(19) RU (11) 2236699 (13) C1
(51) 7 G 06 F 17/30

ФЕДЕРАЛЬНАЯ СЛУЖБА ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ, ПАТЕНТАМ И ТОВАРНЫМ ЗНАКАМ

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ

к патенту Российской Федерации

1

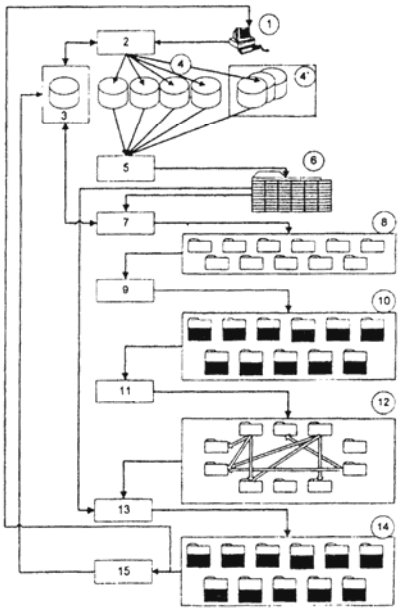
(21) 2003105262/09 (22) 25.02.2003
(24) 25.02.2003
(46) 20.09.2004 Бюл. № 26
(72) Баранов А.В. (RU)
(73) Открытое акционерное общество "Телепортал. Ру" (RU)
(56) RU 2167450 C2, 20.05.2001. US 6185550 B1, 06.02.2001. RU 8819 U1, 16.12.1998. US 6460034 B1, 01.10.2002.
Адрес для переписки: 111250, Москва, ул. Авиамоторная, 53, ЗАО "Патентный поверенный", пат.пов. Г.Н.Андрущак

(54) СПОСОБ ПОИСКА И ВЫБОРКИ ИНФОРМАЦИИ С ПОВЫШЕННОЙ РЕЛЕВАНТНОСТЬЮ

(57) Изобретение относится к средствам поиска и идентификации документов по их описаниям, находящимся в различных базах данных и информационных ресурсах с различными стандартами формирования документов. Технический результат заключается в сокращение объема информации, выводимой на дисплей пользовательского терминала по запросу пользователя, и уменьшение интеллектуальных трудозатрат на анализ полученной информации и принятия решения. Способ заключается в том, что проводят сортировку по отдельным папкам всех однородных документов из различных баз данных, определяют рейтинги каждого документа внутри папки, затем находят число совпадений признаков отдельных документов в различных папках и определяют окончательный рейтинг каждого документа с учетом числа пересечений, сортируют документы в соответствии с этим рейтингом и направляют эти документы на компьютер пользователя. 5 з.п.ф-лы, 1 ил., 3 табл.

2

RU 2236699 C1



RU 2236699 C1

Заявленное изобретение относится к средствам поиска и идентификации документов по их описаниям, находящимся в различных базах данных и информационных ресурсах с различными стандартами формирования документов.

Известны способы идентификации документов по их описаниям, заключающиеся в преобразовании текстов естественного языка в заданных областях знаний в сигналы, пригодные для машинной обработки, формировании запроса в виде выборки ключевых слов и сравнении выборки ключевых слов запроса с тезаурусами текстов, хранящихся в базе данных (см., например, полезную модель RU 8819, патенты РФ №2107942, 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 обработки запросов в системе поиска и выборки информации, описанный в патенте RU 2167450, в соответствии с которым: 1) сохраняют множество объектов в хранилище документов, в котором каждый объект документа определен признаками, заключенными в документе, так что упомянутые объекты, хранимые в документе, определяют общее содержание данного документа; 2) обрабатывают запрос, который включает по меньшей мере один элемент запроса, для выбора по меньшей мере одного документа, релевантного по меньшей мере к упомянутому одному элементу запроса; 3) идентифицируют из множества объектов по меньшей мере один документ; 4) представляют пользователю идентифицированный по меньшей мере один документ, при этом сходство документов оценивают различными способами ранжирования.

Недостатком прототипа является отсутствие оценки объектов и документов по их значимости применительно к заданному элементу запроса, т.е. оценки релевантности.

Равновероятность 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 сокращение объема информации, выводимой на дисплей пользовательского терминала по запросу пользователя, и уменьшение интеллектуальных трудозатрат на анализ полученной информации и принятие решения.

Технический результат достигается за счет того, что способ поиска и выборки информации из баз данных, включающий формирование пользователем на своем рабочем месте по меньшей мере одного поискового запроса, передачу сформированного пользователем запроса в поисковую систему, обработку поисковой системой сформированных пользователем поисковых запросов путем выбора документов из базы данных, дополнительно включает следующие операции: 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.

В частном варианте выполнения заявленного изобретения окончательный рейтинг упомянутого отсортированного (i-го) документа рассчитывают по формуле:

$$Ri = \sum_{\substack{j=1 \\ x_{i,j}=0}}^{n} a_j \frac{1}{x_{i,j}} + l_i + c_i,\ i = 1, m$$

где $X_{i,j}$ - рейтинг i-го документа в j-ой базе данных; a_j - рейтинг j-ой базы данных; l_i - количество рейтингов i-го документа не равных нулю из всех баз данных; c_i - количество совпадений признаков отдельных документов в различных папках.

Еще в одном частном варианте выполнения рейтинг j-ой базы данных варьируется в диапазоне от 0,1 до 1,0.

В другом частном варианте выполнения упомянутыми признаками документов являются авторы, организации, новости, события, все виды научно-технической литературы и патентной документации.

Еще в одном частном варианте выполнения видами научно-технической литературы являются статьи в периодических изданиях, монографии, сборники работ, труды конференций и других научных собраний.

В другом частном варианте выполнения упомянутый окончательный рейтинг документов устанавливают с помощью контрольного тестирования.

Сущность заявленного изобретения поясняется чертежом, на котором представлена блок-схема поисковой системы, реализующей заявленный способ.

Заявленный способ включает следующую последовательность операций:

1) формирование пользователем на своем рабочем месте, представляющем собой любой персональный компьютер, имеющий доступ к различным базам данных, по меньшей мере одного поискового запроса;

2) передачу сформированного пользователем запроса в поисковую систему;

3) обработку поисковой системой сформированных пользователем поисковых запросов путем выбора документов из базах данных;

4) поисковая система сортирует упомянутые выбранные документы по тематикам и формирует папки, каждая из которых содержит упомянутые документы, отсортированные по одной тематике;

5) для каждого отсортированного документа выделяют признаки, характеризующие этот документ;

6) внутри каждой папки поисковая система определяет рейтинг каждого признака, содержащегося в каждом отсортированном документе;

7) после чего поисковая система определяет число совпадений признаков отдельных отсортированных документов одной папки с признаками других документов, содержащихся в других папках;

8) определяет окончательный рейтинг каждого отсортированного документа с учетом числа совпадений признаков и с учетом весового коэффициента базы данных;

9) 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.

Предложенный способ реализуется поисковой системой, которая показана на чертеже.

Система состоит из следующих элементов: рабочее место пользователя (терминал компьютера) 1, блок "Преобразователь запросов" 2, базы данных "Стандарты баз данных" 3, базы данных "Поисковые информационные ресурсы" 4, блок "Интегратор документов" 5, блок "Единое хранилище" 6, блок "Сортировка документов" 7, блок "Папки" 8, блок "Восстановления структуры предметной области" 9, блок "Объекты" 10, блок "Восстановления структуры" 11, блок "Оценки пересечений" 12, блок "Рейтинг" 13, блок "Блок формирования результатов" 14, блок "Формирования рейтингов баз данных" 15.

Рабочее место пользователя (терминал компьютера) 1, как уже было указано выше, представляет собой любой персональный компьютер, например, компании IBM, состоящий из системного блока, к которому подключен монитор, клавиатура и манипулятор типа “мышь”.

Терминал компьютера 1 должен иметь доступ к базам данных 3, 4, которые могут быть как удаленными, так и локальными. Доступ к базам данных можно осуществить посредством подключения терминала 1 к сети глобальной сети Интернет или локальной сети, например, Intranet.

Базы данных 4 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.

База данных "Стандарты баз данных" 3, блоки "Единое хранилище" 6, блок "Папки" 8, блок "Объекты" 10 представляют собой базы данных, хранящиеся в памяти ЭВМ, например, на жестком диске.

Блоки "Преобразователь запросов" 2, "Интегратор документов" 5, "Сортировка документов" 7, "Восстановления структуры предметной области" 9, "Восстановления структуры" 11, "Оценки пересечений" 12, "Рейтинг" 13, "Блок формирования результатов" 14, блок "Формирования рейтингов баз данных" 15 представляют собой обычные 32-битовые машины (Linux, Solaris, FreeBSD, Win32).

Указанное устройство поиска информации работает следующим образом.

Пользователь вводит с терминала компьютера 1 запрос в виде ключевого слова или набора ключевых слов, например "Экологический мониторинг".

Сформированный запрос поступает в блок "Преобразователь запросов" 2, который обрабатывает полученный запрос, как, например, это выполняет поисковая система Fast, реализующая известную логику прямого поиска. Поисковая система Fast разработана и поставляется на рынок норвежской компанией "Fast Search & Transfer ASA".

Далее блок 2 обращается к базе данных "Стандарты баз данных" 3. В блоке 3 хранится информация о структуре и адресах поисковых информационных ресурсов (поисковых машин Интернета и информационных баз данных). Например, формат обращения через Интернет к базе данных патентов США USPTO имеет следующий вид:

"http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2Fsearch-bool.html&r=0&f=S&l=50&TERM1="ключевое слово"&FIELD1=&col=AND&TERM2=&FIELD2=&d=ptxt"

Формат обращения на языке SQL по локальной сети к локальным, корпоративным и другим базам данных, хранящимся на жестком диске или на CD-ROM, имеет следующий стандартный вид:

"DECLARE @FIELD1 VARCHAR(100),@FIELD2
VARCHAR(100),@FIELD3 VARCHAR(100)
SET @FIELD1='%'
SET @FIELD2='%'
SET @FIELD3='%'
SELECT*FROM<TABLE_NAME>
WHERE<FIELD1>LIKE @FIELD1
AND<FIELD2> LIKE @FIELD2
AND<FIELD3> LIKE @FIELD3"

В соответствии с приведенными примерами блок "Преобразователь запросов" 2 формирует различные по своей структуре “вторичные запросы”, которые направляются последовательно в соответствующие информационные ресурсы 4, 4’ в порядке убывания рейтинга баз данных.

Например, если пользователь вводит в систему ключевое слово "Garbage" (мусор), то вторичные запросы могут выглядеть следующим образом:

к базе данных USPTO

"http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=l&u=%2Fnetahtml%2Fsearch-bool.html&r=0&f=S&l=50&TERM1="garbage"&FIELD1=&col=AND&TERM2=&FIELD2=&d=ptxt";

к базе данных "COMPENDEX"

"DECLARE @FIELD1 VARCHAR(100),@FIELD2
VARCHAR(100),@FIELD3 VARCHAR(100)
SET @FIELD1='GARBAGE'
SET @FIELD2='GARBAGE'
SET @FIELD3='GARBAGE'
SELECT*FROM COMPENDEX
WHERE TITLE LIKE @FIELD1
AND CONFERENCE TITLE LIKE @FIELD2
AND ABSTRACT LIKE @FIELD3"

По результатам вторичных запросов из информационных ресурсов 4 в блоке "Интегратор документов" 5 собираются отобранные в информационных ресурсах 4 записи, состоящие из названия, электронного адреса, краткого описания и других данных, определяемых стандартами информационных ресурсов.

Примеры записей, полученных из информационных баз данных:

Из базы данных USPTO:

Inventors: Lieberman; Noah (Boulder, CO)

Assignee:Sun Microsystems, Inc. (Santa Clara, CA)

Appl No: 39101

Current U.S. Class: 709/225; 709/229; 709/2

Intern’l Class: G 06 F 015/173; G 06 F 015/16

Abstract:A content provider manager has been develop for use in an information services such as a portal or desktop application to provide for "pluggable" content that may be modified simply through....

Из базы данных COMPENDEX:

DIALOG №04265680 EI Monthly №EIP95102889590

Title: Cache performance of fast-allocating programs

Author: Goncalves, Marcelo J.R.; Appel, Andrew W.

Corporate Source: Princeton Univ

Conference Title: Conference Record of Conference on Functional Programming

Languages and Computer Architecture

Conference Location: La Jolla, CA, USA

Conference Sponsor: ACM SIGPLAN; ACM SIGARCH; IFIP

Source: Conf Rec Conf Funct Program Lang Comput Archit 1995. ACM. p. 293-305

Publication Year: 1995

Language: English

Conference Number: 43744

Document Type: CA; (Conference Article) Treatment Code: X; (Experimental)

Abstract: We study the cache performance of a set of ML programs, compiled by the Standard ML of New Jersey compiler. We find that more than half of the reads are for objects that have just been allocated...

Descriptors: *Program compilers; Buffer storage; Storage allocation (computer); Computer software; Computer hardware; Performance; Computer architecture

Identifiers: Cache performance; New Jersey compiler; Garbage collection frequency; Runtime systems"

"Интегратор документов" 5 объединяет все отобранные документы в единый массив, который размещается в блоке "Единое хранилище" 6 с сохранением структуры каждого документа.

На этой стадии работы поисковой системы указанный единый массив отобранных документов 6 обладает избыточностью информационных материалов, так как один и тот же документ мог быть отобран в различных поисковых ресурсах или базах данных и неоднократно повторяется в едином массиве 6.

Далее единый массив документов из блока "Единого хранилища" 6 направляется в блок "Сортировки документов" 7, где на основании формальных данных из блока "Стандарты баз данных" 3 производится сортировка отобранных материалов по тематикам и формируются папки, каждая из которых содержит отобранные материалы, отсортированные по одной тематике.

Каждая папка соответствует одной тематике, представляющей реальный объект предметной области: автор, организация, событие, новость, статья, книга и т.д. (см. блок 8 на чертеже).

Отсортированные по папкам блоком 7 материалы помещаются в блок "Папки" 8.

Далее материалы из каждой папки блока "Папки" 8 поочередно передаются в блок "Восстановления структуры предметной области" 9, который предназначен для формирования списков объектов и сортировки списков на основании определения веса каждого объекта. Блок 9 работает следующим образом.

Материалы одной из папок поступают в блок 9 из блока 8. Одновременно в блок 9 поступает информация из блока 3 о структуре документов. В результате сопоставления информации из блока 3 и блока 9, из анализируемых материалов извлекается информация об объекте и его атрибутах, т.е. выделяются признаки документа. Эти атрибуты включают в себя: название объекта, адреса документов, связанных с объектом, а также статистику о порядковых номерах адресов документов в списках поисковых информационных ресурсов.

После обработки всего массива одной папки устанавливается предварительный рейтинг каждого объекта. В таблице 1 приведен пример рейтинга документов в одной папке.

Таблица 1

Пример определения рейтинга авторов

№ п.п.	Ф.И.О	База данных (поисковая система)						Суммарный рейтинг
		Altavista	Yahoo	Amazon	Dialog	Patent	SCI	
1	L.Cotton	7	4	9	-	7	3	30
2	D.Sillivane	2	-	-	12	34	12	60
3	K.Deburg	11	12	14	33	1	1	72
4	J.Smith	12	6	44	2	10	2	76
5	K. Moore	23	17	11	29	5	12	97
.....	.....	.....	.....	.....	.....	.....	.....	.....
154	D.Dennie	125	123	2	-	22	12	284

Списки объектов и их атрибутов хранятся в блоке "Объекты" 10. После завершения работы с одной папкой начинается обработка блоком 9 следующей папки из блока 8. Обработка папок производится последовательно пока не будет обработана последняя папка.

Далее блок "Восстановления структуры" 11 получает списки объектов из блока 10 и присоединяет к этим объектам соответствующие документы, хранящиеся в блоке 6.

Таким образом, в блоке 11 производится предварительное определение релевантности первоначально отобранных в блоке 5 документов.

Далее в блоке "Оценки пересечений" 12 проводится анализ существующих пересечений между отдельными папками. Так, например, два автора L.Cotton и J.Smith (табл. 2) ссылаются в своих статьях на материалы одной и той же конференции (Intl.Conf. of Building Official). Данная конференция находится в списке конференций (табл. 3) под номером 4. Общее число пересечений суммируется по каждому из объектов и учитывается в окончательном расчете рейтинга расчете рейтинга объектов.

Таблица 2

Список ссылок авторов на конференции

№ п.п.	Ф.И.О.	Конференция
1	L.Cotton	Intl. Conference of Building Official
2	D.Sillivane	The United Nation Conference on Trade and Develop.
3	K.Deburg	The Appalachian Trail Conference
4	J.Smith	Intl. Conference of Building Official
5	D.Dennie	The US Conference of Mayors

Таблица 3

Список конференций

№ п.п.	Конференция
1	Intl. Conference of Building Official
2	The United Nation Conference on Trade and Develop.
3	The Appalachian Trail Conference
4	House Republican Conference
5	The US Conference of Mayors
6	JavaOne SM Conference

Блок "Рейтинг" 13 устанавливает окончательный рейтинг каждого объекта, рассчитывая его по формуле:

$$Ri = \sum_{\substack{j=1 \\ x_{i,j} \neq 0}}^{n} a_j \frac{1}{x_{ij}} + l_i + c_i,\ i = \overline{1, m} \qquad [2]$$

где $x_{i,j}$ - рейтинг i-го документа в j-ой базе данных; a_j - рейтинг j-ой базы данных; l_i - количество рейтингов i-го документа не равных нулю из всех возможных баз данных; c_i - количество совпадений признаков отдельных документов в различных папках. Рейтинг j-ой базы данных a_j варьируется в диапазоне от 0,1 до 1,0.

Отсортированные в соответствии с рейтингом документы сохраняются в блоке "Блок формирования результатов" 14. Из блока 14 документы в окончательном виде представляются пользователю на дисплее его компьютера 1.

Блок "Формирования рейтингов баз данных" 15 на основе информации из блока 14 устанавливает рейтинг использованных баз данных. Рейтинг базы данных тем выше, чем больше наиболее релевантных документов отобрано из этой базы данных. Окончательный рейтинг баз данных устанавливают с помощью контрольного тестирования, которое основано на измерении времени отклика базы данных на запрос. Чем больше время отклика, тем ниже рейтинг базы данных.

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

ФОРМУЛА ИЗОБРЕТЕНИЯ

1. Способ поиска и выборки информации из баз данных, включающий формирование пользователем на своем рабочем месте 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, ***отличающийся*** тем, что 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.

2. Способ по п.1, ***отличающийся*** тем, что окончательный рейтинг упомянутого отсортированного (i-го) документа рассчитывают по формуле

$$R_i = \sum_{\substack{j=1 \\ x_{i,j}=0}}^{n} a_j \frac{1}{x_{i,j}} + l_i + c_i, \quad i = 1, m,$$

где $x_{i,j}$ - рейтинг i-го документа в j-й базе данных;

a_j - рейтинг j-й базы данных;

l_i - количество рейтингов i-го документа не равных нулю из всех баз данных;

c_i - количество совпадений признаков отдельных документов в различных папках.

3. Способ по п.2, ***отличающийся*** тем, что рейтинг j-й базы данных a_j варьируется в диапазоне от 0,1 до 1,0.

4. Способ по одному из предыдущих пунктов, ***отличающийся*** тем, что упомянутыми признаками документов являются авторы, организации, новости, события, все виды научно-

технической литературы и патентной документации.

5. Способ по п.4, ***отличающийся*** тем, что видами научно-технической литературы являются статьи в периодических изданиях, монографии, сборники работ, труды конференций и других научных собраний.

6. Способ по одному из предыдущих пунктов, ***отличающийся*** тем, что окончательный рейтинг баз данных устанавливают с помощью контрольного тестирования.

Заказ 26 Подписное
ФИПС, Рег. ЛР № 040921
Научно-исследовательское отделение
по подготовке официальных изданий
Федерального института промышленной собственности
Бережковская наб., д.30, корп.1, Москва, Г-59, ГСП-5, 123995

Отпечатано на полиграфической базе ФИПС
Отделение по выпуску официальных изданий