РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 305 314** (13) **C1**

(51) МПК
G06F 17/30 (2006.01)

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21), (22) Заявка: **2006122998/09, 28.06.2006**

(24) Дата начала отсчета срока действия патента:
28.06.2006

(45) Опубликовано: **27.08.2007 Бюл. № 24**

(56) Список документов, цитированных в отчете о поиске: **RU 2236699 A1, 20.09.2004. RU 2144697 C1, 20.01.2000. RU 2231117 C2, 10.07.2002. RU 2096825 C1, 20.11.1997.**

Адрес для переписки:
111250, Москва, ул. Авиамоторная, 53, ЗАО "Патентный Поверенный", пат.пов. Г.Н.Андрущак, рег. № 189

(72) Автор(ы):
Баранов Алексей Владимирович (RU),
Брянцев Игорь Николаевич (RU),
Жевлаков Игорь Михайлович (RU),
Ковалев Олег Петрович (RU),
Рязанкина Надежда Ивановна (RU)

(73) Патентообладатель(и):
Общество с ограниченной ответственностью "Центр Компьютерного моделирования" (RU)

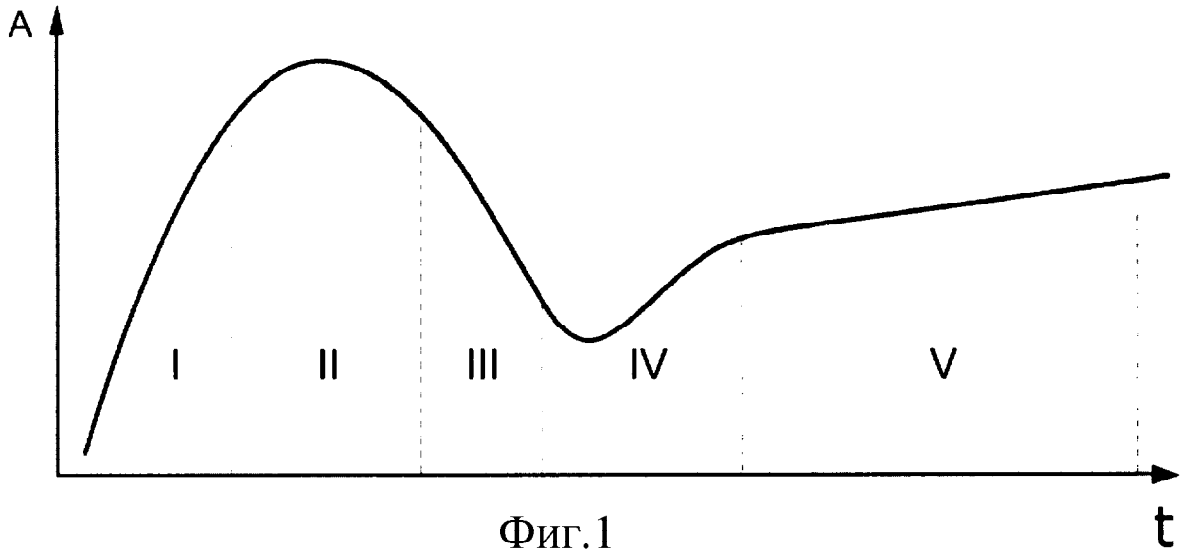
(54) СПОСОБ ПОИСКА И ВЫБОРКИ ИНФОРМАЦИИ ИЗ РАЗЛИЧНЫХ БАЗ ДАННЫХ

(57) Реферат:

Заявленное изобретение относится к способу поиска и идентификации документов по их описаниям, находящимся в различных базах данных и информационных ресурсах с различными стандартами формирования документов. Технический результат заключается в повышении точности поиска и проведения анализа полученной информации. Технический результат достигается за счет того, что сформированные пользователем поисковые запросы передаются в поисковую систему сервера, которая производит обработку упомянутых запросов путем выбора документов из различных баз данных, поисковая система объединяет все выбранные документы в единый список, сортирует упомянутые выбранные документы по тематикам, формирует папки, которые содержат упомянутые документы, одной тематики, снова сортируют упомянутые отсортированные документы с учетом окончательного рейтинга. После чего на основе пользовательского запроса определяют разделы будущего отчета, с помощью поисковой системы определяют текстовые признаки начала и завершения разделов, проводят разметку текста выбранных с наибольшими показателями окончательного рейтинга документов, внутри каждого раздела выделяют сегменты текста, проводят сортировку сегментов в соответствии с датой публикации, подготавливают итоговый отчет, в котором сегменты текста, отсортированные в соответствии с датой публикации оригинального документа, объединены в единый текстовый массив, после чего передают на пользовательский терминал через телекоммуникационные средства связи итоговый отчет. 4 ил., 1 табл.

RU 2305314 C1

RU 2305314 C1



Фиг.1

RU 2305314 C1

RU 2305314 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(19) **RU** (11) **2 305 314** (13) **C1**

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

(12) ABSTRACT OF INVENTION

(21), (22) Application: **2006122998/09, 28.06.2006**

(24) Effective date for property rights: **28.06.2006**

(45) Date of publication: **27.08.2007 Bull. 24**

Mail address:
111250, Moskva, ul. Aviamotornaja, 53, ZAO "Patentnyj Poverennyj", pat.pov. G.N.Andrushchak, reg. № 189

(72) Inventor(s):
Baranov Aleksej Vladimirovich (RU),
Brjantsev Igor' Nikolaevich (RU),
Zhevlakov Igor' Mikhajlovich (RU),
Kovalev Oleg Petrovich (RU),
Rjazankina Nadezhda Ivanovna (RU)

(73) Proprietor(s):
Obshchestvo s ogranichennoj otvetstvennost'ju "Tsentr Komp'juternogo modelirovanija" (RU)

(54) METHOD FOR FINDING AND SELECTING INFORMATION IN VARIOUS DATABASES

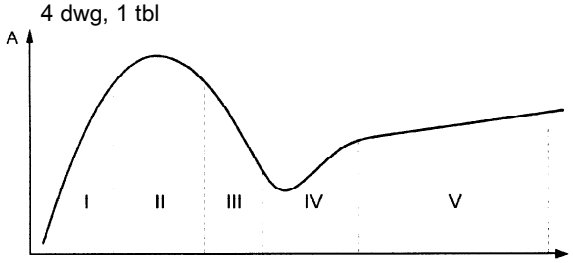
(57) Abstract:

FIELD: technology for finding and identifying documents based on their descriptions, present in various databases and information resources with different document creation standards.

SUBSTANCE: in accordance to the invention, search requests formed by user are dispatched into search system of server, which processes aforementioned requests by selecting documents from various databases, searching system combines all selected documents in single list, sorts aforementioned selected documents based on topics, creates folders, which contains aforementioned documents of one topic, then aforementioned sorted documents are sorted again with consideration of final rating. After that on basis of user request sections of future report are determined, by means of searching system, text signs of beginning and end of sections are determined, text of documents selected for greatest final rating is marked up, inside each section text segments are selected, segments are sorted according to publishing data, final report is prepared, in which text segments, sorted according to publishing date of original document, are combined in single text array, after that final report is dispatched to user terminal through telecommunication communication means.

EFFECT: increased precision of searching and analyzing of received information.

4 dwg, 1 tbl



Фиг.1

RU 2305314 C1

RU 2305314 C1

Изобретение относится к средствам поиска и идентификации документов по их описаниям, находящимся в различных базах данных и информационных ресурсах с различными стандартами формирования документов, а также может использоваться при наполнении базы данных фрагментами изначально неструктурированных текстов.

Известны способы идентификации документов по их описаниям, заключающиеся в преобразовании текстов естественного языка в заданных областях знаний в сигналы, пригодные для машинной обработки, в формировании запроса в виде выборки ключевых слов и в сравнении выборки ключевых слов запроса с тезаурусами текстов, хранящихся в базе данных (см. патенты РФ №2107942, 2167450, патент США №6460034, поисковая база данных Яндекс).

Недостатком известных способов является ограниченность одной базой данных с известным стандартом формирования.

Наиболее близким аналогом, принятым за прототип, является способ поиска и выборки информации из баз данных, описанный в патенте RU 2236699, в соответствии с которым осуществляют: формирование пользователем на своем рабочем месте, представляющем собой любой персональный компьютер, имеющий доступ к различным базам данных, по меньшей мере одного поискового запроса; передачу сформированного пользователем запроса в поисковую систему; обработку поисковой системой сформированных пользователем поисковых запросов путем выбора документов из баз данных; причем поисковая система сортирует упомянутые выбранные документы по тематикам и формирует папки, каждая из которых содержит упомянутые документы, отсортированные по одной тематике; для каждого отсортированного документа выделяют признаки, характеризующие этот документ; внутри каждой папки поисковая система определяет рейтинг каждого признака, содержащегося в каждом отсортированном документе; после чего поисковая система определяет число совпадений признаков отдельных отсортированных документов одной папки с признаками других документов, содержащихся в других папках; определяет окончательный рейтинг каждого отсортированного документа с учетом числа совпадений признаков и с учетом весового коэффициента базы данных; после чего поисковая система снова сортирует упомянутые отсортированные документы с учетом окончательного рейтинга и направляет отсортированные в соответствии с окончательным рейтингом документы на рабочее место пользователя.

Недостатком прототипа является отсутствие структурной обработки и анализа полученных документов по их значимости применительно к заданному элементу запроса. Равноценность всех выбранных объектов и документов приводит к росту объема отобранной информации и росту информационного шума, что в конечном счете увеличивает затраты интеллектуального труда на обработку отобранной информации пользователем.

Кроме того, в случае работы с множеством хранилищ документов с различными стандартами формирования документов идентификация объектов становится трудно выполнимой.

Техническим результатом заявленного изобретения является расширение функциональных возможностей путем повышения точности поиска и проведения анализа полученной информации.

Технический результат достигается за счет того, что в способе поиска и выборки информации из различных баз данных, включающем формирование пользователем на пользовательском компьютере, по меньшей мере, одного поискового запроса, передачу сформированного пользователем запроса через телекоммуникационные средства связи в поисковую систему сервера, обработку поисковой системой сформированных пользователем поисковых запросов путем выбора документов из различных баз данных, дополнительно поисковая система объединяет все выбранные документы в единый список, сортирует упомянутые выбранные документы по тематикам, формирует папки, каждая из которых содержит упомянутые документы, отсортированные по одной тематике, для каждой папки определяет параметры оценки релевантности, для каждого отсортированного

документа каждой папки, выделяет признаки, характеризующие каждый документ, внутри каждой папки определяет рейтинг каждого признака, содержащегося в каждом отсортированном документе, определяет окончательный рейтинг каждого выбранного документа с учетом рейтинга каждого признака документа и параметра оценки релевантности папки, снова сортирует упомянутые отсортированные документы с учетом окончательного рейтинга, отсортированные в соответствии с окончательным рейтингом документы запоминает в памяти поисковой машины, выбирает заданное пользователем в запросе количество документов с наибольшими показателями окончательного рейтинга, осуществляет структурную обработку выбранного количества документов для подготовки итогового отчета, формирует график, показывающий зависимости количества отсортированных в соответствии с окончательным рейтингом документов от текущего времени, передает на пользовательский терминал через телекоммуникационные средства связи итоговый отчет и сформированный график.

При анализе тенденций и мониторинге той или иной предметной области, основанных на анализе неструктурированной информации (книги, статьи, обзоры и т.д.), пользователь сталкивается со следующими основными проблемами:

1. Большой объем информации. В традиционных предметных областях объем публикаций измеряется сотнями тысяч и даже миллионами единиц публикаций. Такой объем материалов пользователь не может физически прочитать. Таким образом, выводы о состоянии предметной области ему приходится делать на основании выборки. При этом ни один пользователь не может быть до конца уверен в том, что не существуют значимые материалы, опровергающие его точку зрения, которые он просто не нашел.

2. Релевантность материалов. Практически все поисковые системы (Google, Yahoo, Yandex и др.) выдают результаты поиска в виде единого списка. В этих списках по формальным принципам (число ссылок, концентрация слова, посещаемость и т.д.) определяется релевантность (ценность) материалов. Однако в повседневной деятельности пользователь никогда не использует эти принципы при сравнении разнородных предметов. Книгу не сравнивают с человеком, а компанию с событием, патент с новостями. В одних случаях для пользователя важна одна информация, например статья, в других случаях это может быть иная информация - материалы конференции или патент.

3. Неполнота информации. Как правило, пользователь имеет дело с набором разнородной информации (текстами): статьями, книгами, материалами конференций, новостями и т.д. Отобранные для работы материалы часто относятся к разным периодам времени развития предметной области, написаны разными авторами, представляющими различные организации, с различными интересами, целями, задачами, уровнем развития и компетенции. Кроме того, эти тексты различаются по своей структуре, глубине, терминологии, объему данных. Различия могут носить принципиальный характер, точки зрения могут противоречить одна другой даже у одного автора, позиция которого часто меняется от работы к работе в зависимости от времени.

Заявленный способ направлен на решение вышеуказанных проблем.

Сущность заявленного изобретения заключается в том, что после формирования пользователем на пользовательском компьютере поискового запроса и передачу его через телекоммуникационные средства связи в поисковую систему сервера, обработки поисковой системой сформированного пользователем поискового запроса путем выбора документов из различных баз данных осуществляют следующие этапы.

1 этап. Отбор информации.

Первоначально пользователь в различных поисковых системах, например Yandex, Yahoo, Google и др., отбирает произвольное количество документов по порядку, который предоставляет ему поисковая система в соответствии с ее собственными правилами. Как отмечалось выше, в этом списке ссылок находятся разнородные по своей природе материалы: статьи, патенты, компании, новости и т.д. Очевидно, что в реальной жизни мы не сравниваем человека и патент или книгу с компанией. Ценность того или иного материала зависит от различных субъективных факторов, интересов пользователя в

данном конкретном случае.

В способе предлагается:

Объединить с помощью поисковой системы все найденные материалы в единый список без учета порядка их появления в списках различных поисковых систем;

Отсортировать с помощью поисковой системы документы по тематикам, а именно собрать в отдельные папки (группы) однородные по своей природе материалы (все книги в папку «Книги», все патенты в папку «Патенты», информацию о специалистах в папку «Персоналии» и т.д.).

Количество папок может быть произвольное, но в каждую папку должны входить однородные документы, существующие как объекты в реальной жизни.

Пример набора папок:

Популярные материалы (Введение в тему)

Новости

Новостные источники

События

Организации

Персоналии

Порталы

Периодические издания

Книги

Обзоры

Определить для каждой папки параметры оценки релевантности (важности) того или иного материала.

Например, папка «Книги». Параметрами оценки являются: индекс цитирования, оценка читателей в Интернет-магазинах, год издания, известность автора, язык произведения и др.

В папке «Компании» параметрами могут быть: известность компании, специализация в данной предметной области, капитализация, регион, страна и т.д.

Число параметров, которое может ввести пользователь, не ограничено. Однако на практике количество важных параметров редко превышает число 7. Эмпирическим путем установлено, что в большинстве практически значимых случаев оптимальным числом параметров является 4.

Каждому параметру в каждой папке присваивается определенная степень важности, называемая весом параметра и принимающая значения от 0 до 1.

Устанавливаются для каждого параметра в каждой папке целочисленные значения уровня в зависимости от его реального значения.

Например, для папке «Книги».

Параметр «индекс цитирования»

более или равно 10-5 баллов,

более 2, но меньше 10-4 балла,

1-2-3 балла;

0-2 балла.

Параметр «Язык публикации»

«Русский» - 5 баллов;

«Английский» - 4 балла;

«Немецкий» - 3 балла;

«Японский» - 2 балла;

«Арабский» -1 балл.

Параметр «Год издания»

2004-2006 - 5 баллов;

2000-2003 - 4 балла;

1995-1999 - 3 балла

1980-1995 - 2 балла;

до 1980 - 1 балл.

Каждый пользователь может настраивать систему параметров и присваивать веса в соответствии со своими предпочтениями. Уровень параметра для каждого документа в папке напрямую зависит от реальных значений параметра для данного документа.

Определить рейтинг i-того документа в j-той папке по формуле:

$$R_i^j = \sum_{k=1}^{m_j} a_k^j p_k^{ij} + c_i^j$$

m_j - число параметров в j-той папке;

a_k^j - значение веса k-того параметра в j-той папке;

p_k^{ij} - реально определяемое значение уровня k-того параметра для i-того элемента в j-той папке;

c_i^j - количество групп, в которых упоминается i-тый документ в j-той папке.

Далее в каждой папке отбираются только первые несколько документов с максимальными в папке показателями релевантности.

Отсортированные поисковой системой в соответствии с окончательным рейтингом документы запоминают в памяти поисковой машины.

Первоначальный отбор документов, сортировка их по папкам (группам), вычисление релевантности каждого элемента и выбор конечного количества документов с большими показателями релевантности завершают 1 этап работы.

2 этап. Структурная обработка.

Отсортированные поисковой системой в соответствии с окончательным рейтингом документы запоминают в памяти поисковой машины. Для реализации 2 этапа необходимо:

1. Определить разделы будущего итогового отчета. Например, Цели, Задачи, Прогнозы, Текущие результаты и другие.

2. Определить ключевые слова, характерные для каждого из разделов.

3. Определить текстовые признаки начала и завершения раздела.

4. Провести разметку текста отобранных на первом этапе документов в соответствии с разделами. Выделить сегменты текста, перенести эти сегменты в базу данных в соответствующий им раздел.

5. Внутри каждого раздела провести сортировку сегментов. Сортировка сегментов проводится в соответствии с датой публикации оригинального документа. Сегменты, в которых присутствуют взаимные цитаты, размещаются последовательно друг за другом, начиная с сегмента с более ранней датой.

6. Осуществлять пополнение базы данных новыми документами (осуществлять мониторинг).

7. Сформировать отчет о проблеме.

В любой момент времени пользователь может сформировать итоговый отчет о проблеме. При формировании отчета все сегменты раздела объединяются в единый текстовый массив с ссылками на первоисточники.

Вышеуказанный способ отбора информации и ее структурная обработка могут также быть использованы при наполнении базы данных фрагментами изначально неструктурированных текстов.

3 этап. Анализ трендов.

Большинство предметных областей развиваются в соответствии с G-кривой (фиг.1.). Эту кривую впервые опубликовала аналитическая корпорация Gartner Group в 1995 году.

G-кривая отображает зависимость количества информационных сообщений документов о технологии от времени и уровня развития технологии. Под информационными сообщениями подразумеваются любые упоминания технологии в СМИ, литературе, Интернете и других источниках информации. Текст каждого упоминания считается документом.

Кривая условно делится на пять участков:

1 участок - старт технологий (появление новой концепции).

На этом участке происходит быстрый рост количества документов. Это объясняется

значительным увеличением инвестиций в данную технологию, ростом рекламы, ростом числа участников в разработке и продвижении технологии.

2 участок - пик чрезмерных ожиданий.

На этом участке прекращается рост числа публикаций, снижается объем рекламы, прекращается рост числа участников.

3 участок - разочарование.

Снижается количество публикуемых документов, снижается объем рекламы, многие участники покидают бизнес. Из предметной области уходят инвестиции. Затраты направляются на поиск качественных изменений технологии.

4 участок - уклон просвещения.

Найдены качественные решения проблем. Инвестиции начинают расти. Увеличивается объем документов. Нарастает объем рекламы.

5 участок - плато продуктивности.

Разработана технология с оптимальными параметрами. Число документов стабильно нарастает. Оптимизируются рекламные бюджеты. Растет объем инвестиций.

Приступая к анализу новой для себя предметной области, пользователь сталкивается с тем, что не может изначально определить, какому участку G-кривой соответствует данный информационный материал (документ) (фиг.2).

Для решения этой задачи было бы хорошо сравнить информационные материалы, относящиеся к разным интервалам времени (фиг.3). Однако на практике реализовать эту задачу сложно, так как большинство информационных материалов имеют различную структуру и полноту описания. Например, в одних отчетах обсуждаются цели, задачи и перспективы развития, в других - проблемы и технические характеристики, в третьих рассматривается инвестиционная привлекательность проекта. Определить тренд развития по отдельно взятым материалам в финансовой, технической, научной областях, относящихся к различным временным промежуткам, очень сложно. Кроме того, в отобранных материалах численные значения каких-либо характеристик изучаемой технологии могут отсутствовать вообще.

Способ позволяет построить G-кривую, выявить тренд при наличии обрывочных данных вербального описания характеристик состояния развития технологии.

Для решения задачи по построению G-кривой необходимо проделать следующие операции:

1. В каждом разделе, созданном на 2 этапе работы, проводится оценка соответствующих характеристик. Как отмечалось выше, все собранные сегменты оригинальных документов внутри раздела размещаются последовательно в соответствии с датой публикации оригинального документа. Для формализации вербальных характеристик технологии проводится попарное сравнение рядом стоящих сегментов.

Если качественные характеристики предметной области усиливаются, то в таблицу результатов обработки заносится +1, если ослабляются, то -1. После заполнения таблицы по ее данным строится G-кривая (фиг.4).

2. Для упрощения процесса формализации используется таблица правил:

№ п.п.	Правило формализации	Значение
1.	Число пунктов в последующем сегменте документа увеличилось	+1
2.	Текст документов повторяется	-1
3.	Появились даты событий	+1
4.	Указаны численные параметры	+1
5.	Присутствуют слова «переносятся», «не удовлетворяют», «не соответствуют»	-1
...	.....	
n	Если А..., то В...	

Если условия Правил из таблицы имеют противоположное значение, то и численное значение сегмента меняется на противоположное.

Пользовательский компьютер представляет собой любой персональный компьютер, например, компании IBM, состоящий из системного блока, к которому подключен монитор, клавиатура и манипулятор типа "мышь".

Пользовательский компьютер должен иметь доступ к базам данных, которые могут быть как удаленными, так и локальными. Доступ к базам данных можно осуществить посредством подключения пользовательского компьютера через телекоммуникационные средства связи к поисковой системе сервера глобальной сети Интернет или локальной сети, например Intranet.

Базы данных могут быть как однородными, каждая из которых содержит документы только по одной тематике, например патентная база данных, так и неоднородными, которые содержат документы по разным тематикам, например Яндекс.

Базы данных записаны в памяти компьютера или сервера, например, на жестком диске.

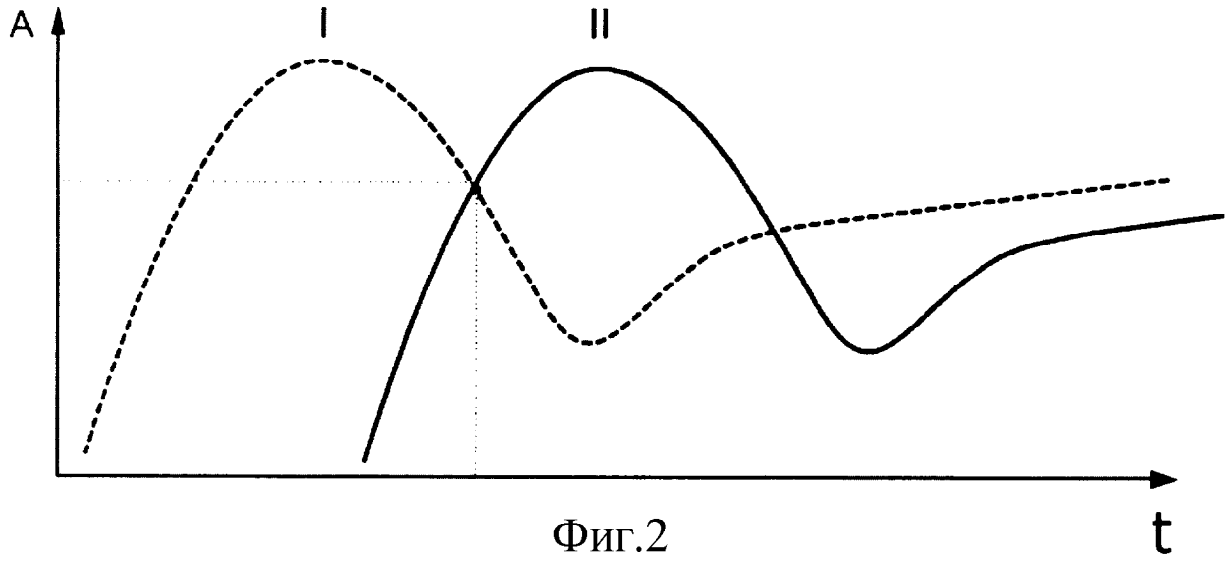
Поисковая машина представляет собой обычную 32-битовую машину (например, Linux, Solaris, Free BSD, Win32).

В качестве поисковой системы используется, например, поисковая система Fast, реализующая известную логику прямого поиска. Поисковая система Fast разработана и поставляется на рынок норвежской компанией "Fast Search & Transfer ASA".

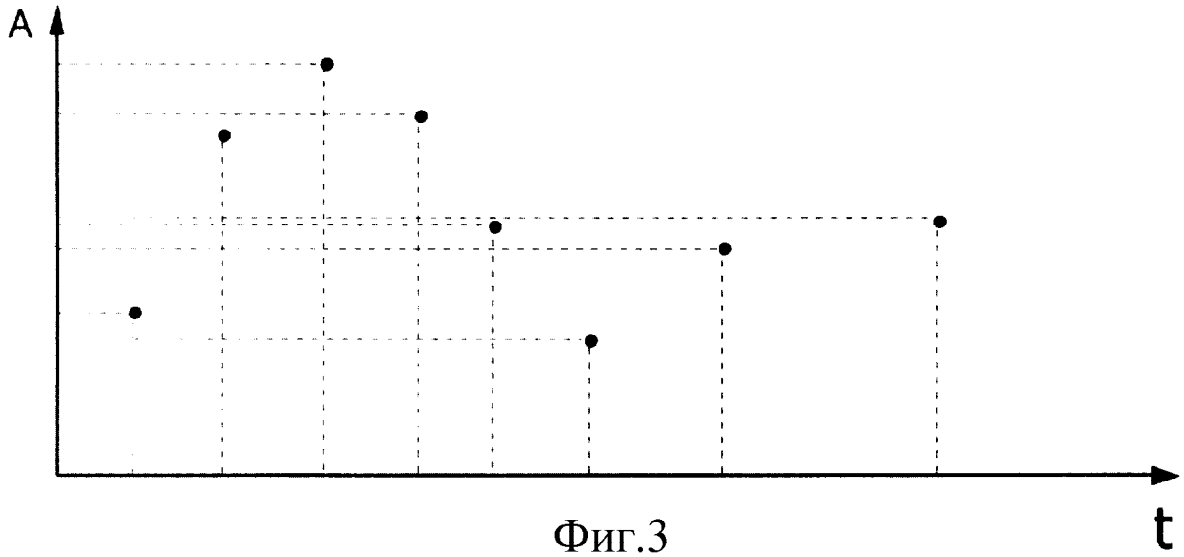
Применение способа позволяет также сократить машинное время поиска, повысить релевантность выборки документов запросу, снизить затраты интеллектуального труда при анализе выборки документов.

Формула изобретения

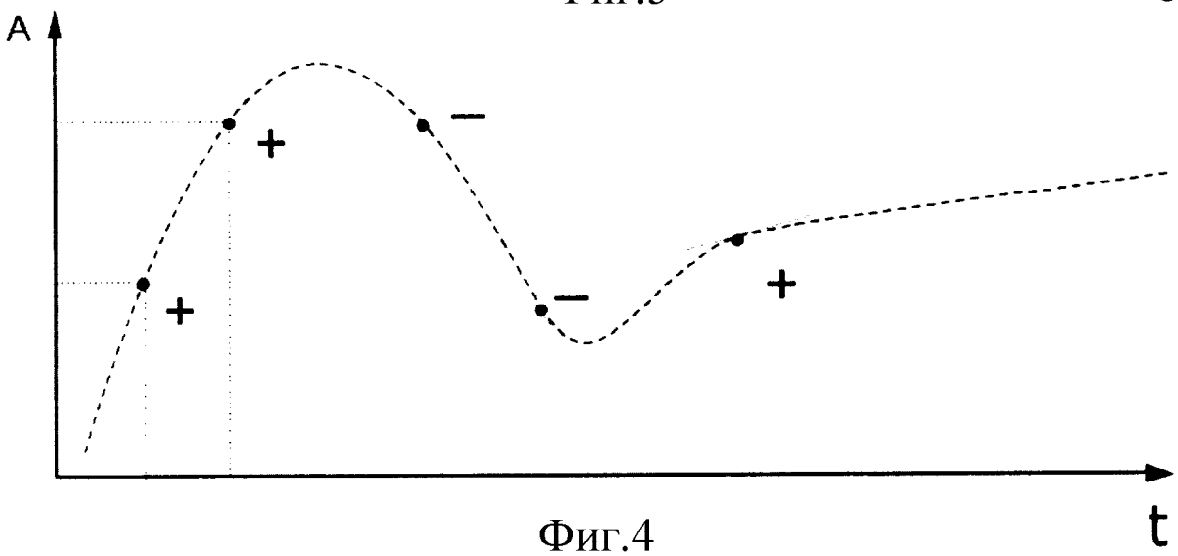
Способ поиска и выборки информации из различных баз данных, включающий формирование пользователем на пользовательском компьютере, по меньшей мере, одного поискового запроса, передачу сформированного пользователем запроса через телекоммуникационные средства связи в поисковую систему сервера, обработку поисковой системой сформированных пользователем поисковых запросов путем выбора документов из различных баз данных, поисковая система объединяет все выбранные документы в единый список, сортирует упомянутые выбранные документы по тематикам, формирует папки, каждая из которых содержит упомянутые документы, отсортированные по одной тематике, для каждой папки определяет параметры оценки релевантности, для каждого отсортированного документа каждой папки, выделяет признаки, характеризующие каждый документ, внутри каждой папки определяет рейтинг каждого признака, содержащегося в каждом отсортированном документе, определяет окончательный рейтинг каждого выбранного документа с учетом рейтинга каждого признака документа и параметра оценки релевантности папки, снова сортирует упомянутые отсортированные документы с учетом окончательного рейтинга, отсортированные в соответствии с окончательным рейтингом документы запоминает в памяти поисковой машины, выбирает заданное пользователем в запросе количество документов с наибольшими показателями окончательного рейтинга, отличающийся тем, что осуществляют структурную обработку выбранного количества документов, содержащую этапы, на которых пользователь в поисковом запросе определяет разделы будущего итогового отчета и ключевые слова, характерные для каждого из разделов, с помощью поисковой системы определяют текстовые признаки начала и завершения разделов, проводят разметку текста выбранных с наибольшими показателями окончательного рейтинга документов, внутри каждого раздела выделяют сегменты текста, проводят сортировку сегментов в соответствии с датой публикации, подготавливают итоговый отчет, в котором сегменты текста, отсортированные в соответствии с датой публикации оригинального документа, объединены в единый текстовый массив, после чего передают на пользовательский терминал через телекоммуникационные средства связи итоговый отчет.



Фиг.2



Фиг.3



Фиг.4