ИНФОРМАТИКА

2004 **июль-сентябрь** **№3**

УДК 681.3.07

В.Ф. Быченков

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ ПОИСКА ИНФОРМАЦИИ НА ОСНОВЕ РЕЛЯЦИОННОЙ АЛГЕБРЫ

Рассмотрено проектирование запросов бинарного поиска и поиска с сортировкой по релевантности в реляционных базах данных для векторной модели документов. Для формулировки запросов использован язык реляционной алгебры. Приведены результаты экспериментальной проверки прототипа информационно-поисковой системы.

Введение

Теория поиска информации и автоматизированных информационно-поисковых систем (ИПС) разрабатывается на протяжении более полувека и продолжает развиваться в настоящее время вместе с развитием информационных систем. В качестве примеров приложения теории можно привести глобальную информационную систему Интернет; коллекции документов различного масштаба, тематической направленности и назначения; проблемно-ориентированные базы данных и системы управления корпоративными знаниями (см., например, [1 – 4]). В настоящее время широкое распространение получает использование технологий реляционных баз данных для создания ИПС корпоративного и более низкого уровня. Объектами исследований являются способы индексирования и алгоритмы поиска, структуры данных и выполняемые операции ([1 – 8] и др.).

Настоящая работа посвящена индексированию и поиску документов в информационно-поисковых массивах (ИП-массивах), основанных на платформе реляционных систем управления базами данных (СУБД) и векторной модели ИП-массива документов. Приведены структуры данных и сформулированные на языке реляционной алгебры запросы выполнения бинарного поиска и поиска с сортировкой по релевантности. Поиск информации в реляционных базах данных при координатном взвешенном индексировании рассматривается как задача векторной оптимизации. Приведены результаты экспериментальной проверки системы.

1. Математическая модель задачи поиска информации

Векторная модель ИП-массива включает в себя словарь терминов (ключевых слов) в виде множества мощности M и множество документов мощности N. Математическая модель документа представляет собой вектор размерности M, элементами (координатами) которого являются веса терминов в данном документе. Таким образом, ИП-массив представляется прямоугольной матрицей размерности $M \times N$, строками которой являются поисковые образы N документов. Элементы матрицы являются весами терминов, образующих поисковый образ документа (ПОД).

Веса терминов могут принимать значения из двоичного множества {0, 1} или множества действительных чисел. В первом случае значение 1 соответствует наличию термина в документе, значение 0 – отсутствию термина. Во втором случае нулевое значение веса соответствует отсутствию термина в документе, при ненулевом значении веса термина его величина определяет относительную значимость (информационный вес) термина в документе.

Широко распространено определение значимости i-го термина следующим соотношением [9]:

$$x_{ij} = f_{ij} \times \log_2(N/N_i), \qquad (1)$$

где f_{ij} – частота термина i в документе j; N_i – количество документов в ИП-массиве, в которых присутствует i-й термин. Первый сомножитель в выражении (1) в литературе часто обозначает-

ся как *tf* (*term frequency*), второй сомножитель обозначается как *idf* и представляет собой обратную частоту документа (*inverse document frequency*). Данное определение сохраняет свое значение и в настоящее время (см., например, [10]), хотя делаются попытки предложить и другие определения [11].

Определение значимости термина как его частоты в документе возможно при автоматическом или автоматизированном индексировании электронных документов. При ручном индексировании документов, в том числе и на бумажном носителе, для указания значимости терминов могут использоваться экспертные оценки в баллах. Подобный подход позволяет использовать одинаковые модели документа в ИП-массиве и запроса, а это дает возможность перед выполнением поиска ввести запрос в ИП-массив для определения релевантности найденных документов относительно запроса. Ниже рассматривается именно этот способ индексирования и поиска документов с экспертным определением значимости терминов.

Поиск документов в ИП-массиве может выполняться в два этапа [9]. На первом этапе формируется множество релевантных документов (бинарный поиск), на втором выполняются вычисление релевантности и сортировка документов по убыванию релевантности. В соответствии с наиболее часто используемыми видами запросов [12, с. 53] будем рассматривать запросы поиска «на включение», когда релевантный документ, проиндексированный терминами Td, содержит все термины Tq запроса одновременно – $Tq \subseteq Td$, и запросы поиска «на пересечение», когда релевантный документ содержит хотя бы один термин запроса – $Tq \cap Td \neq \varnothing$. Очевидно, что множество возвращаемых запросом «на пересечение» документов будет включать в качестве подмножества документы, возвращаемые запросом «на включение», если последнее не является пустым.

Для определения релевантности документа на матрице ИП-массива необходимо задать функционал, значения которого отражают сходство документа и запроса и позволят ранжировать документы по релевантности. В качестве такого функционала может выступать косинус угла между векторами документа D и запроса Q, евклидово расстояние между векторами документа и запроса или евклидова норма вектора документа, координаты которого «взвешены» умножением на веса терминов запроса. Мерой релевантности $Sim(Q,D_j)$ для *j*-го документа в последнем случае может служить степень приближения нормы вектора документа к норме вектора запроса. В дальнейшем используется именно этот способ оценки релевантности на основе формулы (2) из работы [13], которая применительно к многомерному случаю и при нормировке к максимальному значению в процентах может быть записана в виде

$$Sim(Q,D_j) = 100 \times \sqrt{\frac{\sum_{i=1}^{M} q_i^2 d_i^2}{\sum_{i=1}^{M} q_i^2}} / Sim(Q,D_j)_{max}, \qquad (2)$$

где q_i – балльная оценка значимости *i*-го термина в запросе;

d_i – значимость *i*-го термина в документе.

Определим d_i следующим выражением:

$$d_i = W_i^2 \times idf_i, \qquad (3)$$

где W_i – балльная оценка значимости *i*-го термина в документе.

Вторая степень оценки W_i в выражении (3) призвана снизить возможность взаимной компенсации весов терминов в выражении (2). Применение монотонно возрастающих функций

различного вида для учета экспертных оценок неметрических критериев в составе обобщенного критерия является обычной практикой при решении задач векторной оптимизации [14, с. 133]. Конкретный вид этой функции может зависеть, по-видимому, от количества терминов в поисковом запросе.

Выражения (2), (3) являются основными расчетными соотношениями, использующимися в остальной части работы.

2. Схема базы данных для задачи поиска

Схема базы данных для рассматриваемой задачи поиска информации в ИП-массиве содержит четыре отношения, описания атрибутов которых приведены в таблице:

$Документ\,(DocID, \ldots, Description)$ – отношение с атрибутами ИП-массива;
$Словарь\,(WordID, Key\_Word)$ – словарь терминов (ключевых слов);
$ПОД\,(WordID, DocID, Word\_Value)$ – поисковый образ документа;
$ПОЗ\,(WordID, WW)$ – поисковый образ запроса.

Таблица

Атрибуты отношений схемы базы данных

Имя атрибута	Описание атрибута
DocID	Идентификатор документа
Description	Библиографическое описание документа
WordID	Идентификатор ключевого слова
Key_Word	Ключевое слово
Word_Value	Балльная оценка значимости ключевого слова
WW	Квадрат балльной оценки ключевого слова запроса, деленный на знаменатель подкоренного выражения в формуле (2)

Многоточием в отношении $Документ\,(DocID, \ldots, Description)$ обозначены атрибуты, не используемые в данной работе, в частности элементы библиографической записи, по которым будет осуществляться атрибутивный поиск, формироваться библиографическое описание и т. п. Атрибуты $DocID, WordID$ являются первичными ключами отношений $Документ$ и $Словарь$ соответственно, отношение $ПОД$ имеет составной первичный ключ из атрибутов $WordID, DocID$.

Рассматриваемая схема базы данных отличается от использованных в работах [7, 8] наличием отношения $ПОЗ\,(WordID, WW)$, что позволяет исключить зависимость структуры поисковых запросов от количества ключевых слов в запросе, а от используемых в работах [5, 6, 8] – максимальным упрощением, чтобы предельно наглядно показать построение поисковых запросов. Упрощения связаны с исключением параллельного выполнения нескольких поисковых запросов, отказом от применения атрибутов терминов, ограничением только указанными выше типами запросов и т. п.

3. Запросы поиска документов

Ниже приведены запросы для выполнения бинарного поиска и поиска с сортировкой по релевантности, сформулированные на языке реляционной алгебры применительно к описанной в предыдущем разделе схеме базы данных. Использованная нотация в основном соответствует [15]. Для улучшения читаемости выражений:

– полный список имен атрибутов отношений заменен на символ звездочки или опущен там, где это не может вызвать затруднений;

– вместо операции расширенной проекции использованы две традиционные операции: проекции π и расширения λ;

– операция реляционного деления обозначена символом ÷ (см., например, работу [16] и список литературы к ней);

– в операциях внутреннего соединения всегда используются одноименные атрибуты соединяемых отношений;

– перенос выражения на другую строку обозначен символом подчеркивания в конце строки аналогично тому, как это принято в языке программирования VBA;

– в именах запросов использован префикс *qry* по соглашению Реддика для VBA;

– списки имен атрибутов в подстрочных индексах заключены в скобки.

3.1. Бинарный поиск «на включение»

Задача бинарного поиска «на включение» не отличается от задачи поиска решения по таблице решений с ограниченными входами [16], поэтому отношение с кодами искомых документов $qryR(DocID)$ может быть записано в виде

$$qryR\ (DocID)\ := \pi_{(WordID,\ DocID)}(ПОД\ (*))\ \div\ \pi_{(WordID)}(ПОЗ\ (*))$$

Результат выборки документов $qryList\ (*)$ из отношения $Документ\ (*)$ при бинарном поиске может быть описан следующим образом:

$$qryList\ (*) := \sigma_C(Документ\ (*))$$

где условие выборки ***C*** определяется выражением

$$Документ.DocID\ \text{IN}\ \sigma(qryR\ (DocID))$$

Здесь IN – оператор принадлежности к списку значений.

3.2. Бинарный поиск «на пересечение»

Задача бинарного поиска «на пересечение» отличается от рассмотренной выше задачи бинарного поиска «на включение» способом формирования отношения с кодами искомых документов:

$$qryR\ (DocID)\ := \pi_{(PubID)}(\sigma_C\ ПОД\ (*))$$

где условие выборки ***C*** определяется выражением

$$ПОД.WordID\ \text{IN}\ \ \pi_{(WordID)}(ПОЗ\ (*))$$

3.3. Поиск с сортировкой по релевантности

Как отмечено выше, поиск с сортировкой по релевантности предполагает выполнение предварительно бинарного поиска с получением соответствующего отношения $qryR\ (DocID)$. Вычисление релевантности и формирование результирующего списка документов выполняются следующей последовательностью запросов.

Шаг 1. Получаем количество документов в ИП-массиве:

$$qryN\ (N) := \gamma_{(\text{COUNT}(DocID)\rightarrow N)}(\pi_{(DocID)}(Документ\ (*)))$$

Шаг 2. Определяем входимость ключевых слов запроса *Ni* в документы ИП-массива:

$$qryNi\ (WordID, WW,\ N,\ Ni) := \gamma_{(WordID,\ WW,\ N,\ \text{COUNT}(DocID)\rightarrow Ni)} -$$

$$(\pi_{(WordID,\ WW,\ N,\ DocID)}((ПОЗ\ \triangleright\triangleleft\ ПОД) \times qryN\ (N))))$$

Декартово умножение на $qryN\,(N)$ позволяет включить количество документов ИП-массива N в отношение $qryNi\,(*)$ в виде отдельного столбца.

Шаг 3. Вычисляем idf и слагаемые подкоренного выражения (2) для каждого термина запроса без учета балльных оценок значимости терминов в документе:

$$qryIDF\,(WordID, WW, N, Ni, IDF, IDF2) := \lambda_{(*,\ IDF,\ IDF2)}(qryNi\,(*))$$

где вычисляемые поля в операции расширения

$$\mathrm{Log}(N / Ni)\,/\mathrm{Log}(2) \to IDF$$

$$IDF * IDF * WW \to IDF2$$

Шаг 4. Определяем все ключевые слова для выбранных документов:

$$qryKW\_R(DocID, WordID, Word\_Value) := \pi_{(DocID,\ WordID,\ Word\_Value)}(qryR \triangleright\!\triangleleft ПОД)$$

Шаг 5. Вычисляем слагаемые подкоренного выражения (2):

$$qryRank\_i\,(DocID, WordID, Word\_Value, IDF2, Rank\_i) :=$$

$$:= \pi_{(DocID,\ WordID,\ Word\_Value,\ IDF2,\ Rank\_i)}(\lambda_{(*,\ Rank\_i)}(qryKW\_R\,(*) \triangleright\!\triangleleft qryIDF\,(*)))$$

где вычисляемое поле в операции расширения

$$IDF2 * (Word\_Value ** 4) \to Rank\_i$$

Шаг 6. Вычисляем подкоренное выражение в формуле (2):

$$qryRank\,(DocID, Rank) := \gamma_{(DocID,\ \mathrm{SUM}(Rank\_i) \to Rank)}(\pi_{(DocID,\ Rank\_i)}\, qryRank\_i\,(*))$$

Шаг 7. Формируем ранжированный и отсортированный по атрибутам $Rank$, $Description$ (соответственно по убыванию и по возрастанию) список выбранных документов:

$$qryList\_Ranked\,(*) := \tau_{(Rank,\ Description)}(qryRank\,(*) \triangleright\!\triangleleft Документ\,(*))$$

В силу монотонности функции квадратного корня значения поля $Rank$ использованы для сортировки записей по релевантности. Для отображения в отчете значения релевантности вычисление квадратного корня и нормировка в формуле (2) могут выполняться как с помощью дополнительных запросов, так и в полях отчета средствами базового языка программирования.

4. Практическая реализация системы и обсуждение результатов

Рассмотренная модель системы поиска проверена экспериментально, для чего использована персональная библиотечная информационная система Bi_Docs [2, 17], содержащая в настоящее время более 5 000 записей научно-технических и учебно-методических публикаций по информационным технологиям.

Построение шкалы балльных оценок ключевых слов связано со структурированием текста публикации и является самостоятельной научной задачей. В работе определение значимости ключевых слов в публикации и запросе при индексировании выполнялось путем экспертной оценки релевантного ключевому слову фрагмента текста с использованием следующей упрощенной шкалы: 1 – упоминание ключевого слова; 2 – тезисы; 3 – реферат; 4 – сводный реферат, доклад; 5 – статья, раздел учебника или производственного издания с описательно-повествовательным содержанием; 6 – статья, раздел учебника, производственного издания или монографии с формально-логическим содержанием; 7 – учебник, производственное издание; 8 – монография.

Использованная шкала, очевидно, не является универсальной, и необходимы предварительная оценка и накопление опыта для определения целесообразности ее применения в конкретной предметной области и для конкретной цели использования ИП-массива. Для практического применения подобных шкал необходима разработка методики с целью снижения разброса оценок, что является обычной практикой при индексировании документов (см., например, [4, 18]). Тем не менее в процессе проведенных экспериментов подобная шкала значительно облегчила индексирование и поиск публикаций и ее применение может рассматриваться как полезное дополнение к выбору ключевых слов. Присвоенные в соответствии с данной шкалой балльные оценки использовались в качестве W_i и q_i.

Сортировка по релевантности обеспечивает существенное улучшение представления списка отобранных публикаций. Для запросов поиска «на включение», когда каждая из отобранных публикаций содержит все ключевые слова запроса, значение релевантности определяется балльными оценками ключевых слов в публикации. Для запросов поиска «на пересечение» значение релевантности убывает по мере уменьшения в публикации количества ключевых слов запроса и балльных оценок их значимости. Возможность изменения весов ключевых слов запроса позволяет перемещать отобранные публикации по списку [18], а наличие запроса в составе списка позволяет ориентироваться в назначении весов ключевым словам запроса и публикаций. Интерфейс прототипа системы предусматривает ввод до девяти ключевых слов запроса, однако рассмотренная модель системы поиска не накладывает ограничений на количество ключевых слов ни в индексе документов, ни в запросах.

Полученные результаты для запроса «Найти учебную литературу по управлению проектами создания автоматизированных информационных систем (АИС) с применением модели жизненного цикла Microsoft Solutions Framework (MSF) на основе архитектуры АИС» показаны на рисунке. Библиографическое описание публикаций в поле *Description* (Описание) формировалось в соответствии с ГОСТ 7.1-84 и ГОСТ 7.82-2001 из элементов библиографической записи посредством запросов, построение которых рассмотрено в работе [2].

5 августа 2004 г.		Страница 1 из 16	***Система Bi_Docs***
Всего: 269		*Ключевые слова:* Управление проектами, 7; АИС, 7; MSF, 7; Архитектура, 7;	
№	*Sim*	*Описание*	
1.	100.00	_Запрос от 05.08.2004 14:46:40 № б/н "======= Текущий запрос =======".	
2.	100.00	Анализ требований и создание архитектуры решений на основе Microsoft .NET: Учебный курс MCSD (Exam 70-300) / Microsoft Corporation. - М.: Изд.-торг. дом "Русская Редакция", 2004. - 416 с.	
3.	100.00	Брандт Д. Architectures. Экзамен - экстерном (экзамен 70 - 100). - СПб.: Питер, 2001. - 432 с.	
4.	100.00	Принципы проектирования и разработки программного обеспечения. Учебный курс MCSD: Пер. с англ. - 2-е изд., испр. / Microsoft Corporation. - М.: Изд.-торг. дом "Русская Редакция", 2002. - 736 с.	
5.	100.00	Принципы проектирования и разработки программного обеспечения. Учебный курс MCSD: Пер. с англ. / Microsoft Corporation. - М.: Изд.-торг. дом "Русская Редакция", 2000. - 608 с.	
6.	90.15	Быченков В.Ф. Проектный менеджмент и развитие информационных систем: Учебное пособие. - Мн.: Академия управления при Президенте Республики Беларусь, 2003. - 195 с.	
7.	76.80	Ройс, Уокер. Управление проектами по созданию программного обеспечения. - М.: Издательство "ЛОРИ", 2002. - 424 с.	
8.	63.00	Йордон Э. Управление сложными Интернет-проектами. - М.: Издательство "ЛОРИ", 2003. - 354 с.	
9.	60.84	Управление проектами. Толковый англ.-рус. словарь-справочник / В.Д.Шапиро, Н.Г.Ольдерогге, А.А.Юркевич. Под ред. В.Д.Шапиро. - М.: Высшая школа, 2000. - 379 с.	
10.	60.84	Фатрелл Р.Т., Шафер Д.Ф., Шафер Л.И. Управление программными проектами: достижение оптимального качества при минимуме затрат: Пер. с англ. - М.: Издательский дом "Вильямс", 2003. - 1136 с.	

Рис. Пример экранного изображения фрагмента отчета при поиске с сортировкой по релевантности

Общий список найденных публикаций для данного примера составил 269 позиций. При поиске «на включение» с теми же ключевыми словами запрос возвращает лишь первые шесть позиций показанного на рисунке списка. Позиции 1 – 6 имеют в своих индексах все термины запроса с балльными оценками 7, за исключением позиции 6, для которой термин MSF имеет оценку 6. Индекс позиции 8 имеет два термина запроса с оценками 7. Позиции 7, 9 и 10 по термину «управление проектами» проиндексированы как монографии (балльная оценка 8) и по этой причине находятся весьма близко к началу списка, хотя и не полностью отвечают сформулированному запросу. Назначение балльной оценки, равной единице, перечисленным ключевым словам для любой из позиций 1 – 6 перемещало ее на позицию 93 с релевантностью 2,04.

Балльная оценка значимости ключевых слов запроса изменялась от 1 до 100. Увеличение балльной оценки ключевого слова запроса до максимального значения позволяет перенести в начало списка публикации, содержащие в своем индексе данное ключевое слово.

Таким образом, поиск с сортировкой по релевантности наиболее полезен для запросов «на пересечение», при которых система выдает значительно большее количество публикаций, чем при поиске «на включение». Расположение найденных публикаций в списке полностью определяется балльными оценками ключевых слов документов и запроса и согласуется с интуитивными представлениями о соответствии найденных документов запросу. В то же время возможность изменять балльные оценки ключевых слов запроса позволяет рассматривать задачу поиска как задачу векторной оптимизации с позиции предпочтений конкретного пользователя системы.

Заключение

Поиск информации в реляционных базах данных с сортировкой результатов по релевантности может рассматриваться как задача векторной оптимизации на основе обобщенного аддитивного критерия. Успешное решение задачи не всегда полностью зависит от исследователя, поскольку результаты поиска определяются качеством индексирования документов. Тем не менее построение правильной стратегии поиска позволяет повысить его результативность. Большое значение имеют возможность гибкого изменения стратегии, например совместное использование поиска по ключевым словам, атрибутам библиографической записи и классификаторам; использование обратной связи по релевантности; возможность изменения балльных оценок терминов запроса и др., а также представление результатов поиска, основными требованиями к которому являются полнота, компактность и упорядоченность. Вывод на экран общего количества выбранных записей, порядкового номера, значения релевантности и библиографического описания документа облегчает корректировку стратегии поиска. Кроме того, подобные возможности позволяют улучшить и качество индексирования документов.

Документирование запросов в форме выражений реляционной алгебры позволило избежать излишней преждевременной детализации, имеющей место при использовании языка SQL, а также привязки к конкретному диалекту языка. Этот уровень абстракции особенно привлекателен при проектировании системы на этапе, когда СУБД еще не выбрана. Приведенные в работе выражения реляционной алгебры проверены экспериментально путем реализации в среде СУБД MS Access на языке SQL MS Access и обеспечивают выполнение всех необходимых вычислений при минимальном использовании средств базового языка программирования. Для проиндексированных в соответствии с приведенной шкалой балльных оценок публикаций результаты поиска не противоречили интуитивному представлению о релевантности публикаций. Рассмотренное в работе построение запросов может быть использовано и при автоматическом или автоматизированном индексировании документов.

Для полноценного использования языка реляционной алгебры в проектировании систем необходимо совершенствование его нотации для обеспечения компактности и однозначности трактовки выражений. Здесь могут быть полезны приемы, используемые языками программирования. Важным направлением является формирование и использование типовых высокоуровневых конструкций, например операции реляционного деления. Решение задач проектирования информационных систем позволит выявить и другие целесообразные приемы компактной и наглядной записи выражений реляционной алгебры.

Список литературы

1. Храмцов П. Моделирование и анализ работы информационно-поисковых систем Internet // Открытые системы. – 1996. – № 6. – С. 46–56.

2. Быченков В.Ф. Автоматизация работы с библиографической НТИ в учебном процессе // Информатизация образования. – 2002. – № 1. – С. 64–78.

3. Игнатова И., Резонтов К., Чаплыгин Ю. Создание тематических подборок электронных ресурсов // Информационные ресурсы России. – 2002. – № 7. – С. 7–9.

4. Белоозеров В.Н., Кулькова Г.В. Лингвистическое обеспечение корпоративной информационно-поисковой системы // НТИ. Сер. 1. – 2004. – № 3. – С. 14–18.

5. Putz S. Using a Relational Database for an Inverted Text Index / Xerox Palo Alto Research Center, 1991. http://www.n3labs.com/pdf/putz91using–inverted–DBMS.pdf.

6. Sysoev T. Indexing and search services in Integrated System of Information Resources of the Russian Academy of Sciences / Computing Center, Russian Academy of Sciences // Proceedings of the Spring Young Researcher's Colloquium on Database and Information Systems SYRCoDIS. – St.-Perersburg, Russia, 2004. http://syrcodis.citforum.ru/2004/sysoev.pdf.

7. Игумнов Е. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных, 2001. http://www.citforum.ru/database/articles/search_sys.shtml.

8. Аграновский А.В., Арутюнов Р.Э., Телеснин Б.А. Использование SQL для индексации и обработки текстовых документов // Информационные технологии. – 2004. – № 5. – С. 14–19.

9. Salton G. Another Look at Automatic Text-Retrieval Systems // Communications of the ACM. – 1986. – V. 29. – № 7. – P. 648–656.

10. Толчеев В.О. Модели и методы классификации текстовой информации // Информационные технологии. – 2004. – № 5. – С. 6–14.

11. Козлов А.В., Мальцева С.В. Методы повышения эффективности автоматического индексирования документов // Автоматизация и современные технологии. – 2004. – № 6. – С. 43–46.

12. Блюменау Д.И. Информационный анализ / синтез для формирования вторичного потока документов. – СПб.: Профессия, 2003. – 240 с.

13. Salton G., Fox E.A., Wu H. Extended Boolean Information Retrieval // Communications of the ACM. – 1983. – V. 26. – № 11. – P. 1022–1036.

14. Брахман Т.Р. Многокритериальность и выбор альтернатив в технике. – М.: Радио и связь, 1984. – 268 с.

15. Гарсиа-Молина Г., Ульман Дж., Уидом Дж. Системы баз данных: Пер. с англ. – М.: Издательский дом «Вильямс», 2003. – 1088 с.

16. Быченков В.Ф. Моделирование таблиц решений средствами реляционных СУБД // Информатизация образования. – 2004. – № 1. – С. 73–82.

17. Быченков В.Ф. Информационное моделирование на основе библиографической НТИ в учебном процессе // Новые информационные технологии = New Information Technologies (NITe–2002): Мат. V Междунар. науч. конф. Минск, 29–31 окт. 2002 г. В 2-х т. Т. 1. – Мн.: БГЭУ, 2002. – С. 284–288.

18. Солтон Дж. Динамические библиотечно-информационные системы. – М.: Мир, 1979. – 557 с.

Поступила 20.09.04

УП «НИИЭВМ»,
Минск, Богдановича, 155
e-mail: vlad2@niievm.by

V.F. Bachenkov

MODELING OF INFORMATION RETRIEVAL PROCESSES ON THE BASE OF RELATIONAL ALGEBRA

Creating the information retrieval queries for binary and relevance sorted search in relational databases is discussed for vector document model. Relational algebra is used to formulate the queries. The experimental results of testing information retrieval system prototype are given.