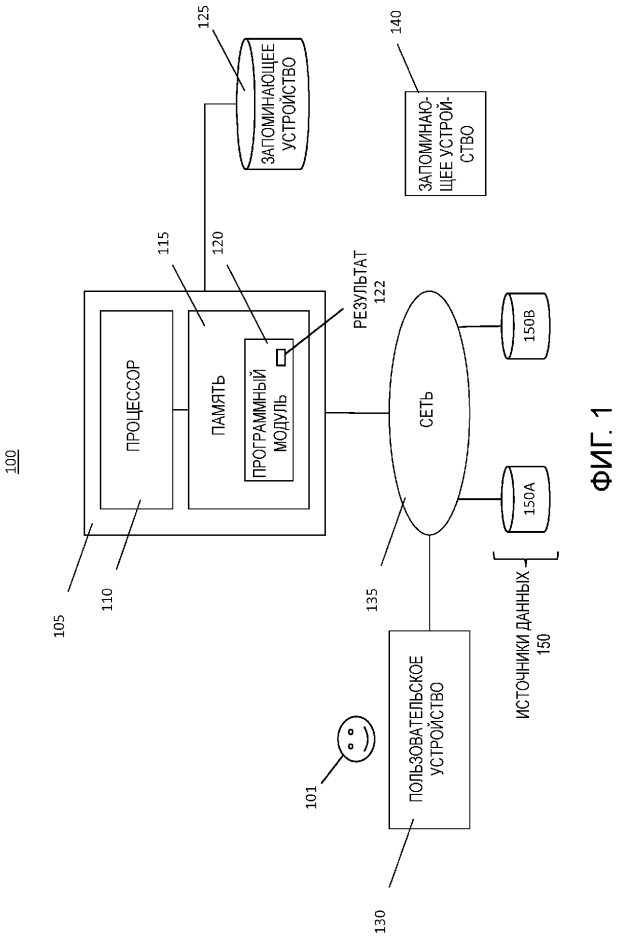
14. Запоминающее устройство по п. 11, в котором упомянутые инструкции, которые предписывают упомянутому процессору анализировать упомянутые данные в аспекте, выбранном из группы, состоящей из извлечения сущности, устранения семантической неоднозначности, анализа отношения, извлечения языка, лингвистического преобразования и базовых метаданных.

15. Запоминающее устройство по п. 11, в котором упомянутая характеристика запутанности выбрана из группы, состоящей из сарказма, неологизма, вариации грамматики, неправильно фразированного текста, пунктуации, многоязычных данных, орфографии, обфускации, шифрования, контекста и использования комбинации информационных материалов.

1

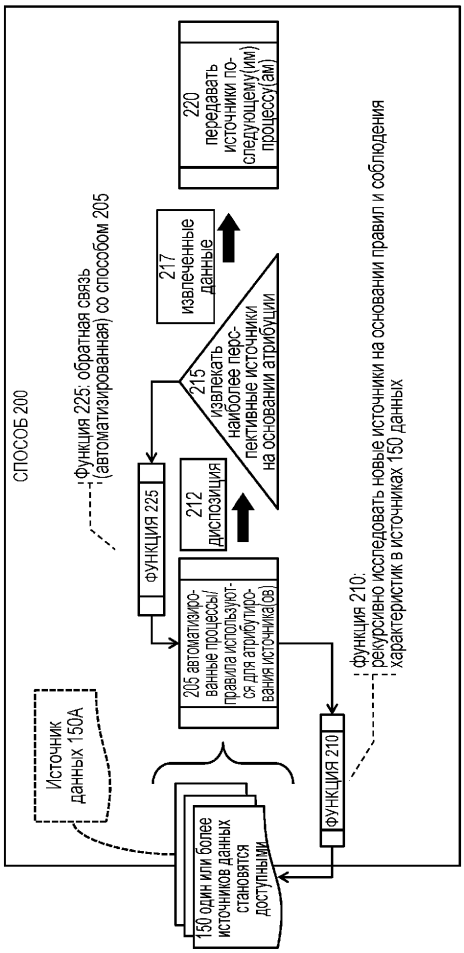
541091

1/4



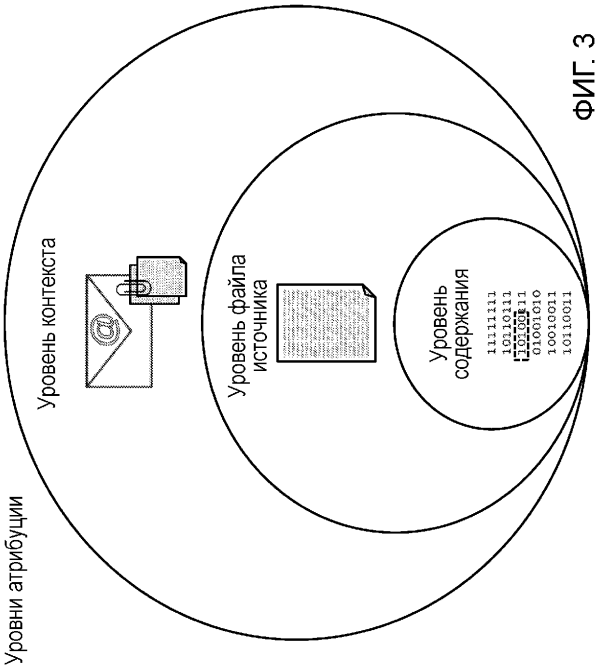
ФИГ. 1

2


СПОСОБ 200
Функция 225: обратная связь (автоматизированная) со способом 205
Источник данных 150A
150 один или более источников данных становятся доступными
205 автоматизированные процессы/правила используются для атрибутирования источника(ов)
ФУНКЦИЯ 225
212 диспозиция
215 извлекать наиболее перспективные источники на основании атрибуции
217 извлеченные данные
220 передавать источники последующему(им) процессу(ам)
ФУНКЦИЯ 210
функция 210: рекурсивно исследовать новые источники на основании правил и соблюдения характеристик в источниках 150 данных

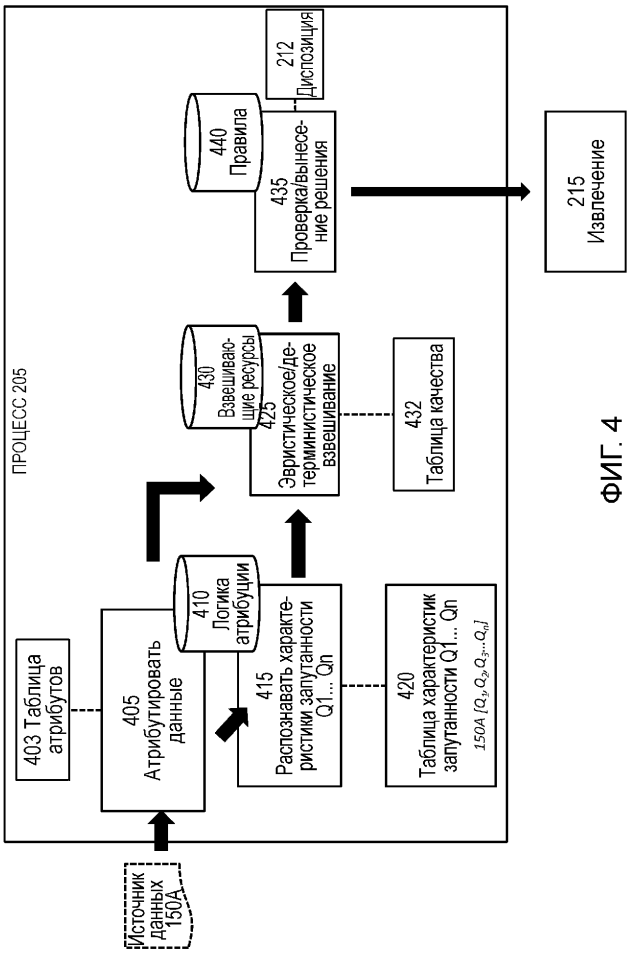
ФИГ. 2

3


Уровни атрибуции
Уровень контекста
Уровень файла источника
Уровень содержания
11111111
10110111
10100111
01001010
10010011
10110011

ФИГ. 3

4


ПРОЦЕСС 205
Источник данных 150A
403 Таблица атрибутов
405 Атрибутировать данные
410 Логика атрибуции
415 Распознавать характеристики запутанности Q1... Qn
420 Таблица характеристик запутанности Q1... Qn
150A [Q2, Q3...Qn]
425 Эвристическое/детерминистическое взвешивание
430 Взвешивающие ресурсы
432 Таблица качества
435 Проверка/вынесение решения
440 Правила
212 Диспозиция
215 Извлечение

ФИГ. 4