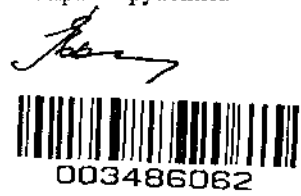
На правах рукописи



ВАСИНА ЕЛЕНА НИКОЛАЕВНА

ИССЛЕДОВАНИЕ И РАЗРАБОТКА МОДЕЛЕЙ И АЛГОРИТМОВ СТРУКТУРНО-ЛОГИЧЕСКОЙ ОБРАБОТКИ ИНФОРМАЦИИ В ДОКУМЕНТАЛЬНЫХ ИНФОРМАЦИОННО-АНАЛИТИЧЕСКИХ СИСТЕМАХ

05.25.05 – Информационные системы и процессы,
правовые аспекты информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Москва 2009

Работа выполнена на кафедре Информатики ГОУ ВПО «Российская экономическая академия им. Г.В. Плеханова».

Научный руководитель:	доктор технических наук, профессор Максимов Николай Вениаминович
Официальные оппоненты:	доктор технических наук, доцент Косяченко Станислав Анатольевич
	кандидат физико-математических наук Куприянов Вячеслав Михайлович
Ведущая организация:	ФГУ "Федеральный институт промышленной собственности Федеральной службы по интеллектуальной собственности, патентам и товарным знакам"

Защита диссертации состоится «16» декабря 2009 г. в 14-00 часов на заседании диссертационного совета Д 212.198.02 в Российском государственном гуманитарном университете по адресу 125993, г. Москва, Миусская пл., 6.

С диссертацией можно ознакомиться в библиотеке Российского государственного гуманитарного университета.

Автореферат разослан «16» ноября 2009 г.

Ученый секретарь
диссертационного совета Меркулов В.Н.

Общая характеристика работы

Актуальность избранной темы. Современный уровень развития информационных технологий делает доступными в реальном масштабе времени информационные ресурсы самого разного объема и содержания. Для облегчения работы с большими объемами информации разрабатываются разнообразные формы и способы ее представления, а также методы поиска, что выражается, например, в создании систем, индивидуально настраиваемых самим пользователем.

Принципиально важным фактором, определяющим направление развития современных информационных систем, является то, что взаимодействие пользователей с информационными ресурсами происходит в режиме «информационного самообслуживания», когда пользователь, по существу, уже не разделяет свою деятельность на информационную и основную.

Соответственно, тенденции развития документальных АИС заключаются в постепенном расширении традиционных функций и активном подключении к поисковым механизмам аналитических возможностей, т.е. в переходе к документальным информационным системам следующего поколения – интегральным информационно-аналитическим системам, которые сочетают функции создания базы данных, анализа ее лексического и документального содержания, синтеза и оптимизации лингвистических структур (словарей, рубрикаторов, тезаурусов), совместно с БД образующих информационную модель предметной области. Это означает, что пользователь создает по существу новый, проблемно-ориентированный, самостоятельно обновляемый и пополняемый информационный ресурс, включающий помимо подборок документов также и метаинформацию.

В связи с этим проблема исследования и моделирования как процессов информационного поиска в документальных информационно-аналитических системах, так и методов и алгоритмов построения средств, формирующих информационное пространство пользователя согласно его потребностям, является актуальной.

Степень разработанности проблемы. Проблемам моделирования поисковых процессов в информационных системах посвящены труды зарубежных ученых Chen Hsinchun, Salton G., Rijsbergen C.J.. Среди отечественных ученых, труды которых могут рассматриваться в качестве теоретической базы диссертации, выделяются: Белоногов Г.Г., Гиляревский Р.С., Романенко А.Г., Попов И.И., Максимов Н.В.

Современные достижения информационных технологий ставят новые задачи в области развития возможностей информационных систем, поэтому дальнейшие исследования данного научного направления представляются целесообразными.

Объектом исследования являются процессы автоматизированного поиска и анализа документальных баз данных, определяемых как машиночитаемые массивы информации, представленной в различной форме и на различном уровне (в том числе в виде комплекса баз данных первичной, вторичной и справочной информации), и рассматриваемых совместно со средствами доступа к ним.

Предметом исследования являются:

– комплекс лингвистических и технологических средств автоматизированных информационно-поисковых систем, обеспечивающих эффективность процессов поиска информации в документальных БД;

– технологии и алгоритмы управления информационными ресурсами, организующие информационное пространство пользователя.

Целью исследования является разработка комплекса моделей, алгоритмов, методов и средств систематизации документальной информации, ориентированных на совершенствование технологий и механизмов поиска информации в документальных информационных ресурсах, а также анализа структуры и динамики предметных областей.

Данная цель конкретизируется следующими **задачами:**

- определение основных принципов функционирования АИПС, ориентированных на задачи анализа информационных потоков;

- системный анализ взаимосвязи информационных объектов в процессах генерации и поиска информации;

- определение понятия и построение модели интегрального рабочего пространства пользователя;

- разработка модели когнитивного рубрикатора предметной области, как основного компонента рабочего пространства;

- разработка программных средств поддержки когнитивного рубрикатора пользователя;

- разработка метода автоматической классификации документов, основанного на применении когнитивного рубрикатора.

Методы исследования. Основные результаты получены и обоснованы с использованием методов теории вероятностей, теории множеств, линейной алгебры, системного анализа и компьютерного моделирования.

Экспериментально-статистической базой исследования послужили базы данных реферативно-библиографической информации ВИНИТИ РАН «Информатика», ВНТИЦентра «Информационные карты НИР и ОКР» и «Информационные карты диссертаций».

Нормативную базу исследования составили такие стандарты, как ГОСТ 7.0-99 Система стандартов по информации, библиотечному и издательскому делу; ГОСТ 7.74-96 Информационно-поисковые языки, термины и определения; ГОСТ 7.77-98 СИБИД Межгосударственный рубрикатор научно-технической информации. Структура, правила использования и ведения.

Научная новизна работы.

Разработаны модели и алгоритмы структурно-логической обработки информации, основанные на введенных понятиях рабочего пространства пользователя и когнитивного рубрикатора, обеспечивающих управляемую навигацию в локальных и распределенных информационных ресурсах.

Обоснована структура когнитивного рубрикатора как операционного средства рабочего пространства пользователя, интегрально отражающего видение предметной области на знаковом, понятийном и предметном уровнях.

На защиту выносятся следующие положения:

- понятие рабочего пространства пользователя, включающего информационные и процедурные компоненты, управляющие навигацией в локальных и распределенных документальных информационных ресурсах;
- понятие когнитивного рубрикатора, включающего систематическую и объектную составляющие и динамически отражающего когнитивное состояние пользователя по отношению к состоявшемуся знанию;
- модель когнитивного рубрикатора как операционного средства рабочего пространства пользователя, интегрально связывающего представления пользователя с информационными ресурсами предметной области на знаковом, понятийном и предметных уровнях;
- математическая модель классификации текстовых документов, динамически соотносящая найденные документы с разделами когнитивного рубрикатора пользователя.

Теоретическая и практическая значимость работы. Отдельные положения работы представляют собой вклад в теорию и практику информационного поиска, использованы при разработке конкретных прикладных программных комплексов управления документальными информационными ресурсами на примере реализации подсистем документальной информационно-аналитической системы xIRBIS[1] и могут быть рекомендованы к дальнейшему применению при разработке и развитии документальных информационных систем.

Отдельные положения могут быть использованы в теоретических курсах и лабораторных практикумах в учебном процессе вузов при подготовке бака-

[1] Документальная информационно-аналитическая система xIRBIS - программа для ЭВМ. Свидетельство №2008611511 от 25.03.2008г. Государственный реестр программ для ЭВМ, 2008г.

лавров, дипломированных специалистов и магистров по специальностям «Информационные системы (по областям)» и «Прикладная информатика (по областям)».

Внедрение результатов. При непосредственном участии автора разработана и применяется для создания промышленных информационных ресурсов документальная ИАС xIRBIS.

Результаты диссертационной работы внедрены в ВИНИТИ РАН, ИНИОН РАН, ВНТИЦентре РФ.

Публикации и апробация работы. По материалам исследований опубликовано 11 печатных работ, в том числе три работы в издании, входящем в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук, утвержденный Высшей аттестационной комиссией Министерства образования и науки Российской Федерации.

Результаты работы докладывались на следующих научных конференциях:

Международная конференция под эгидой международной федерации по информации и документации (МФД) – НТИ 96. Информационные продукты, процессы и технологии, Москва, 20-21 ноября 1996;

5-я международная конференция НТИ-2000. Информационное общество, информационные ресурсы и технологии телекоммуникации, Москва, 22-24 ноября 2000 г.";

Научно-практическая конференция «Информационные технологии в экономике XXI века», посвященная 100-летию РЭА им. Г.В. Плеханова, Москва, февраль 2006 г.;

7-я Международная конференция. НТИ-2007;

Научная сессия МИФИ-2008. 25-27 янв. 2008;

Научная сессия МИФИ-2009. XXIII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании».

Диссертационное исследование соответствует паспорту специальности 05.25.05 – Информационные системы и процессы, правовые аспекты информатики, пункту 1: Методы и модели описания, оценки, оптимизации информационных процессов и информационных ресурсов, а также средства анализа и выявления закономерностей в информационных потоках.

Объем и структура диссертации. Диссертация состоит из введения, четырех глав, заключения, библиографического списка и приложений. Диссертация содержит 10 таблиц и 47 рисунков. Общий объем работы составляет 139 страниц машинописного текста.

Содержание работы

Во введении обоснована актуальность рассматриваемой проблемы, определены цель, задачи, предмет, объект исследования, раскрыта его теоретическая и методологическая база, сформулирована научная новизна, теоретическая и практическая значимость.

В первой главе определены структурно-методологические основы информационно-поисковых систем, рассматриваемых как составляющая совокупной системы основной/информационной деятельности в процессах создания нового знания. Принимается[2], что формализованная схема научного поиска - процесса генерирования информации с использованием ранее полученных знаний с точки зрения общей теории систем включает следующие этапы:

1) поиск и извлечение информационных блоков из информационной среды;

2) упорядоченная или случайная комбинаторная проверка ценности этих блоков;

3) расширение знаний за счет тех комбинаций информационных компонентов, которые образуют целостную систему понятий, и их использование для решения задачи основной деятельности;

[2] *Дружинин В.В.* Проблемы системологии. / Дружинин В.В., Конторов Д.С. – М.: Советское радио, 1976. – 296 с.

4) публикация информации – представление нового «личного» знания субъекта в такой форме, которая обеспечит его «узнаваемость» и, соответственно, повторное использование внутри и вне когнитивной системы.

Если в процессе генерации новой информации используется АИПС, то её работа, рассматриваемая как замещающая часть соответствующего участка основной деятельности, будет включать следующие основные функции:

1) поиск - процесс отбора из информационных ресурсов документов, каждый из которых представляет, по крайней мере, один информационный компонент или его образ;

2) комбинаторное построение на основе множества характеристических признаков кластеров информационных компонентов и определение степени «целостности» этих кластеров уже как новых информационных компонентов;

3) упорядочение кластеров в порядке убывания их «ценности» с целью сокращения объема просматриваемой субъектом выборки в предположении, что мера ценности соответствует вероятности содержания в кластере искомого нового.

При этом, применение системного подхода[3] к построению комбинаций информационных компонентов, который любой объект представляет как систему в системе объектов того же рода, позволит:

- представить объект как совокупность типизированных элементов, связанных некоторыми отношениями, в совокупности образующими единство, для которого характерно появление свойств, не присущих составляющим;
- представить систему этих однородных объектов в виде классификации, что дает возможность выделять в явной форме, в том числе и новые характеристические признаки, определять способы выделения подсистем, а на основе свойств соответствия и симметрии обнаруживать связи с другими системами классификации.

[3] Урманцев Ю.А. Общая теория систем: Состояние, приложения и перспективы развития. Сборник «Система, Симметрия, Гармония", – М., Мысль, 1988, с.38-124

Именно этот подход методологически связывает относительно самостоятельные и, в тоже время, взаимообуславливающие объекты и процессы основной и информационной деятельности в цикле генерации новых знаний.

Процесс поиска является итеративным, каждая итерация которого включает два действия: 1) построение кластера документов и 2) построение терминологической системы – некоторого явного представления контекста этого кластера документов.

На уровне интерфейса среди средств поиска и работы с документами должны быть выделены технологические объекты и инструменты, чтобы облегчить пользователю переключение с задачи своей информационно-поисковой деятельности (сбора информации для решения задачи) на «вспомогательную» информационно-управляющую – оценку своих поисковых действий и состояний.

Такими объектами могут быть словари поисковой системы, тематические словники, тезаурусы, представляющие информативную лексику предметной области. Эти объекты, являясь технологически вспомогательными, используются на разных этапах поиска и обеспечивают возможность более или менее адекватного выражения информационной потребности пользователя. Однако эффективность их использования для отражения индивидуальных особенностей информационной потребности достаточно низка, поскольку, вследствие усредненной природы, они представляют предметную область в целом.

Схема и механизмы поиска в диалоговой АИПС должны строиться в предположении, что любая нетривиальная реальная информационная потребность не может быть удовлетворена одним или несколькими сразу найденными документами, а требует проведения серии поисков и выделения полезных фрагментов информации на каждой стадии развития запроса.

Объект, хранящий информацию о процессе поиска, имеет линейную структуру и в различных АИПС носит разные названия – «протокол поиска», «история поиска» и т.д. (в дальнейшем будет использоваться термин «протокол»).

Протокол, как технологический объект поискового процесса, позволяет представить результаты этого процесса в виде объединения подмножеств документов, каждое из которых построено в соответствии с критерием отбора и характеризуется степенью соответствия информационной потребности.

Для обеспечения соответствия объектов физического и логического уровней вводится промежуточный *интерфейсный уровень* представления процесса поиска.

Объекты этого уровня (и характер их представления, например, упорядочение) структурно будут соответствовать логическому уровню, и каждый из них будет представлять элементы (поисковые образы запросов, словники, результаты поиска), относящиеся к соответствующему предмету поиска, но физически полученные, возможно, на разных этапах.

Динамически создаваемые пользователем иерархически организованные структуры должны отражать его персональное видение предметной области (ПрО). Причем, каждый такой объект представляет как общепринятое, так и индивидуальное видение ПрО. Интегральность такого представления достигается за счет того, что оно реализуется объектами как уровня ресурсов (подборками документов, ссылками на ассоциированные ресурсы и т.д.), так и уровня терминологии (тезаурусами, рубрикаторами, словниками).

С точки зрения такого пользователя, «интегрально» осуществляющего основную и информационную деятельность, можно говорить о некоторой среде, которая может быть названа «интегральным рабочим пространством пользователя».

Структура такай среды, рассмотренная по отношения к ПрО, где выделены *«состоявшееся» обобществленное и опубликованной знание, проблемная* ситуация и гипотетическое или генерируемое знание представлена на рис. 1.

Рабочее пространство (РП) пользователя – среда, включающая информационные и процедурные компоненты, используемые (и порождаемые) в процессе информационной деятельности, направленной на решение задач основной деятельности (ОД).

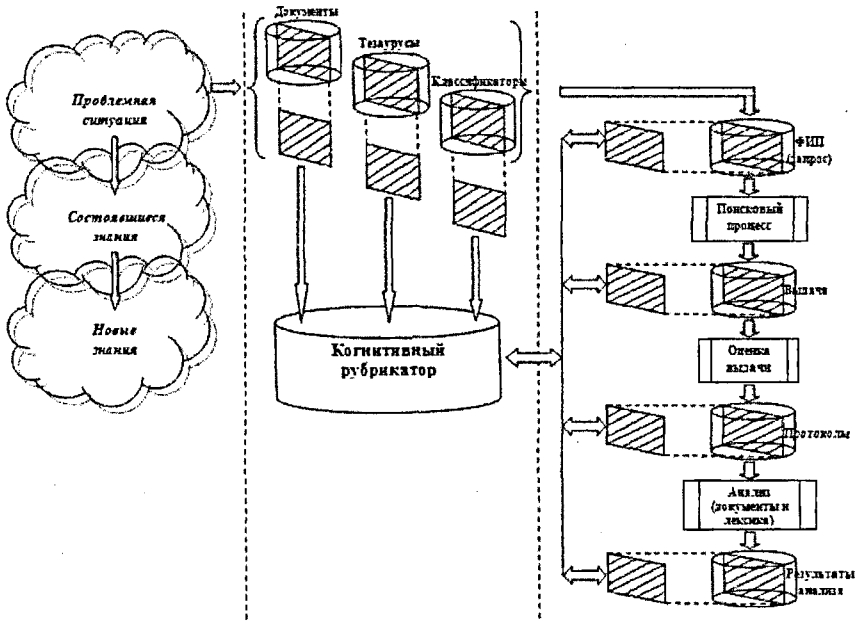


Рис. 1 Структура рабочего пространства пользователя

РП реализуется в виде совокупности когнитивного рубрикатора и справочника ресурсов.

Когнитивный рубрикатор РП фиксирует структурно-семантическое состояние и динамику когнитивного процесса (поиска информации в информационных ресурсах для решения задачи ОД) и отражает сходство и различия состоявшегося (опубликованного) и нового (генерируемого пользователем) знания на знаковом (лексическом), понятийном и документальном уровнях. КР, представленный в форме иерархической классификации, отражает через систему классификационных признаков аспекты и «позицию» поисковой задачи (прежде всего в сфере ОД).

Справочник ресурсов реализует «физическую» (по аналогии с теорией БД) составляющую, обеспечивающую доступ к экземплярам хранения: идентификацию и поиск документа в локальных и внешних информационных ресурсах по признакам содержания и адресам хранения. По-существу, справочник

для каждого документа содержит один или несколько идентификаторов, представляющих способ (или адрес) доступа к экземпляру хранения, а также один или несколько поисковых образов, представляющих на разных ИПЯ или с разной степенью полноты и точности смысловое содержание. При этом идентификаторы доопределяются метаданными, отражающими характер и способ их построения.

Соответственно, КР как технологический (и интерфейсный) объект, используемый для организации поиска информации для решения задачи ОД, должен удовлетворять следующим требованиям:

1) иметь средства явной систематизации информации (связывать найденные документы с разделами классификации предопределенным способом);

2) отражать соотношение нового и состоявшегося знания;

3) позволять фиксировать соотношение вновь вводимой классификационной схемы с существующими системами представления знаний (отраслевыми и общенаучными классификаторами, рубрикаторами, тезаурусами и т.д.);

4) представлять как статику (определения, соотносящие объект исследования с объектами состоявшегося знания) через декларативное определение, так и динамику использования определений и их компонентов.

Исходя из семиотической модели, такое представление ПрО будет включать компоненты знакового, понятийного и предметного уровней.

Во второй главе представлен комплекс математических моделей, ориентированных на исследование процессов в документальных информационно-поисковых системах.

В качестве основного математического аппарата используется линейная модель информационных объектов и процессов в документальных АИПС, в рамках которой все описываемые объекты представляются векторами, а процессы – операциями линейной алгебры.

Основным объектом модели является универсальный массив документальной информации $\mathbf{L}_0$ (в линейном представлении матрица размерности $D \times n_0$, где D – количество терминов, n_0 – количество документов):

$$\mathbf{L}_0 = \begin{pmatrix} b_{11} & \cdots & b_{1j} & \cdots & b_{1n_0} \\ \vdots & & \vdots & & \vdots \\ b_{i1} & \cdots & b_{ij} & \cdots & b_{in_0} \\ \vdots & & \vdots & & \vdots \\ b_{D1} & \cdots & b_{Dj} & \cdots & b_{Dn_0} \end{pmatrix}$$

Подобные матрицы известны под названием матрицы «термин-документ». На основе $\mathbf{L}_0$ могут быть построены также матрицы «термин-термин» и «документ-документ».

Простейшая матрица «термин-термин»

$$\mathbf{F} = (f_{ij}) = \mathbf{L}_0 \times \mathbf{L}_0^{\mathrm{T}}, \text{ где } f_{ij} = \sum_{k=1}^{n_0} b_{ik} \cdot b_{jk}$$

Компоненты матрицы **F** являются коэффициентами ассоциации (статистической меры связи) терминов. Но основе **F** могут быть построены корреляционная и ковариационная матрицы, а также многочисленные варианты коэффициентов близости векторов, которые могут применяться как меры ассоциации терминов.

Линейная модель протокола. Согласно линейной модели поиска результат фиксируется в протоколе и описывается бинарным вектором:

$$q_i = (q_{i1} \ldots q_{ij} \cdots q_{in_0}), \text{ где } q_{ij} = \begin{cases} 1, \text{если j - й документ входит в i - й результат} \\ 0 \text{ - в противном случае} \end{cases}$$

Тогда теоретико-множественный образ протокола поиска, представляющего собой совокупность результатов, есть матрица размерности $M \times n_0$, где M – количество зафиксированных результатов в протоколе.

При выполнении очередного процесса поиска результат-строка добавляется к матрице, увеличивая ее размерность по строкам на единицу.

Рассмотрим выполнение логических операций над результатами поиска - AND (И), OR (ИЛИ), XOR (ИСКЛЮЧАЮЩЕЕ ИЛИ) и NOT (НЕ), поставив в соответствие каждой логической операции правило ее выполнения с использованием матрицы Q:

$$q_l o_k q_m = (q_{lj} o_k q_{mj}, j = \overline{1, n_0})$$

где o_k из множества бинарных логических операций:

$o_k \in O, O = \{o_1, o_2, ..., o_s\}$

Для унарной операции NOT это правило реализуется следующим образом:

$$\neg q_i = \left(\neg q_{ij}\right), j = \overline{1, n_0}$$

После выполнения операции формируется результирующий вектор $q_k = q_i o_k q_m$, который становится $(M+1)$-й строкой матрицы.

Линейная модель когнитивного рубрикатора. Линейное представление отдельной рубрики r_i на документальном уровне - бинарный вектор:

$$r_i = \left(r_{i1} ... r_{ij} \cdots r_{in_0}\right), \text{где } r_{ij} = \begin{cases} 1, \text{если j - й документ входит в i - ую рубрику} \\ 0 \text{ - в противном случае} \end{cases}$$

Следовательно, теоретико-множественный образ КР, представляющего собой совокупность рубрик, это матрица «рубрика-документ» **R** размерности $T \times n_0$, где T – количество рубрик в рубрикаторе:

$$\mathbf{R} = \begin{pmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n_0} \\ \vdots & & \vdots & & \vdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{in_0} \\ \vdots & & \vdots & & \vdots \\ r_{T1} & \cdots & r_{Tj} & \cdots & r_{Tn_0} \end{pmatrix}$$

Логические операции над документами рубрик интерпретируются так же, как и логические операции над документами поисковых результатов, хранящихся в протоколе.

Матрица $\mathbf{K} = \mathbf{R} \times \mathbf{R}^T$ отражает степень взаимного пересечения рубрик на документальном уровне:

$$\mathbf{K} = \begin{pmatrix} k_{11} & \cdots & k_{1j} & \cdots & k_{1T} \\ \vdots & & \vdots & & \vdots \\ k_{i1} & \cdots & k_{ij} & \cdots & k_{iT} \\ \vdots & & \vdots & & \vdots \\ k_{T1} & \cdots & k_{Tj} & \cdots & k_{TT} \end{pmatrix}, \text{где} \quad k_{ij} = \sum_{k=1}^{n_0} r_{ik} \cdot r_{kj}$$

Рассматривая КР на понятийном уровне, представим его ассоциативной матрицей S - «рубрика-термин». размерности $T \times D$ (T –количество рубрик, D – количество терминов в словаре информационного массива документов:

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1j} & \cdots & s_{1D} \\ \vdots & & \vdots & & \vdots \\ s_{i1} & \cdots & s_{ij} & \cdots & s_{iD} \\ \vdots & & \vdots & & \vdots \\ s_{T1} & \cdots & s_{Tj} & \cdots & s_{TD} \end{pmatrix}$$

где s_{ij} - коэффициент близости j- термина и i – рубрики.

Каждый столбец матрицы соответствует отдельному термину и описывает множество рубрик, содержащих его. Строка матрицы соответствует рубрике и представляет собой вектор мер значимости терминов для рубрики.

Матрица **S** может быть построена на основании частотной матрицы «рубрика-термин» $\mathbf{F}_R = (f_{ij}) = \mathbf{R} \times \mathbf{L}_0^{\mathrm{T}}$, где $f_{ij} = \sum_{k=1}^{n_0} r_{ik} \cdot b_{kj}$ является частотой j-го термина в i-ой рубрике.

Математическая модель метода классификации. Строку матрицы S в рамках КР интерпретируем как описание рубрики. Используем представление документа в виде бинарного вектора

$$l = (b_1 \ldots b_i \ldots b_D), \text{ где } b_i = \begin{cases} 1, \text{ если } i \text{ - й термин входит в документ} \\ 0 \text{ - в противном случае} \end{cases}$$

Операция классификации документа выражается в матричном умножении: $\mathbf{S} \times \mathbf{l} = \tilde{\mathbf{l}}$. Элементы результирующего вектора $\tilde{\mathbf{l}} = \{\tilde{b}_i\}$, где $\tilde{b}_i = \sum_{j=1}^{D} s_{ij} b_j$ характеризуют исходный документ с точки зрения близости к рубрикам: чем больше его величина, тем больше документ соответствует рубрике.

Тем самым, определив максимальный из $\tilde{b}_i$, получим рубрику, которой принадлежит классифицируемый документ, т.е. искомая рубрика r такова, что

$$r : b_r = \max_i \tilde{b}_i$$

Для оценки классификации по нескольким рубрикам применяются усредненные (макроусреднение и микроусреднение) показатели полноты *(r)* и точности *(p)*:

$$r_{macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{a_i + c_i}, \quad p_{macro} = \frac{1}{n}\sum_{i=1}^{n}\frac{a_i}{a_i + b_i}, \tag{1}$$

$$r_{micro} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n}(a_i + c_i)}, \quad p_{micro} = \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n}(a_i + b_i)} \tag{2}$$

где a_i – число правильно рубрицированных документов для i-ой рубрики,

b_i – число неправильно рубрицированных документов,

c_i – число неправильно отвергнутых документов.

Рассмотренные представления в линейной форме различных объектов и процессов, осуществляющихся как в документальных АИПС, так и в системах следующего поколения – интегральных информационно-аналитических системах, образуют линейную модель рабочего пространства пользователя и охватывают:

– процессы формирования документальной БД;

– оптимизацию структуры тезаурусов[4] и рубрикаторов;

– процессы поиска документов[5];

– оценку качества поиска;

– анализ структуры потока публикаций предметной области;

– анализ структуры лексики предметной области.

В третьей главе представлены результаты проведенных экспериментальных исследований применения когнитивного рубрикатора для процесса классификации документов и структуризации предметной области научных исследований.

Для проверки представленного во второй главе метода классификации документов был проведен эксперимент на базах данных реферативно-библиографической информации ВИНИТИ РАН «Информатика» и ВНТИЦ

[4] Попов И.И. Моделирование и оптимизация автоматизированных информационных систем и технологий управления документальными информационными ресурсами. Дисс. на соискание ученой степени доктора техн. наук – М.: РГГУ, 1996.

[5] Голицына О.Л. Моделирование и разработка средств и технологий поиска документальной информации. Дисс. на соискание ученой степени кандидата техн. наук. – М.: РГГУ, 2004.

«Информационные карты НИР и ОКР» (ИК) и «Информационные карты диссертаций» (ИКД), под управлением ИАС xIRBIS. В качестве меры близости в данном методе использовались статистические коэффициенты, вычисление которых строится на частотных характеристиках терминов и рубрик (коэффициенты корреляции, Андерберга и Юла). Результаты работы метода сравнивались с классификацией, проведенной экспертами.

Результаты экспериментов для БД «Информатика» (ВИНИТИ)

Документы базы данных "Информатика" (73693 док.) размечены экспертами по рубрикатору ВИНИТИ 395 рубриками.

Цель первого эксперимента состояла в том, чтобы определить влияние терминов документа, имеющих отрицательную корреляцию с рубрикой, на качество классификации.

Сводная таблица результатов по коэффициентам отражает усредненные полноту и точность процесса классификации документов без учета отсутствующих в рубрике терминов документа (Этап 1) и после добавления этих терминов в расчет суммарных коэффициентов для рубрик (Этап 2) (см. таблицу 1).

Таблица .1

Результаты 1 и 2 этапа исследований

Оценка классификации	Корреляция				Коэффициент Юла				Коэффициент Андерберга			
	Полнота %		Точность %		Полнота %		Точность %		Полнота %		Точность %	
	1 этап	2 этап	1 этап	2 этап	1 этап	2 этап	1 этап	2 этап	1 этап	2 этап	1 этап	2 этап
макроусреднение	86	91	82	87	86	88	72	76	86	86	80	82
микроусреднение	85	88	88	88	86	89	86	88	88	91	86	88

Показатели по всем коэффициентам улучшаются на втором этапе. Следовательно, для повышения качества классификации следует учитывать термины с отрицательной корреляцией.

В следующем эксперименте были применены различные ограничения для уменьшения размерности матрицы близости.

Для ограничения количества терминов, которые используются при проведении классификации, для каждой рубрики определялись значимые термины. Исследованы следующие варианты использования полученного словника.

1) Используются термины словника с частотой *больше средней частоты в рубрике.* Словники рубрик ограничивались терминами, частота которых больше средней частоты в рубрике (термин значим, если его частота встречаемости в рубрике больше средней частоты);

2) Используются термины словника, имеющие вес *больше среднего весового коэффициента.* Термин считается значимым, если его весовой коэффициент $W_i = (Log(n_0 / f_i)) \times f_{ij}$ больше средней весового коэффициента по текущей рубрике. Здесь: n_0 - общее число документов информационного массива, f_i - частота i-го термина, f_{ij} - частота i-го термина в j рубрике.

Таблица 2.

Результаты эксперимента с использованием ограничений на словники

Определение значимости термина для рубрики	Макроусреднение		Микроусреднение		Соотношение с полным словником %
	Полнота %	Точность %	Полнота %	Точность %	
Без ограничения	91	87	88	88	100
Больше сред. частоты в рубрике	86	88	86	86	12
Больше сред. вес. коэффициента	81	84	79	72	15

Результаты эксперимента для БД «Информатика» с использованием коэффициента близости - корреляция и применением различных ограничений для формирования словников рубрик содержатся в таблице 2. Результаты представлены усредненными (макроусреднение и микроусереднение) показателями полноты и точности по всем рубрикам согласно (1) и (2). Последний столбец в таблице показывает, какое количество терминов словника считаются значимыми для рубрики. Полученные результаты позволяют сделать вывод о целесообразности применения ограничения для словников по средней частоте в рубрике.

Результаты эксперимента для БД ИК и ИКД (ВНТИЦ)

Для проведения эксперимента на БД ИК (123347 док.) и ИКД (38527 док.) использовалась рубрикация документов по трехуровневому рубрикатору ГРНТИ. Результаты, представленные в таблице 3, свидетельствуют о предпочтении использования коэффициента корреляции для классификации документов методом, использующим матрицу близости.

Таблица 3.

Результаты эксперимента на БД ИК и БД ИКД

Коэффициент близости	БД ИК				БД ИКД			
	Макроусреднение		Микроусреднение		Макроусреднение		Микроусреднение	
	Полнота %	Точность %	Полнота %	Точность %	Полнота %	Точность %	Полнота %	Точность %
Корреляция	92	75	86	78	74	65	71	66
Андерберга	86	68	76	70	75	66	72	70
Юла	83	66	74	69	66	60	64	65

Результаты проведенных экспериментов позволяют сделать выводы о возможности применения предлагаемого метода для автоматической классификации документов документальных БД.

В третьей главе также описан пример применения когнитивного рубрикатора для проведения анализа ПрО и представлены полученные результаты.

В четвертой главе рассмотрены документальная ИАС xIRBIS и реализованные в ней механизмы управления лингвистическими и документальными ресурсами, включая интерфейсные средства построения и ведения когнитивного рубрикатора, справочник ресурсов и процедуры, реализующие процесс автоматической классификации документов.

Основные функции ИАС xIRBIS:

- формирование структурированного описания предметной области;
- мультиагентный поиск в локальных и распределенных ИР документальной информации и информационных источников;
- формирование терминологических систем предметной области (словари терминов, классификации, тезаурусы, онтологии);
- анализ состояния и динамики научных направлений на основе статистического анализа информационных потоков и лексики предметных областей.

Результаты диссертационной работы реализованы в следующих подсистемах ИАС xIRBIS.

Подсистема статистического анализа документальных потоков и лексики обеспечивает для результатов тематического поиска и для документального ресурса в целом статистический анализ с применением компонентов

деловой графики и представление документального потока в виде временного ряда с последующим анализом с целью выявления характеристических свойств и общих тенденций. Для результатов поиска, представленных в виде тематических частотных словарей лексики, реализованы возможности отображения в форме таблиц и диаграмм с поддержкой функций сортировки, редактирования и вывода.

Подсистема анализа и ведения объектов лингвистического обеспечения ориентирована в основном на создание и поддержку пользовательского лексического пространства в рамках интересующей его предметной области и обеспечивает:

- формирование когнитивного рубрикатора;
- формирование тематических словников по результатам вычислений мер тематической близости, основанных на частотных характеристиках, и соотнесение их с рубриками пользовательского рубрикатора;
- построение иерархических словарных структур, которые в дальнейшем могут быть использованы в качестве мини-тезаурусов при формировании поискового запроса;
- формирование на основе когнитивных рубрикаторов и тематических словников специализированных матриц тематической близости, применяющихся при реализации процедуры автоматической классификации документов.

Логическая модель рубрикатора представляет собой иерархическую древовидную структуру, элементами которой являются рубрики (см. рис.2). Порожденные элементы называют подрубриками. Каждая рубрика может иметь одну или несколько подрубрик. Все рубрики имеют одинаковый набор атрибутов, которые отражают три различных уровня представления ПрО.

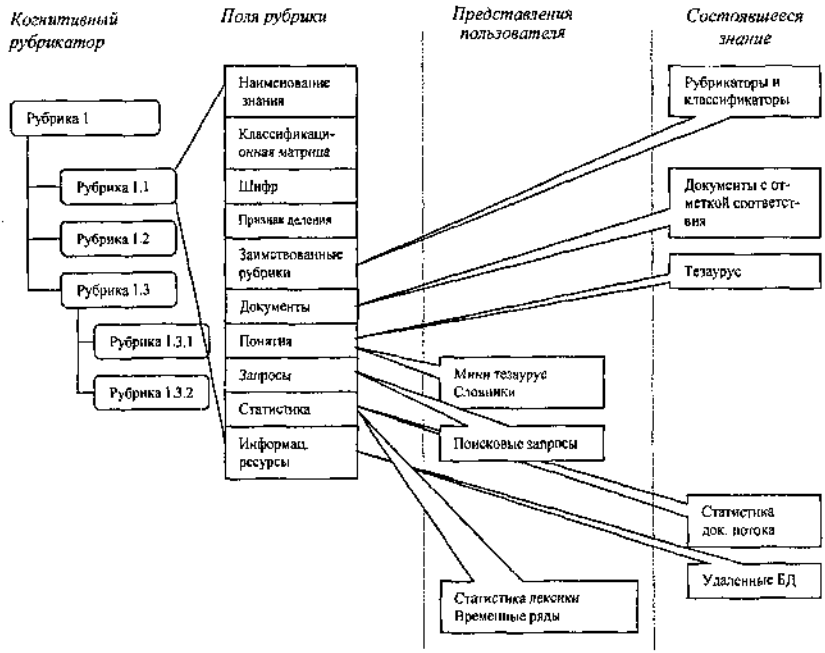


Рис 2 Структура когнитивного рубрикатора

Уровень знаков:

Наименование. Шифр. Каждая рубрика имеет наименование и шифр, которые составляют идентификатор рубрики. Наименование отражает тематику рубрики и является обязательным полем, шифр определяет ее место в иерархии рубрикатора.

Заимствованные рубрики. Пользователь может заимствовать рубрику из официального рубрикатора, что в дальнейшем даст возможность воспользоваться классификационными признаками заимствованной рубрики.

Уровень понятий:

Признак деления. Это поле содержит правило деления рубрики на подрубрики.

Понятия. Свое представление об исследуемой тематике пользователь может выразить с помощью составленных им словников, а также воспользоваться доступными тематическими тезаурусами, найдя в них подходящие дескрипторные статьи.

Статистика. Статистические словники, построенные на основе документов рубрики, позволят проанализировать лексику, применяемую к данной проблематике. Для этой же цели служат распределения и временные ряды употребления лексики, которые отражают изменения в терминологии.

Предметный уровень:

Документы, присоединенные к рубрике, имеют различную степень соответствия исследуемой тематике, которая фиксируется соответствующим признаком.

Поисковые запросы, представленные в рубрике, отражают, во-первых, лексику текущей рубрики, а, во-вторых, правила включения документов в рубрику, сформулированные на ИПЯ.

Информационные ресурсы. Ссылки на внешние ресурсы указывают на соответствующие данной тематики удаленные БД, Internet-сайты и т.д.

Классификационная матрица. Поле, содержащее ссылку на матрицу близости «рубрика-термин» для проведения автоматической классификации с использованием рубрикатора.

Все **операции,** выполняемые над рубрикатором, делятся на две группы. К первой группе относятся операции, связанные с изменением структуры рубрикатора: добавление новой рубрики, удаление и перемещение рубрики, заимствование рубрики из другого рубрикатора.

Вторая группа – это операции, связанные с объектами уровня ресурсов (добавление документа в рубрику, добавление ссылки на ИР) и уровня терминологии (присоединение дескрипторной статьи тезауруса, присоединение словника и т.д.).

В четвертой главе также представлены укрупненные алгоритмы, реализующие создание и ведение рубрикатора и протокола, описаны алгоритмы автоматической классификации документов, а также их программная реализация.

Основные результаты и выводы

В рамках исследования информационно-поисковых систем и создания моделей, методов и средств поиска и анализа данных в документальных информационно-аналитических системах получены следующие результаты.

1. На основе анализа процесса информационного поиска введено понятие рабочего пространства пользователя, которое включает информационные и процедурные компоненты, обеспечивающие управляемую навигацию в локальных и распределенных документальных информационных ресурсах.

2. Введено понятие когнитивного рубрикатора, реализующего структурированную форму представления предметной области и включающего систематическую (классификационную схему) и объектную (документы, запросы, словники, статьи тезаурусов) составляющие, что позволяет динамически отражать на уровне сходства и различий когнитивное состояние пользователя по отношению к состоявшемуся знанию.

3. На основе семиотического подхода разработана логическая модель когнитивного рубрикатора как операционного средства рабочего пространства пользователя, интегрально связывающего представления пользователя с информационными ресурсами предметной области на знаковом, понятийном и документальном уровнях.

4. В рамках структуры когнитивного рубрикатора разработана математическая модель классификации, позволяющая динамически соотносить найденные документы с разделами когнитивного рубрикатора пользователя.

5. В результате экспериментальных исследований, проведенных на материале реферативных баз данных научно-технической информации ВИНИТИ РАН и ВНТИЦентра, получены данные, подтверждающие работоспособность предложенного метода классификации документов.

6. Разработан алгоритм применения когнитивного рубрикатора для проведения автоматической классификации документов и описан пример использования КР для проведения комплексного анализа предметной области.

7. Разработан комплекс программных средств, реализующих физическое представление когнитивного рубрикатора и автоматическую классификацию документов, а также интерфейсные средства, используемые для создания и поддержки объектов и процессов рабочего пространства пользователя.

Список опубликованных работ по теме диссертации:

Научные статьи в журналах и изданиях, выпускаемых в Российской Федерации, включенных в перечень ВАК:

1. Борисова Л.Ф., Васина Е.Н., Максимов Н.В. и др. Системы и технологии распределенной обработки научно-технической информации в ВИНИТИ // НТИ. – Сер. 1, -2003, – №10, 1,0 п.л. (авт. 0,2 п.л.).
2. Васина Е.Н., Голицына О.Л., Максимов Н.В. Архитектура АИПС: технологии и средства поиска в документальных информационных ресурсах. //НТИ Сер.1, 2007, №5, 1,1 п.л. (авт. 0,3 п.л.).
3. Васина Е.Н., Голицына О.Л., Максимов Н.В. Вопросы проектирования автоматизированной системы подготовки и выпуска информационных изданий // НТИ. – Сер. 1, 1986. -№5, 0,8 п.л. (авт. 0,3 п.л.).

Научные статьи и труды в других изданиях:

4. Бебчук М.Б., Буров М.А., Васина Е.Н., и др. Средства документального поиска в распределенных гетерогенных информационных ресурсах // 7-я Международная конф. НТИ-2007, Сб. трудов. – М.: ВИНИТИ, 2007, 0,7 п.л. (авт. 0,2 п.л.).
5. Васина Е.Н., Голицына О.Л., Максимов Н.В. и др. Документальная информационно-аналитическая система xIRBIS -программа для ЭВМ. Свидетельство №2008611511 от 25.03.2008г. Государственный реестр программ для ЭВМ, 2008г.
6. Васина Е.Н., Голицына О.Л., Максимов Н.В. Архитектура аналитической информационно-поисковой системы. // Научная сессия МИФИ-2008. 25-27 янв. 2008., – М.: МИФИ, 2008, 0,05 п.л. (авт. 0,01 п.л.).
7. Васина Е.Н., Голицына О.Л., Максимов Н.В. и др. Интегральная информационная система поддержки научных исследований и процессов управления научными кадрами // Научная сессия МИФИ – 2009. XXIII выставка-конференция «Телекоммуникации и новые информационные технологии в образовании». Сборник научных трудов. – М.: МИФИ, 2009, 0,05 п.л. (авт. 0,02 п.л.).
8. Васина Е.Н., Голицына О.Л., Максимов Н.В. Оптимизация поисковых стратегий в документальных базах данных. Проблемы и перспективы.// Инновационные технологии когнитивного управления в экономике, менеджменте и образования : межвузовский сборник научных трудов. Сер. «Бизнес-информатика». Вып. 1. – М.: ГОУ ВПО «РЭА им.Г.В.Плеханова», 2008, 0,3 п.л. (авт. 0,1 п.л.).

9. Васина Е.Н., Голицына О.Л., Максимов Н.В., Попов И.И., Резниченко П.И. Электронный обучающий узел "Информационные ресурсы Internet" //Материалы 5-й международной конференции "НТИ-2000. Информационное общество, информационные ресурсы и технологии телекоммуникации, Москва, 22-24 ноября 2000 г.", М.: ВИНИТИ, 2000, 0,1 п.л. (авт. 0,02)

10. Васина Е.Н., Голицына О.Л., Максимов Н.В., Попов И.И., Резниченко П.И. Информационные ресурсы документальных баз данных. // НТИ-96: Международн. конференция. Москва, 20-21 ноября 1996г. – М.: ВИНИТИ, 1996, 0,2 п.л. (авт. 0,05 п.л.).

11. Васина Е.Н., Партыка Т.Л., Попов И.И., Информационные системы бухгалтерского учета: Учеб. пособие. М.:ФОРУМ:ИНФРА-М, 2006, 26,6, п.л. (авт. 5,0 п.л.).

Заказ № 296. Объем 1 п.л. Тираж 100 экз.
Отпечатано в ООО «Петроруш».
г.Москва, ул.Палиха 2а.тел.250-92-06
www.postator.ru