МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное учреждение высшего профессионального образования
«ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
ТЕХНОЛОГИЧЕСКИЙ ИНСТИТУТ В г. ТАГАНРОГЕ
(филиал федерального университета)

На правах рукописи

ТИЕК ЛЕНГ

РАЗРАБОТКА И ИССЛЕДОВАНИЕ НЕЧЕТКИХ МОДЕЛЕЙ ДИНАМИЧЕСКОГО УПОРЯДОЧЕНИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ В ХРАНИЛИЩАХ ДАННЫХ С УЧЕТОМ ИХ ВОСТРЕБОВАННОСТИ ПОТРЕБИТЕЛЯМИ ИНФОРМАЦИИ

Специальность:

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Таганрог – 2011

Работа выполнена на кафедре математического обеспечения и применения ЭВМ факультета автоматики и вычислительной техники Федерального государственного автономного образовательного учреждения высшего профессионального образования «Южный Федеральный Университет» Технологический институт в г. Таганроге (ТТИ Южного федерального университета).

НАУЧНЫЙ РУКОВОДИТЕЛЬ:

доктор технических наук, профессор Вишняков Юрий Муссович

ОФИЦИАЛЬНЫЕ ОППОНЕНТЫ:

доктор технических наук, профессор Чернухин Юрий Викторович;

кандидат технических наук, Спиридонов Олег Борисович.

ВЕДУЩАЯ ОРГАНИЗАЦИЯ:

ОАО «ТАНТК им. Г.М. Бериева», г.Таганрог.

Защита диссертации состоится « 17 » июня 2011 г. в 14^{20} на заседании диссертационного совета (Д 212.208.21) при Южном федеральном университете по адресу: 347928 г.Таганрог, пер.Некрасовский, 44, ауд. Д-406.

Просим Вас прислать отзыв, заверенный гербовой печатью учреждения, по адресу: 347928, Ростовская область, г. Таганрог, пре. Некрасовский, 44, Технологический институт Южного федерального университета в г. Таганроге. Ученому секретарю диссертационного совета Д 212.208.21 Чернову Николаю Иванову.

С диссертацией можно ознакомиться в Зональной научной библиотеке ЮФУ по адресу: 344007, Ростовская обл. г., Ростов-на-Дону, ул. Пушкинская, 148.

Автореферат разослан «14» мая 2011 г.

Ученый секретарь
диссертационного совета Д 212.208.21,
доктор технических наук, профессор Чернов Н.И.

Актуальность проблемы. В настоящее время в связи с развитием глобальных информационных коммуникаций все большее значение приобретает доступ к информационным ресурсам, представленным в электронном виде. В связи с этим появляется острая необходимость создания различного рода электронных хранилищ данных (ХД), которые обладали бы свойствами адаптации к запросам потребителей и подстраивались под их потребности. Это может быть достигнуто, если в хранилищах данных будут предусмотрены соответствующие механизмы адаптации и динамического упорядочения информационных ресурсов (ИР).

Сегодня разработка хранилищ ИР, в которых предусматриваются выше названые функции, является сложной и до конца не решенной задачей, а ее исследованию посвящен ряд научных работ авторитетных исследователей. Среди них следует отметить работы: Уильяма Инмона (Liam Inmon), Дугласа Хэкниа (Douglas Hackney), Доринна Хосса (Dorinny Hoss), Вишнякова Ю.М.. Однако, общее решение данной проблемы до сих пор не найдено. Это обстоятельство послужило основанием для формулировки темы диссертационного исследования, которое направлено на разработку и исследование адаптационных механизмов хранилищ данных, ориентированных на потребности потребителей информации.

Целью работы является разработка и исследование нечетких моделей динамического упорядочивания информационных ресурсов в хранилищах данных с учетом их востребованности потребителями информации.

Основные задачи диссертационного исследования:

1. Провести сравнительный анализ эффективности известных подходов к хранению информационных ресурсов и доступу к ним, а также провести исследование факторов информационных запросов, влияющих на качество поиска информационных ресурсов в хранилищах данных.

2. Исследовать частотные характеристики востребованности информационных ресурсов хранилища данных и изучить вопросы использования данных частотных характеристик в качестве параметров востребованности информационных ресурсов со стороны потребителей информации.

3. Разработать нечеткую классификацию информационных ресурсов на основе их востребованности потребителями информации, нечеткую модель группы потребителей на основе их интересов к информационным ресурсам и модель учета интересов потребителей информационных ресурсов в упорядочении индекс-каталогов.

4. Провести экспериментальное исследование основных теоретических положений диссертационного исследования.

Объект исследований: нечеткие модели динамического упорядочения информационных ресурсов в хранилищах данных с учетом их востребованности потребителями информации.

Методы исследования основываются на нечеткой математике, комбинаторике, теории информационных систем, методах классификации, информационного поиска, а также методах обработки результатов экспериментов.

Научная новизна работы заключается в следующем:

1. На основе сравнительного анализа известных подходов к хранению информационных ресурсов и доступу к ним, а так же анализа факторов информацион-

ных запросов, влияющих на качество поиска информации в хранилищах данных, показано, что учет интересов потребителей может существенно повысить качество выдачи информационных ресурсов.

2. Предложено использовать частотную характеристику информационного ресурса в качестве характеристики его востребованности со стороны потребителей информации и на ее основе проводить динамическое упорядочение индекс-каталогов.

3. Разработаны нечеткая классификация информационных ресурсов на основе их востребованности потребителями информации, нечеткая модель группы потребителей на основе их интересов к информационным ресурсам и модель учета интересов потребителей информационных ресурсов в упорядочении индекс-каталогов, которые совместно образуют механизм эффективного доступа к информационным ресурсам хранилища данных, учитывающий их востребованность со стороны потребителей информации.

4. Проведено экспериментальное исследование основных теоретических положений диссертации, для которого разработана математическая модель представления запросов потребителей информации и построен моделирующий программный комплекс, результаты проведенных экспериментов на котором подтвердили основные теоретические положения.

Основные положения, выносимые на защиту:

1. Нечеткая частотная характеристика информационного ресурса, которая представляет его востребованность со стороны потребителей информации и используется для динамического упорядочения индекс-каталогов.

2. Нечеткие классификации информационных ресурсов на основе их востребованности потребителями информации, нечеткая модель группы потребителей на основе их интересов к информационным ресурсам и модель учета интересов потребителей информационных ресурсов в упорядочении индекс-каталогов, которые совместно реализуют механизм эффективного доступа к информационным ресурсам хранилища данных, учитывающий их востребованность со стороны потребителей информации.

Практическая ценность диссертационного исследования состоит в том, чт разработанные нечеткие модели динамического упорядочения информационны ресурсов в хранилищах данных с учетом их востребованности потребителями информации позволяют реализовать механизм эффективного доступа разных катего рий потребителей к информационным ресурсам с учетом их интересов, которы может быть использован в электронных хранилищах информационных ресурсо различного назначения.

Достоверность результатов подтверждается корректным использование методов нечеткой математики, комбинаторики, теории информационных систем методов классификации, информационного поиска, а также методов обработки ре зультатов экспериментов.

Использование результатов работы. Результаты диссертационного исследования используются в ряде научно-исследовательских работ, выполненных в международной лаборатории ELDIC, и учебном процессе по дисциплине “Организация электронных архивов данных” магистерской программы “Интеллектуальные

системы" по направлению 230100 "Информатика и вычислительная техника" факультета автоматики и вычислительной техники Таганрогского технологического института Южного федерального университета.

Разработанный механизм доступа к информационным ресурсам хранилища данных, учитывающий их востребованность со стороны потребителей информации, реализован программно и использован при разработке электронной библиотеки международной лаборатории ELDIC, а также в научных исследованиях факультета автоматики и вычислительной техники Таганрогского технологического института Южного федерального университета.

Апробация результатов работы. Основные результаты работы неоднократно докладывались и обсуждались на конференциях и семинарах различного уровня, в том числе на IV Всероссийской научной конференции молодых ученых, аспирантов и студентов "Техническая кибернетика, радиоэлектроника и системы управления", Таганрог, 2006; Всероссийской научной школа-семинар студентов, аспирантов и молодых ученых "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки", Таганрог, 2007; Всероссийской научной школа-семинар студентов, аспирантов и молодых ученых "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки", Таганрог, 2008; V Всероссийской конференции студентов, аспирантов и молодых ученых Технологии Microsoft в теории и практике программирования, Таганрог, 2008; VI Всероссийской научной конференции представлены доклады и сообщения студентов, аспирантов и молодых ученых вузов России по информационным технологиям, системную анализу и управлению "Информационные технологии, системный анализ и управление", Таганрог, 2008; Известия ЮФУ, "Технические науки", Тематический выпуск "Интеллектуальный САПР", Таганрог, 2008; Всероссийской научной школа-семинар молодых ученых, аспирантов и студентов "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки", Таганрог, 2009.

Публикации. По материалам диссертации автором опубликовано 13 печатных работ, в том числе одна статья в издании из списка, рекомендованного ВАК, в которых отражены основные результаты диссертационного исследования.

Структура и объем работы. Материал основной части работы изложен на 144 страницах машинописного текста. Работа состоит из введения, четырех разделов, заключения и списка литературы из 132 наименований, содержит 53 рисунки, 21 таблицу и 2 приложения на 20 страницах.

Краткое содержание работы

Во введении обоснована актуальность проблемы, сформированы цели и задачи диссертационного исследования.

В первом разделе определены основные понятия, используемые в диссертационном исследовании, проведен анализ существующих подходов к созданию эффективных методов хранения ИР и доступа к ним. Рассмотрены и проанализированы иерархические и фасетные методы классификации и предложено учитывать интересы потребителей информации в упорядоченности динамических индекс-каталогах. Рассмотрены особенности организации индекс-каталогов, кластеризация потребителей ИР по интересам и учету этих интересов в упорядоченности индекс-каталогов. Пока-

зано, что в качестве характеристики, отражающей интересы потребителей информа ции, может быть использована частота обращений к ИР.

Во втором разделе проведен анализ влияния факторов информационных запросов на качество поиска и информационную выдачу. Здесь выявлена зависимость точности поиска от числа слов в поисковом запросе, а также проанализирована процедуры сравнения слов.

Для решения поставленных задач была построена асимптотическая функция точности поиска из предположения, что поисковое пространство содержит только релевантные запросам ИР. Перове допущение, если на поисковый запрос выдается один ИР, то предполагается, что точность поиска соответствует β, если n ИР, то точность соответствует α. Эти значения определены как крайние точки шкалы точности и из практического опыта для них выбраны следующие значения: $n = 30$, $\beta = 1, \alpha = 0{,}03$.

Второе допущение состояло в том, что точность поиска (y) связана с числом ИР (x) линейной зависимостью вида:

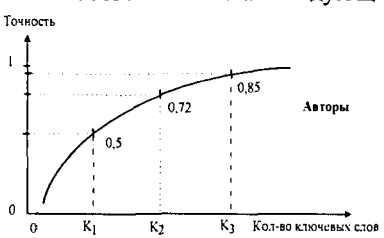
$$y = k \cdot x + b. \tag{1}$$

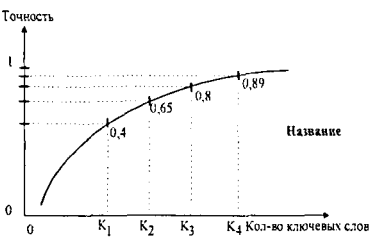
Для определения параметров k и b составлена система уравнений вида:

$$\begin{cases} \beta = k \times 1 + b; \\ \alpha = k \times n + b, \end{cases} \tag{2}$$

и для частного случая $n = 30$, $\beta = 1$, $\alpha = 0{,}03$ решения системы уравнений имеют следующий вид: $b = 1{,}03$ и $k = -0{,}03$.

Далее определялась точность поиска в зависимости от числа слов в поисковом запросе. Так, средние результаты точности по реквизиту “Авторы” и "Название" в зависимости от числа используемых ключевых слов в поисковом запросе представлены соответственно на следующих графиках.





Из анализа зависимости, который подтвержден экспериментами, следует, чт для обеспечения приемлемой точности поиска вполне достаточно в поисковом за просе использовать по два слова из каждого реквизита ИР.

Исследующее исследование точности связывалось с комбинаторной оценко числа сравниваемых букв словах. Пусть задан некоторый алфави $A = \{a_{r1}, a_{r2}, \dots, a_m\}$ и некоторая цепочка α из букв данного алфавита. Разделим це почку α на две подцепочки α_p - голова и α_q - хвост таким образом, что $n = p + q$.

Множество всех цепочек M длины n, которое можно построить из букв дан ного алфавита, имеет вид:

$$M = A^1 \cup A^2 \cup \dots \cup A^n, \tag{3}$$

мощность данного множества (3) без учета пустой цепочки представляется в виде:

$$|M| = m^1 + m^2 + \dots + m^n \tag{4}$$

Если обозначить через подмножество $M(a_{r1}a_{r2}\dots a_{rp})$ совокупность всех цепочек, имеющих одно и ту же голову $\alpha_p = a_{r1}a_{r2}\dots a_{rp}$, то все цепочки множества M, имеющие не пустой хвост цепочки, неразличимы по голове α_p.

Комбинаторно мощность множества $M(a_{r1}a_{r2}\dots a_{rp})$ определяется следующим образом:

$$\left|M(a_{r1}a_{r2}\dots a_{rp})\right| = m^0 + m^1 + \dots + m^q \quad (5)$$

При этом m^0 соответствует случаю пустого хвоста у головы $a_{r1}a_{r2}\dots a_{rp}$, m^1 соответствует однобуквенному хвосту, … и m^q соответствует хвосту из q букв.

Введем следующий коэффициент неразличимости цепочек:

$$K_{неразл} = \frac{Q}{N} \quad (6)$$

Здесь N соответствует числу всех цепочек до длины n включительно, а Q - числу неразличимых цепочек.

Неразличимую цепочку определим следующим образом. Выделим голову $a_{r1}a_{r2}\dots a_{rp}$ *из* p *символов некоторой цепочки* α. Очевидно, что *любая цепочка* $\beta \in M$, которая имеет длину меньше или равную p, является различимой по отношению к голове $a_{r1}a_{r2}\dots a_{rp}$, т.е. условие различимости записывается в виде: $|\beta| \le p$. В тоже время условие $|\beta| > p$ является противоположным и представляет условие неразличимости.

В диссертации показывается, что коэффициент неразличимости для цепочек из p символов определяется соотношением вида:

$$K_{неразл} = \frac{m^p \times (m^1 + m^2 + \dots + m^{n-p})}{m^1 + m^2 + \dots + m^n} \quad (7)$$

Из данного соотношения можно получить следующее выражение коэффициента неразличимости для голов цепочек из p символов:

$$K_{неразл} = \frac{m^p \times (m^{n-p} - 1)}{m^n - 1} = 1 - \frac{m^p - 1}{m^n - 1} \quad (8)$$

Далее в диссертации проводится сравнительный анализ комбинаторных оценок коэффициента неразличимости и результаты его экспериментального моделирования для русского языка ($m = 33$) в предположении, что длина сравниваемых слов не превышает 10. Сравнительный анализ теоретических и экспериментальных данных показывают практических полное их совпадение.

В третьем разделе определяются базовые теоретико-множественные понятия и соотношения, которые составляют основу динамического упорядочения ИР. Здесь разрабатываются нечеткая классификация ИР на основе их востребованности потребителями ИР, нечеткая модель группы потребителей на основе их интересов к ИР и модель учета интересов потребителей ИР в упорядочении индекс-каталогов, которые совместно образуют механизм эффективного доступа к хранилищу данных.

Пусть индекс-каталог содержит следующее множество ИР:

$$I = \{a_1, a_2, a_3, \dots, a_n\}, \quad (9)$$

где: a_j – имя ИР, а j – его порядковый номер.

Пусть за некоторый, достаточно длительный период времени T к *j-ому* ИР происходит обращения потребителей. Обозначим через f_t текущую частоту обращения к данному ИР в момент времени t. Отметим, что за период времени T текущая частота обращения к данному ИР может изменять свои значения. Обозначим

через f_{max} – максимальное значение текущей частоты обращений f_t за период T для такого ИР индекс-каталога, к которому идет наибольшее число обращений:

$$f_{max} = \max_{T}(f_t) \tag{10}$$

Введем понятие нормализованного веса ИР, который обозначим через w и определим следующим образом:

$$w = \frac{f_t}{f_{max}}. \tag{11}$$

Очевидно, что значение веса ИР всегда лежит в интервале $0 \le w \le 1$.

Введем для каждого ИР индекс-каталога характеристический вектор вида: $\langle a_i, M_i, w_i \rangle$, в котором a_i – имя i-го ИР, M_i – число обращений к нему за период T, w_i – вес ИР. Все характеристические вектора конкретного индекс-каталога соберем в одну таблицу, которую назовем характеристической таблицей векторов ИР данного индекс-каталога. Очевидно, что таблица характеристических векторов всегда отражает текущее состояние обращений к ИР.

Теперь рассмотрим нечеткую классификацию ИР на основе понятия нечеткой частоты востребованности. Для этого разобьем условно множество ИР на n подмножеств в соответствии с частотой обращения к ИР. Без потери общности рассуждения можно принять $n = 3$, что реально соответствует практическому случаю. Тогда множество ИР индекс-каталога разбивается на три подмножества A, B и C в соответствии со значением текущей частоты обращения к ИР. Будем считать, что подмножество A, B и C содержат ИР, к которым обращаются редко, средне и часто. Данные подмножества можно представить в виде:

$$\begin{aligned} A &= \{a_i;\ f(a_i) \in [r_1 .. r_2)\}; \\ B &= \{b_i;\ f(b_i) \in [r_2 .. r_3)\}; \\ C &= \{c_i;\ f(c_i) \in [r_3 .. r_4]\}; \\ & r_1 < r_2 < r_3 < r_4. \end{aligned} \tag{12}$$

Здесь параметрические переменные r_1, r_2, r_3 и r_4 задают разбиение всего интервала нормализованных частот на непересекающиеся подинтервалы.

В процессе обращения потребителей информации к индекс-каталогу частоты обращения к ИР могут меняться и поэтому принадлежность элементов подмножеств может перераспределяться в соответствии с представленной ниже схемы.

Если множество ИР индекс-каталога рассматривать как базовую шкалу, то на ней можно построить нечеткие множества ИР вида “Редко”, “Средне” и “Часто”:

$$\begin{aligned} &\text{"}Редко\text{"} = \{a_i;\ \mu_i(a_i); w_i \in [r_1 .. r_2)\}; \\ &\text{"}Средне\text{"} = \{b_i;\ \mu_i(b_i); w_i \in [r_2 .. r_3)\}; \\ &\text{"}Часто\text{"} = \{ac_i;\ \mu_i(c_i); w_i \in [r_3 .. r_4]\}; \end{aligned} \tag{13}$$

где: ИР a_i, веса которых $w_i \in [r_1 .. r_2)$ составляют множество-носитель нечеткого множества “Редко”; ИР b_i, веса которых $w_i \in [r_2 .. r_3)$ составляют множество-носитель нечеткого множества “Средне”; ИР c_i, веса которых $w_i \in [r_3 .. r_4)$ составляют множество-носитель нечеткого множества “Часто”.

Свяжем частоту обращения к ИР с лингвистической переменной (ЛП), которую определим следующим образом:

$$\langle \alpha, T, X, G, M \rangle, \quad (14)$$

где: α – имя ЛП, в качестве которой выступает нормализованная частота обращения к ИР; T – терм-множество ЛП, которое представляет следующее множество ее значений: T = {"очень редко", "редко", "средне", "часто", "очень часто"}; X – базовое множество, которое является областью определения термов и представляет собой интервал значений [0..1] нормализованных частот обращений к ИР; G – процедура образования новых термов с помощью связок "и", "или" и модификаторов типа "очень", "не", "очень" и др.; M – процедура задания на $X = [0..1]$ нечетких подмножеств A_1 = "очень редко", A_2 = "редко", A_3 = "средне", A_4 = "часто" и A_5 = "очень часто" а также нечетких множеств для термов из $G(T)$ в соответствии с правилами трансляции нечетких связок и модификаторов "и", "или", "не", "очень", а также операций над нечеткими множествами.

Вместе с рассмотренными выше базовыми значениями ЛП α, которое представляет терм-множество T ={"очень редко", "редко", "средне", "часто", "очень часто"}, можно в последующем определять значения ЛП α в виде нечетких чисел.

Очевидно, что используя частотные свойства ИР, можно их учитывать при управлении содержимым ХД. Так, невостребованные ресурсы могут удаляться, к некотором ресурсам может применяться условное удаление, наиболее востребованным ресурсам может приписываться особый статус, а в случае уменьшения их востребованности к ним могут применяться не жесткие санкции по удалению. Учитывая данные обстоятельства, можно построить правила редактирования ХД.

Рассмотрим нечеткие модели групп потребителей на основе интересов к ИР. В общем случае все группы потребителей ИР можно представить в виде множества:

$$Us = \{Name_1, Name_2, \dots, Name_j, \dots, Name_n\}, \quad (15)$$

в котором список членов обобщенной группы имеет вид:

$$Name_j = \{name_1^j, name_2^j, \dots, name_i^j, \dots, name_m^j\}. \quad (16)$$

Введем для каждого члена некоторой группы $Name$ характеристику p, принимающую значения на интервале [0..1], и свяжем ее с интересом потребителя. Тогда группу $Name$ можно представить в виде нечеткого множества следующего вида:

$$\widetilde{Name} = \{(Name_1, p_1), (Name_2, p_2), \dots, (Name_i, p_i), \dots, (Name_m, p_m)\} \quad (17)$$

С учетом выше рассмотренного нечеткое множество всех потребителей ИР можно представить в виде:

$$\widetilde{Us} = \{\widetilde{Name}_1, \widetilde{Name}_2, \dots, \widetilde{Name}_j, \dots, \widetilde{Name}_n\}. \quad (18)$$

Определим для нечеткого множества группы потребителей $Name$ абсолютный суммарный вес интересов в виде суммы интересов всех ее членов:

$$P_{name} = \sum_{i=1}^{m}(p_i) \quad (19)$$

По этой аналогии введем абсолютный суммарный вес интересов всего множества потребителей Us:

$$P_{Us} = \sum_{j=1}^{n}\left(P_{Name_j}\right) \quad (20)$$

Проведем нормирование веса отдельной группы потребителей $Name_i$:

$$k_{Name_i} = \frac{P_{Name_i}}{P_{Us}}, \quad (21)$$

а для отдельного *j-го* члена $Name_i$ группы потребителей его нормированный вес будет иметь вид:

$$k(Name_{ij}) = \frac{P_{Name_{ij}}}{P_{Us}}, \quad (22)$$

Таким образом, в формализованной нечеткой модели группы потребителей основу представляет интегральный вес интересов к ИР.

Теперь рассмотрим, каким образом следует учитывать этот интерес групп потребителей при упорядочении ИР в индекс-каталоге. Пусть существует *M* видов ИР некоторой предметной области, с которой работают все группы потребителей. Не трудно заметить, что у одной и той же группы потребителей степень востребованности разных видов ИР отличается. Очевидно, что для групп потребителей, интересы которых выше, должно отдаваться предпочтения и их интересы должны учитываться в первую очередь. Учтем этот интерес на основе текущей частоты обращения к ИР. Выделим некоторый отдельный ИР в индекс-каталоге и будем считать, что к нему за некоторый интервал времени происходит обращения всех групп потребителей. Очевидно, что частота обращений к данному ИР будет складываться из частот обращений каждого потребителя в виде:

$$f_t^{Name} = \sum_{i=1}^{n} f_t^{Name_i}, \quad (23)$$

где каждый член суммы определяется следующим образом:

$$f_t^{Name_j} = \sum_{i=1}^{m_j} f_t^{name_i^j}. \quad (24)$$

здесь m_j число членов j-ой группы потребителей.

Теперь введем текущую частоту обращения к ИР группы потребителей с учетом ее интересов, получим:

$$\tilde{f}_t^{Name_i} = k_{Name_i} * f_t^{Name_i}. \quad (25)$$

С учетом данного выражения общая частота обращения к ИР может быть представлена в виде:

$$\tilde{f}_t^{Name} = \sum_{i=1}^{n} \tilde{f}_t^{Name_i}. \quad (26)$$

Именно эта частота является характеристикой востребованности информационного ресурса и должно учитываться при упорядочении ИР в индекс-каталоге.

Далее общие теоретические результаты поясним на примере конкретных групп потребителей «Ученики» - *Pup*, «Студенты»- *St*, «Инженеры» - *Eng*, «Учителя» - *Te* и «Преподаватели»- *Pr*. Так, для этих групп интересы потребителей их интересы выражаются следующими соотношениями:

$$\tilde{f}_t^{Pup} = k_{Pup} * f_t^{Pup}; \quad (27)$$

$$\tilde{f}_t^{St} = k_{St} * f_t^{St}; \quad (28)$$

$$\tilde{f}_t^{Eng} = k_{Eng} * f_t^{Eng}; \quad (29)$$

$$\tilde{f}_t^{Te} = k_{Te} * f_t^{Te}; \quad (30)$$

$$\tilde{f}_t^{Pr} = k_{Pr} * f_t^{Pr}, \quad (31)$$

а общая частота обращения к ИР с учетом интересов всех потребителей примет вид:

$$\tilde{f}_t^{Us} = \tilde{f}_t^{Pup} + \tilde{f}_t^{St} + \tilde{f}_t^{Eng} + \tilde{f}_t^{Te} + \tilde{f}_t^{Pr}. \quad (32)$$

Экспериментальные данные по данным групп потребителей представлены в табл. 1 и 2.

Таблица 1. Нормированные весовые коэффициенты для членов групп потребителей

User	Pup	St	Eng	Te	Pr
i	$k(pup_i)$	$k(st_i)$	$k(eng_i)$	$k(te_i)$	$k(pr_i)$
1	0.028	0.030	0.028	0.022	0.025
2	0.030	0.023	0.023	0.023	0.028
3	0.024	0.032	0.020	0.031	0.029
4	0.022	0.029	0.030	0.029	0.031
5	0.017	0.025	0.018	0.025	0.026
6	0.015	0.015	0.032	0.028	0.024
7	0.012	0.013	0.022	0.031	0.033
8	0.023	0.017	0.025	0.032	0.031

Таблица 2. Нормированные весовые коэффициенты для различных групп потребителей

k_{Pup}	k_{St}	k_{Eng}	k_{Te}	k_{Pr}
0,170	0,184	0,198	0,221	0,227

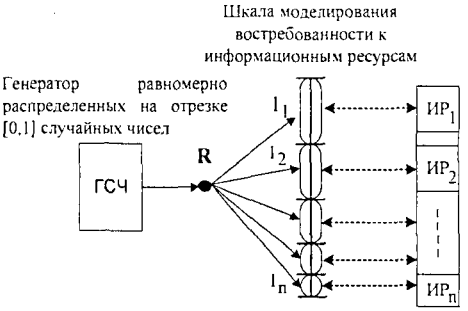
Таким образом, в разделе разработаны: нечеткая классификация ИР на основе их востребованности, нечеткая модель группы потребителей на основе их интересов к ИР и модель учета интересов потребителей ИР в упорядочении индекс-каталогов, которые совместно образуют механизм эффективного доступа к ХД.

В четвертом разделе описывается структура моделирующего программного комплекса, схемы экспериментов и их результаты.

Ниже приводится общая схема экспериментов,

в которой генератор запросов моделирует запросы групп потребителей; блок обработки данных отвечает за результаты экспериментов и управляет динамическим упорядочением индекс-каталогов, что показано на схеме в виде обратных связей от блока обработки данных к индекс-каталогам.

Для моделирования интересов потребителей и формирования от них запросов разработан механизм, которой представлен на следующей схеме:



Здесь ГСЧ – генератор случайных чисел, который формируют в конкретный момент времени случайное число R ∈ [0..1] (Randomize). Совокупность сгенерированных случайных чисел имеет равномерную плотность распределения на отрезке [0..1].

Разобьем отрезок [0,1] на множество L непересекающихся подотрезков:

$$L = \{l_1, l_2, \dots, l_n\}, \tag{33}$$

длины которых определим следующим образом:

$$\begin{gathered} l_1 = [0..a_1] = a_1 - 0, \\ l_2 = (a_1..a_2] = a_2 - a_1, \\ \dots\dots\dots\dots\dots \\ l_i = (a_{i-1}..a_i] = a_i - a_{i-1}, \\ \dots\dots\dots\dots\dots \\ l_n = (a_{n-1}..1] = 1 - a_{n-1}. \end{gathered} \tag{34}$$

Из способа построения множества отрезков L справедливо соотношение:

$$(\sum_{i=1}^{n} l_i) = 1; \tag{35}$$

Введем следующую схему интерпретации запроса к ИР индекс-каталога. Пусть генератор случайных чисел генерирует некоторое число R. Если $R \in l_i$, то будем считать, что произошел запрос к *i-ому* ИР. Пусть за некоторый период времени T генератор случайных чисел выдал N чисел, а m чисел попали в отрезок l_i. Тогда востребованность *i-го* ИР будем представлять в виде:

$$f_i = \frac{m}{N}. \tag{36}$$

Если $T \to \infty$, то в силу законов больших чисел востребованность f_i будет стремиться к постоянному значению и зависеть только от длины отрезка l_i. Поэтому моделирование отличающихся востребованностей ИР в индекс-каталоге можно реализовать через формирование различных длин отрезков множество L. Очевидно, что при одинаковых длинах отрезков востребованности всех ИР индекс-каталога одинаковые. Если $l_i < l_j$, то из двух ИР согласно схеме $ИР_j$ более востребован, чем $ИР_i$.

В наших экспериментах формирование разных востребованностей ИР (длин отрезков L) основано на следующих соотношениях:

$$l_i = l_{i-1} + d; \tag{37}$$

$$l_{i-1} < l_i, \tag{38}$$

где d является некоторой константой.

Назовем следующую величину:

$$k = \frac{l_n}{l_1}, \tag{39}$$

коэффициентом различения отрезков множества L. Здесь l_n – длина *n-го* подотрезка, а l_1 – длина первого подотрезка. Теперь задача заключается в том, что бы для некоторого n и k построить множество отрезков L.

Опуская математические выкладки, приведем общее решение, имеющее вид:

$$\begin{cases} l_1 = \frac{n-1}{k-1} * d; \\ d = \frac{2(k-1)}{n*(n-1)*(k+1)} \end{cases} . \quad (40)$$

В случае $k = 3$ данные решения приводятся к виду:

$$\begin{cases} l_1 = \frac{n-1}{2} * d; \\ d = \frac{1}{n*(n-1)} \end{cases} \quad (41)$$

С целью изучения текущего состояния таблицы характеристических векторов ИР проводилось моделирование динамики упорядочения ИР. Для этого был создан индекс-каталог, для него построена шкала востребованности (*n=75, k=3)*, а также вычислены значения параметров $d \approx 0{,}00018$ и $l_1 \approx 0{,}00666$. Ниже приводится пример фрагмента характеристической таблицы, в которую сведены экспериментальные данные по востребованности ИР при разном числе запросов.

Таблица характеристических векторов ИР для индекс-каталога (n=75, k=3)

№ ПП	Имя ИР	Длины отрезки		M=3334	M=5451	M=7582	M=28045
	$ИР_i$	l_i		M_i			
1	$ИР_{13}$	l_{13}	0,01998	100	130	200	1000
2	$ИР_3$	l_3	0,0198	98	128	198	987
3	$ИР_{33}$	l_{33}	0,01962	97	127	196	966
4	$ИР_{12}$	l_{12}	0,01944	95	125	194	941
5	$ИР_{43}$	l_{43}	0,01926	94	124	190	921
…	…	…	…	…	…	…	…
22	$ИР_{22}$	l_{22}	0,01602	65	96	131	584
23	$ИР_{52}$	l_{52}	0,01584	64	95	128	562
24	$ИР_{11}$	l_{11}	0,01566	63	94	126	532
…	…	…	…	…	…	…	…
55	$ИР_{50}$	l_{50}	0,01188	28	58	81	156
56	$ИР_{29}$	l_{29}	0,0117	27	56	79	145
57	$ИР_{18}$	l_{18}	0,01152	25	55	77	132
58	$ИР_{47}$	l_{47}	0,01134	24	…	75	125
59	$ИР_{36}$	l_{36}	0,01116	17	33	74	110
60	$ИР_{25}$	l_{25}	0,01098	19	30	68	99
61	$ИР_{54}$	l_{54}	0,0108	23	28	64	95
…	…	…	…	…	…	…	…
65	$ИР_{39}$	l_{39}	0,00792	7	23	31	35
66	$ИР_{13}$	l_{13}	0,00774	9	10	29	29
67	$ИР_{61}$	l_{61}	0,00756	1	…	28	24
68	$ИР_5$	l_5	0,00738	0	6	26	17

…	…	…		…	18	10	…
74	$ИР_{29}$	l_{29}	0,00684	0	0	7	3
75	$ИР_{14}$	l_{14}	0,00666	0	0	14	2

Нетрудно заметить, что при увеличении числа запросов достоверность востребованности ИР возрастает и уже при М=7582 становится полностью определенной. Так, при числе обращений М=3334 востребованности ИР, занявших последние строки таблицы с номерами 59 – 75, доверять нельзя. При М=5451 уже востребованности ИР, занявших последние строки таблицы с номерами 65–75, доверять нельзя, а при М=7582 востребованности всех ИР можно уже доверять.

Для экспериментального моделирования интересов потребителей и учету этих интересов при упорядочении тематического индекс-каталога были построены весовые коэффициенты интересов разных групп потребителей, которые представлены следующей таблицей.

Таблица характеристических весовых коэффициентов групп потребителей							
		Students	Pupils	Engineers	Teachers	Professors	User
№	ИР	w_{st_i}	w_{pup_i}	w_{eng_i}	w_{te_i}	w_{pr_i}	w_{Us_i}
1	A	1,000	0,727	0,833	0,943	0,720	0,845
2	B	0,950	0,673	0,583	0,857	1,000	0,813
3	G	0,717	1,000	0,771	0,714	0,560	0,752
4	K	0,200	0,855	0,958	0,800	0,920	0,747
5	M	0,400	0,909	1,000	0,571	0,800	0,736
6	J	0,450	0,618	0,479	1,000	0,600	0,629
7	H	0,633	0,800	0,542	0,429	0,440	0,569
8	I	0,600	0,455	0,729	0,486	0,160	0,486
9	D	0,883	0,509	0,188	0,343	0,200	0,425
10	C	0,917	0,091	0,333	0,286	0,320	0,389
11	L	0,283	0,236	0,396	0,514	0,480	0,382
12	E	0,817	0,309	0,125	0,171	0,360	0,356
13	F	0,750	0,345	0,042	0,086	0,120	0,269
14	N	0,100	0,400	0,271	0,257	0,280	0,262
15	O	0,033	0,018	0,021	0,029	0,040	0,028

Здесь каждому ИР поставлен в соответствие весовой коэффициент в каждой отдельной категории потребителей, а также интегральный весовой коэффициент для категории *User*, который учитывает запросы всех категорий потребителей. Весовые коэффициенты *i*-го ИР отдельных категорий потребителей обозначены через $w_{st_i}, w_{pup_i}, w_{eng_i}, w_{te_i}, w_{pr_i}$, а через w_{Us_i} - интегральный весовой коэффициент всех категорий потребителей.

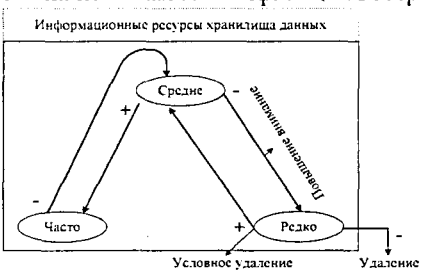
Вид текущего динамического упорядочения ИР, которое учитывает интегральный интерес различных групп потребителей, приведен ниже в таблице:

Таблица упорядоченности ИР, учитывающая интегральный интерес групп потребителей				
ИР	Востребованность "*w*"	Функция принадлежности элементов нечеткого множества *«Часто»* $\mu_{часто}$	Функция принадлежности элементов нечеткого множества *«Средне»* $\mu_{средне}$	Функция принадлежности элементов нечеткого множества *«Редко»* $\mu_{редко}$
A	0.845	0.845	0,31	0,155
B	0.813	0.813	0,374	0,187

G	0.752	0.752	0,496	0,248
M	0.747	0.747	0,506	0,253
K	0.736	0.736	0,528	0,264
J	0.629	0.629	0,742	0,371
H	0.569	0.569	0,862	0,431
I	0.486	0.486	0,972	0,514
D	0.425	0.425	0,85	0,574
C	0.389	0.389	0,778	0,575
L	0.382	0.382	0,764	0,618
E	0.356	0.356	0,712	0,644
F	0.269	0.269	0,538	0,731
N	0.262	0.262	0,524	0,738
O	0.028	0.028	0,056	0,912

В диссертации представлены подробно результаты всех экспериментов по динамическому упорядочению индекс-каталогов.

Управление содержимым ХД необходимо для минимизации его объема при сохранении качества обслуживания различных категорий потребителей, поэтому в основу управления положена нечеткая классификация востребованности ИР:



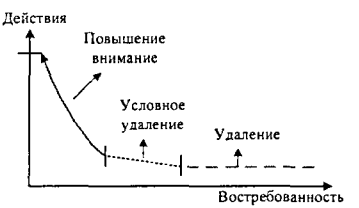
Для управления ИР ХД построены правила работы, основу которых образуют составные высказывательные формы:

IF $L_{St_i}\&L_{pup_i}\&L_{eng_i}\&L_{te_i}\&L_{pr_i} \le$ Редко **THEN** $ИР_i$ удалить из ХД;

IF $L_{St_i}\&L_{pup_i}\&L_{eng_i}\&L_{te_i}\&L_{pr_i} \ge$ Редко **THEN** $ИР_i$ условно удалить из ХД; (42)

IF Редко$<L_{St_i}\&L_{pup_i}\&L_{eng_i}\&L_{te_i}\&L_{pr_i} \le$ Средне **THEN** повысить внимание к ИР.

Правила учитывают вычисленные значения ЛП, а их графическая интерпретация имеет вид:



Очевидно, что по аналогии с рассмотренной схемой можно построить более сложные стратегии управления, например, учитывающие историю востребованности ресурса и формирующие прогноз на востребованность.

В заключении формулируются основные результаты диссертационного исследования.

Список опубликованных работ по теме диссертации в издании ВАК

1. Тиек Ленг. Использование структурного подхода при разработке систем интеграции информационных ресурсов. // Известия ЮФУ. Технические науки. Тематический выпуск «Интеллектуальный САПР» – Таганрог. Изд-во ТТИ ЮФУ, 2008, № 9 (86) – с. 171–175.

Основные публикации по теме диссертации

2. Тиек Ленг. Использование платформы XML для представления информация в электронных библиотеках. // Труды IV Всероссийской научной конференции молодых ученых, аспирантов и студентов «Техническая кибернетика, радиоэлектроника и системы управления» – Таганрог: ТРТУ, 2006г. – с. 43–46.
3. Тиек Ленг. Интеграция распределенных данных для создания развитой динамической библиотеки поддержки знаний. // Всероссийская научная школа-семинар студентов, аспирантов и молодых ученых «Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки» – Таганрог: ТТИ ЮФУ, 2007г. – с. 50–56.
4. Тиек Ленг. Теоретическое построение интеллектуальной системы поиска в хранилище данных. // Известия ЮФУ. Технические науки. Тематический выпуск «Интеллектуальный САПР». – Таганрог: Изд-во ТТИ ЮФУ, 2007г. – № 2 (77). – с. 116–119.
5. Тиек Ленг. Распределённая поисковая система сбора и хранения информации Технологии Microsoft в теории и практике программирования: труды V-ой Всероссийской конференции студентов, аспирантов и молодых ученых. // Южный регион, Таганрог, 13-14 марта 2008 г. – Таганрог: Изд-во ТТИ ЮФУ, 2008. – с. 59–62.
6. Тиек Ленг. Развитие динамической библиотеки поддержки знаний при интегрировании распределенных данных. // Технологии Microsoft в теории и практике программирования: труды V-ой Всероссийской конференции студентов, аспирантов и молодых ученых. Южный регион, Таганрог, 13-14 марта 2008г. – Таганрог: Изд-во ТТИ ЮФУ, 2008. – с. 80–85.
7. Тиек Ленг. Интеграция неоднородных информационных электронных ресурсов в лаборатории Eldic. // Всероссийская научная школа-семинар студентов, аспирантов и молодых ученых «Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки» – Таганрог: ТТИ ЮФУ, 2008г. – с. 99–104.
8. Тиек Ленг. Разработка глобального хранилища данных и средств интеграции в него информационных ресурсов для системы интеграции // Всероссийская научная школа-семинар молодых ученых, аспирантов и студентов "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки" – Таганрог: Изд-во ТТИ ЮФУ, 2009. – с. 79–82.

Технологический институт Южного федерального университета в г. Таганроге

347928, Ростовская область
г. Таганрог, пер. Некрасовский 44.