РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 666 277**(13) **C1**

(51) МПК
G06F 17/27 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 17/2785 (2006.01); G06F 17/30707 (2006.01); G06F 17/3071 (2006.01)

(21)(22) Заявка: **2017131334, 06.09.2017**

(24) Дата начала отсчета срока действия патента:
06.09.2017

Дата регистрации:
06.09.2018

Приоритет(ы):
(22) Дата подачи заявки: **06.09.2017**

(45) Опубликовано: **06.09.2018** Бюл. № **25**

Адрес для переписки:
127273, Москва, а/я 20, ООО "Аби Продакшн"

(72) Автор(ы):
Инденбом Евгений Михайлович (RU),
Колотиенко Сергей Сергеевич (RU)

(73) Патентообладатель(и):
Общество с ограниченной ответственностью "Аби Продакшн" (RU)

(56) Список документов, цитированных в отчете о поиске: **RU 2595594 C2, 27.08.2016. RU 2210809 C2, 20.08.2003. US 2007/0073533 A1, 29.03.2007. CN 105787088 A, 20.07.2016. CN 106570170 A, 19.04.2017. CN 106326346 A, 11.01.2017.**

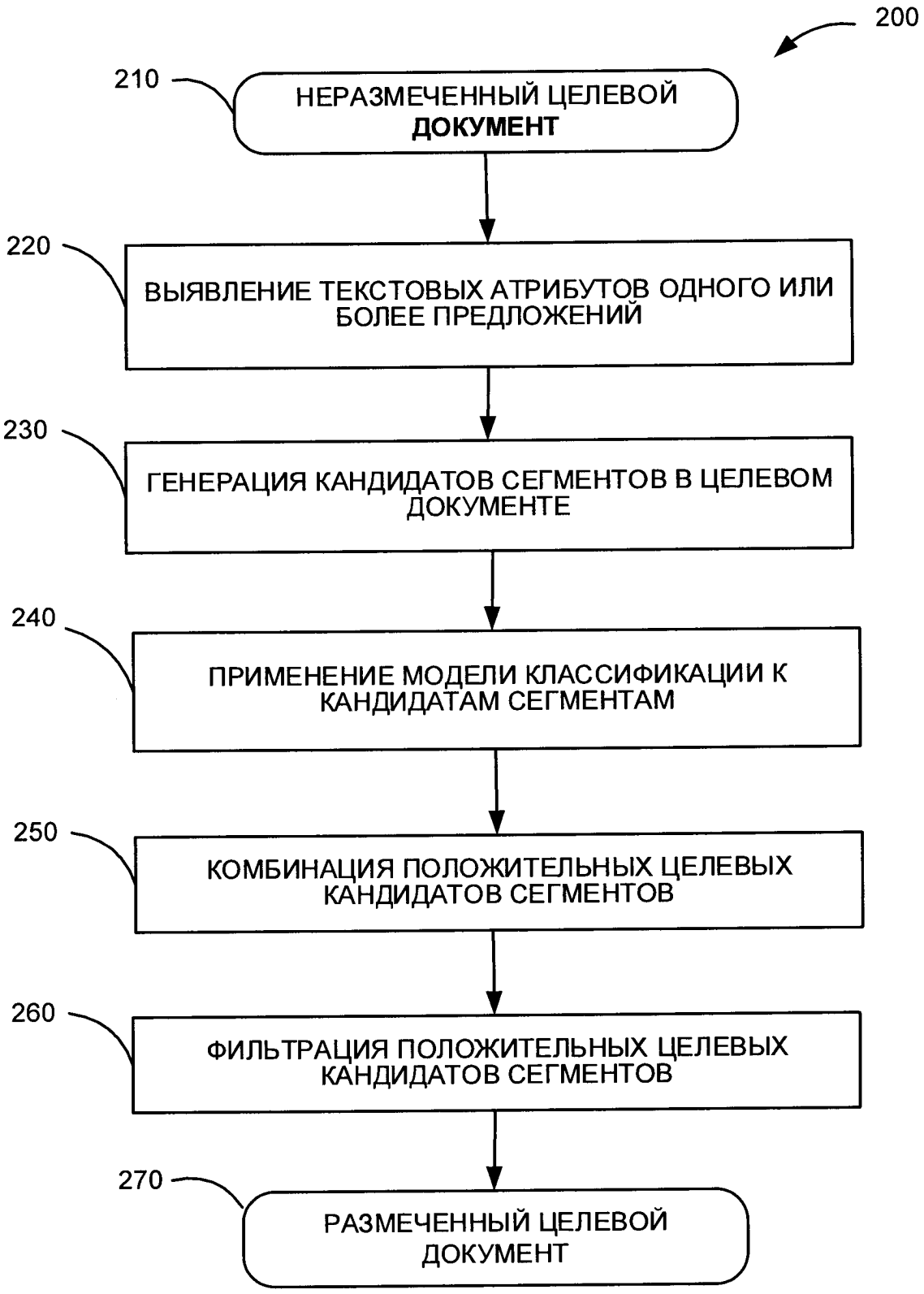
(54) СЕГМЕНТАЦИЯ ТЕКСТА

(57) Реферат:

Изобретение в целом относится к вычислительным системам, а точнее к системам и способам обработки естественного языка. Техническим результатом является повышение эффективности извлечения информации за счет сокращения времени предобработки документов и повышение точности извлекаемой информации. В способе автоматической сегментации текстового документа выполняют сегментацию для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, принадлежащих к типам сегментов из множества типов сегментов. Выявляют атрибуты целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов. Анализируют атрибуты целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента. Причем первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте. Анализируют текст первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов. 3 н. и 15 з.п. ф-лы, 4 ил.

RU 2666277 C1

RU 2666277 C1



Фиг. 2

RU 2666277 C1

RU 2666277 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 666 277**(13) **C1**

(51) Int. Cl.
G06F 17/27 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 17/2785 (2006.01); *G06F 17/30707* (2006.01); *G06F 17/3071* (2006.01)

(21)(22) Application: **2017131334, 06.09.2017**

(24) Effective date for property rights:
06.09.2017

Registration date:
06.09.2018

Priority:
(22) Date of filing: **06.09.2017**

(45) Date of publication: **06.09.2018** Bull. № **25**

Mail address:
127273, Moskva, a/ya 20, OOO "Abi Prodakshn"

(72) Inventor(s):
Indenbom Evgenij Mikhajlovich (RU),
Kolotienko Sergej Sergeevich (RU)

(73) Proprietor(s):
Obshchestvo s ogranichennoj otvetstvennostyu "Abi Prodakshn" (RU)

(54) TEXT SEGMENTATION

(57) Abstract:

FIELD: computer equipment.

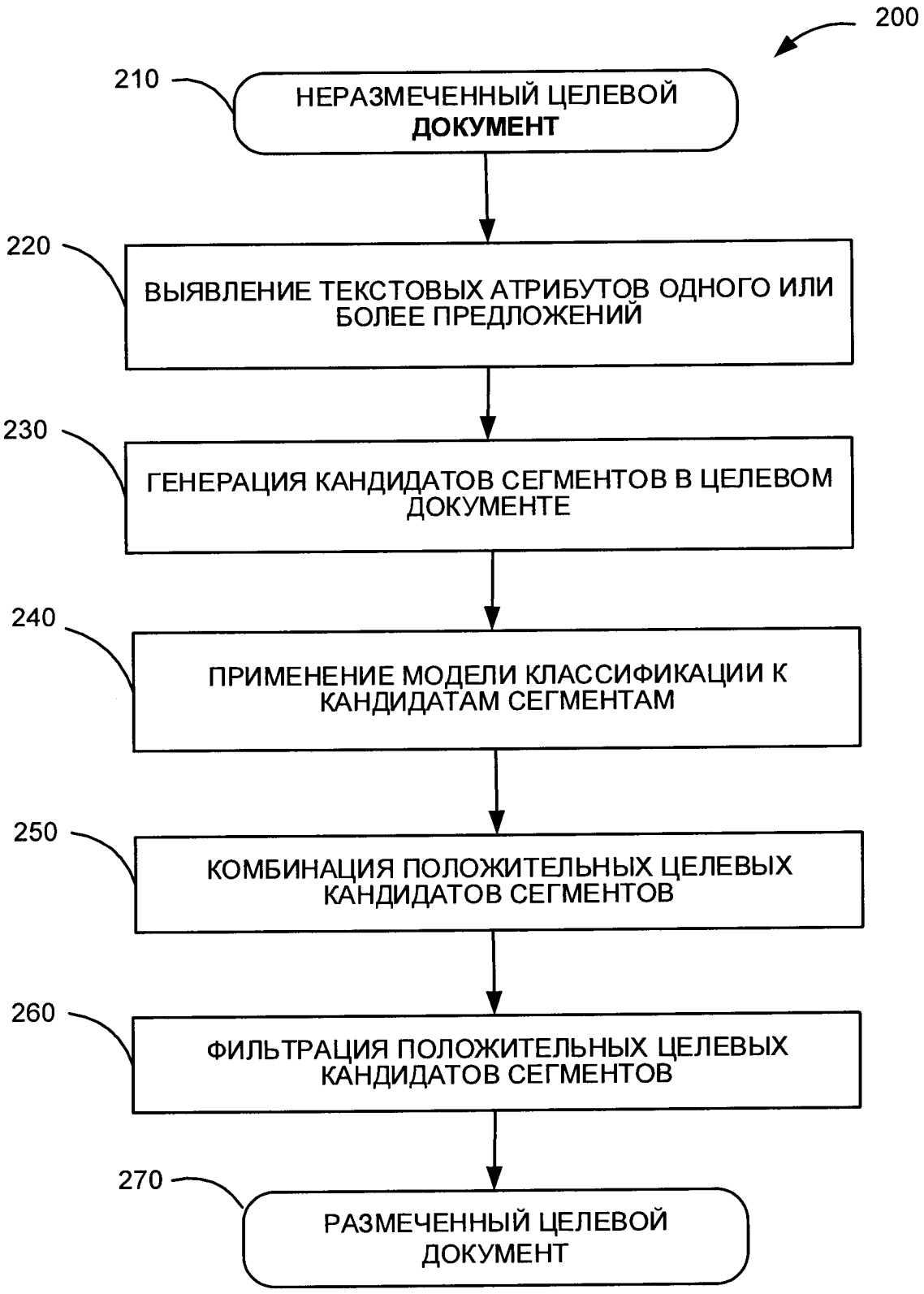
SUBSTANCE: invention, in general, relates to computer systems, or specifically to natural language processing systems and methods. In the method of automatic segmentation of a text document, segmentation is performed to mark out an unmarked target text to obtain a plurality of target candidate segments belonging to the types of segments from the plurality of types of segments. Attributes of the target text in the first target candidate segment are identified from the set of target candidate segments. Attributes of the target text in the first target candidate segment are analyzed using the first classifier of the segment type from the plurality of classifiers to determine the first target candidate segment as having the first type of the segment. And the first classifier of segment type was trained to define segments as corresponding to the first type of segments on the marked text. Text of the first target candidate segment is analyzed based on assigning the first target candidate segment to the first type of segments.

EFFECT: technical result is higher efficiency of information retrieval by reducing time of pre-processing of documents and higher accuracy of the information retrieved.

18 cl, 4 dwg

RU 2666277 C1

RU 2666277 C1



Фиг. 2

RU 2666277 C1

RU 2666277 C1

ОБЛАСТЬ ТЕХНИКИ

[0001] Настоящее изобретение в целом относится к вычислительным системам, а точнее - к системам и способам обработки естественного языка.

УРОВЕНЬ ТЕХНИКИ

[0002] Извлечение информации является одной из важных операций автоматизированной обработки текстов на естественном языке. Извлечение информации из текстов на естественном языке может быть затруднено многозначностью, которая является неотъемлемой особенностью естественных языков. Точное и своевременное извлечение информации, в свою очередь, может требовать значительных ресурсов. Извлечение информации можно оптимизировать за счет правил извлечения, с помощью которых идентифицируется конкретная информация в этих документах.

РАСКРЫТИЕ ИЗОБРЕТЕНИЯ

[0003] В соответствии с одним или более вариантами реализации настоящего изобретения пример способа сегментации текста может включать: выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов; выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов; анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте; выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов. В некоторых вариантах реализации обучение первого классификатора типа сегментов на размеченном тексте дополнительно включает: выявление атрибутов текста в размеченном тексте; создание множества сегментов-кандидатов в размеченном тексте; создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов; обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте.

[0004] В соответствии с одним или более вариантами реализации настоящего изобретения пример системы сегментации текста может включать: память и процессор, соединенный с запоминающим устройством, в котором процессор выполнен с возможностью выполнения следующих действий: выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов; выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов; анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте; выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов. В некоторых вариантах

реализации обучение первого классификатора типа сегментов на размеченном тексте дополнительно включает: выявление атрибутов текста в размеченном тексте; создание множества сегментов-кандидатов в размеченном тексте; создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов; обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте.

[0005] В соответствии с одним или более вариантами реализации настоящего изобретения пример постоянного машиночитаемого носителя данных, предназначенный для сегментации текста, может включать исполняемые команды, которые при выполнении их вычислительной системой приводят к следующим действиям вычислительной системы: выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов; выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов; анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте; выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов. В некоторых вариантах реализации обучение первого классификатора типа сегментов на размеченном тексте дополнительно включает: выявление атрибутов текста в размеченном тексте; создание множества сегментов-кандидатов в размеченном тексте; создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов; обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте. Технический результат от внедрения системы сегментации документов на основе выделения наиболее существенных признаков сегментов состоит в повышении эффективности извлечения информации за счет сокращения времени предобработки документов и в повышении точности извлекаемой информации.

КРАТКОЕ ОПИСАНИЕ ЧЕРТЕЖЕЙ

[0006] Настоящее изобретение иллюстрируется с помощью примеров, а не способом ограничения, и может быть лучше понято при рассмотрении приведенного ниже описания предпочтительных вариантов реализации в сочетании с чертежами, на которых:

[0007] На Фиг. 1 приведена блок-схема одного иллюстративного примера способа обучения параметров классификатора для выявления сегментов текста внутри документа;

[0008] Фиг. 2 иллюстрирует применение модели классификации 160 для разметки неразмеченного целевого документа.

[0009] Фиг. 3 иллюстрирует пример документа, содержащего разные типы сегментов.

[00010] На Фиг. 4 приведена схема примера вычислительной системы, реализующей методы настоящего изобретения.

ОПИСАНИЕ ПРЕДПОЧТИТЕЛЬНЫХ ВАРИАНТОВ РЕАЛИЗАЦИИ

[00011] Ниже описаны способы и системы сегментации документов, обучаемой на размеченном наборе документов. Извлечение данных может осуществляться с помощью применения правил извлечения. Однако, таких правил может быть слишком много, и

их последовательный перебор требует большого отрезка времени. Извлечение данных может быть оптимизировано, если разделить документ на определенные смысловые сегменты, и тогда, к каждому такому сегменту применять только ограниченный набор правил. В свою очередь, применение различных наборов правил для различных сегментов документа может предусматривать выполнение затратных операций для определения типа сегмента документа, прежде, чем появится возможность выбора конкретного правила извлечения. В некоторых вариантах реализации документы могут содержать «разметку», с помощью которой маркируются или иным образом определяются подлежащие извлечению сегменты текста в документе. Использование разметки может сократить объем обработки, необходимый для извлечения данных, однако выявление и разметка сегментов зачастую могут требовать значительного объема ручного труда.

[00012] Варианты реализации настоящего изобретения устраняют отмеченные выше и другие недостатки путем создания системы, способной быстро и точно производить автоматическую разметку сегментов внутри документа, используя процесс обучения, который позволяет системе создавать классификаторы, способные находить в документе и размечать сегменты определенных типов.

[00013] В иллюстративном примере система разметки получает целевой документ на естественном языке без какой-либо разметки. Под целевым документом на естественном языке понимается документ, содержащий текстовый контент (например, текстовый документ, документ в формате текстового редактора, изображение документа после оптического распознавания символов (OCR)). Затем система разметки документа может применять к целевому документу процесс классификации для разметки сегментов определенных типов.

[00014] Классификаторы, которые используются в процессе классификации, обучаются выявлять в документе сегменты определенного типа. Обучение проводится на размеченном наборе документов и позволяет системе быстро и эффективно выявлять в документе сегменты текста, снижая количество продукционных правил, применяемых к этому сегменту, и таким образом оптимизируя скорость и качество извлечения фактов из этого документа.

[00015] Различные аспекты упомянутых выше способов и систем подробно описаны ниже в этом документе с помощью примеров, а не способом ограничения.

[00016] На Фиг. 1 приведена блок-схема одного из иллюстративных примеров способа обучения параметров функций классификатора, используемых для выявления сегментов текста в целевых документах в соответствии с одним или более вариантами реализации настоящего изобретения. Способ 100 и (или) каждая из его отдельных функций, процедур, подпрограмм или операций может быть реализована с помощью одного или более процессоров вычислительной системы {например, вычислительной системы 400 на Фиг. 4), в которой реализован этот способ. В некоторых вариантах осуществления способ 100 может выполняться в одном потоке обработки. При альтернативном подходе способ 100 может осуществляться с использованием двух или более потоков обработки, при этом в каждом потоке реализована одна или более отдельных функций, процедур, подпрограмм или действий этого способа. В одном из иллюстративных примеров потоки обработки, в которых реализован способ 100, могут быть синхронизированы (например, с использованием семафоров, критических секций и (или) других механизмов синхронизации потоков). При альтернативном подходе потоки обработки, реализующие способ 100, могут выполняться асинхронно по отношению друг к другу.

[00017] На шаге 110 блок-схемы вычислительная система, реализующая способ,

может получать размеченный текст 110 на естественном языке (например, документ или совокупность документов). В одном из иллюстративных примеров вычислительное устройство может получить текст 110 на естественном языке в виде электронного документа, который может быть получен путем сканирования или за счет применения иного способа получения изображения с бумажного документа с последующим выполнением оптического распознавания символов (OCR) для получения текста документа. В другом иллюстративном примере вычислительная система может получить текст 110 на естественном языке в виде одного или более форматированных файлов, например, файлов системы электронной обработки текста, сообщений электронной почты, файлов цифровых данных и т.д. Размеченным текстом называется текст, содержащий информацию о разметке для размеченных сегментов, т.е. в котором явно выделен явно по меньшей мере один сегмент. В некоторых вариантах реализации сегмент представляет собой часть текста, которая содержит одно или более полных предложения, так что начальной точкой сегмента может являться начало предложения, и конечной точкой сегмента - может являться конец предложения. В некоторых вариантах реализации сегмент может содержать несколько предложений и абзацев. Каждый размеченный сегмент в размеченном тексте 110 на естественном языке связан с одним или более типами сегментов. В одном варианте реализации, типы сегментов могут включать, например, такие типы, как "заголовок", "текст", "таблица", "блок подписей", "стороны", "условия контракта", "условия оплаты", "порядок расторжения", "применимое законодательство" и т.д.

[00018] Фиг. 3 иллюстрирует пример размеченного документа 110, содержащего текст на естественном языке. В этом примере 110 текста размечены сегменты 310, 320, 330, 340, 350, 360, 370, 380, 390. Для каждого сегмента отмечены его начальная координата (311, 321, 331, 341, 351, 361, 371, 381, 391) и его конечная координата (312, 322, 332, 342, 352, 362, 372, 382, 392). Сегменты 310 и 320 имеют тип "заголовок". Сегмент 330 является сегментом типа "таблица". А сегменты 340 и 350 помечены как сегменты типа "текст". Сегмент 360 выделен как сегмент типа "стороны". Сегмент 370 является сегментом типа "цена", сегмент 380 размечен типом "оплата", а сегмент 390 является сегментом типа "дата".

[00019] В некоторых вариантах реализации информация о разметке размеченного сегмента включает информацию, описывающую сегмент. Эта информация может в некоторых вариантах реализации включать начальную точку размеченного сегмента, конечную точку размеченного сегмента и тип сегмента. В других вариантах реализации информация о разметке может включать начальную точку размеченного сегмента, длину размеченного сегмента и тип сегмента. В некоторых вариантах реализации размеченный текст на естественном языке может содержать множество сегментов одинакового типа. Однако размеченные сегменты одинакового типа не перекрываются. Могут существовать части размеченного текста на естественном языке, которые не принадлежат ни одному размеченному сегменту, то есть размеченные сегменты могут не покрывать весь текст. Сегменты разных типов могут пересекаться.. Кроме того, сегмент одного типа может быть внутри сегмента другого типа.

[00020] На шаге 120 вычислительная система может выявлять текстовые атрибуты для предложений текста 110 на естественном языке. Атрибутами для предложения являются текстовые характеристики этого предложения и (или) других предложений, примыкающих к рассматриваемому предложению. Атрибуты могут включать внутренние атрибуты, такие как определенное слово, имеющееся внутри предложения, или граничные атрибуты, такие как слово или знак пунктуации, находящийся рядом с

этим предложением. Положение предложения в тексте относительно других предложений также может быть одним из атрибутов.

[00021] На шаге 130 вычислительная система может создавать набор сегментов-кандидатов для каждого типа сегмента. В некоторых вариантах реализации набор сегментов-кандидатов представляет собой набор всех сочетаний соседних предложений в тексте, включая состоящие из одного предложения, одного абзаца, все сочетания 2 соседних предложений, 3 соседних предложений и т.д.

[00022] В некоторых вариантах реализации система может использовать больше селектирующих критериев для создания набора сегментов-кандидатов, например, используя классификатор для выявления кандидатов начала и кандидатов конца для сегментов-кандидатов. В некоторых вариантах реализации эти классификаторы обучаются на полученном размеченном тексте на естественном языке. В других вариантах реализации классификатор обучается заранее.

[00023] В других вариантах реализации система может устанавливать ограничение на длину сегментов-кандидатов. В некоторых вариантах реализации максимальная длина сегмента-кандидата определяется заранее. В некоторых вариантах реализации максимальная длина сегмента-кандидата определяется исходя из анализа полученного размеченного текста на естественном языке и размеченных в нем сегментов. Признаками сегмента может являться комбинация внутренних атрибутов входящих в него предложений, а также краевых атрибутов от крайних предложений.

[00024] На шаге 140 вычислительная система может создавать обучающую выборку для каждого типа сегментов. В одной из реализаций для создания обучающей выборки для определенного типа сегментов система создает подмножество сегментов-кандидатов из набора кандидатов-сегментов, созданного на шаге 130, и присваивает каждому сегменту-кандидату в подмножестве значение 1 или 0. Сегменту-кандидату приписывается значение 1, если размеченный текст на естественном языке содержит размеченный сегмент определенного типа с таким же местоположением, как этот сегмент-кандидат. Все остальные сегменты-кандидаты в обучающей выборке обозначаются как 0. В некоторых вариантах реализации такие обучающие выборки сегментов-кандидатов создаются для каждого типа сегментов. В некоторых вариантах реализации обучающие выборки создаются для некоторого подмножества типов сегментов. В некоторых вариантах реализации пользователь может указать, для каких типов сегментов нужны обучающие выборки.

[00025] На шаге 150 вычислительная система может обучать классификаторы вида "один против всех" для каждого типа сегментов. Для этих классификаторов могут использоваться различные модели машинного обучения. В некоторых вариантах реализации классификаторы представляют собой классификаторы на основе модели линейного метода опорных векторов (SVM). В других вариантах реализации используются классификаторы на основе случайного леса (random forest). В некоторых вариантах реализации для разных типов сегментов используются классификаторы различных типов. При обучении классификатора для определенного типа сегментов система использует обучающую выборку, созданную для этого типа сегментов на шаге 140, и текстовые атрибуты, выявленные на шаге 120. В некоторых вариантах реализации используются все выявленные атрибуты типа сегмента. В других вариантах реализации в обучении используются только те атрибуты типа сегмента, которые присутствуют в соответствующей обучающей выборке.

[00026] Группа таких обученных классификаторов, каждый из которых соответствует только одному типу сегментов, образует модель классификации 160, которую можно

использовать позднее для разметки сегментов в произвольном документе.

[00027] На Фиг. 2 показано, как модель классификации 160 может использоваться для разметки неразмеченного целевого документа 210.

[00028] На Фиг. 2 приведена блок-схема одного иллюстративного примера способа разметки неразмеченного документа с использованием модели классификации в соответствии с одним или более вариантами реализации настоящего изобретения. Способ 200 и (или) каждая из его отдельных функций, процедур, подпрограмм или операций может быть реализована с помощью одного или более процессоров вычислительной системы (например, вычислительной системы 400 на Фиг. 4), в которой реализован этот способ. В некоторых реализациях способ 200 может быть реализован в одном потоке обработки. В качестве альтернативы способ 200 может быть реализован с помощью двух или более потоков обработки, при этом каждый поток выполняет одну или более отдельных функций, стандартных программ, подпрограмм или операций данного способа. В иллюстрирующем примере реализующие способ 200 потоки обработки могут быть синхронизированы (например, с помощью семафоров, критических секций и (или) других механизмов синхронизации потоков). В качестве альтернативы реализующие способ 200 потоки обработки могут выполняться асинхронно по отношению друг к другу.

[00029] На шаге 210 вычислительная система, реализующая способ, может получать неразмеченный целевой документ 210, содержащий текст на естественном языке, который в соответствии со способом 200 размечается с использованием модели классификации 160. В одном из иллюстративных примеров вычислительное устройство может получить целевой текст 210 на естественном языке в виде электронного документа, который может быть получен путем сканирования или за счет применения иного способа получения изображения с бумажного документа с последующим выполнением оптического распознавания символов (OCR) для получения текста документа. В некоторых вариантах реализации целевой текст 210 не содержит какой-либо разметки, определяющей размеченные сегменты текста. В других вариантах реализации целевой текст 210 содержит определенную разметку сегментов, которая дополняется и (или) заменяется разметкой сегментов, создаваемой по способу 200.

[00030] На шаге 220 вычислительная система может выявлять текстовые атрибуты для некоторых предложений в неразмеченном тексте 210 на естественном языке, аналогично шагу 120. В некоторых вариантах реализации система может выявлять атрибуты текста для каждого предложения целевого текста 210.

[00031] На шаге 230 вычислительная система может создавать набор сегментов-кандидатов для текста 210. Аналогично шагу 130 в некоторых вариантах реализации набор сегментов-кандидатов представляет собой набор всех сочетаний соседних предложений в тексте, включая состоящие из одного предложения, одного абзаца, все сочетания 2 соседних предложений, 3 соседних предложений, и т.д. Как и на шаге 130, в некоторых вариантах реализации система устанавливает ограничение на длину сегментов-кандидатов. В некоторых вариантах реализации максимальная длина сегмента-кандидата определяется заранее. В других вариантах реализации максимальная длина сегмента-кандидата определяется другими средствами.

[00032] На шаге 240 вычислительная система может применять модель классификации 160 к набору сегментов-кандидатов, созданному на шаге 230. Другими словами, система использует классификаторы, обученные на шаге 150, для выявления сегментов определенного типа в наборе сегментов-кандидатов, созданном на шаге 230. Каждый отдельный классификатор в модели 160, соответствующий определенному типу

сегментов, сортирует сегменты-кандидаты из набора сегментов-кандидатов неразмеченного целевого текста 210. В результате сегменты этого определенного типа из набора сегментов-кандидатов классифицируются как положительные сегменты-кандидаты этого типа. Каждый положительный сегмент-кандидат связывается с типом сегмента классификатора, который отметил его как положительный.

[00033] В некоторых вариантах реализации система применяет к набору сегментов-кандидатов все классификаторы модели классификации 160. В других вариантах реализации системой или пользователем выбирается подмножество типов сегментов и соответствующих им классификаторов.

[00034] На шаге 250 вычислительная система может объединять все положительные сегменты-кандидаты всех типов для всех примененных на шаге 240 классификаторов. В некоторых вариантах реализации система создает предварительно размеченный целевой текст на естественном языке, который включает разметку для всех положительных сегментов-кандидатов, созданных всеми классификаторами на шаге 240.

[00035] На шаге 260 вычислительная система может фильтровать объединенный набор сегментов, созданный на шаге 250. В некоторых вариантах реализации фильтрация включает объединение двух или более перекрывающихся положительных сегментов-кандидатов одного типа с образованием одного сегмента, покрывающего все перекрывающиеся положительные сегменты-кандидаты. В других вариантах реализации, если два или более положительных сегментов-кандидатов одного типа перекрываются, выбирается сегмент с более высокой степенью уверенности классификации оставляется, а другие перекрывающие сегменты исключаются из рассмотрения.

[00036] В результате способ 200 создает размеченный целевой текстовый документ 270, который содержит разметку сегментов, аналогичную разметке сегментов в размеченном тексте 110. Кроме того, разметка размеченного целевого текста 270 может содержать информацию о степени уверенности в классификации размеченных сегментов.

[00037] В некоторых вариантах реализации система дополнительно обрабатывает целевой текст, разрешая противоречия в типах сегментов. Система выявляет в целевом тексте противоречивые сегменты, которые были определены как принадлежащие к двум или более различным типам сегментов. В некоторых вариантах реализации система разрешает эту неоднозначность, выполняя семантический анализ этих сегментов.

[00038] В некоторых вариантах реализации разметка в размеченном целевом тексте 270 используется при обработке естественного языка, применяемой к целевому тексту, например, извлечению данных, для оптимизации наборов правил извлечения для размеченного сегмента, в соответствии с типом сегмента.

[00039] На Фиг. 4 показан иллюстративный пример вычислительной системы 400, которая может исполнять набор команд, которые вызывают выполнение вычислительной системой любого отдельно взятого или нескольких способов настоящего изобретения. Вычислительная система может быть соединена с другой вычислительной системой по локальной сети, корпоративной сети, сети экстранет или сети Интернет. Вычислительная система может работать в качестве сервера или клиента в сетевой среде «клиент/сервер» либо в качестве однорангового вычислительного устройства в одноранговой (или распределенной) сетевой среде. Вычислительная система может быть представлена персональным компьютером (ПК), планшетным ПК, телевизионной приставкой (STB), карманным ПК (PDA), сотовым телефоном или любой вычислительной системой, способной выполнять набор команд (последовательно или иным образом), определяющих операции, которые должны быть выполнены этой

вычислительной системой. Кроме того, несмотря на то что показана только одна вычислительная система, термин «вычислительная система» также может включать любую совокупность вычислительных систем, которые отдельно или совместно выполняют набор (или несколько наборов) команд для выполнения одной или более методик, обсуждаемых в настоящем документе.

[00040] Пример вычислительной системы 400 включает процессор 502, основное запоминающее устройство 504 (например, постоянное запоминающее устройство (ПЗУ) или динамическое оперативное запоминающее устройство (ДОЗУ)) и устройство хранения данных 518, которые взаимодействуют друг с другом по шине 530.

[00041] Процессор 502 может быть представлен одной или более универсальными вычислительными системами, например, микропроцессором, центральным процессором и т.д. В частности, процессор 502 может представлять собой микропроцессор с полным набором команд (CISC), микропроцессор с сокращенным набором команд (RISC), микропроцессор с командными словами сверхбольшой длины (VLIW), процессор, реализующий другой набор команд или процессоры, реализующие комбинацию наборов команд. Процессор 502 также может представлять собой одну или более вычислительных систем специального назначения, например заказную интегральную микросхему (ASIC), программируемую пользователем вентильную матрицу (FPGA), процессор цифровых сигналов (DSP), сетевой процессор и т.п. Процессор 502 реализован с возможностью выполнения команд 526 для осуществления рассмотренных в настоящем документе операций и функций.

[00042] Вычислительная система 400 может дополнительно включать устройство сетевого интерфейса 522, устройство визуального отображения 510, устройство ввода символов 512 (например, клавиатуру) и устройство ввода в виде сенсорного экрана 514.

[00043] Устройство хранения данных 518 может содержать машиночитаемый носитель данных 524, в котором хранится один или более наборов команд 526 и в котором реализованы одна или более методик или функций, рассмотренных в настоящем документе. Команды 526 также могут находиться полностью или по меньшей мере частично в основной памяти 504 и (или) в процессоре 502 во время выполнения их в вычислительной системе 1000, при этом оперативное запоминающее устройство 504 и процессор 502 также представляют собой машиночитаемый носитель данных. Команды 526 также могут передаваться или приниматься по сети 516 через устройство сетевого интерфейса 522.

[00044] В некоторых вариантах реализации изобретения набор команд 526 может содержать команды способов 100, 400 для восстановления текстовых аннотаций, связанных с информационными объектами, в соответствии с одним или более вариантами реализации настоящего изобретения. Несмотря на то что машиночитаемый носитель данных 524 показан в примере на Фиг. 20 в виде одного носителя, термин «машиночитаемый носитель» следует понимать в широком смысле, подразумевающем один носитель или несколько носителей (например, централизованную или распределенную базу данных и (или) соответствующие кэши и серверы), в которых хранится один или более наборов команд. Термин «машиночитаемый носитель данных» также следует понимать как включающий любой носитель, который может хранить, кодировать или переносить набор команд для выполнения машиной и который обеспечивает выполнение машиной любой одной или более методик настоящего изобретения. Поэтому термин «машиночитаемый носитель данных» относится, помимо прочего, к твердотельным запоминающим устройствам, а также к оптическим и

магнитным носителям.

[00045] Способы, компоненты и функции, описанные в этом документе, могут быть реализованы с помощью дискретных компонентов оборудования либо они могут быть встроены в функции других компонентов оборудования, например ASICS (специализированная заказная интегральная схема), FPGA (программируемая логическая интегральная схема), DSP (цифровой сигнальный процессор) или аналогичных устройств. Кроме того, способы, компоненты и функции могут быть реализованы с помощью модулей встроенного программного обеспечения или функциональных схем аппаратного обеспечения. Способы, компоненты и функции также могут быть реализованы с помощью любой комбинации аппаратного обеспечения и программных компонентов, либо исключительно с помощью программного обеспечения.

[00046] В приведенном выше описании изложены многочисленные детали. Однако любому специалисту в этой области техники, ознакомившемуся с этим описанием, должно быть очевидно, что настоящее изобретение может быть осуществлено на практике без этих конкретных деталей. В некоторых случаях хорошо известные структуры и устройства показаны в виде блок-схем без детализации, чтобы не усложнять описание настоящего изобретения.

[00047] Некоторые части описания предпочтительных вариантов реализации изобретения представлены в виде алгоритмов и символического представления операций с битами данных в запоминающем устройстве компьютера. Такие описания и представления алгоритмов представляют собой средства, используемые специалистами в области обработки данных, что обеспечивает наиболее эффективную передачу сущности работы другим специалистам в данной области. В контексте настоящего описания, как это и принято, алгоритмом называется логически непротиворечивая последовательность операций, приводящих к желаемому результату. Операции подразумевают действия, требующие физических манипуляций с физическими величинами. Обычно, хотя и необязательно, эти величины принимают форму электрических или магнитных сигналов, которые можно хранить, передавать, комбинировать, сравнивать и выполнять другие манипуляции. Иногда удобно, прежде всего для обычного использования, описывать эти сигналы в виде битов, значений, элементов, символов, терминов, цифр и т.д.

[00048] Однако следует иметь в виду, что все эти и подобные термины должны быть связаны с соответствующими физическими величинами и что они являются лишь удобными обозначениями, применяемыми к этим величинам. Если явно не указано обратное, принимается, что в последующем описании термины «определение», «вычисление», «расчет», «получение», «установление», «определение», «изменение» и т.п.относятся к действиям и процессам вычислительной системы или аналогичной электронной вычислительной системы, которая использует и преобразует данные, представленные в виде физических (например, электронных) величин в реестрах и запоминающих устройствах вычислительной системы, в другие данные, также представленные в виде физических величин в запоминающих устройствах или реестрах вычислительной системы или иных устройствах хранения, передачи или отображения такой информации.

[00049] Настоящее изобретение также относится к устройству для выполнения операций, описанных в настоящем документе. Такое устройство может быть специально сконструировано для требуемых целей, либо оно может представлять собой универсальный компьютер, который избирательно приводится в действие или дополнительно настраивается с помощью программы, хранящейся в памяти компьютера.

Такая компьютерная программа может храниться на машиночитаемом носителе данных, например, помимо прочего, на диске любого типа, включая дискеты, оптические диски, CD-ROM и магнитно-оптические диски, постоянные запоминающие устройства (ПЗУ), оперативные запоминающие устройства (ОЗУ), СППЗУ, ЭППЗУ, магнитные или оптические карты и носители любого типа, подходящие для хранения электронной информации.

[00050] Следует понимать, что приведенное выше описание призвано иллюстрировать, а не ограничивать сущность изобретения. Специалистам в данной области техники после прочтения и уяснения приведенного выше описания станут очевидны и различные другие варианты реализации изобретения. Исходя из этого область применения изобретения должна определяться с учетом прилагаемой формулы изобретения, а также всех областей применения эквивалентных способов, на которые в равной степени распространяется формула изобретения.

(57) Формула изобретения

1. Способ автоматической сегментации текстового документа, включающий:

выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов;

выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов;

анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте;

выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов.

2. Способ по п. 1, отличающийся тем, что:

первый классификатор типа сегмента представляет собой классификатор вида "один против всех".

3. Способ по п. 1, отличающийся тем, что:

целевой сегмент-кандидат состоит из одного или более предложений.

4. Способ по п. 1, дополнительно включающий:

фильтрацию классифицированных сегментов-кандидатов.

5. Способ по п. 1, дополнительно включающий:

выявление противоречивых целевых сегментов, где противоречивыми целевыми сегментами считаются сегменты из множества целевых сегментов-кандидатов, классифицированные двумя или более классификаторами типов сегментов как принадлежащие к двум или более типам сегментов;

выполнение семантического анализа противоречивых сегментов;

классификацию противоречивых предложений, как принадлежащих к сегментам одного типа из множества типов сегментов, исходя из семантического анализа противоречивых сегментов.

6. Способ по п. 1, отличающийся тем, что обучение первого классификатора типа сегментов на размеченном тексте включает:

выявление атрибутов текста в размеченном тексте;

создание множества сегментов-кандидатов в размеченном тексте;

создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов;

обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте.

7. Система автоматической сегментации текстового документа, включающая:

память;

процессор, соединенный с запоминающим устройством, в котором процессор выполнен с возможностью выполнения следующих действий:

выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов;

выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов;

анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте;

выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов.

8. Система по п. 7, отличающаяся тем, что:

первый классификатор типа сегмента представляет собой классификатор вида "один против всех".

9. Система по п. 7, отличающаяся тем, что:

целевой сегмент-кандидат состоит из одного или более предложений.

10. Система по п. 7, дополнительно включающая: фильтрацию классифицированных сегментов-кандидатов.

11. Система по п. 7, дополнительно включающая:

выявление противоречивых целевых сегментов, где противоречивыми целевыми сегментами считаются сегменты из множества целевых сегментов-кандидатов, классифицированные двумя или более классификаторами типов сегментов как принадлежащие к двум или более типам сегментов;

выполнение семантического анализа противоречивых сегментов;

классификацию противоречивых предложений, как принадлежащих к сегментам одного типа из множества типов сегментов, исходя из семантического анализа противоречивых сегментов.

12. Система по п. 7, где обучение первого классификатора типа сегментов на размеченном тексте включает:

выявление атрибутов текста в размеченном тексте;

создание множества сегментов-кандидатов в размеченном тексте;

создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов;

обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте.

13. Постоянный машиночитаемый носитель данных, предназначенный для сегментации текста, включающий исполняемые команды, которые при выполнении их вычислительной системой приводят к следующим действиям вычислительной системы:

выполнение обрабатывающим устройством сегментации для разметки неразмеченного целевого текста для получения множества целевых сегментов-кандидатов, таких, что один или более сегментов-кандидатов принадлежат к одному или более типам сегментов из множества типов сегментов;

выявление атрибутов целевого текста в первом целевом сегменте-кандидате из множества целевых сегментов-кандидатов;

анализ атрибутов целевого текста в первом целевом сегменте-кандидате с использованием первого классификатора типа сегмента из множества классификаторов типов сегментов для определения первого целевого сегмента-кандидата как имеющего первый тип сегмента из множества типов сегментов, при том, что первый классификатор типа сегмента был обучен определять сегменты как соответствующие первому типу сегментов на размеченном тексте;

выполнение анализа текста первого целевого сегмента-кандидата исходя из отнесения первого целевого сегмента-кандидата к первому типу сегментов.

14. Постоянный машиночитаемый носитель данных по п. 13, отличающийся тем, что:

первый классификатор типа сегмента представляет собой классификатор вида "один против всех".

15. Постоянный машиночитаемый носитель данных по п. 13, отличающийся тем, что:

целевой сегмент-кандидат состоит из одного или более предложений.

16. Постоянный машиночитаемый носитель данных по п. 13, дополнительно включающий:

фильтрацию классифицированных сегментов-кандидатов.

17. Постоянный машиночитаемый носитель данных по п. 13, дополнительно включающий:

выявление противоречивых целевых сегментов, где противоречивыми целевыми сегментами считаются сегменты из множества целевых сегментов-кандидатов, классифицированные двумя или более классификаторами типов сегментов как принадлежащие к двум или более типам сегментов;

выполнение семантического анализа противоречивых сегментов;

классификацию противоречивых предложений, как принадлежащих к сегментам одного типа из множества типов сегментов, исходя из семантического анализа противоречивых сегментов.

18. Постоянный машиночитаемый носитель данных по п. 13, отличающийся тем, что обучение первого классификатора типа сегментов на размеченном тексте включает:

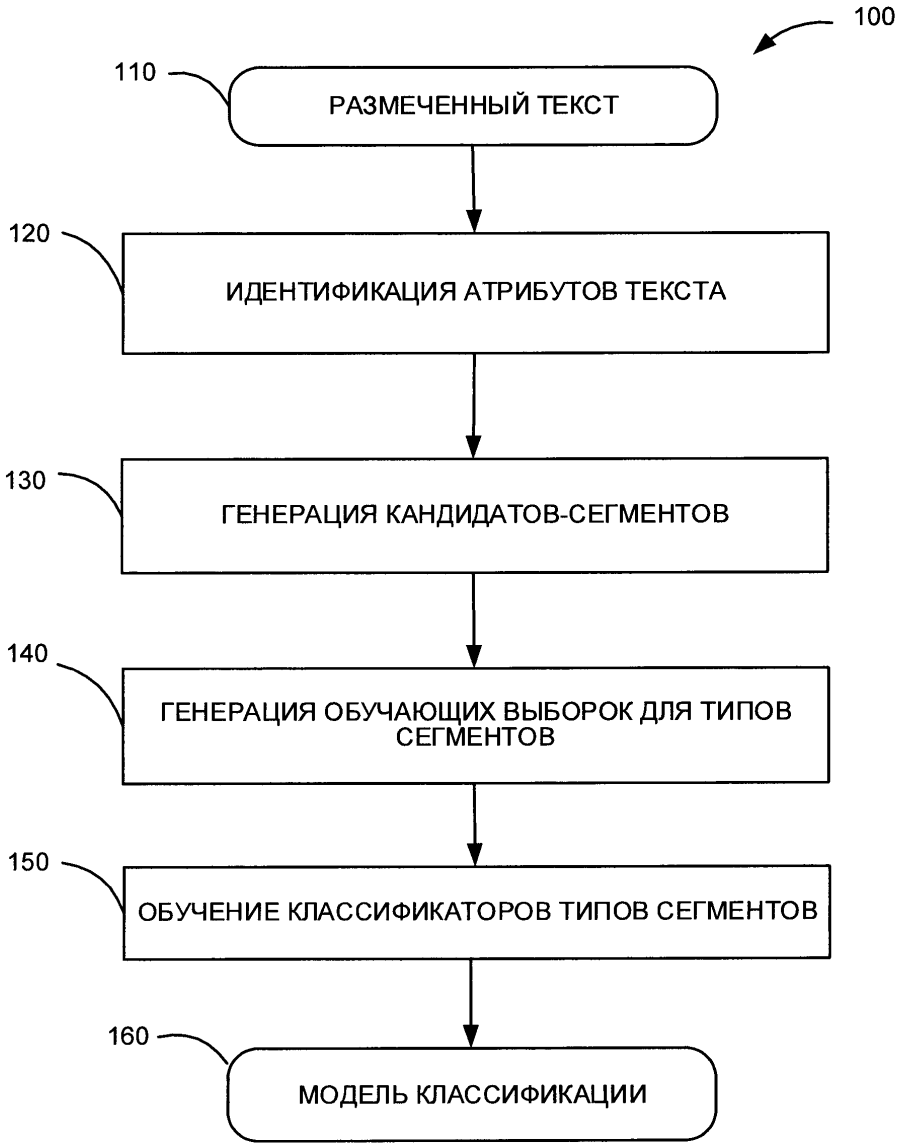
выявление атрибутов текста в размеченном тексте;

создание множества сегментов-кандидатов в размеченном тексте;

создание обучающей выборки первого типа для первого типа сегментов из множества сегментов-кандидатов;

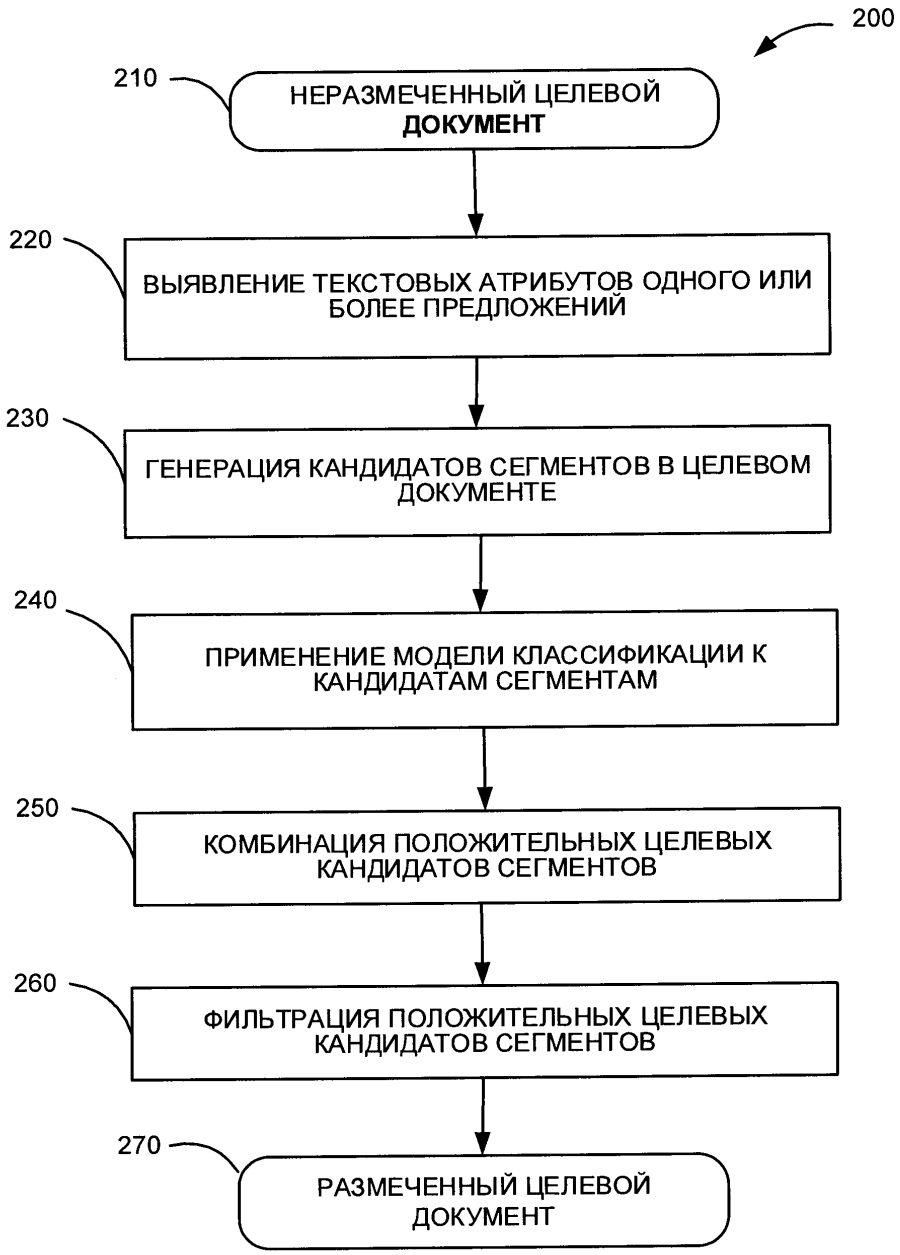
обучение классификатора первого типа сегментов на обучающей выборке первого типа с использованием атрибутов текста в размеченном тексте.

1


100
110
РАЗМЕЧЕННЫЙ ТЕКСТ
120
ИДЕНТИФИКАЦИЯ АТРИБУТОВ ТЕКСТА
130
ГЕНЕРАЦИЯ КАНДИДАТОВ-СЕГМЕНТОВ
140
ГЕНЕРАЦИЯ ОБУЧАЮЩИХ ВЫБОРОК ДЛЯ ТИПОВ СЕГМЕНТОВ
150
ОБУЧЕНИЕ КЛАССИФИКАТОРОВ ТИПОВ СЕГМЕНТОВ
160
МОДЕЛЬ КЛАССИФИКАЦИИ

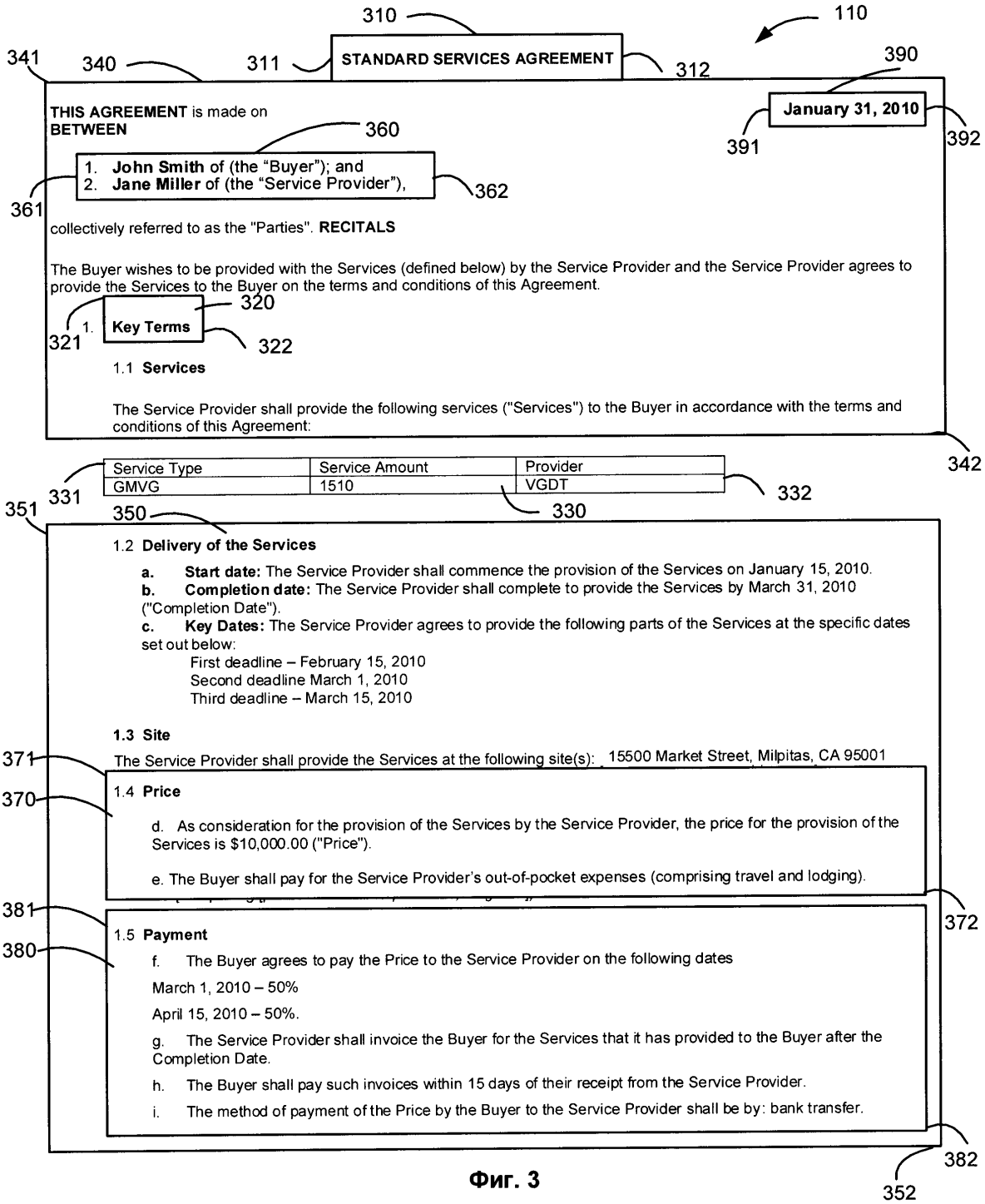
Фиг. 1

2


200
210
НЕРАЗМЕЧЕННЫЙ ЦЕЛЕВОЙ ДОКУМЕНТ
220
ВЫЯВЛЕНИЕ ТЕКСТОВЫХ АТРИБУТОВ ОДНОГО ИЛИ БОЛЕЕ ПРЕДЛОЖЕНИЙ
230
ГЕНЕРАЦИЯ КАНДИДАТОВ СЕГМЕНТОВ В ЦЕЛЕВОМ ДОКУМЕНТЕ
240
ПРИМЕНЕНИЕ МОДЕЛИ КЛАССИФИКАЦИИ К КАНДИДАТАМ СЕГМЕНТАМ
250
КОМБИНАЦИЯ ПОЛОЖИТЕЛЬНЫХ ЦЕЛЕВЫХ КАНДИДАТОВ СЕГМЕНТОВ
260
ФИЛЬТРАЦИЯ ПОЛОЖИТЕЛЬНЫХ ЦЕЛЕВЫХ КАНДИДАТОВ СЕГМЕНТОВ
270
РАЗМЕЧЕННЫЙ ЦЕЛЕВОЙ ДОКУМЕНТ

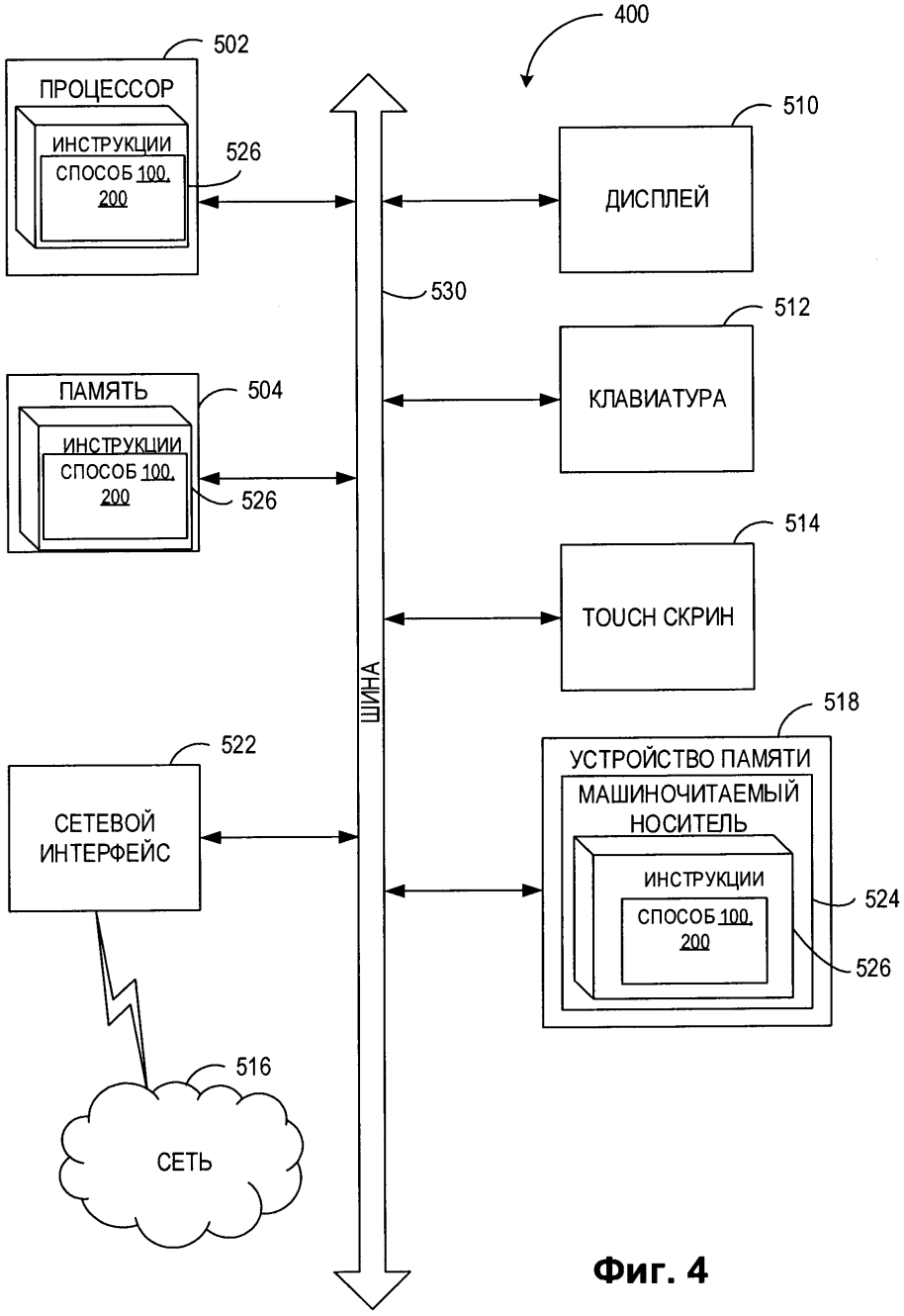
Фиг. 2

3


310
110
341
340
311
STANDARD SERVICES AGREEMENT
312
390
THIS AGREEMENT is made on
January 31, 2010
391
392
BETWEEN
360
1. John Smith of (the "Buyer"); and
2. Jane Miller of (the "Service Provider"),
362
361
collectively referred to as the "Parties". RECITALS
The Buyer wishes to be provided with the Services (defined below) by the Service Provider and the Service Provider agrees to provide the Services to the Buyer on the terms and conditions of this Agreement.
320
1. Key Terms
321
322
1.1 Services
The Service Provider shall provide the following services ("Services") to the Buyer in accordance with the terms and conditions of this Agreement:
342
Service Type | Service Amount | Provider
GMVG | 1510 | VGDT
331
332
330
351
350
1.2 Delivery of the Services
a. Start date: The Service Provider shall commence the provision of the Services on January 15, 2010.
b. Completion date: The Service Provider shall complete to provide the Services by March 31, 2010 ("Completion Date").
c. Key Dates: The Service Provider agrees to provide the following parts of the Services at the specific dates set out below:
First deadline – February 15, 2010
Second deadline March 1, 2010
Third deadline – March 15, 2010
1.3 Site
371
The Service Provider shall provide the Services at the following site(s): 15500 Market Street, Milpitas, CA 95001
370
1.4 Price
d. As consideration for the provision of the Services by the Service Provider, the price for the provision of the Services is $10,000.00 ("Price").
e. The Buyer shall pay for the Service Provider's out-of-pocket expenses (comprising travel and lodging).
372
381
380
1.5 Payment
f. The Buyer agrees to pay the Price to the Service Provider on the following dates
March 1, 2010 – 50%
April 15, 2010 – 50%.
g. The Service Provider shall invoice the Buyer for the Services that it has provided to the Buyer after the Completion Date.
h. The Buyer shall pay such invoices within 15 days of their receipt from the Service Provider.
i. The method of payment of the Price by the Buyer to the Service Provider shall be by: bank transfer.
382
352

Фиг. 3

4


400
502
ПРОЦЕССОР
ИНСТРУКЦИИ
СПОСОБ 100, 200
526
510
ДИСПЛЕЙ
530
512
КЛАВИАТУРА
ПАМЯТЬ
504
ИНСТРУКЦИИ
СПОСОБ 100, 200
526
514
TOUCH СКРИН
ШИНА
518
УСТРОЙСТВО ПАМЯТИ
МАШИНОЧИТАЕМЫЙ НОСИТЕЛЬ
ИНСТРУКЦИИ
СПОСОБ 100, 200
524
526
522
СЕТЕВОЙ ИНТЕРФЕЙС
516
СЕТЬ

Фиг. 4