РОССИЙСКАЯ ФЕДЕРАЦИЯ

(19) **RU** (11) **2 266 560** (13) **C1**

(51) МПК[7] **G 06 F 17/30**

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ,
ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(21), (22) Заявка: **2004113072/09, 28.04.2004**

(24) Дата начала действия патента: **28.04.2004**

(45) Опубликовано: **20.12.2005 Бюл. № 35**

(56) Список документов, цитированных в отчете о поиске: **RU 2167450 C2, 20.05.2001. RU 2138076 C1, 20.09.1999. SU 1837327 A1, 30.08.1993. US 4839853 A, 13.01.1989. US 5819261 A, 06.10.1998. US 5926811 A, 20.07.1999. US 6006221 A, 21.12.1999. US 6247010 B1, 12.06.2001. US 2003/0217072 A1, 20.11.2003. JP 10-207910 A, 07.08.1998. WO 02/073331 A2, 19.09.2002. WO 03/027902 A1, 03.04.2003.**

Адрес для переписки:
105037, Москва, городок им. Баумана, 1, стр.1, ФГУП ИПР "Информэлектро", для С.В. Попова

(72) Автор(ы):
Попов С.В. (RU)

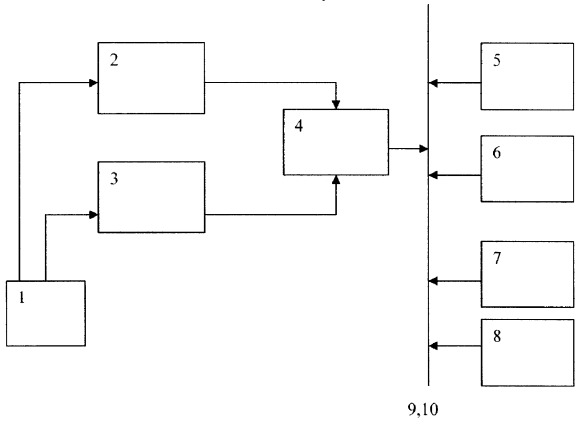
(73) Патентообладатель(ли):
Федеральное государственное унитарное предприятие "Институт промышленного развития "Информэлектро" (RU)

RU 2266560 C1

(54) СПОСОБ ПОИСКА ИНФОРМАЦИИ В ПОЛИТЕМАТИЧЕСКИХ МАССИВАХ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВ

(57) Реферат:

Изобретение относится к области информационных технологий. Его использование при поиске информации в больших документальных базах данных обеспечивает технический результат в виде сокращения времени поиска нужной информации за счет сокращения количества рекурсий (повторений запросов). Способ заключается в том, что терминам вектора запроса присваивают порядковые номера, осуществляют поиск с занесением в память компьютера номеров документов хотя бы с одним термином вектора запроса, заносят в память компьютера количество терминов, совпавших с терминами запроса, и порядковые номера совпавших терминов, сортируют в памяти компьютера документы по классам с равным количеством совпавших терминов. Технический результат достигается тем, что вводится новый критерий выдачи документов, позволяющий пользователю получать релевантные документы, наполненные новыми терминами, необходимыми для проведения дальнейших рекурсий. Эффективность способа при этом не зависит от того, на каком естественном языке написаны тексты в базе данных. 3 з.п. ф-лы, 2 ил.



Фиг. 1

RU 2266560 C1

RUSSIAN FEDERATION

(19) **RU** (11) **2 266 560** (13) **C1**

(51) Int. Cl.[7] **G 06 F 17/30**

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY,
PATENTS AND TRADEMARKS

(12) ABSTRACT OF INVENTION

(21), (22) Application: **2004113072/09, 28.04.2004**

(24) Effective date for property rights: **28.04.2004**

(45) Date of publication: **20.12.2005 Bull. 35**

Mail address:
105037, Moskva, gorodok im. Baumana, 1, str.1, FGUP IPR "Informehlektro", dlja S.V. Popova

(72) Inventor(s):
Popov S.V. (RU)

(73) Proprietor(s):
Federal'noe gosudarstvennoe unitarnoe predprijatie "Institut promyshlennogo razvitija "Informehlektro" (RU)

(54) METHOD UTILIZED TO SEARCH FOR INFORMATION IN POLY-TOPIC ARRAYS OF UNORGANIZED TEXTS

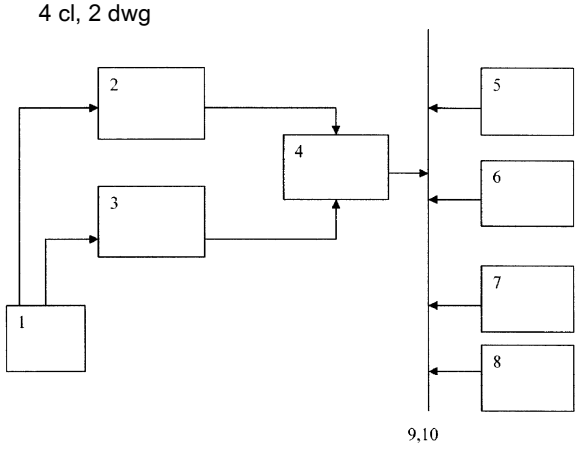
(57) Abstract:

FIELD: computer science.

SUBSTANCE: method includes assigning order names to query vector terms, performing search with recording numbers of documents in computer memory with at least one term of query vector, number of terms, matched by query terms, is recorded in computer memory, as well as order numbers of matching terms, documents are sorted in computer memory in accordance to classes with even number of matching terms. Additionally introduced is new criterion of documents dispensing, allowing for user to receive relevant documents, filled with new terms, necessary to perform further recursions. Efficiency of method does not depend on natural language of texts in database.

EFFECT: when used to search for information in extensive document databases, it is possible to reduce time required for finding needed information due to lesser number of recursive (repeated) queries.

4 cl, 2 dwg



Фиг. 1

RU 2266560 C1

RU 2266560 C1

Изобретение относится к области информационных технологий, в частности к способам поиска информации в больших документальных базах данных (БД).

Известен способ поиска информации путем анализа взаимной встречаемости терминов запроса и терминов в найденных документах, а также анализа мер сходства векторов документов, представленных на различных языках, так называемое семантическое векторное совпадение (US 6006221, G 06 F 17/30, опубл. 21.12.1999).

Недостатком данного способа является сложность операций по построению и преобразованию (суммирование, нормализация) векторов.

Известен способ автоматизированного поиска информации с расширением запроса путем построения статистического тезауруса (US 5926811, G 06 F 17/30, опубл. 20.07.1999).

Недостатком указанного способа является то, что тезаурусы требуют частого обновления.

Наиболее близким аналогом к заявляемому способу поиска информации является способ поиска информации (US 4839853, G 06 F 15/40, опубл. 13.01.1989) с использованием латентно-семантической структуры. Согласно этому способу из найденных в ответ на первоначальный запрос пользователя документов выделяются значимые для данной темы термины, затем этим терминам присваиваются веса значимости, после этого строится вектор запроса и все документы исходной БД ранжируются по степени сходства с этим вектором на основании соответствующей меры близости - косинус угла между вектором запроса и вектором найденного документа.

По своей сути описанный способ является рекурсивным, то есть потенциально позволяющим на основе статистического анализа последующих выдач документов строить все более развитые векторы запросов.

Недостатком этого способа является его низкая производительность вследствие того, что значения мер близости векторов запросов и документов (ранги документов) уменьшаются для каждой БД и каждого запроса слишком быстро, и следовательно, вместо "плавного" рекурсивного наращивания полноты поиска системам приходится выдавать пользователям только весьма небольшое множество документов самых высоких рангов, предварительно установив жесткое пороговое значение меры близости. Другими словами, настоящей рекурсии не получается из-за того, что все последующие (развитые) векторы запросов слишком зависят от лексического состава выдачи, полученной в ответ на первый, зачастую весьма неэффективный запрос пользователя. Это приводит к тому, что значительно увеличивается время, затрачиваемое на проведение поиска.

Решаемой изобретением задачей является устранение указанного недостатка и усовершенствование информационно-поисковой системы (ИПС). Достигаемый технический результат заключается в сокращении времени поиска нужной информации за счет сокращения количества рекурсий (повторений запросов).

Указанный технический результат достигается тем, что вводится новый критерий выдачи документов, позволяющий пользователю получать релевантные документы, наполненные новыми терминами, необходимыми для проведения дальнейших рекурсий (повторений запросов).

А именно, в способе поиска информации с использованием информационно-поисковой системы, в котором терминам вектора запроса присваивают порядковые номера, затем осуществляют поиск с занесением в память компьютера номеров найденных документов, в которых присутствует хотя бы один термин вектора запроса, затем заносят в память компьютера количество совпавших терминов с терминами запроса и порядковые номера совпавших терминов, затем сортируют в памяти компьютера документы по классам с равным количеством совпавших терминов, согласно данному изобретению осуществляют формирование внутри всех классов - подклассов индекса i классов индекса j, характеризующихся полным совпадением номеров терминов, затем определение количества документов (n_{ij}) в подклассах индекса i классов индекса j, затем определение количества документов (n_j) класса j, затем определение вероятности принадлежности документа к подклассу i, при условии его принадлежности к классу j, как:

$$P_{ij} = \frac{n_{ij}}{nj},$$

затем определение критерия выдачи для каждого класса как:

$$H_j = -\sum_i P_{ij} \ln P_{ij}, \qquad (1)$$

и далее расширение запроса, если в документах класса H_{jmax}, содержатся новые термины, которые относятся к тематике поиска. H_{jmax} - максимальное значение параметра характеризующего критерий выдачи классов документов.

Другой дополнительной особенностью данного способа может являться то, что в ИПС загружаются документы, представленные на естественном языке. При этом в ИПС для осуществления предлагаемого способа используется входной формат ASC11.

Еще одной дополнительной особенностью данного способа может являться то, что формирование классов и подклассов документов осуществляется автоматически.

Еще одной дополнительной особенностью данного способа может являться то, что количество терминов вектора запроса находится в диапазоне от 10 до 1000.

В данном случае под вектором запроса понимается набор ключевых слов, классификационных индексов, фраз или просто слов без присвоения им весов значимости.

Наиболее сложной задачей информационного поиска является обнаружение информации, обозначение которой пользователю неизвестно. Поэтому, прежде чем получить нужный документ, необходимо найти дескрипторы - слова, классификационные индексы, имена и.т.п., по которым информация может быть найдена.

Это отнюдь не простая задача. Даже слова естественного языка не всегда легко подобрать для проведения эффективного сеанса поиска. Индексы различных классификаций и рубрикаторов, марки, названия фирм могут быть и вовсе не известны пользователю системы. Поэтому необходим механизм обнаружения таких терминов, по которым может быть найдена лексически удаленная, но необходимая пользователю информация.

Простейшим способом расширения запроса является отбор новых потенциально полезных терминов из документов, найденных в ответ на данный запрос.

Если пользователь выбрал набор терминов t_1, t_2, t_3...t_k, то необходимо установить правило, по которому ему будут выдаваться другие документы из исходного поискового массива, содержащие эти термины. Обычная логика подсказывает, что чем больше терминов из выбранных содержит документ, тем выше вероятность, что его содержание соответствует тематике первоначального запроса, и, следовательно, этот документ должен быть выдан в первую очередь. С другой стороны, такой документ лексически похож на те документы, из которых были выбраны термины t_1, t_2, t_3...t_k, и следовательно, слишком мала вероятность того, что в этом документе могут быть найдены дополнительные, полезные термины для дальнейшего расширения запроса и продолжений рекурсивного поиска.

Если произвести разбиение исходного поискового массива на классы документов с равным количеством терминов, совпавших с набором t_1, t_2, t_3...t_k и использовать в качестве критерия выдачи класса с индексом j количество совпавших терминов, то число отобранных новых полезных терминов на каждом шаге итерации будет в среднем в 2 раза меньше, чем при использовании критерия H_j(1), при одинаковом количестве просмотренных релевантных документов.

Изобретение поясняется чертежами.

Заявленный способ может быть реализован с помощью системы поиска информации

На фигуре 1 представлена функциональная схема системы поиска информации.

На фигуре 2 представлена блок схема алгоритма заявленного способа.

Система содержит блок формирования запроса 1, первый выход которого связан с входом блока памяти номеров документов 2, выход которого связан с первым входом блока поиска и сортировки 4, выход которого через соответствующие шины данных 9 и шины управления 10 связан с процессором 5, блоком воспроизведения 7, базой данных 6 и

контроллером 8, причем второй вход блока поиска и сортировки 4 связан с выходом блока памяти номеров терминов 3, вход которого связан со вторым выходом блока формирования запроса 1.

Система для поиска информации согласно изобретению работает следующим образом.

Блок формирования запроса 1 может представлять собой стандартный блок ввода-вывода данных с клавиатурой и мышью, с возможностью отображения вводимой информации на экране блока воспроизведения 7, т.е. это может быть дисплей, экран монитора и.т.п. В то же время блок формирования запроса 1 может быть выполнен в виде формирователя сообщения о выборе базы данных для проведения поиска, которое передается в контроллер 8 для запуска программы поиска в выбранной базе данных.

Поиск осуществляется следующим образом.

При включении системы пользователю с помощью блока воспроизведения 7 предлагается меню, которое отображается на экране, на котором, в частности, представлен перечень названий имеющихся баз данных системы. Далее с помощью блока формирования запроса 1 пользователь формирует первоначальный запрос, сообщение об этом сразу попадает в контроллер 8.

Далее пользователю системы предлагаются документы, выданные на первоначальный запрос, которые отображаются на экране, в которых ему предлагается выбрать новые термины, которые по его мнению могут относиться к интересующей его тематической области, причем терминам запроса присваивают порядковые номера с занесением их в блок памяти номеров документов 2 и далее в блок поиска и сортировки 4, который через шину данных 9 отправляет запрос в базу данных 6.

С помощью блока воспроизведения 7 пользователь может ознакомиться с документами, найденными на запрос.

Далее номера документов, содержащие термины, совпавшие с терминами запроса, заносятся в блок памяти номеров документов 2, после чего в блоке поиска и сортировки 4 осуществляют сортировку документов по классам с равным количеством совпавших терминов.

Далее внутри классов формируют подклассы, характеризующиеся полным совпадением номеров совпавших терминов. Затем процессор 5 проводит расчет характеристики H_j для каждого класса документов.

Используя такую характеристику, пользователь системы может специальной командой с помощью блока формирования запроса 1 дополнить терминами (из документов класса с H_{jmax}) первоначальный запрос. Дальнейший поиск может быть также проведен с использованием сохраненных запросов в блоке памяти номеров терминов 3 и состоящих только из терминов, содержащихся в документах класса с H_{jmax}.

По дополненному запросу ИПС позволяет найти необходимую пользователю, но лексически удаленную от первоначального запроса информацию.

Указанная последовательность действий повторяется до тех пор, пока в найденных документах класса с H_{jmax} будут встречаться новые термины, относящиеся к исследуемой тематике.

Опыты показывают, что указанный технический результат может быть достигнут только взаимосвязанной совокупностью всех существенных признаков заявленного изобретения, отраженных в формуле изобретения. Указанные в ней отличия дают основание сделать вывод о новизне данного технического решения, а совокупность испрашиваемых притязаний в связи с их не очевидностью - об изобретательском уровне, что было показано выше. Соответствие критерию "промышленная применимость" предложенного способа доказывается как его реализацией, так и отсутствием в заявленных притязаниях каких-либо практически трудно реализуемых в промышленных масштабах признаков.

Формула изобретения

1. Способ поиска информации с использованием информационно-поисковой системы, заключающийся в том, что терминам вектора запроса присваивают порядковые номера,

затем поиск осуществляют с занесением в память компьютера номеров документов, в которых присутствует хотя бы один термин вектора запроса, затем заносят в память компьютера количество совпавших терминов с терминами запроса и порядковые номера совпавших терминов, затем в памяти компьютера документы сортируют по классам с равным количеством совпавших терминов, отличающийся тем, что внутри каждого класса формируют подклассы индекса i класса индекса j, характеризующиеся полным совпадением номеров терминов, затем определяют количество документов (n_{ij}) в подклассе индекса i класса индекса j, затем определяют количество документов (n_j) класса j, затем определяют вероятность принадлежности документа к подклассу i при условии его принадлежности к классу j, как

$$P_{ij} = \frac{n_{ij}}{n_j},$$

затем определяют критерий выдачи для каждого класса как

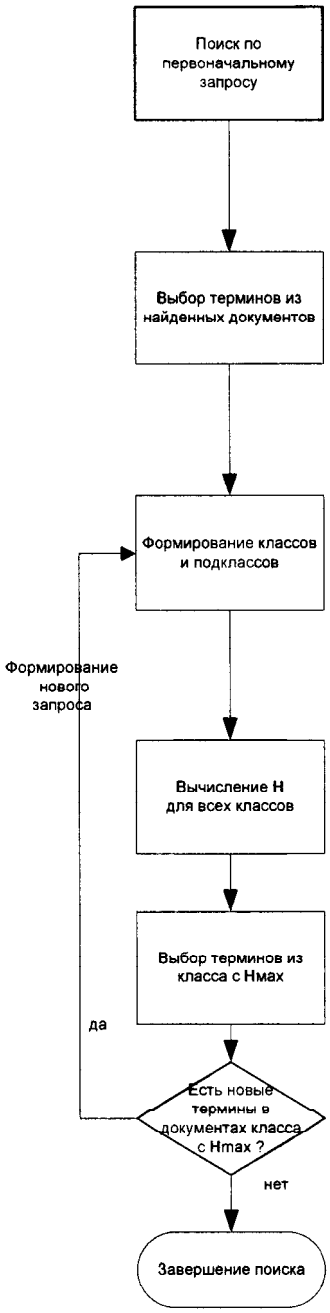
$$H_j = -\sum_i P_{ij} \ln P_{ij},$$

и далее расширяют запрос, если в документах класса с H_{jmax} содержатся новые термины, которые относятся к тематике поиска.

2. Способ по п. 1, отличающийся тем, что в информационно-поисковую систему загружаются документы, представленные на естественном языке.

3. Способ по п. 2, отличающийся тем, что формирование классов и подклассов документов осуществляется автоматически.

4. Способ по п. 1, или 2, или 3, отличающийся тем, что количество терминов вектора запроса находится в диапазоне от 10 до 1000.


Поиск по первоначальному запросу
Выбор терминов из найденных документов
Формирование классов и подклассов
Формирование нового запроса
Вычисление H для всех классов
Выбор терминов из класса с Hmax
да
Есть новые термины в документах класса с Hmax ?
нет
Завершение поиска

Фиг. 2