РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(19) **RU** (11) **2 399 959** (13) **C2**

(51) МПК
G09B 19/00 (2006.01)
G06F 17/27 (2006.01)

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21), (22) Заявка: **2008142648/12, 29.10.2008**

(24) Дата начала отсчета срока действия патента:
29.10.2008

(43) Дата публикации заявки: **10.05.2010**

(45) Опубликовано: **20.09.2010** Бюл. № **26**

(56) Список документов, цитированных в отчете о поиске: **US 2007/0073533 A1, 29.03.2007. RU 2273879 C2, 10.04.2006. US 7346493 B2, 18.03.2008. US 7305336 B2, 04.12.2007. US 7191115 B2, 13.03.2007.**

Адрес для переписки:
119606, Москва, пр-кт Вернадского, 84, корп.2, ЗАО "Авикомп Сервисез"

(72) Автор(ы):
Хорошевский Владимир Фёдорович (RU),
Клинцов Виктор Петрович (RU)

(73) Патентообладатель(и):
Закрытое акционерное общество "Авикомп Сервисез" (RU)

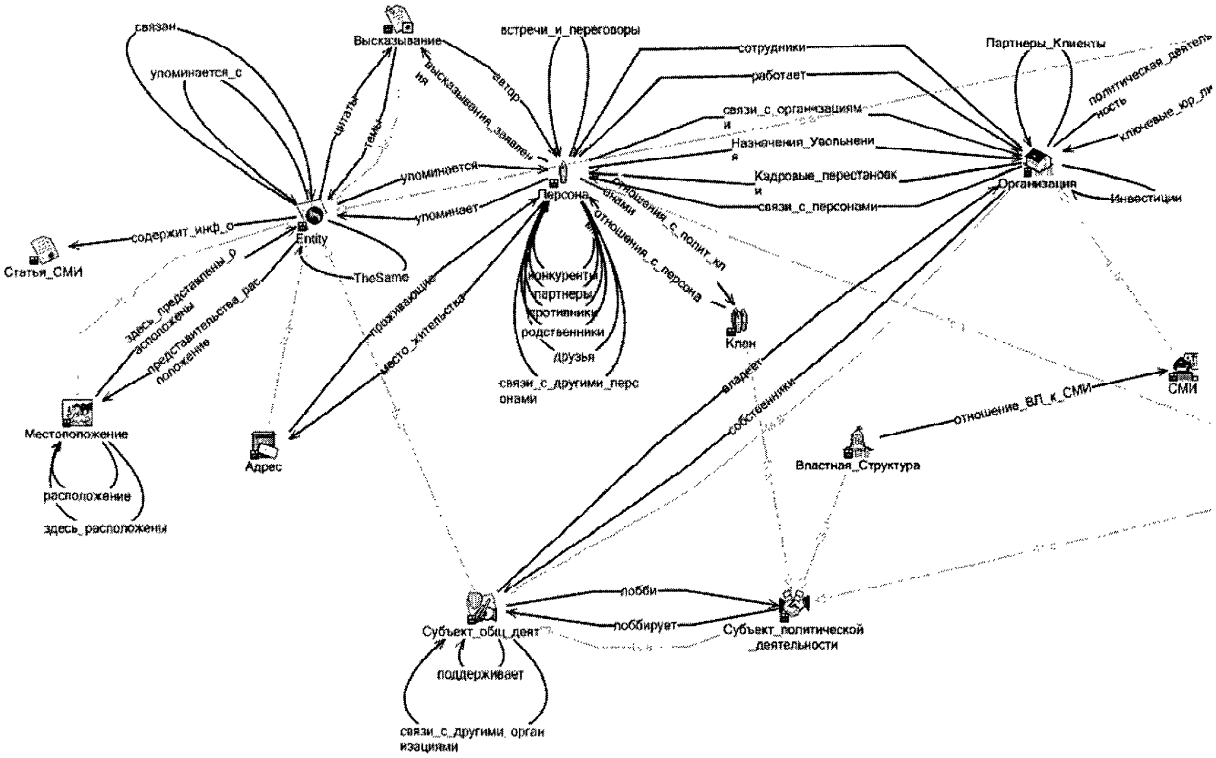
(54) СПОСОБ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ ПУТЕМ ЕГО СЕМАНТИЧЕСКОЙ ИНДЕКСАЦИИ, СПОСОБ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ КОЛЛЕКЦИИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ ПУТЕМ ИХ СЕМАНТИЧЕСКОЙ ИНДЕКСАЦИИ И МАШИНОЧИТАЕМЫЕ НОСИТЕЛИ

(57) Реферат:

Изобретение относится к области информационных технологий. Текст сегментируют в электронной форме на элементарные единицы. Выявляют устойчивые словосочетания, формируют предложения. Выявляют семантически значимые объекты и семантически значимые отношения между ними. Формируют для каждого семантически значимого отношения множество триад, в которых единственная триада первого типа соответствует связи, устанавливаемой семантически значимым отношением между двумя семантически значимыми объектами. Каждая из триад второго типа соответствует значению конкретного атрибута одного из этих семантически значимых объектов. Каждая из триад третьего типа соответствует значению конкретного атрибута самого семантически значимого отношения. Индексируют на множестве сформированных триад все связанные семантически значимыми отношениями семантически значимые объекты по отдельности. Запоминают в базе данных сформированные триады и полученные индексы вместе со ссылкой на исходный текст, из которого сформированы эти триады. Техническим результатом изобретения является повышение точности и скорости поиска релевантных фактов и документов. 4 н. и 8 з.п. ф-лы, 16 табл., 7 ил.

RU 2399959 C2

RU 2399959 C2



Фиг. 2

RU 2399959 C2

RU 2399959 C2

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(19) **RU** (11) **2 399 959** (13) **C2**

(51) Int. Cl.
G09B 19/00 (2006.01)
G06F 17/27 (2006.01)

(12) ABSTRACT OF INVENTION

(21), (22) Application: **2008142648/12, 29.10.2008**

(24) Effective date for property rights:
29.10.2008

(43) Application published: **10.05.2010**

(45) Date of publication: **20.09.2010 Bull. 26**

Mail address:
119606, Moskva, pr-kt Vernadskogo, 84, korp.2, ZAO "Avikomp Servisez"

(72) Inventor(s):
Khoroshevskij Vladimir Fedorovich (RU),
Klintsov Viktor Petrovich (RU)

(73) Proprietor(s):
Zakrytoe aktsionernoe obshchestvo "Avikomp Servisez" (RU)

(54) METHOD FOR AUTOMATIC TEXT PROCESSING IN NATURAL LANGUAGE THROUGH SEMANTIC INDEXATION, METHOD FOR AUTOMATIC PROCESSING COLLECTION OF TEXTS IN NATURAL LANGUAGE THROUGH SEMANTIC INDEXATION AND COMPUTER READABLE MEDIA

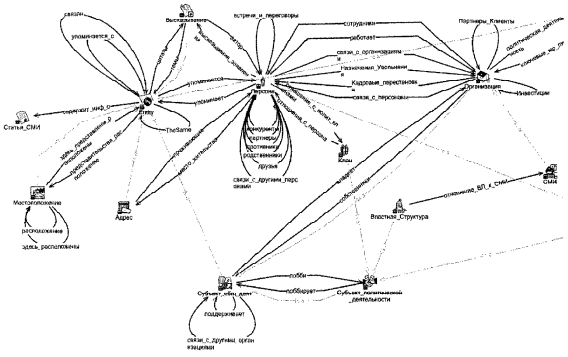
(57) Abstract:

FIELD: information technology.

SUBSTANCE: text is segmented in electronic form to elementary units. Fixed collocations are identified and sentences are formed. Semantically significant objects and semantically significant relationships between then are identified. Several triads are formed for each semantically significant relationship, in which a single first type triad corresponds to the link set by the semantically significant relationship between two semantically significant objects. Each second type triad corresponds to the value of a specific attribute of one of these semantically significant objects. Each third type triad corresponds to the value of a specific attribute of the semantically significant relationship itself. All semantically significant objects which are linked by semantically significant relationships are separately indexed into several formed triads. The formed triads and the obtained indices together with the link to the initial text from which said triads were formed are stored in a database.

EFFECT: more accurate and faster searching for relevant facts and documents.

12 cl, 9 dwg, 16 tbl, 1 ex



Фиг. 2

RU 2399959 C2

RU 2399959 C2

Область техники, к которой относится изобретение

Настоящее изобретение относится к области информационных технологий, а именно к способам автоматизированной обработки текста на естественном языке путем его семантической индексации, а также к машиночитаемым носителям, содержащим соответствующие программы, и может применяться для упорядочивания и накопления информации по конкретно заданным предметным областям с целью семантической навигации по документам и коллекциям документов, а также высокоточного и быстрого поиска релевантных информационным потребностям пользователя фактов и документов.

Уровень техники

В настоящее время известны различные способы автоматизированной индексации текстов на естественных языках.

Так, в патенте ЕАПВ №002016 (опубл. 22.01.2001) описан способ, в котором во фрагментах текстового документа определяют уникальные блоки информации и используют их для последующей обработки и поиска. В патенте РФ №2268488 (опубл. 20.01.2006), выданном на основе заявки PCT WO 01/06414, раскрыт способ, в котором кодируют слова, фразы, идиомы, предложения и даже идеи для последующей числовой обработки. В патенте РФ №2273879 (опубл. 10.04.2006) приведен способ, в котором проводят морфологический и синтаксический анализ текста с последующей индексацией найденных единиц. В способе по патенту США №6871174 (опубл. 22.03.2005) определяют сходство текстов по текстовым фрагментам. Недостаток всех этих способов состоит в том, что в них не учитывается семантическая неоднозначность слов и выражений естественного языка.

В патенте США №6189002 (опубл. 13.02.2001) раскрыт способ, в котором текст разбивают на абзацы и слова, которые преобразуют в векторы упорядоченных элементов. Каждый элемент вектора соответствует абзацу, найденному применением заданной функции к числу появлений в этом абзаце слова, соответствующего этому элементу. Текстовый вектор рассматривается как семантический профиль документа. Однако, с учетом многообразия абзацев, данный способ требует огромного массива запомненных данных и не различает семантической неоднозначности слов и выражений.

Учет семантической неоднозначности осуществляется во многих известных способах. Например, в патенте РФ №2242048 (опубл. 10.12.2004), патентах США №№6871199 (опубл. 22.03.2005), 7024407 (опубл. 04.04.2006) и 7383169 (опубл. 03.06.2008), заявках на патент США №№2007/0005343 и 2007/0005344 (обе опубл. 04.01.2007), 2008/0097951 (опубл. 24.04.2008), выложенных заявках Японии №№05-128149 (опубл. 25.05.1993), 06-195374 (опубл. 15.07.1994), 10-171806 (опубл. 26.06.1998) и 2005-182438 (опубл. 07.07.2005), в заявке ЕПВ №0853286 (опубл. 15.07.1998) описаны способы, в которых тем или иным образом устраняется неоднозначность встречающихся в текстах слов и (или) выражений. Однако все эти способы имеют лишь частное применение и не затрагивают полноценной семантической индексации текста.

Наиболее близкий к заявленной группе изобретений способ семантической индексации текста или коллекции текстов на естественном языке раскрыт в заявке на патент США №2007/0073533 (опубл. 29.03.2007). В этом способе в сегментированном тексте определяют функциональную структуру для каждого участка текста и, в каждой функциональной структуре, находят триады, характеризующие предикатные члены, на основе правил переноса линеаризации. Затем выделяют из каждого участка

текста такие признаки как: именованная сущность, тождество по референту, лексическая статья, семантико-структурное отношение, атрибутивная и меронимическая информация. Далее определяют для каждого участка текста, на основе найденных структур конституэнтов, канонизированные представления триад, характеризующих предикатные члены, и выявленных признаков, и определяют структурный индекс на основе канонизированного представления участка текста. Этот способ обеспечивает хорошие результаты, но все же несколько ограничен вследствие того, что в виде триад линеаризуют фрагменты предикатно-аргументной структуры, полученные при синтаксическом анализе. Кроме того, этот способ ориентирован только на поисковые задачи, а не на задачи навигации по массиву документов.

Сущность изобретения

Цель настоящего изобретения состоит в расширении арсенала способов автоматизированной обработки текста на естественном языке путем его семантической индексации за счет использования методов автоматизированного лингвистического анализа и последующего использования его результатов для построения семантических индексов, что обеспечивает семантическую навигацию по документам и коллекциям документов, а также высокоточный и быстрый поиск релевантных информационным потребностям пользователя фактов и документов, особенно в применении к текстам на высоко флективных языках.

Достижение этой цели и получение указанного технического результата обеспечиваются с помощью способа автоматизированной обработки текста на естественном языке путем его семантической индексации и способа автоматизированной обработки коллекции текстов на естественном языке путем их семантической индексации согласно признакам независимых пунктов, соответственно 1 и 6, приложенной формулы изобретения. Варианты обоих способов раскрываются в соответствующих зависимых пунктах этой формулы изобретения.

Краткое описание чертежей

Изобретение поясняется описанием конкретного примера его выполнения и прилагаемыми чертежами, где:

на фиг.1 приведена условная блок-схема, поясняющая заявленные способы;

на фиг.2 приведен фрагмент спецификации предметной области;

на фиг.3a, 3b приведены спецификация правила для выделения семантически значимых объектов типа «Персона» и соответствующая ей логическая схема обработки сигналов;

на фиг.4a, 4b приведены спецификация правила для выделения семантически значимых отношений типа «работать» и соответствующая ей логическая схема обработки сигналов;

на фиг.5 приведен фрагмент графического представления результатов обработки текста;

на фиг.6 показана общая схема сохранения результатов обработки одного текста;

на фиг.7 приведены спецификация левой части правила для объединения семантически значимых объектов типа «Персона» и соответствующая ей логическая схема обработки сигналов.

Подробное описание изобретения

Предлагаемые способы позволяют эффективно осуществлять смысловую индексацию текстов на естественном языке, как для целей дальнейшей семантической навигации по документам и коллекциям документов, так и для поисковых целей.

Способ автоматизированной обработки текста на естественном языке путем его семантической индексации по первому объекту настоящего изобретения и способ автоматизированной обработки коллекции текстов на естественном языке путем их семантической индексации по второму объекту настоящего изобретения могут быть реализованы практически в любой вычислительной среде, к примеру на персональном компьютере, подключенном к внешним базам данных. Этапы осуществления этих способов иллюстрируются на фиг.1.

Все дальнейшие пояснения даются в применении к русскому языку, который является одним из самых высоко флективных языков, хотя заявляемые способы применимы к семантической индексации текстов на любых естественных языках.

Прежде всего подлежащий индексации текст необходимо представить в электронной форме для последующей автоматизированной обработки. Этот этап на фиг.1 условно обозначен ссылочной позицией 1 и может быть выполнен любым известным способом, например сканированием текста с последующим распознаванием с помощью общеизвестных средств типа ABBYY FineReader. Если же текст поступает на индексацию из электронной сети, к примеру из Интернета, то этап его представления в электронной форме выполняется заранее, до размещения этого текста в сети.

Преобразованный в электронную форму текст поступает на обработку, в процессе которой сначала этот текст сегментируется на элементарные единицы первого уровня, именуемые в общеизвестной литературе токенами (Token). Токеном может быть любой текстовый объект из следующего множества: слова, состоящие каждое из последовательности букв и, возможно, дефисов; последовательность пробелов; знаки препинания; числа. Иногда сюда же относят такие последовательности символов как A300, i150b и т.п. Выделение токенов всегда осуществляется по достаточно простым правилам, например, как в упомянутой заявке на патент США №2007/0073533. На фиг.1 этот этап условно обозначен ссылочной позицией 2.

Кроме того, для каждого токена, представляющего собой слово, на основе морфологического анализа формируются соответствующие элементарные единицы второго уровня, именуемые в общеизвестной литературе и далее морфами. При этом для каждого слова выявляется его нормализованная словоформа. К примеру, для слова «иду» нормализованной словоформой будет «идти», для слова «красивого» нормализованной словоформой будет «красивый», а для слова «стеной» нормализованная словоформа - «стена». Кроме того, для каждой словоформы указывается часть речи, к которой относится данное слово, и его морфологические характеристики. Естественно, что для разных частей речи эти характеристики различны. К примеру, для существительных и прилагательных это род (мужской - женский - средний), число (единственное - множественное), падеж; для глаголов это вид (совершенный - несовершенный), лицо, число (единственное - множественное); и т.д. Таким образом, для заданного слова его морфом является нормализованная словоформа + морфологические характеристики, в том числе часть речи. Одно и то же слово может иметь несколько морфов. Например, слово «стекло» имеет два морфа - один для существительного среднего рода и один для глагола в прошедшем времени.

Специалистам должно быть понятно, что операции этого и последующих этапов осуществляются с запоминанием промежуточных результатов, например в оперативном запоминающем устройстве (ОЗУ).

Следующий этап, условно обозначенный на фиг.1 ссылочной позицией 3, состоит в том, что на множестве полученных элементарных единиц первых двух уровней

(токенов и морфов) выявляют словосочетания. Это действие выполняется путем преобразования элементарных единиц, т.е. токенов и морфов, в последовательности, которые сопоставляются с последовательностями нормализованных слов и их характеристик в заранее запомненных в базе данных словарях, где слова приведены с указанием морфологических и синтаксических связей между ними. При совпадении очередной сопоставляемой последовательности с соответствующей словарной последовательностью эта очередная сопоставляемая последовательность считается словосочетанием и в таком качестве запоминается в базе данных как элементарная единица третьего уровня.

На следующем этапе, обозначенном на фиг.1 ссылочной позицией 4, формируют предложения, соответствующие участкам индексируемого текста. Обычно это реальные предложения, оканчивающиеся точкой, но в некоторых случаях бывает удобно трактовать в качестве предложения какие-то части обычных предложений, скажем, отдельные элементы при перечислении. Поэтому данная операция может дать на выходе предложения, не всегда совпадающие с предложениями индексируемого текста в традиционном понимании.

Указанная выше последовательность этапов определяется тем, что выявление устойчивых словосочетаний до формирования предложений позволяет в некоторых случаях снять определенные неоднозначности еще до этапов детального анализа текста. Так, например, фиксация устойчивого словосочетания «МГУ им. М.В.Ломоносова» в тексте «…С 1992 г. по 1997 г. Иванов учился в МГУ им. М.В.Ломоносова…» позволяет снять ложный конец предложения после слова «им», которое в общем случае является местоимением, а в данном - сокращением от слова «имени».

После этапа 4 выполняют многоступенчатый семантико-синтаксический анализ. Условно на фиг.1 этот анализ разбит на этапы, обозначенные ссылочными позициями 5-11. Упомянутый многоступенчатый семантико-синтаксический анализ выполняют путем обращения к сформированным в базе данных лингвистическим и эвристическим правилам в заранее заданной лингвистической среде. Такой средой может быть, например, лингвистическая среда, упомянутая в вышеуказанном патенте РФ №2242048, или среда, раскрытая в упомянутой заявке на патент США №2007/0073533, либо любая иная лингвистическая среда, определяющая соответствующие правила, которые позволяют устранять синтаксические и семантические неоднозначности слов и выражений реального текста. Лингвистические и эвристические правила в выбранной среде именуются далее правилами. В процессе упомянутого многоступенчатого семантико-синтаксического анализа выявляют семантически значимые объекты (ссылочная позиция 5 на фиг.1) и их атрибуты (ссылочные позиции 7 и 9 на фиг.1).

Выявление семантически значимых объектов, которые считаются элементарными единицами четвертого уровня, производится в предложении на множестве элементарных единиц первого, второго и (или) третьего уровней. При этом для каждого семантически значимого объекта с помощью упомянутых правил формируют морфологические атрибуты из морфологических атрибутов тех элементарных единиц второго и/или третьего уровней (т.е. морфов и/или словосочетаний), которые составляют данный семантически значимый объект. Кроме того, для каждого семантически значимого объекта с помощью упомянутых правил формируют семантические атрибуты из семантических атрибутов элементарных единиц второго и/или третьего уровней, которые составляют данный семантически

значимый объект. Этап формирования указанных атрибутов условно обозначен на фиг.1 ссылочной позицией 7. А на этапе, обозначенном на фиг.1 ссылочной позицией 9, каждому семантически значимому объекту присваивают соответствующий тип из предметной онтологии по тематике той предметной области, к которой относится индексируемый текст. Под онтологией в данном случае понимается спецификация конкретной предметной области, которая хранится в соответствующей базе данных.

Для каждого семантически значимого объекта, т.е. элементарной единицы четвертого уровня, с присвоенным ему типом находят соответствующую ему анафорическую ссылку (если она есть), считающуюся элементарной единицей пятого уровня. Например, в предложении «Капитану Леклеру не суждено было ступить на родную землю: он умер от горячки в открытом море.» анафорической ссылкой к семантически значимому объекту «Леклер» (слово «Леклеру» в тексте) будет местоимение «он», тогда как объект «Леклер» будет антецедентом для этой анафоры. Этот этап нахождения анафорической ссылки условно обозначен на фиг.1 ссылочной позицией 11.

После этого каждый выявленный семантически значимый объект сохраняют в соответствующей памяти вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами. Анафорическую ссылку запоминают вместе с типом и атрибутами семантически значимого объекта, который является антецедентом этой анафорической ссылки, а также с указанием тождества по референции между этим семантически значимым объектом и его анафорической ссылкой.

После выполнения этапов, обозначенных на фиг.1 ссылочными позициями 5-7-9-11, на основе элементарных единиц первого, второго, третьего, четвертого и/или пятого уровней находят с помощью упомянутых правил семантически значимые отношения между семантически значимыми объектами (этап 6). Семантически значимые отношения могут связывать семантически значимые объекты как внутри одного предложения, так и в пределах всего индексируемого текста.

Для каждого семантически значимого отношения на этапе, обозначенном на фиг.1 ссылочной позицией 8, с помощью упомянутых правил находят морфологические атрибуты из элементарных единиц второго уровня (т.е. морфов), составляющих данное отношение, а также семантические атрибуты из составляющих данное семантически значимое отношение элементарных единиц первого, второго, третьего и/или четвертого уровней.

На этапе, обозначенном на фиг.1 ссылочной позицией 10, каждому семантически значимому отношению присваивают соответствующий тип из хранящейся в базе данных предметной онтологии по тематике той предметной области, к которой относится индексируемый текст. После этого каждое семантически значимое отношение сохраняют в соответствующей памяти вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами.

На этапе, обозначенном на фиг.1 ссылочной позицией 12, сохраненные семантически значимые объекты и семантически значимые отношения используют для формирования триад. При этом в пределах индексируемого текста для каждого из выявленных семантически значимых отношений, связывающих определенные семантически значимые объекты, формируют множество триад трех типов. Единственная триада первого типа соответствует связи, устанавливаемой семантически значимым отношением между двумя семантически значимыми объектами. Каждая из множества триад второго типа соответствует значению

конкретного атрибута одного из этих семантически значимых объектов, а каждая из множества триад третьего типа соответствует значению конкретного атрибута самого семантически значимого отношения. Если обозначить два семантически значимых объекта через O_i и O_j, а связывающее их семантически значимое отношение через R_{ij}, то триаду первого типа можно условно представить (изобразить) как $O_i \rightarrow R_{ij} \rightarrow O_j$. Каждая из триад второго типа может быть представлена как $O_i \rightarrow A_{im} \rightarrow V_{im}$ или $O_j \rightarrow A_{jn} \rightarrow V_{jn}$, где A_{im} и A_{jn} являются соответствующими атрибутами, а V_{im} или V_{jn} являются соответственно значениями этих атрибутов. Аналогично, каждую из множества триад третьего типа можно представить как $R_{ij} \rightarrow A_{ijk} \rightarrow V_{ijk}$, где A_{ijk} является соответствующим атрибутом, а V_{ijp} является значением этого атрибута. В этих записях индексы i, j, k, m, n и p представляют собой целые числа.

Затем на этапе, обозначенном на фиг.1 ссылочной позицией 13, выполняют индексацию текста. Для этого на множестве сформированных триад индексируют все связанные семантически значимыми отношениями семантически значимые объекты по отдельности, все пары вида «семантически значимый объект - семантически значимое отношение» и все триады вида «семантически значимый объект - семантически значимое отношение - семантически значимый объект» с учетом атрибутов соответствующих семантически значимых объектов и/или семантически значимых отношений. Сформированные на этапе 12 триады и полученные на этапе 13 индексы вместе со ссылкой на исходный текст, из которого сформированы эти триады, сохраняют в базе данных (этап 15 на фиг.1; этап 14 при этом пропускается). Перед этим сначала выполняют (не показано на фиг.1) свертку объектов, которые связаны отношениями тождества по референции, в единый объект, множество атрибутов которого является объединением атрибутов всех объектов, связанных друг с другом отношениями тождества по референции. Это делается для того, чтобы сократить объем памяти в базе данных, требуемый для сохранения таких объектов, а также для того, чтобы интегрировать в рамках одного объекта информацию, полученную из всего текста.

Способ автоматизированной обработки коллекции текстов на естественном языке путем их семантической индексации согласно второму объекту настоящего изобретения выполняется точно так же, как и уже рассмотренный способ автоматизированной обработки текста на естественном языке путем его семантической индексации согласно первому объекту настоящего изобретения, но в этом случае после этапа 13 индексации и перед этапом 15 сохранения в базе данных осуществляют еще один этап. На этом этапе, обозначенном на фиг.1 ссылочной позицией 14 и выполняемом по существу одновременно с этапом 15, при сохранении в базе данных сформированных триад и полученных семантических индексов очередного текста из коллекции текстов выполняют следующее. С помощью сформированных в базе данных лингвистических и эвристических правил в заранее заданной лингвистической среде сравнивают вновь выявленные семантически значимые объекты и семантически значимые отношения с уже имеющимися в базе данных семантически значимыми объектами и семантически значимыми отношениями. В случае идентификации одинаковых семантически значимых объектов и/или семантически значимых отношений дублирующую информацию в базе данных не запоминают, а к соответствующим семантически значимым объектам и/или семантически значимым отношениям добавляют ссылки на очередные тексты, в которых они присутствуют, и ссылки на текстовые фрагменты в пределах каждого из очередных текстов, из которых они выделены. Благодаря этому индексация коллекции

текстов происходит практически так же, как и индексация первого текста этой коллекции (или первого текста, проиндексированного данным способом), что позволяет существенно упростить всю процедуру индексации, сократить требуемый объем памяти и интегрировать в рамках одного объекта информацию, полученную из разных текстов.

Для специалистов очевидно, что упоминавшиеся на отдельных этапах запоминающие устройства могут на деле быть как разными устройствами, так и одним запоминающим устройством достаточного объема. Точно так же отдельные базы данных, упоминавшиеся на соответствующих этапах, могут быть не только физически раздельными базами данных, но и единственной базой данных. Более того, упомянутые запоминающие устройства (памяти) могут быть выполнены на той же самой единственной базе данных, либо объединяться с одной из упомянутых баз данных. Специалистам также понятно, что заявленные в настоящем изобретении способы выполняются в соответствующей вычислительной среде под управлением соответствующих программ, которые записаны на машиночитаемых носителях, предназначенных для непосредственного участия в работе компьютера. Поэтому объектами настоящего изобретения являются также и машиночитаемые носители с такими программами.

Пример

Для иллюстрации осуществления заявленного способа автоматизированной обработки текстов на естественном языке путем их семантической индексации рассмотрим следующий пример. Пусть имеется совокупность русских текстов из электронной коллекции литературной классики, представленной на Интернет-сайте http://www.litra.ru. Таким образом, можно считать, что преобразование текстов в электронную форму, обозначенное на фиг.1 ссылочной позицией 1, уже выполнено.

Типичным примером таких текстов является следующий фрагмент из романа А. Дюма «Граф Монте-Кристо»:

27 февраля 1815 г. Марсель из очередного плавания возвращается трехмачтовый корабль «Фараон». Капитану Леклеру не суждено было ступить на родную землю: он умер от горячки в открытом море. Молодой моряк Эдмон Дантес принял на себя командование. Владелец корабля Моррель предлагает Дантесу официально вступить в должность капитана корабля «Фараон»…

В соответствии с заявленным способом автоматизированной обработки текстов на естественном языке путем их семантической индексации используют предварительно созданную спецификацию предметной области, в рамках которой будет осуществляться обработка коллекции документов и построение семантического индекса. Фрагмент такой спецификации представлен на фиг.2. Подобные спецификации готовятся людьми-экспертами, которые на основании своего опыта и знаний фиксируют перечень типов объектов и перечень типовых отношений между ними, существенных для данной предметной области.

В приведенном примере основными типами объектов являются «ФизЛицо», «Организация», «Местоположение» и некоторые другие. Типовые отношения между ними делятся на два класса - общие, характерные для любых предметных областей, например отношение «БЫТЬ_ПРИМЕРОМ» (is_a), фиксирующее иерархию объектов типа «потомок-предок», и специальные - специфичные для выбранной предметной области, например в приведенном примере это типовые отношения «работать_в», «владеть», «высказывания_заявления» и тому подобные.

Кроме того, людьми-экспертами предварительно строится и множество правил,

причем каждое правило содержит в левой части шаблон поиска примеров объектов и/или примеров отношений между ними, а в правой части - операторы фиксации в тексте найденных по шаблону примеров объектов и/или примеров отношений между ними. В дальнейшем с помощью таких правил, подготовленных людьми-лингвистами, в обрабатываемых текстах автоматически выявляют конкретные сведения, соответствующие спецификации предметной области.

Кроме спецификации предметной области и правил в соответствии с изложенными выше способами используются словари общей и специальной лексики.

В соответствии с заявленным способом автоматизированной обработки текстов на естественном языке путем их семантической индексации сначала осуществляют сегментацию текста на элементарные единицы - токены и морфологический анализ токенов-слов (ссылочная позиция 2 на фиг.1). В результате выполнения этого этапа исходный текст трансформируется во множество токенов и морфов, которые представлены в Таблице 1 и Таблице 2, соответственно.

Далее после сегментации текста на токены и морфологического анализа токенов-слов осуществляют выделение устойчивых словосочетаний с помощью общих и специальных словарей (ссылочная позиция 3 на фиг.1). В результате выполнения этого этапа исходный текст, кроме элементарных единиц первого и второго уровней, дополняется множеством единиц третьего уровня - устойчивыми словосочетаниями. Фрагмент этого множества для нашего примера представлен в Таблице 3.

После выполнения вышеуказанных этапов осуществляют фрагментацию обрабатываемого текста на предложения (ссылочная позиция 4 на фиг.1). В результате выполнения этого этапа сформированные выше множества дополняются множеством предложений, представленным в Таблице 4.

Таким образом, после выполнения всех рассмотренных выше этапов обрабатываемый текст будет сегментирован на предложения, каждое из которых размечено множествами аннотаций элементарных единиц первого, второго и третьего уровней.

Вслед за этим в соответствии с заявленным способом автоматизированной обработки текстов на естественном языке путем их семантической индексации выявление семантически значимых объектов (элементарных единиц четвертого уровня) производится в каждом предложении на множестве элементарных единиц первого, второго и (или) третьего уровней с помощью упомянутых правил. Так, например, в предложении «Молодой моряк Эдмон Дантес принял на себя командование.» рассматриваемого текста с помощью правила, спецификация которого представлена на фиг.3a, а соответствующая ей схема обработки сигналов - на фиг.3b, выделяется семантически значимый объект «Эдмон Дантес». Другие семантически значимые объекты выделяются с помощью правил, аналогичных представленному на фиг.3a, b. В результате выполнения этапов, обозначенных на фиг.1 ссылочными позициями 5-7-9, в исходном тексте выделяют элементарные единицы четвертого уровня (семантически значимые объекты с их атрибутами). Фрагмент множества таких единиц для рассматриваемого примера представлен в Таблице 5.

После этого в пределах всего обрабатываемого текста в процессе выполнения этапа, обозначенного на фиг.1 ссылочной позицией 11, находят местоимения, которые могут быть анафорическими ссылками на соответствующие семантически значимые объекты, и для местоимений, которые действительно таковыми являются, фиксируют тождество по референции между соответствующим семантически значимым объектом

и его анафорической ссылкой (элементарной единицей пятого уровня). Для рассматриваемого примера полученное множество кореференций анафорических ссылок представлено в Таблице 6.

После выполнения предыдущих этапов на множестве выделенных элементарных единиц первого, второго, третьего, четвертого и пятого уровней с помощью упомянутых правил находят семантически значимые отношения между семантически значимыми объектами. Так, например, в предложении «Владелец корабля Моррель предлагает Дантесу официально вступить в должность капитана корабля «Фараон»…» рассматриваемого текста с помощью правила, спецификация которого представлена на фиг.4a, а соответствующая схема обработки сигналов - на фиг.4b, выделяют именованное отношение «работать». В результате выполнения этапов, обозначенных на фиг.1 ссылочными позициями 6-8-10, в исходном тексте выделяют множество семантически значимых отношений между семантически значимыми объектами, фрагмент которого для нашего примера представлен в Таблице 7.

Таким образом, после выполнения всех рассмотренных выше этапов обработки исходный текст будет размечен множеством аннотаций, соответствующих семантически значимым объектам с их атрибутами и семантически значимым отношениям с их атрибутами между семантически значимьми объектами. Для нашего примера графическое представление фрагмента результата обработки текста показано на фиг.5.

Следующий этап, обозначенный на фиг.1 ссылочной позицией 12, является техническим и выполняется для формирования триад, соответствующих сохраненным семантически значимым объектам и семантически значимым отношениям. Фрагмент множества таких триад для нашего примера представлен в Таблице 8. По сути дела, сформированное множество триад составляет исходные данные для построения семантического индекса обработанного на предыдущих этапах текста.

На этапе, обозначенном на фиг.1 ссылочной позицией 13, строят семантический индекс следующим образом: сначала из множества триад, полученных на предыдущем этапе, формируют подмножества триад, каждое из которых соответствует одному семантически значимому объекту с его атрибутами, и каждое полученное подмножество триад используют как вход для одного из стандартных индексаторов, например широко известного свободно распространяемого индексатора Lucene, индексатора поисковой машины Яндекс, индексатора Google или любого другого индексатора, с выхода которого получают уникальный для заданного подмножества триад индекс. Аналогичную последовательность действий выполняют для всех подмножеств триад, соответствующих парам вида «семантически значимый объект - семантически значимое отношение» и триадам вида «семантически значимый объект - семантически значимое отношение - семантически значимый объект» с учетом атрибутов соответствующих семантически значимых объектов и/или семантически значимых отношений, получая множество соответствующих уникальных индексов, которые в совокупности и составляют семантический индекс текста. Фрагмент семантического индекса для рассматриваемого примера представлен в Таблицах 9-11.

На этапе, обозначенном на фиг.1 ссылочной позицией 15, сформированные на этапе 12 триады и полученные на этапе 13 индексы вместе со ссылкой на исходный текст, из которого сформированы эти триады, сохраняют в базе данных, а этап 14, в случае обработки одного текста, пропускают. Общая схема сохранения всех полученных на предыдущих этапах результатов представлена на фиг.6.

На фиг.6 в качестве первого этапа (51) формируют множество непрерывных

цепочек триад для отношения «The same». На следующем этапе (52) проверяют, является ли полученное на предыдущем этапе множество цепочек триад пустым. Если это множество не пустое, то последовательно на следующих этапах (53-56) формируют множество объектов для очередной цепочки (53), свертывают это множество объектов в единый объект (54), имеющий объединенное множество атрибутов (без повторений), сохраняют полученный единый объект с его атрибутами (55) и удаляют множество обработанных объектов очередной цепочки (56). Если же множество цепочек триад на этапе 52 оказывается пустым (изначально или в результате выполнения этапов 53-55), далее на этапе 57 формируют общее множество триад, полученных на всех предыдущих этапах, на этапе 58 дополняют сформированное общее множество триад семантическими индексами и ссылками на исходный текст, после чего на этапе 59 и сохраняют дополненное множество триад в базе данных.

В соответствии с заявленным способом автоматизированной обработки коллекции текстов на естественном языке путем их семантической индексации обработка каждого следующего текста, включая построение его семантического индекса, осуществляется путем выполнения в точности тех же этапов, что и для одного текста. Однако в этом случае после этапа 13 индексации и перед этапом 15 сохранения в базе данных осуществляют еще один этап, обозначенный на фиг.1 ссылочной позицией 14 - этап объединения результатов обработки очередного текста с результатами обработки предыдущих текстов, уже сохраненных в базе данных, который выполняют следующим образом.

Вновь выявленные в очередном индексируемом тексте семантически значимые объекты и семантически значимые отношения сравнивают с уже имеющимися в базе данных семантически значимыми объектами и семантически значимыми отношениями путем проверки совпадения их семантических индексов и, в случае положительного результата такого сравнения, соответствующие объекты и отношения из дальнейшей обработки исключают, сохраняя при этом в уже присутствующем в базе данных объекте и/или отношении ссылку на тот текст и тот фрагмент этого текста, в котором выявлены объекты и/или отношения, исключенные из дальнейшей обработки. В случае отрицательного результата сравнения семантических индексов с помощью заранее сформированных в базе данных лингвистических и эвристических правил выявляют подобие между новыми объектами и/или отношениями и теми объектами и/или отношениями, которые уже присутствуют в базе данных и, в случае положительного результата, расширяют уже существующие в базе данных описания объектов и/или отношений новыми данными, после чего перестраивают соответствующие семантические индексы и добавляют новые семантические индексы в качестве вторичных к уже существующим и, кроме того, сохраняют в уже присутствующем в базе данных объекте и/или отношении ссылку на тот текст и тот фрагмент этого текста, в котором выявлены объекты и/или отношения, после чего соответствующие объекты и отношения из дальнейшей обработки исключают. В противном случае вновь выявленные именованные объекты и именованные отношения с их семантическими индексами добавляют в базу данных.

Так, например, если в качестве следующего, по отношению к уже рассмотренному примеру, обрабатывался текст «Дантес приговаривается к пожизненному заточению в замке Иф, политической тюрьме среди моря, неподалеку от Марселя… За долгие годы отсутствия Дантеса в судьбах тех, кто был повинен в его страданиях, тоже произошли значительные перемены. Данглар - богатый банкир. Де Вильфор - королевский

прокурор… В банке Данглара граф Монте-Кристо открывает «неограниченный кредит». Данглар ставит под сомнение финансовые возможности графа. Граф иронизирует: «Для вас - может быть, но не для меня». «Моей кассы еще никто не считал!» - уязвлен Данглар. «В таком случае я - первый, кому это предстоит», - обещает ему Монте-Кристо. …Прежде чем испустить дух, аббат Бузони сообщает, что Монте-Кристо и Эдмон Дантес - одно лицо…», то после выполнения этапов 1-13 в нем будут выявлены, например, такие семантически значимые объекты и отношения между ними, как «замок Иф», «Марсель», «Дантес», «Де Вильфор», «Монте-Кристо», «высказывания_заявления» и др., а также будет сформирован его семантический индекс, фрагмент которого представлен в Таблицах 12-14.

Далее в соответствии с заявленным способом автоматизированной обработки коллекций текстов путем их семантической индексации на этапе 14 будет, в частности, выявлен объект «Марсель», семантический индекс которого полностью совпадает с семантическим индексом объекта «Марсель», уже присутствующего в базе данных, и, кроме того, будет выявлено (путем применения правила, спецификация которого приведена на фиг.7a, а соответствующая схема обработки сигналов - на фиг.7b) подобие между новым объектом «Эдмон Дантес» и объектом «Дантес», а также тождество по референции между объектом «Эдмон Дантес» и объектом «Монте-Кристо», уже присутствующими в базе данных, после чего существующее описание объекта «Эдмон Дантес» в базе данных будет расширено за счет новой информации и дополнительного семантического индекса, что показано в Таблицах 15 и 16.

Таким образом, настоящее изобретение обеспечивает расширение арсенала способов индексации текстов на естественных языках за счет использования методов их автоматизированного лингвистического анализа и последующего использования его результатов для построения семантических индексов, основное отличие которых от известных способов индексации в том, что индексируются не ключевые слова и словосочетания, а семантически значимые понятия и отношения между ними, что обеспечивает семантическую навигацию по документам и коллекциям документов, а также высокоточный и быстрый поиск релевантных информационным потребностям пользователя фактов и документов, особенно в применении к текстам на высоко флективных языках.

Таблица 1. Результаты токенизации текста примера

Тип элемента	Нач. позиция	Кон. позиция	Атрибуты
Token	0	2	{тип=number, длина=2, строка=27}
Token	3	10	{тип=word, кодировка=cyr, длина=7, орфо=lowercase, строка=февраля}
Token	11	15	{тип=number, длина=4, строка=1815}
Token	16	17	{тип=word, кодировка=cyr, длина=1, орфо=lowercase, строка=г}
Token	17	18	{тип=punctuation, длина=1, строка=.}
Token	19	20	{тип=word, кодировка=cyr, длина=1, орфо=lowercase, строка=в}
Token	21	28	{тип=word, кодировка=cyr, длина=7, орфо=upperInitial, строка=Марсель}
.....			
Token	190	197	{тип=word, кодировка=cyr, длина=7, орфо=upperInitial, строка=Молодой}
Token	198	203	{тип=word, кодировка=cyr, длина=5, орфо=lowercase, строка=моряк}
Token	204	209	{тип=word, кодировка=cyr, длина=5, орфо=upperInitial, строка=Эдмон}
Token	210	216	{тип=word, кодировка=cyr, длина=6, орфо=upperInitial, строка=Дантес}
Token	217	223	{тип=word, кодировка=cyr, длина=6, орфо=lowercase, строка=принял}
Token	224	226	{тип=word, кодировка=cyr, длина=2, орфо=lowercase, строка=на}
Token	227	231	{тип=word, кодировка=cyr, длина=4, орфо=lowercase, строка=себя}
Token	232	244	{тип=word, кодировка=cyr, длина=12, орфо=lowercase, строка=командование}
Token	244	245	{тип=punctuation, длина=1, строка=.}
Token	246	254	{тип=word, кодировка=cyr, длина=8, орфо=upperInitial, строка=Владелец}
Token	255	262	{тип=word, кодировка=cyr, длина=7, орфо=lowercase, строка=корабля}
Token	263	270	{тип=word, кодировка=cyr, длина=7, орфо=upperInitial, строка=Моррель}
Token	271	281	{тип=word, кодировка=cyr, длина=10, орфо=lowercase, строка=предлагает}
Token	282	289	{тип=word, кодировка=cyr, длина=7, орфо=upperInitial, строка=Дантесу}
Token	290	300	{тип=word, кодировка=cyr, длина=10, орфо=lowercase, строка=официально}
Token	301	309	{тип=word, кодировка=cyr, длина=8, орфо=lowercase, строка=вступить}
Token	310	311	{тип=word, кодировка=cyr, длина=1, орфо=lowercase, строка=в}
Token	312	321	{тип=word, кодировка=cyr, длина=9, орфо=lowercase,

			строка=должность}
Token	322	330	{тип=word, кодировка=cyr, длина=8, орфо=lowercase, строка=капитана}
Token	331	338	{тип=word, кодировка=cyr, длина=7, орфо=lowercase, строка=корабля}
Token	339	340	{тип=punctuation, длина=1, position=startpunct, строка=«}
Token	340	346	{тип=word, кодировка=cyr, длина=6, орфо=upperInitial, строка=Фараон}
Token	346	347	{тип=punctuation, длина=1, position=endpunct, строка=»}
Token	347	348	{тип=punctuation, длина=1, строка=.}
Token	348	349	{тип=punctuation, длина=1, строка=.}
Token	349	350	{тип=punctuation, длина=1, строка=.}

Таблица 2. Результаты морфологического анализа текста примера

Тип элемента	Нач. позиция	Кон. позиция	Атрибуты
Morph	0	2	{CAS=prp, POS=NUM, base=27}
Morph	0	2	{CAS=nom, POS=NUM, base=27}
Morph	3	10	{CAS=gen, GEND=m, NMB=sg, POS=N, base=февраль}
Morph	11	15	{CAS=gen, POS=NUM, base=1815}
Morph	11	15	{CAS=nom, POS=NUM, base=1815}
Morph	16	18	{CAS=nom, GEND=m, NMB=sg, POS=N, base=г}
Morph	19	20	{GCAS=prp, POS=PREP, base=в}
Morph	19	20	{GCAS=acc, POS=PREP, base=в}
Morph	21	28	{CAS=nom, GEND=m, NMB=sg, POS=N, base=марсель}
Morph	21	28	{CAS=acc, GEND=f, NMB=sg, POS=N, base=Марсель}
Morph	21	28	{CAS=nom, GEND=m, NMB=sg, POS=N, base=Марсель}
Morph	21	28	{CAS=prp, GEND=f, NMB=sg, POS=N, base=Марсель}
.....			
Morph	29	31	{GCAS=gen, POS=PREP, base=из}
Morph	32	42	{CAS=acc, GEND=m, NMB=sg, POS=A, base=очередной}
Morph	32	42	{CAS=gen, GEND=m, NMB=sg, POS=A, base=очередной}
Morph	43	51	{CAS=acc, GEND=n, NMB=pl, POS=N, base=плавание}
Morph	43	51	{CAS=gen, GEND=n, NMB=sg, POS=N, base=плавание}
Morph	43	51	{CAS=nom, GEND=n, NMB=pl, POS=N, base=плавание}
Morph	52	64	{NMB=sg, POS=V, PRS=3, REPR=fin, TNS=pres, base=возвращаться}
Morph	52	64	{MD=ind, NMB=sg, POS=V, PRS=3, REPR=fin, TNS=pres, TRANS=vt, VOX=pass, base=возвращать}
Morph	65	77	{CAS=nom, GEND=m, NMB=sg, POS=A, base=трехмачтовый}
Morph	65	77	{CAS=acc, GEND=m, NMB=sg, POS=A, base=трехмачтовый}
Morph	78	85	{CAS=nom, GEND=m, NMB=sg, POS=N, base=корабль}
Morph	78	85	{CAS=acc, GEND=m, NMB=sg, POS=N, base=корабль}
Morph	87	93	{CAS=acc, GEND=m, NMB=sg, POS=N, base=фараон}
Morph	87	93	{CAS=nom, GEND=m, NMB=sg, POS=N, base=фараон}
Morph	96	104	{CAS=dat, GEND=m, NMB=sg, POS=N, base=капитан}
Morph	105	112	{AGGROTYPE=unknown, string=Леклеру}
Morph	113	115	{POS=PCL, base=не}

Morph	116	123	{AGGROTYPE=unknown, string=суждено}
Morph	124	128	{GEND=n, NMB=sg, POS=V, REPR=fin, TNS=past, base=быть}
Morph	129	136	{POS=V, REPR=inf, base=ступить}
Morph	137	139	{GCAS=prp, POS=PREP, base=на}
Morph	140	146	{CAS=acc, GEND=f, NMB=sg, POS=A, base=родной}
Morph	147	152	{CAS=acc, GEND=f, NMB=sg, POS=N, base=земля}
Morph	154	156	{AGGROTYPE=unknown, string=он}
Morph	157	161	{GEND=m, NMB=sg, POS=V, REPR=fin, TNS=past, base=умереть}
Morph	162	164	{GCAS=gen, POS=PREP, base=от}
Morph	165	172	{CAS=gen, GEND=f, NMB=sg, POS=N, base=горячка}
Morph	165	172	{CAS=nom, GEND=f, NMB=pl, POS=N, base=горячка}
Morph	173	174	{GCAS=acc, POS=PREP, base=в}
Morph	173	174	{GCAS=prp, POS=PREP, base=в}
Morph	175	183	{CAS=prp, GEND=m, NMB=sg, POS=V, REPR=part, TNS=past, base=открыть}
Morph	175	183	{CAS=prp, GEND=m, NMB=sg, POS=A, base=открытый}
Morph	184	188	{CAS=acc, GEND=n, NMB=sg, POS=N, base=море}
Morph	184	188	{CAS=nom, GEND=n, NMB=sg, POS=N, base=море}
Morph	184	188	{CAS=prp, GEND=n, NMB=sg, POS=N, base=море}

.....

Morph	190	197	{CAS=nom, GEND=m, NMB=sg, POS=A, base=молодой}
Morph	190	197	{CAS=prp, GEND=f, NMB=sg, POS=N, base=молодая}
Morph	198	203	{CAS=nom, GEND=m, NMB=sg, POS=N, base=моряк}
Morph	204	209	{CAS=nom, GEND=m, NMB=sg, POS=N, base=Эдмон}
Morph	210	216	{CAS=nom, GEND=m, NMB=sg, POS=N, base=Дантес}
Morph	217	223	{GEND=m, NMB=sg, POS=V, REPR=fin, TNS=past, base=принять}
Morph	224	226	{GCAS=prp, POS=PREP, base=на}
Morph	224	226	{GCAS=acc, POS=PREP, base=на}
Morph	227	231	{CAS=acc, GEND=mfn, NMB=sg, POS=N, base=себя}
Morph	227	231	{CAS=gen, GEND=mfn, NMB=sg, POS=N, base=себя}
Morph	232	244	{CAS=acc, GEND=n, NMB=sg, POS=N, base=командование}
Morph	232	244	{CAS=nom, GEND=n, NMB=sg, POS=N, base=командование}
Morph	246	254	{CAS=nom, GEND=m, NMB=sg, POS=N, base=владелец}
Morph	255	262	{CAS=gen, GEND=m, NMB=sg, POS=N, base=корабль}
Morph	263	270	{AGGROTYPE=unknown, string=Моррель}
Morph	271	281	{NMB=sg, POS=V, PRS=3, REPR=fin, TNS=pres, base=предлагать}
Morph	282	289	{CAS=dat, GEND=m, NMB=sg, POS=N, base=Дантес}
Morph	290	300	{GEND=n, NMB=sg, POS=A, base=официальный}
Morph	301	309	{POS=V, REPR=inf, base=вступить}
Morph	310	311	{GCAS=prp, POS=PREP, base=в}

Morph	310	311	{GCAS=acc, POS=PREP, base=в}
Morph	312	321	{CAS=acc, GEND=f, NMB=sg, POS=N, base=должность}
Morph	312	321	{CAS=nom, GEND=f, NMB=sg, POS=N, base=должность}
Morph	322	330	{CAS=gen, GEND=m, NMB=sg, POS=N, base=капитан}
Morph	322	330	{CAS=acc, GEND=m, NMB=sg, POS=N, base=капитан}
Morph	331	338	{CAS=gen, GEND=m, NMB=sg, POS=N, base=корабль}
Morph	340	346	{CAS=nom, GEND=m, NMB=sg, POS=N, base=фараон}

Таблица 3. Результаты выявления в тексте устойчивых словосочетаний

Тип элемента	Нач. позиция	Кон. позиция	Атрибуты
Lookup	3	10	{CAS=gen, GEND=m, NMB=sg, POS=N, base=февраль, majorType=time, minorType=month}
Lookup	16	18	{CAS=dat, GEND=m, NMB=sg, POS=N, base=г, majorType=locKey, minorType=inhab}
Lookup	16	18	{CAS=acc, GEND=m, NMB=sg, POS=N, base=г., majorType=title}
Lookup	16	18	{CAS=nom, GEND=m, NMB=pl, POS=N, base=г., majorType=locKey, minorType=inhab}
Lookup	16	18	{CAS=nom, GEND=m, NMB=pl, POS=N, base=г., majorType=title}
Lookup	21	28	{CAS=nom, GEND=m, NMB=sg, POS=N, base=Марсель, majorType=location6 minorType=inhab}
.....			
Lookup	96	104	{CAS=dat, GEND=m, NMB=sg, POS=N, base=капитан, majorType=jbt_military}
Lookup	124	128	{GEND=n, NMB=sg, POS=V, REPR=fin, TNS=past, base=быть, majorType=ToBe_relKey}
Lookup	124	128	{GEND=n, NMB=sg, POS=V, REPR=fin, TNS=past, base=быть, majorType=Function_Verb, minorType=PersPers}
Lookup	140	146	{CAS=acc, GEND=f, NMB=sg, POS=A, base=родной, majorType=PersOrg_relKey, minorType=relative_attrib}
Lookup	147	152	{CAS=acc, GEND=f, NMB=sg, POS=N, base=Земля, majorType=location}
Lookup	157	161	{GEND=m, NMB=sg, POS=V, REPR=fin, TNS=past, base=умереть, majorType=bio}
.....			
Lookup	184	188	{CAS=acc, GEND=n, POS=N, base=море, majorType=locKey, minorType=water}
Lookup	184	188	{CAS=nom, GEND=n, NMB=sg, POS=N, base=море, majorType=locKey, minorType=water}
Lookup	217	223	{GEND=m, NMB=sg, POS=V, REPR=fin, TNS=past, TRANS=vt, base=принять, majorType=meetwith}
Lookup	246	254	{CAS=nom, GEND=m, NMB=sg, POS=N, base=владелец, majorType=jobTitle}
Lookup	246	254	{CAS=nom, GEND=m, NMB=sg, POS=N, base=владелец, majorType=PersOrg_relKey, minorType=Own}
Lookup	322	330	{CAS=gen, NMB=sg, POS=N, base=капитан, majorType=jbt_military}

Таблица 4. Результаты сегментации текста на предложения

Тип элемента	Нач. позиция	Кон. позиция	Атрибуты
Sentence	0	95	{}
Sentence	96	189	{}
Sentence	190	245	{}
Sentence	246	350	{}

Таблица 5. Результаты выявления в тексте именованных сущностей

Тип элемента	Нач. позиция	Кон. позиция	Атрибуты
Date	0	18	{$view_name$=27 февраля 1815 г., day=27, month=февраля, year=1815}
Location	21	28	{ COUNTRY=Франция, base=Марсель, minor-Type=town }
PrivStructures	78	94	{$view_name$= корабль «Фараон»}
JobTitle	96	104	{base=капитан}
Person	105	112	{FAMIL=Леклер}
Person	154	156	{$view_name$=он}
.....			
Person	204	216	{FAMIL=Дантес, FNAME=Эдмон}
JobTitle	246	254	{base=владелец}
Person	263	270	{ FAMIL=Моррель }
Person	282	289	{FAMIL=Дантес}
JobTitle	322	330	{base=капитан}
PrivStructures	331	347	{$view_name$= корабль «Фараон»}

Таблица 6. Результаты выявления в тексте кореференций

Тип отношения	Атрибуты
TheSame	{from=3013 (Эдмон Дантес), to=3030 (Дантес)}
TheSame	{from=3644 (корабль «Фараон»), to=3645 (корабль «Фараон»)}
TheSame	{from=3646 (он), to=3004 (Леклер)}

Таблица 7. Результаты выявления в тексте семантически значимых отношений

Тип отношения	Атрибуты
BeEmployeeOf	{имя=работает, должность=капитан, от=3030 (Дантес), к=3645 (корабль «Фараон»)}
.....	
BeOwnerOf	{имя=владеет, от=3647 (Моррель), к=3645 (корабль «Фараон»)}

Таблица 8. Триадное представление результатов обработки текста

Объект	Атрибут	Значение
Марсель	is_a	Местоположение
корабль «Фараон»	is_a	Частная_Структура
Леклер	is_a	Персона
Леклер	Пол	m
Леклер	Статус_Роль	капитан
Леклер	Фамилия	Леклер
Эдмон Дантес	is_a	Персона
Эдмон Дантес	Пол	m
Эдмон Дантес	Статус_Роль	капитан
Эдмон Дантес	Имя	Эдмон
Эдмон Дантес	Инициал	Э.
работает	is_a	Relation
работает	от	3030 (Дантес)
работает	к	3645 (корабль «Фараон»)
владеет	is_a	Relation
владеет	от	3647 (Моррель)
владеет	к	3645 (корабль «Фараон»)

Таблица 9. Фрагмент семантического индекса текста (семантически значимые объекты)

Сущность (объект с атрибутами)	Семантический индекс (уникальный код)
Персона Леклер *Пол* m *Статус_Роль* капитан *Фамилия* Леклер	a_2a7aebf65ce54a03_b14a4639a641808d
Персона Эдмон Дантес *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* Дантес *Статус_Роль* капитан	a_7986d80e781ce95b_253defa0906d86a3
Персона Моррель *Пол* m *Фамилия* Моррель	a_7986d80e781ce95b_253defa0906d93a
Частная_Структура корабль «Фараон»	a_2d0a16bffbc7dd8f_d77ae55aab932dc5
Местоположение Марсель	a_d86dbb670872b609_78aed5dc39a8e2b5

Таблица 10. Фрагмент семантического индекса текста
(пары вида «семантически значимый объект – семантически значимое отношение»)

Пара (объект – отношение)	Семантический индекс (уникальный код)
.....	
Персона Эдмон Дантес *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* Дантес *Статус_Роль* капитан **Отношение** работает	b_348eda6c3d9a407_3ebd0c4b59f6bdd0
Персона Моррель *Пол* m *Фамилия* Моррель **Отношение** владеет	b_4d188f5919b61bd6_a7d4a195ada52738
.....	

Таблица 11. Фрагмент семантического индекса текста (триады вида «семантически значимый объект – семантически значимое отношение – семантически значимый объект»)

Триада (объект – отношение – объект)	Семантический индекс (уникальный код)
.....	
Персона Эдмон Дантес *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* Дантес *Статус_Роль* капитан **Отношение** работает **Частная_Структура** корабль «Фараон»	c_348eda6c3d9a407_253defa0906d86a3
Персона Моррель *Пол* m *Фамилия* Моррель **Отношение** владеет **Частная_Структура** корабль «Фараон»	c_4d188f5919b61bd6_86ad15ca5e326c5b
.....	

Таблица 12. Фрагмент семантического индекса нового текста (именованные сущности)

Сущность (объект с атрибутами)	Семантический индекс (уникальный код)
.....	
Местоположение замок Иф	a_2d0a16bffbc7dd8f_d77ae55aab932dc5
Местоположение Марсель	a_d86dbb670872b609_78aed5dc39a8e2b5
Персона Дантес *Пол* m *Фамилия* Дантес	a_7cd1659a3047c3bc_6843f8d1284fdcf0
Персона Данглар *Пол* m *Фамилия* Данглар	a_7cd1659a3047c3bc_6843f8d1284fdce0
Персона Монте-Кристо *Пол* m *Фамилия* Монте-Кристо	a_b5b8f5d8c84d5db1_1fec7102e0d017ab
Персона Эдмон Дантес *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* Дантес	a_7986d80e781ce95b_253defa0906d86a3
Высказывание Моей кассы еще никто не считал!	a_7986d80e781ce95b_253defa0906d86ac
Высказывание В таком случае я — первый, кому это предстоит	a_7986d80e781ce95b_253defa0906d86ad
.....	

Таблица 13. Фрагмент семантического индекса нового текста (пары вида «именованная сущность – именованное отношение»)

Пара (объект – отношение)	Семантический индекс (уникальный код)
.....	
Персона Монте-Кристо *Пол* m *Фамилия* Монте-Кристо **Отношение** высказывания_заявления	b_1fec7102e0d017ab_ b5b8f5d8c84d5db1
Персона Данглар *Пол* m *Фамилия* Данглар **Отношение** высказывания_заявления	b_b5b8f5d8c84d5db1_1fec7102e0d017ab
.....	

Таблица 14. Фрагмент семантического индекса нового текста (триады вида «именованная сущность – именованное отношение – именованная сущность»)

Триада (объект – отношение – объект)	Семантический индекс (уникальный код)
.....	
Персона Монте-Кристо *Пол* m *Фамилия* Монте-Кристо **Отношение** высказывания_заявления **Высказывание** В таком случае я — первый, кому это предстоит	c_7986d80e781ce95b_ b5b8f5d8c84d5db1
Персона Данглар *Пол* m *Фамилия* Данглар **Отношение** высказывания_заявления Моей кассы еще никто не считал!	c_b5b8f5d8c84d5db1_7986d80e781ce95b
.....	

Таблица 15. Фрагмент семантического индекса именованных сущностей для коллекции из двух текстов до объединения подобных объектов

Сущность (объект с атрибутами)	Семантический индекс (уникальный код)
.....	
Персона Дантес *Пол* m *Фамилия* Дантес	a_7cd1659a3047c3bc_6843f8d1284fdcf0
Персона Эдмон Дантес *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* Дантес	a_7986d80e781ce95b_253defa0906d86a3
Персона Монте-Кристо *Пол* m *Фамилия* Монте-Кристо	a_b5b8f5d8c84d5db1_1fec7102e0d017ab
.....	

Таблица 16. Фрагмент семантического индекса именованных сущностей для коллекции из двух текстов после объединения подобных объектов

Сущность (объект с атрибутами)	Семантический индекс (уникальный код)
.....	
Персона [Эдмон Дантес, Дантес, Монте-Кристо] *Имя* Эдмон *Инициал* Э. *Пол* m *Фамилия* [Дантес, Монте-Кристо]	a_7986d80e781ce95b_253defa0906d86a3, a_7cd1659a3047c3bc_6843f8d1284fdcf0, a_b5b8f5d8c84d5db1_1fec7102e0d017ab
.....	

Формула изобретения

1. Способ автоматизированной обработки текста на естественном языке путем его семантической индексации, содержащий этапы, на которых:

представляют индексируемый текст в электронной форме для последующей автоматической и/или автоматизированной обработки;

сегментируют текст в электронной форме на элементарные единицы первого уровня, которые выбирают из группы, состоящей из слов в виде последовательностей букв или букв и дефисов, чисел, знаков препинания и последовательностей пробелов;

для каждой элементарной единицы текста, представляющей собой слово, на основе

морфологического анализа формируют элементарные единицы второго уровня, включающие для заданного слова нормализованную словоформу и его морфологические характеристики;

на множестве полученных элементарных единиц первых двух уровней в процессе лингвистического анализа выявляют в тексте устойчивые словосочетания, которые являются элементарными единицами третьего уровня;

формируют предложения, соответствующие участкам индексируемого текста, каждое их которых размечено множествами аннотаций элементарных единиц первого, второго и третьего уровней;

выполняют многоступенчатый семантико-синтаксический анализ путем обращения к сформированным в базе данных лингвистическим и эвристическим правилам в заранее заданной лингвистической среде, в процессе которого выявляют в каждом предложении с выявленными словосочетаниями на множестве элементарных единиц первого, второго и/или третьего уровней семантически значимые объекты, которые являются элементарными единицами четвертого уровня, фиксируют тождество по референции между соответствующим семантически значимым объектом и его анафорической ссылкой, которая является элементарной единицей пятого уровня, после чего на множестве выделенных элементарных единиц первого, второго, третьего, четвертого и пятого уровней с помощью указанных правил определяют семантически значимые отношения между семантически значимыми объектами, являющимися элементарными единицами шестого уровня;

формируют в пределах индексируемого текста для каждого из выявленных семантически значимых отношений, связывающих соответствующие семантически значимые объекты, множество триад, причем единственная триада первого типа соответствует связи, устанавливаемой семантически значимым отношением между двумя семантически значимыми объектами, каждая из триад второго типа соответствует значению конкретного атрибута одного из этих объектов, а каждая из триад третьего типа соответствует значению конкретного атрибута самого семантически значимого отношения;

индексируют на множестве сформированных триад все связанные семантически значимыми отношениями семантически значимые объекты по отдельности, все пары вида "семантически значимый объект - семантически значимое отношение" и все триады вида "семантически значимый объект - семантически значимое отношение - семантически значимый объект" с учетом атрибутов соответствующих семантически значимых объектов и/или семантически значимых отношений;

сохраняют в базе данных сформированные триады и полученные индексы вместе со ссылкой на исходный текст, из которого сформированы эти триады.

2. Способ по п.1, в котором в процессе упомянутого лингвистического анализа при формировании словосочетаний преобразуют в каждом предложении последовательности элементарных единиц первого и/или второго уровней с помощью обращения к сохраненным в базе данных словарям и морфологическим связям в упомянутые словосочетания - элементарные единицы третьего уровня.

3. Способ по п.1, в котором в процессе упомянутого многоступенчатого семантико-синтаксического анализа выполняют этапы, на которых формируют с помощью упомянутых правил для каждого семантически значимого объекта семантические атрибуты из атрибутов элементарных единиц второго и/или третьего уровней, составляющих данный семантически значимый объект; присваивают каждому семантически значимому объекту соответствующий тип из хранящейся в базе данных

предметной онтологии по тематике каждой предметной области, к которой относится индексируемый текст; сохраняют в памяти каждый семантически значимый объект вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами.

4. Способ по п.1, в котором определяют с помощью упомянутых правил для каждого семантически значимого отношения морфологические атрибуты из составляющих данное семантически значимое отношение элементарных единиц второго уровня; находят с помощью упомянутых правил для каждого семантически значимого отношения семантические атрибуты из элементарных единиц первого, второго, третьего и/или четвертого уровней; присваивают каждому семантически значимому отношению соответствующий тип из хранящейся в базе данных предметной онтологии по тематике той предметной области, к которой относится индексируемый текст; сохраняют в памяти каждое семантически значимое отношение вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами.

5. Способ по п.1, в котором перед сохранением в базе данных сформированных триад и полученных индексов осуществляют свертку каждой группы объектов, связанных отношениями тождества по референции, в единый объект, множество атрибутов которого является объединением атрибутов объектов данной группы, связанных отношениями тождества по референции.

6. Способ автоматизированной обработки коллекции текстов на естественном языке путем их семантической индексации, содержащий все этапы способа по п.1 в применении к очередному индексируемому тексту, после чего при запоминании в базе данных сформированных триад и полученных индексов очередного текста осуществляют сравнение с помощью сформированных в базе данных лингвистических и эвристических правил в заранее заданной лингвистической среде вновь выявленных семантически значимых объектов и семантически значимых отношений с уже имеющимися в базе данных семантически значимыми объектами и семантически значимыми отношениями и в случае идентификации одинаковых объектов и/или отношений дублирующую информацию в базе данных не запоминают, а к соответствующим семантически значимым объектам и/или семантически значимым отношениям добавляют ссылки на текстовые фрагменты в пределах каждого из очередных текстов, из которых они выделены.

7. Способ по п.6, в котором в процессе упомянутого лингвистического анализа при формировании словосочетаний преобразуют в каждом предложении последовательности элементарных единиц первого и/или второго уровней с помощью обращения к сохраненным в базе данных словарям и морфологическим связям в упомянутые словосочетания - элементарные единицы третьего уровня.

8. Способ по п.6, в котором в процессе упомянутого многоступенчатого семантико-синтаксического анализа выполняют этапы, на которых присваивают каждому семантически значимому объекту соответствующий тип из хранящейся в базе данных предметной онтологии по тематике каждой предметной области, к которой относится индексируемый текст; сохраняют в памяти каждый семантически значимый объект вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами.

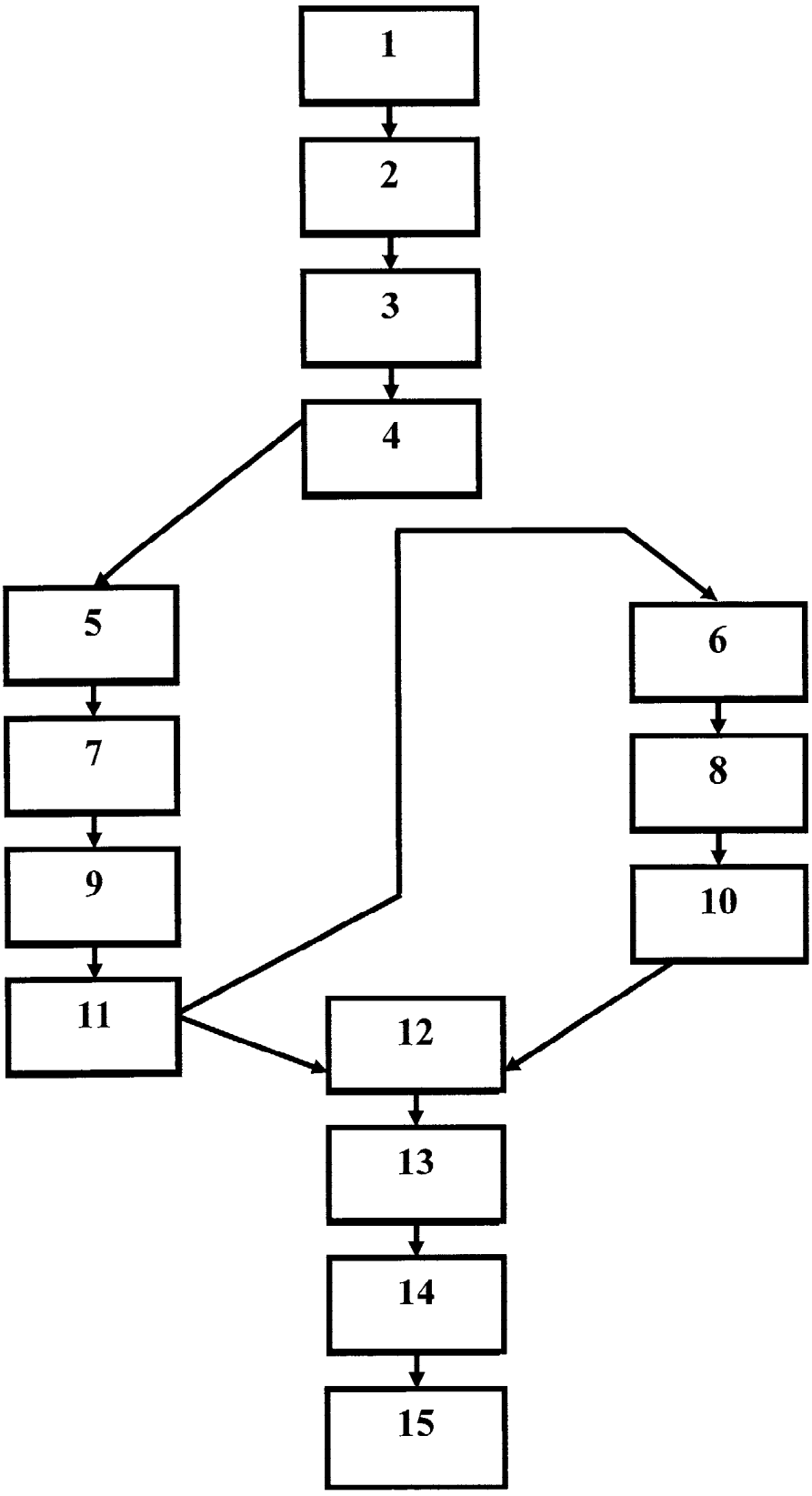
9. Способ по п.6, в котором с помощью упомянутых правил находят для каждого семантически значимого отношения морфологические атрибуты из составляющих данное семантически значимое отношение элементарных единиц второго уровня и

семантические атрибуты из элементарных единиц первого, второго, третьего и/или четвертого уровней; присваивают каждому семантически значимому отношению соответствующий тип из хранящейся в базе данных предметной онтологии по тематике той предметной области, к которой относится индексируемый текст; сохраняют в памяти каждое семантически значимое отношение вместе с присвоенным ему типом и найденными для него морфологическими и семантическими атрибутами.

10. Способ по п.6, в котором перед сохранением в базе данных сформированных триад и полученных индексов осуществляют свертку каждой группы объектов, связанных отношениями тождества по референции, в единый объект, множество атрибутов которого является объединением атрибутов объектов данной группы, связанных отношениями тождества по референции.

11. Машиночитаемый носитель, предназначенный для непосредственного участия в работе компьютера и содержащий программу для осуществления способа по п.1.

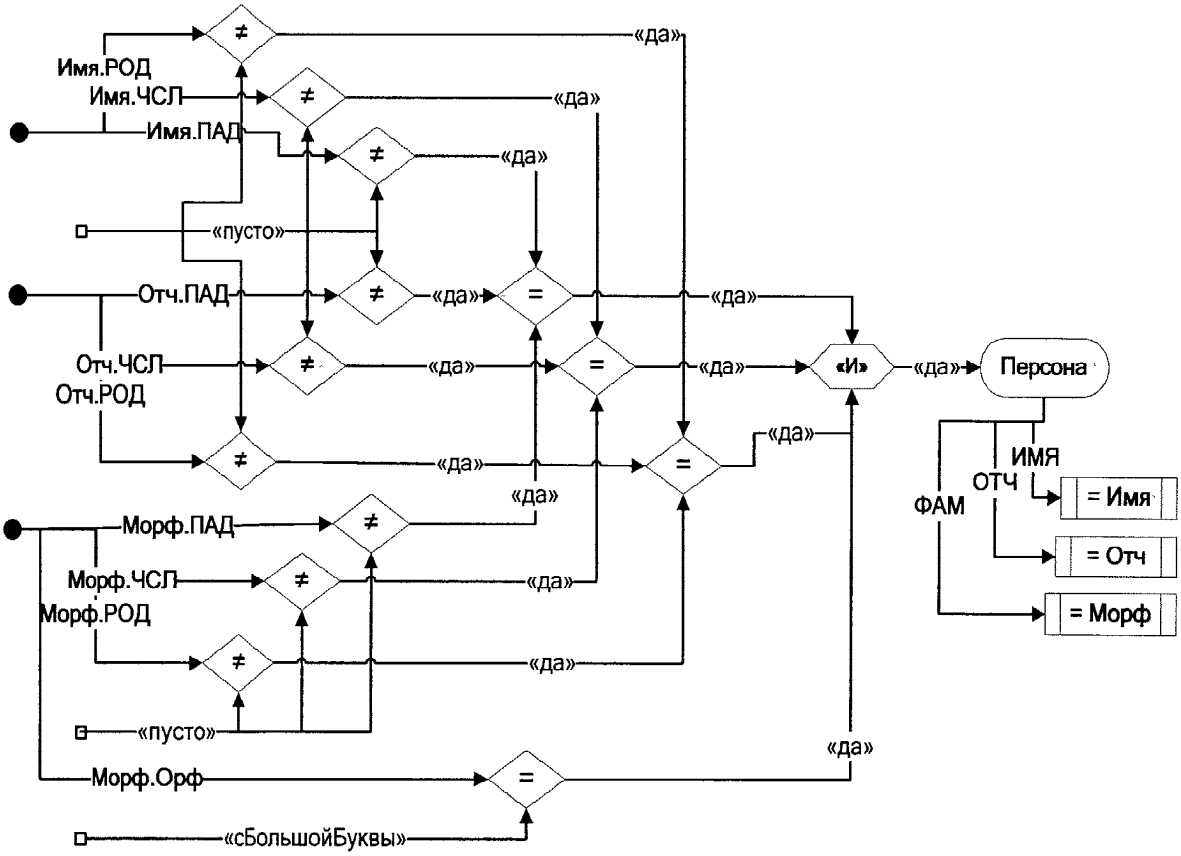
12. Машиночитаемый носитель, предназначенный для непосредственного участия в работе компьютера и содержащий программу для осуществления способа по п.6.



Фиг. 1

```
(
   {(Имя.РОД != null):a, (Имя.ЧСЛ != null):b, (Имя.ПАД != null):c}):x +
   (({Отч.РОД == :a, Отч.ЧСЛ == :b, Отч.ПАД == :c }):y)? +
   ({Морф.РОД == :a, Морф.ЧСЛ == :b, Морф.ПАД == :c, Морф.Орф == Upperinitial
}):z
) -->
    Персона = {ИМЯ= x, ОТЧ= y, ФАМ=z}
```

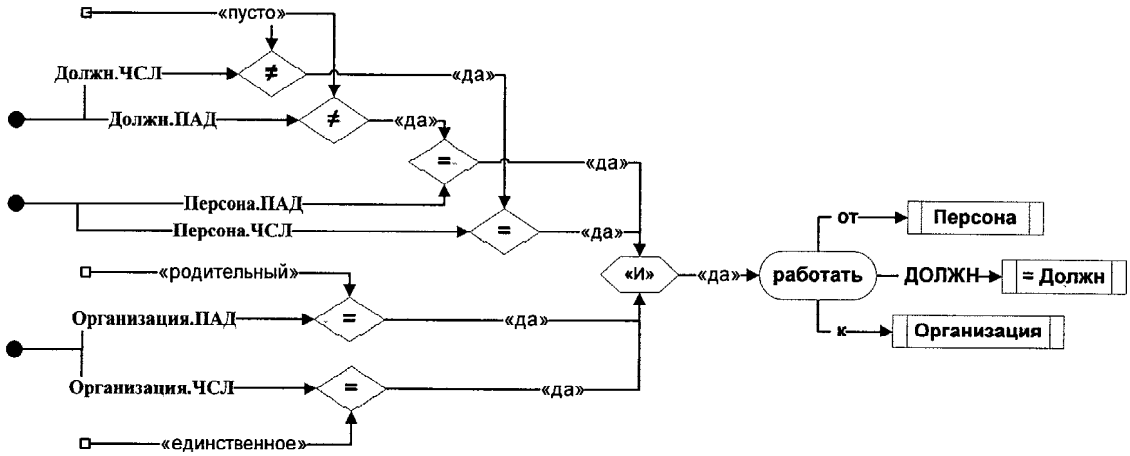
Фиг. 3a



Фиг. 3b

```
(
   {(Должн.ЧСЛ != null):a, (Должн.ПАД != null):b }):x +
   ({Организация.ЧСЛ == sg, Организация.ПАД == gen }):y +
   ({Персона.ЧСЛ == :a, Персона.ПАД == :b }):z
) -->
    работать = {от = z, к = y, ДОЛЖН = x}
```

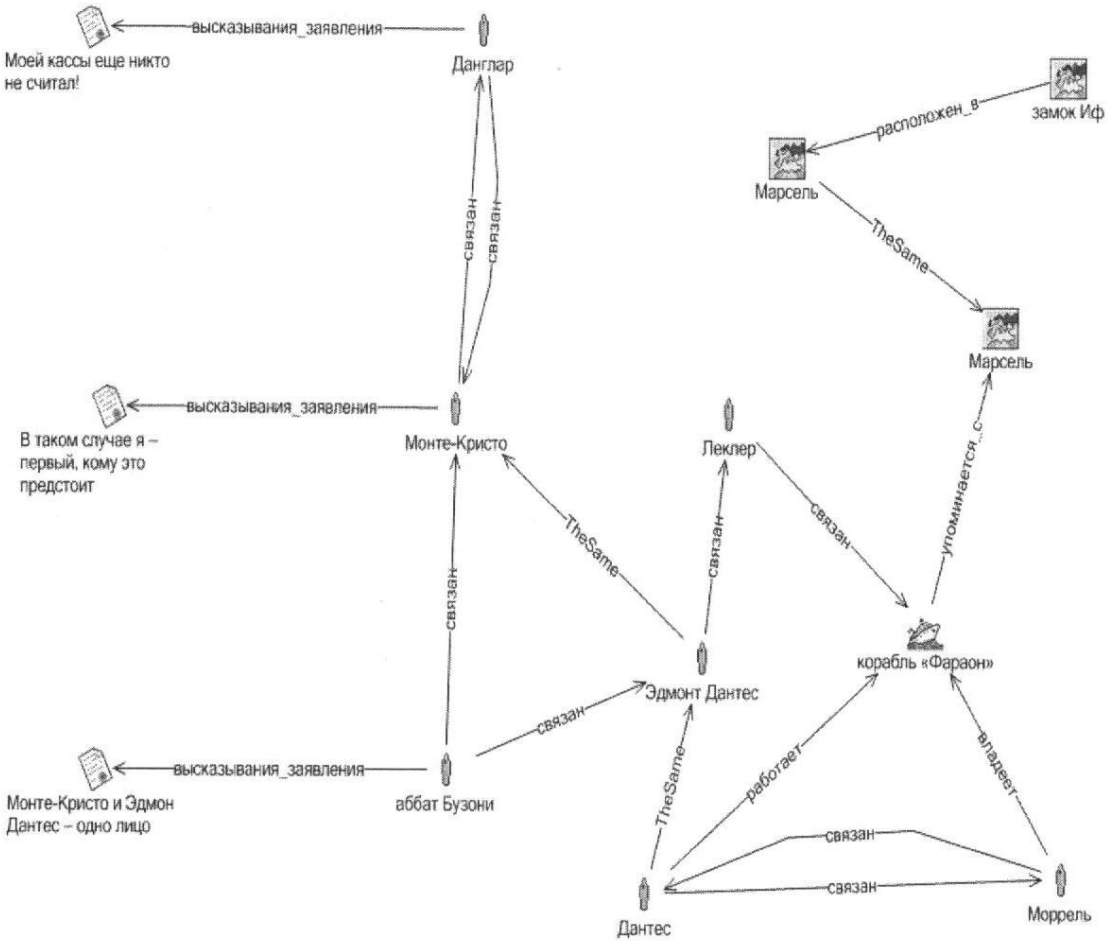
Фиг. 4a


«пусто»
Должн.ЧСЛ
≠
«да»
Должн.ПАД
≠
«да»
=
«да»
Персона.ПАД
Персона.ЧСЛ
=
«да»
«родительный»
Организация.ПАД
=
«да»
Организация.ЧСЛ
=
«да»
«единственное»
«И»
«да»
работать
от
Персона
ДОЛЖН
= Должн
к
Организация

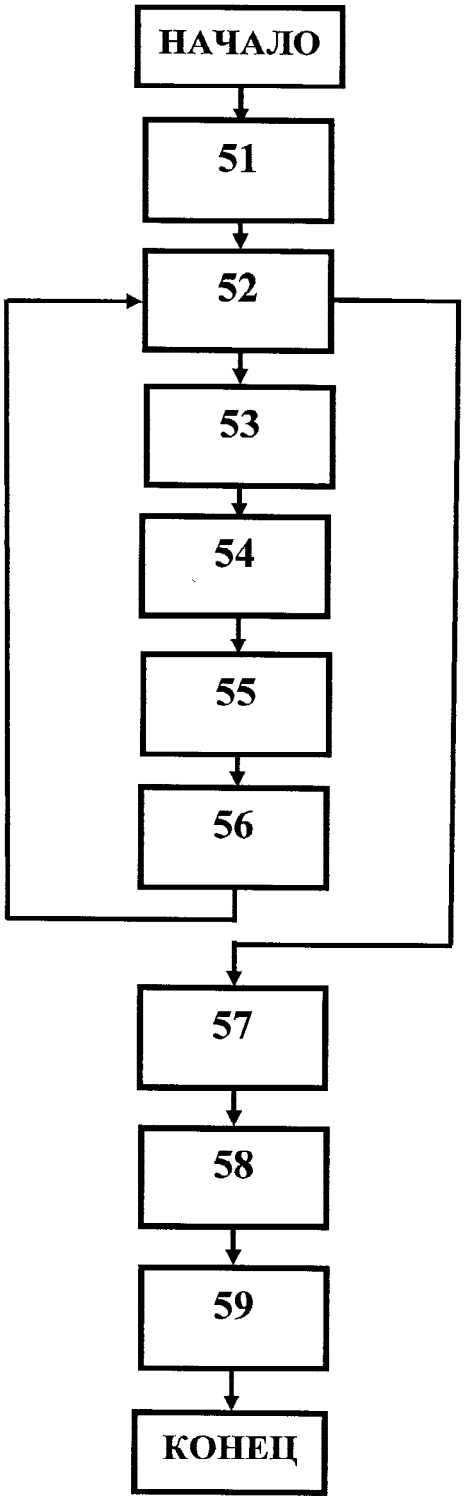
Фиг. 4b

27 февраля 1815 г. в Марсель из очередного плавания возвращается трехмачтовый корабль «Фараон». Капитану Леклеру не суждено было ступить на родную землю: он умер от горячки в открытом море. Молодой моряк Эдмон Дантес принял на себя командование. Владелец корабля Моррель предлагает Дантесу официально вступить в должность капитана корабля «Фараон»...

Дантес приговаривается к пожизненному заточению в замке Иф, политической тюрьме среди моря, неподалеку от Марселя... «Моей кассы еще никто не считал!» – уязвлен Данглар. «В таком случае я – первый, кому это предстоит», – обещает ему Монте-Кристо. ...Прежде чем испустить дух, аббат Бузони сообщает, что Монте-Кристо и Эдмон Дантес – одно лицо...



Фиг. 5



Фиг. 6

```
SELECT DISTINCT ?uid ?FamilyName ?FirstName ?EmploymentLabel
      WHERE
      {
            ?uid
                  sofa:__INSTANCEOF_REL onto:Персона;
                  onto:Фамилия ?FamilyName;
                  onto:Имя ?FirstName.
            ?e
                  sofa:__INSTANCEOF_REL onto:работать_в;
                  os:_to ?orgUid;
                  os:_from ?uid.
            ?orgUid
                  sofa:__LABEL_REL ?работает.
      }ORDER BY ?FamilyName
```

Фиг. 7