

**Федеральное агентство по образованию
Московский инженерно-физический институт
(государственный университет)**

**Н.В. Максимов, О.Л. Голицына,
Г.В. Тихомиров, П.Б. Храмцов**

Информационные ресурсы и поисковые системы

**Рекомендовано УМО «Ядерные физика и технологии»
в качестве учебного пособия для студентов
высших учебных заведений**

Москва 2008

УДК 004(075)

ББК 32.81я7

И74

Информационные ресурсы и поисковые системы: учебное пособие / Н.В. Максимов, О.Л. Голицына, Г.В. Тихомиров, П.Б. Храмцов. – М.: МИФИ, 2008. 400 с.

Рассматриваются основы функционирования и архитектура информационных систем и распределенных ресурсов. Особое внимание уделяется информационно-поисковым системам и технологиям, представляемым в виде совокупных человеко-машинных систем, целью которых является построение системы знаний. Приводятся описания промышленных информационных ресурсов, используемых в научных исследованиях и разработках в ядерной отрасли. Приложение включает пример итеративного процесса поиска, а также справочные материалы по лингвистическому обеспечению.

Для студентов специальностей «Прикладная информатика», «Информационные системы», «Прикладная математика и информатика», а также для широкого круга студентов, аспирантов и специалистов, ведущих научно-исследовательскую работу.

Пособие подготовлено в рамках Инновационной образовательной программы МИФИ.

*Рецензент: зам. директора научной библиотеки
МИФИ Стукалова Т.Н.*

ISBN 978-5-7262-1049-0

© Московский инженерно-физический институт
(государственный университет), 2008

...Чтоб умно поступать - одного ума мало.
Ф.М. Достоевский

Введение

Для ядерной отрасли, как и для всех бурно развивающихся областей человеческой деятельности, характерны и постоянно растущий объем информации, и смена приоритетов, и появление новых технологий. Потребители информации могут иметь различные задачи и цели, их поисковые запросы варьируются от вопросов обзорного или познавательного характера до очень подробных аналитических запросов. Поэтому широкое внедрение автоматизированных информационных систем (АИС) и баз данных (БД) в сочетании с сетевыми решениями является магистральным направлением в попытках человека справиться с проблемами получения и переработки больших объемов информации.

Но при этом важно понимать, что информационные системы и ресурсы создаются не ради них самих: они нужны для процесса передачи знаний от человека к человеку и от поколения к поколению. Кроме того, эффективное использование информации невозможно без определенного уровня профессиональной подготовки. Каждый человек достигает этого уровня по своей траектории, но в любом случае он, так или иначе, обращается к специальной литературе.

Учебные пособия, учебники, монографии и периодические издания, как и научно-популярные и исторические материалы, рекомендуются к каждому специальному курсу. Как правило, их можно найти в библиотеке вуза. Со статьями и тезисами докладов, которые можно отнести к специальным изданиям, студенты обычно сталкиваются в процессе учебно-исследовательской работы, при прохождении практики и подготовке диплома. При написании дипломного проекта или магистерской диссертации студенты должны проводить самостоятельный поиск литературы. Отметим, что знание периодических журналов и конференций по специальным вопросам является необходимым элементом эрудиции специалиста.

Перечисленные типы информационных материалов в информатике относятся к так называемым «первичным документам», содержание которых непосредственно используется человеком для обучения, решения научных проблем и т.д. Совершенно понятно (но не для всех очевидно), что для того, чтобы использовать содержание, необходимо знать о существовании этого конкретного документа, а точнее, знать, что содержание, возможно, позволит решить задачу и знать, где он физически находится. В учебном процессе конкретного вуза в рамках конкретной дисциплины эти знания берутся из программы курса, рекомендаций преподавателя, а сами документы, обычно, присутствуют в вузовской библиотеке. В исследовательских задачах такой предопределенности не может быть в принципе: сама природа исследований имеет пионерский характер и потому публикации необходимо *искать*. Более того, заранее известно, что отыскиваемая информация, скорее всего, будет лишь частично относиться к решаемой задаче (т.к. нахождение публикации, полностью описывающей решение задачи, свидетельствует о том, что исследователь повторяет ранее сделанное). Именно в этом случае становится очевидной целесообразность использования *вторичной информации* – документов, которые в краткой форме сообщают о содержании и местах хранения документа-первоисточника.

В последние годы, в связи с бурным развитием и внедрением в повседневную жизнь вычислительной техники и сетевых технологий, основные тенденции и надежды, обоснованно или нет, связаны с Internet-технологиями и, в частности, с гипертекстом. Но, чтобы в итоге правильно понимать концепции автоматизации работы с огромными массивами информации, необходимо изначально рассматривать задачи поиска информации не только как отыскание ссылки на *экземпляр хранения*, но и как процесс поиска и организации накопленных знаний для создания новых.

Для подтверждения этого положения уместно обратиться к истории и начать с публикации Ванневара Буша (Vannevar Bush) в 1945 г. в журнале «The Atlantic Monthly» статьи «As We May Think», которая, как принято считать, открыла «эру гипертекста». Буш предложил *Метех* – концептуальную машину, которая могла бы хранить обширные объемы информации, но, главное, давала бы пользователям возможность самим организовывать необходимые

им информационные структуры, а также прямые связи между близкими по избранной ими тематике текстами или иллюстрациями для дальнейшего их использования в качестве необходимых ссылок¹. Концепция гиперсвязей Ванневара Буша исходила из того, что наши нынешние трудности с оперативным доступом к необходимой информации вызваны в значительной степени искусственностью системы индексации, принятой в хранилищах информации (библиотеках и пр.). Данные размещаются в хранилищах определенной структуры, обычно в систематическом или в алфавитном порядке, либо просто в порядке поступления. Соответственно и информацию в подобных хранилищах можно находить только двигаясь по иерархии такого рода индексов, обычно от верхнего подкласса к расположенному ниже. То есть пользователь должен в каждом отдельном случае строго следовать заданным правилам отыскания пути к определенному элементу информации, и эти правила обычно оказываются громоздкими и трудными для восприятия. Между тем человеческий мозг работает не так — он работает по ассоциации. От одного вызванного из памяти элемента данных он тут же переходит к следующему в том их ряду, который предлагает ему ассоциативный ход мысли в соответствии с некоторой паутиной связей сложной структуры, формирующейся в клетках человеческого мозга.

В 1981 г. Тед Нельсон, который в сборнике трудов 20-й конференции АСМ в 1965 г. ввел в компьютерный слэнг термин «гипертекст»² и «гипермедиа», создает концепцию *Xanadu* — глобально централизованной гипертекстовой базы данных, охватывающей всю открыто доступную в мире письменную информацию для её машинного хранения и коммерческого использования. Таким образом была бы создана «вселенная документов» (*docuverse*) или, иными словами, универсальная библиотека и всемирная гипертекстовая издательская система мультимедийных документов, как меха-

¹ Фрагмент приводится по книге [7], которую мы настоятельно рекомендуем читателям, стремящимся к глубокому пониманию существа информационных технологий и выбору путей их развития.

² Гипертекст — это такой текст, который содержит ветвящиеся разделы и при этом самому читателю предоставляется выбор для следования в процессе чтения той или иной ветви изложения текста.

низм организации действительно свободных, глобально доступных форумов для обсуждений и разноплановых дебатов на них (в том числе и для разрешения споров об авторском праве).

Это требует заметно более мощной адресной структуры данных, чем универсальный указатель ресурса (Universal Resource Locators – URL), используемый в World Wide Web: каждый отдельный байт (символ) в каждом документе нуждается в своем уникальном адресе во всем мире.

Сегодня проект Xanadu – это идеал, который пока невозможно реализовать практически. Тем не менее, разработанная в CERN технология Web, являясь редукцией Xanadu, оказалась не только конструктивной и эффективной, но и создала практические предпосылки для успешного осуществления и самого проекта Xanadu. Дело здесь не в том, что авторы отказались от «улучшения качества жизни человечества в целом», а в том, что несколько краеугольных «камней» базовой концепции Web были реализованы фактически за счет установления вполне конкретных рамок: документы связывались в контексте вполне определенной семантики, и цель связывания состояла в организации эффективной последовательности получения субъектом знаний путем прохождения по ссылкам. При этом, естественно, ссылки строятся по наперед принятому автором сценарию тематического связывания фрагментов текста.

Но и концепция гиперссылок сама по себе известна тысячи лет. Достаточно вспомнить, что библейские рассказы включают в себя отсылки или вставки из иных рассказов, а те, в свою очередь, содержат «гиперссылки» следующего уровня, и так много раз. Однако все довебовские приложения концепции гиперссылок имели один технический недостаток – много сил и времени требуется читателю для поиска фрагмента текста по каждой очередной ссылке. Web устранил такого рода технические ограничения, допуская любую глубину вложенности ссылок, однако породил проблему управляемости «навигацией». Пользователь должен сам «держат» контекст переходов: использовать гиперссылки очень легко, но также легко забыть цель, ради которой обратился в Web.

Гипертекст не является альтернативой технологиям поиска по «традиционным» индексам, как это представляется некоторым пользователям Internet. Это скорее взаимодополняющие технологии: если при поиске по индексам мы следуем по документам в по-

рядке, определенном *общепринятой* (и, обычно, укрупненной) классификацией, то гипертекстовая ссылка имеет *частный*, хотя и вполне конкретный контекст, зачастую определяемый автором в разрез с общепринятой точкой зрения, но при этом не исключено, что точка зрения не совпадает ни с общепринятой, ни с авторской. Кроме того, здесь речь шла, в основном, о *концепции* гипертекста, отражающей в первую очередь особенности *технологии на стороне пользователя*, а не способы и средства, позволяющие эффективно реализовать её в *вычислительной среде*, которые, как будет показано далее, так или иначе, оказываются связанными с понятием «индекс». Кстати, отметим, что и в *Metex*, наряду с гиперссылками, были предусмотрены средства для поиска в архиве в соответствии с обычной схемой индексации.

Наука и знания человека всегда системны, а формы их организаций достаточно консервативны. Именно эта системность и консервативность всегда будет предопределять системность хранилищ знаний, что, собственно, в итоге и определяет эффективность средств и технологий поиска.

Предметом данного пособия являются документальные информационные ресурсы и автоматизированные информационно-поисковые системы (АИПС). В этой области науки и технологии сделано очень много. Развитие информационных технологий и глобализация информационных коммуникаций, как кажется, обеспечили возможность взаимодействия исследователей, минуя информационных посредников. Но парадокс современности состоит в том, что общество, получив технические возможности, в том числе оперативного непосредственного взаимодействия исследователей, обрело при этом такие проблемы, как, например, ограничения авторского и имущественного права на передачу информации. Открытость Internet-среды проявила и другую, ранее почти невероятную проблему. Уже недостаточно найти публикацию: прежде чем использовать её материал, необходимо потратить гораздо больше усилий на то, чтобы убедиться в достоверности и адекватности полученного. Причина проста: в Internet статьи не рецензируются (как, например, в академических журналах), а автор может считать свою работу гениальной и не утруждать себя доказательствами, например, практической проверкой или экспериментом. Нельзя исключать и того, что текст, как и выходные данные публикации, может

быть просто кем-то изменен (и даже многократно), причем без видимых следов редактирования.

Кроме того, пользователь, привыкший к интуитивному освоению программных сред (в основном стандартных средств операционной системы, большинство из которых вследствие узкой функциональной направленности имеет простой интерфейс), может неадекватно оценить состояние и результаты поиска. Зачастую он рискует обрести убежденность в том, что поисковые Internet-машины всемогущи и вездесущи, а их способности извлекать информацию (и даже знания) из текстов и массивов документов бесконечно выше человеческих.

Прежде, чем обратиться к основному материалу пособия, для понимания существа автоматизированной обработки информации уместно было бы провести достаточно простой эксперимент по поиску в Internet какой-нибудь, но жизненно важной информации. Например, вообразим себя жителем славных Петушков, который теплым летним вечером предельно оттягивает момент расставания с приятной московской компанией, и, как «конкретный юзер», ищет время отправления последней электрички.

Начнем с ввода вопроса в «прямой»¹ форме: «Время отправления последней электрички «Москва-Петушки». В ответ будет выдан длинный список ссылок, в котором, наверное, будут ссылки на самые разные расписания, в разной форме и с разным (и даже противоречивым) содержанием, а не единственно затребованное время отправления электропоездов по нужному направлению. Более того, вполне вероятно, что в начале списка окажутся ссылки, связанные с культовой поэмой 1970-х годов.

Совсем другой результат мы получим, если сформулируем запрос следующим образом: «График движения электропоездов Москва-Петушки, суббота, летнее расписание, 2008 г.»». Здесь мы учли, что электричка в расписании официально именуется «электропоезд», понятие «сегодня» не конкретно, а существует день

¹ Эта форма является сокращенным вариантом от вопроса, который мы вполне естественно задали бы сотруднику справочной службы вокзала: «Скажите, пожалуйста, в какое время сегодня отправляется последняя электричка на Петушки?»

недели, расписания имеют летний и зимний вариант. Но и в этом случае ссылки на официальный сайт и действующее расписание не обязательно будут одними из первых, да, впрочем, если и будут, то это не лишит невезучего «счастья» увидеть у кассы объявление об отмене поезда: ведь мы не искали «Изменений к расписанию».

Если мы продолжим эксперимент и обратим эти запросы к разным поисковым машинам, получим опять совершенно разные ответы по каждому из запросов. Но, поскольку во всех случаях речь идет об одном объекте, единственном и уникальном во времени и пространстве электропоезде, естественно должен возникнуть вопрос о различиях в механизмах, генерирующих столь многочисленные и разные ответы.

Конечно, эксперимент не дает ответа на этот вопрос, но позволяет сделать вывод, что процесс поиска – это что-то более сложное, чем простое сравнение машиной байтов строки запроса с байтами Internet-страниц. Почти очевидно, что запрос предварительно преобразуется в нечто, что и будет сравниваться и позволит получить, в частности, по разным выражениям запроса совпадающие ссылки. Аналогично, различие результатов поиска по отдельному выражению, но полученных с помощью разных поисковых машин, позволяет предположить, что поисковые машины работают с разными информационными ресурсами, а найденные документы по-разному упорядочиваются. То есть процесс поиска включает, помимо отбора, также пред- и постобработку.

Рассмотренный пример отражает наши предположения о том, как автоматизированная система «вычисляет ответ», выдаваемый пользователю. Все эти рассуждения относятся к «внутреннему» миру поисковых машин, а нас, «пользователей», в итоге интересует «действенность» получаемых данных, а именно: безусловно «принимать полученную информацию к исполнению», или снова искать, теперь уже подтверждения достоверности полученного.

Именно эти два, как казалось бы, простых вопроса ЧТО и КАК мы ищем, и будут составлять существо дальнейшего рассмотрения.

Задачи поиска информации появились не сегодня, а уже тогда, когда для передачи личного знания человек стал использовать внешние носители вместо личной передачи. Соответственно, методы и принципы организации и поиска появились, как только воз-

никли первые носители и хранилища информации. И что примечательно – сегодня эти методы практически не изменились. Сравните современные методы индексирования с тем, что приведено в [18]: «...уже в VII в. до н. э. на плитках из библиотеки ассирийского царя ... в конце текста приведен ряд библиографических сведений – заглавие, номер плитки, первые слова следующей плитки, имя владельца...» И еще, «...чтобы облегчить опознавание книг, в некоторых каталогах XII в. описания многотомников содержали первые слова первого и следующих томов». Действительно, новое – это хорошо забытое старое!

Особенностью поисковых технологий являются скорее не сами машинные алгоритмы, а то, что для процессов использования информации (как результатов работы системы) характерна непредсказуемость межотраслевых связей: практически любые решения находят иногда существенно более эффективное применение далеко от тех предметных областей, в которых были изначально получены. Ситуация часто выглядит так, что «...по сути дела легче открыть новый факт или создать теорию, чем удостовериться, что они еще не были созданы или выведены»¹.

Действительно, никакая ИПС, сколь бы интеллектуальной она ни была, никакой информационный ресурс не могут автоматически обеспечить исчерпывающую полноту представления научной проблемы². Но это вовсе не означает, что ресурс не надо использовать. Здесь уместно привести фрагмент из [7], цитирующий точку зрения Ф. Джорджа (автора известной книги «Конструкция мозга»): «Мы вовсе не предполагаем, что неточное описание следует предпочитать точному, мы лишь хотим сказать, что если описываемые факты не точны или наше знание о них не полно, то предпочтительнее иметь точное описание степени их неточности». Но мы, как и Г.Р. Громов, все еще часто видим, что эта внешне столь простая и, казалось бы, очевидная точка зрения остается крайне

¹ Воробьев Г.Г. Проблема документальной информации // Кибернетика и документалистика. Механизмы процесса накопления, хранения и поиска научной информации. – М.: Наука. – 1966. С.5-34.

² Это известный парадокс задачи информационного обеспечения: для достижения полноты поиска по проблеме необходимо представить ее в полном виде, т.е. в качестве запроса лучше всего использовать описание решения.

трудной к восприятию пользователями, а главное, многими авторами работ в области информационных технологий.

Образовательные процессы сравнительно консервативны и во многих случаях источники информации считаются известными. Практически во всех вузах есть своя информационная среда, свои электронные библиотеки, банки данных и знаний. В ядерной отрасли такие ресурсы также существуют и о них необходимо не только рассказывать студентам, но и добиваться *необходимости* их использования в учебе и научно-исследовательской деятельности.

Данное пособие представляет собой учебный курс по основам «навигации» в информационном поле специальности, включая описания наиболее представительных электронных ресурсов, необходимых в учебной и научной работе.

Материал курса, представленный в двенадцати главах и пяти приложениях, условно можно отнести к следующим разделам:

- введение в информатику и информационно-поисковые системы;
- лингвистическое обеспечение и обработка текстовой информации;
- распределенная обработка информации и документальные информационные ресурсы.

В первой и второй главах с точки зрения обобщенного поискового процесса в ИПС приводится типология поисковых задач и информационной неопределенности. Рассмотрены концептуальные, технологические и лингвистические основы информационного поиска.

В третьей и четвертой главах рассмотрено лингвистическое обеспечение информационных систем. Даны основные понятия и типология методов описания содержания документов, а также описаны классификационные и дескрипторные языки. Кратко описаны библиотечно-библиографические и патентные классификации, а также отраслевые классификационные системы. Проанализированы различные типы терминологических структур: словари, тезаурусы, онтологии.

Пятая глава посвящена методам и средствам лексического анализа информационных потоков: выделению наиболее информа-

тивных терминов, расчету статистических мер близости терминов, классификации и кластеризации текстовой информации.

В шестой и седьмой главах определены основные понятия, относящиеся к распределенной обработке информации.

В восьмой главе обсуждаются архитектура электронных библиотек и методы доступа к распределенным информационным ресурсам разного типа.

Главы 9, 10, 11, 12 содержат описания промышленных ресурсов, сгруппированные по характеру предоставляемой информации и форме организации деятельности: электронные каталоги, реферативно-библиографические документальные БД, электронные ресурсы издательств, специализированные БД.

Приложение содержит пример итеративного процесса поиска в реальных информационных ресурсах, приведены фрагменты УДК, Рубрикатора ГРНТИ.

Заключение представляет собой резюме, обобщающее основные принципы и особенности итеративного интерактивного поиска информации в условиях ее распределенного хранения и представления.

В списке литературы приведены ссылки на все материалы, которые авторы использовали в процессе подготовки пособия.

РАЗДЕЛ 1. ИНФОРМАЦИОННЫЕ РЕСУРСЫ И АВТОМАТИЗИРОВАННЫЙ ДОКУМЕНТАЛЬНЫЙ ПОИСК

ГЛАВА 1. КОНЦЕПТУАЛЬНЫЕ ОСНОВЫ ИНФОРМАЦИОННОГО ПОИСКА

1.1. О понятиях «информация» и «информационные ресурсы»

Основное внимание в этой главе будет уделено профессиональному информационному поиску – методам, средствам и технологиям отыскания в информационных ресурсах (ИР) документов или их фрагментов, обеспечивающих пользователю обоснованное решение значимых задач в сфере его основной деятельности.

Во многих пособиях по информатике термины «информация», «сведения», «данные» практически не различаются. Для конструктивного определения основного для данного пособия понятия «информация» вернемся к описанному во введении примеру и зададимся вопросом, что же собственно является ответом информационной системы.

В приведенном примере, когда мы обращаемся к сотруднику справочной службы вокзала, ожидаемым, хотя и крайне лаконичным, но абсолютно полным и точным (а главное – разрешающим нашу жизненную проблему) ответом будет: «23 часа 20 минут».

Точно такое же выражение (величину, определяющую время суток) можно было бы обнаружить (или получить) и в других местах, например, в собственной записной книжке, на экране компьютера, в объявлении киносеансов и т.д. Но, естественно, воспринимая эти данные, в итоге *правильные по величине*, мы не отправимся на вокзал. Эти данные (*величина*, а не *значение*) для нас (точнее, нашей задачи) не будут *действенными*. Разница между этими ситуациями вполне очевидна, и можно сделать вывод, что значение одной и той же величины определяется обстоятельствами (*контекстом*) её появления или использования. В примере с сотрудником справочной службы такой контекст содержится в отдельном (от ответа) выражении нашего вопроса и в том, что этот вопрос на-

правлен компетентному лицу. В других случаях контекст может находиться вместе с величиной в содержании развернутого ответа, или явно или неявно определяться структурой и содержанием.

То есть информация может быть определена как данные, связанные с контекстом и обладающие свойством действенности. Более подробно понятие информации рассмотрено в [5].

Отсюда можно сделать важный вывод, который предопределяет подходы как к проектированию систем автоматизированной обработки информации, так и к поиску информации: информационная система, помимо средств преобразования данных, так или иначе имеет средства хранения и обработки контекста¹. Соответственно, при поиске пользователь, если он заинтересован в эффективности результата, должен использовать метаинформационную составляющую, например, явно указывая поисковые поля (ключевые слова, полные тексты, заголовки, авторы), предметную область, грамотно применяя конструкции поискового языка и т.д.

Конструктивность такого определения состоит не столько в том, чтобы декларировать, что *контекст* есть и его надо использовать (обрабатывать), сколько в том, что система *берет* данные (сигналы, величины и т.д.) из бесконечно большого множества данных окружающей среды. Следовательно, *необходимо выбрать* только те, которые соответствуют контексту, т.е. необходимы и достаточны *для решения конкретной задачи*. Очевидно, что данные в этом случае должны обладать, а точнее (вследствие элементарности (атомарности) того, что называется «данное») должны быть связаны с контекстом, который обычно задается в виде набора отличительных признаков, которые, в свою очередь, также представляют собой некоторый набор данных. Далее для некоторой целевой обработки эти данные обрабатываются прикладной программой (данные связываются с методом обработки, являющимся одной из форм задания контекста) и, в итоге, полученный результат (тоже данные) должен быть связан со способом его использования, что и обеспечит действенность информации для «конечного пользователя» в реальности.

¹ При этом контекст – это, естественно, тоже данные, но выполняющие роль метаданных – данных о характере обрабатываемых данных.

Если бы назначением информационных систем было только хранение и поиск данных в массивах записей, то структура системы и базы данных была бы простой. Причина сложности в том, что практически любой объект характеризуется не только параметрами-величинами, но и взаимосвязями частей или состояний. Кроме того, как отмечалось выше, сам по себе отдельный элемент данных (его величина) приобретает смысл (значение) только тогда, когда связан с природой значения (другими элементами данных), что и позволит его интерпретировать.

Поэтому физическому размещению данных (и, соответственно, определению структуры физической записи) должно предшествовать описание логической структуры предметной области – построение *модели* соответствующего фрагмента реального мира, выделяющей только те объекты, которые будут интересны будущим пользователям, и представленные только теми параметрами, которые будут значимы при решении прикладных задач. Такая модель будет иметь очень мало физического сходства с реальностью, но будет полезна как *представление* пользователя о реальном мире. Причем это представление описываться *удобными для пользователя* средствами, задавая его для манипулирования в *неадекватной человеку* жесткой вычислительной среде с числовым представлением информации.

Таким образом, прежде чем описывать физическую реализацию объектов и связей между ними, необходимо определить:

1) способ, с помощью которого внешние пользователи представляют (описывают) объекты и связи;

2) форму и методы внутримашинного представления элементов данных и взаимосвязей;

3) средства, обеспечивающие взаимно однозначные преобразования внешнего и внутримашинного представлений.

Такой подход является компромиссом: за счет *предварительно определяемого множества абстракций*, общих для большинства задач обработки данных, обеспечивается возможность построения *надежных* программ обработки. Пользователь, используя *ограниченное множество формальных, но достаточно знакомых понятий*, выделяя сущности и связи, описывает объекты и связи предметной области; программист, используя такие *типовые абстрактные понятия* (как, например, числа, множества, агрегаты

данных), определяет соответствующие информационные структуры. Система управления данными, используя *двоичные формы представления типизированных данных*, обеспечивает эффективные процедуры хранения и обработки данных.

При любом методе отображения предметной области в машинных базах данных (БД) в основе отображения лежит фиксация (кодирование) понятий и отношений между понятиями. Абстрактное понятие *структуры* ближе всего находится к так называемой концептуальной модели предметной среды и часто лежит в основе последней.

Понятие структуры используется на всех уровнях представления предметной области и реализуется как:

- *структура информации* – схематичная форма (обеспечивающая переход к атрибутивной форме) представления сложных композиционных объектов и связей реальной предметной области (ПрО), выделяемых как актуально необходимые для решения прикладных задач, в общем случае без учета того, будут ли для ее решения использованы средства программирования и вычислительные машины. В случае документального представления ПрО необходимо рассматривать еще и *структуру текста*, обеспечивающую возможность выделения семантически значимых компонентов текста и определение их роли. Эффективность здесь определяется уровнем абстрагирования, а также полнотой и точностью представления свойств с посредством выбранной системы характеристик;

- *структура данных* – атрибутивная форма представления свойств и связей ПрО, ориентированная на выражение описания данных средствами формальных языков (т. е. учитывающая возможности и ограничения конкретных средств с целью сведения описаний к стандартным типам и регулярным связям). Эффективность в этом случае связывается с процессом построения программы («решателя» прикладной задачи) и, в каком-то смысле – с эффективностью работы программиста;

- *структура записей* – целесообразная (учитывающая особенности физической среды) реализация способов хранения данных и организации доступа к ним как на уровне отдельных записей, так и их элементов. Эффективность в этом случае связывается с процессами обмена между устройствами оперативной и внешней памяти и обеспечивается избыточностью данных, искусственно

вводимой для обеспечения функциональной эффективности отдельных операций (например, поиска по ключам).

Структура является общепринятым и удобным инструментом, одинаково эффективно используемым как на уровне сознания человека при работе с абстрактными понятиями, так и на уровне логики машинных алгоритмов. Структура позволяет простыми способами свести многомерность содержательного описания к линейной последовательности записей. При этом сама структура представляется достаточно простыми способами, а в ряде случаев именно описание структуры представляет собой контекст использования данных.

Приведенные определения иллюстрируют, что представления на любом уровне задаются парой: данными – величинами свойств и данными – определяющими характер свойств (метаданные).

«Информационные технологии»; «информационные системы» и «информационные ресурсы» – это другая важная группа широко используемых и также по-разному определяемых понятий. Приведем определения этих понятий, которые авторы считают конструктивными для понимания существа автоматизированной обработки информации.

Согласно определению [12], *«Информационная технология (ИТ) – это представленное в проектной форме (формализованном виде, пригодном для практического использования) концентрированное выражение научных знаний и практического опыта, позволяющее рациональным образом организовать тот или иной достаточно часто повторяющийся информационный процесс»*.

Понятие «Информационные системы» (ИС) наиболее часто трактуется (согласно определению, данному в [13]), как *«..комплекс, состоящий из информационного фонда и процедур: управляющей, обновления, информационного поиска и завершающей обработки, позволяющих накапливать, хранить, корректировать и выдавать информацию»*¹.

¹ Такое определение, безусловно, является конструктивным по своей сути, но, тем не менее, оно не проводит границы между информацией и данными, которые собственно накапливаются, обрабатываются и передаются.

Понятие «Информационные ресурсы» с наиболее общих позиций определяется как «накопленная информация об окружающей действительности».

В принципе, можно утверждать, что информационные технологии являются менее зависимыми от структуры и специфики предметной области, чем информационные системы и ресурсы, однако эта связь всегда существует, если, например, определить *автоматизированную информационную технологию* как целенаправленное и согласованное использование технических средств, программного обеспечения и информационных ресурсов.

Аналогично, информационные системы – это комплексы, реализующие информационные технологии, ориентированные на процедуры сбора, обработки, хранения, поиска, передачи и отображения информации предметной области, а информационные ресурсы – комплексы соответствующих информационных систем, существующих на конкретных социально-экономических уровнях.

Информационные технологии, системы и ресурсы в данном пособии будем рассматривать в контексте процессов информационного обмена. В традиционном цикле информационного обмена основной поток документальной информации идет по цепочке *автор–издательство–библиотека–читатель*, хотя существуют и обходные пути, например: *автор–читатель*; *издательство–читатель* (подписка). Переход в начале 1970-х годов информационных служб на дублирование в машиночитаемой форме сначала *вторичных* (каталоги, реферативные журналы), а затем и *первичных (полнотекстовых)* документов не нарушает в принципе этой общей структуры. Поставщиками содержания в этом цикле являются любые из участников процесса: автор (индивидуальный или коллективный), издательство или библиотека. Поставка может осуществляться и в форме первоисточника, и в форме БД, как на традиционных носителях, так и по каналам электронной связи. Информационные сети, представляющие собой коммуникационную среду для конечного пользователя, часто входят в состав издательств или распространителей БД, образуя интегральные распределенные информационные службы. Электронная инфраструктура при этом образует относительно самостоятельный слой, не являясь зеркальным подобием традиционных коммуникаций.

1.2. Введение в информационный поиск

С развитием информационных коммуникаций поиск информации стал для пользователей делом довольно обыденным, но как показывает анализ выражений запросов и действий пользователей профессиональных баз данных научной информации [5], вряд ли грамотным (с осознанием существа выполняемых действий) и поэтому – малоэффективным. Причиной такого состояния является не только то, что поиск для человека – это естественная, «встроенная» функция и выполняется скорее интуитивно, сколько то, что среды, в которых этот процесс осуществляется, принципиально различаются. Для сознания человека характерна целостность и образность представления, а также ассоциативность отбора, а для ЭВМ – дискретность и точечность (двоичность) представления, а также четкие алгоритмы идентификации и соотнесения объектов. Автоматизированный информационный поиск, таким образом, должен быть интегрирующим процессом, выполняемым в обеих средах, и потому требующим согласования форм представления, методов обработки и средств, обеспечивающих взаимодействие. Для этого наряду с формами и методами представления и обработки информации в машинной среде необходимо достаточно обстоятельно рассмотреть особенности порождения, поиска и использования информации в основной деятельности (ОД) человека. Необходимость такого систематизированного рассмотрения определяется следующими факторами, имеющими как объективную, так и субъективную природу.

1. Целью информационного поиска в большинстве случаев является отыскание документов, содержащих сведения, нужные для решения конкретных управленческих, научных или практических задач, в том числе генерации нового знания. При этом характер информации, способ ее представления может быть самым разным – от объявлений о продаже товаров до интерактивных научных конференций, от технического описания, пригодного для непосредственного применения, до не формализуемой в явном виде совокупности фактов, приводящих к творческому озарению или принятию неординарного решения. В дальнейшем изложении мы будем придерживаться именно профессиональной точки зрения: отыскиваемая информация предназначена для использования в ос-

новой (профессиональной) деятельности и она (точнее, публикации об искомых объектах или каким-то образом связанных с ними), должна быть не только найдена, но также должна быть обоснована её полнота, точность и достоверность.

2. Требования к полноте, точности и достоверности информации, характеру процесса поиска, а в большей степени – к выбору типов и набора информационных ресурсов, а также последующей обработке найденного зависят от характера задачи (и текущего этапа жизненного цикла). Действительно, когда задача сформулирована в сложившейся предметной области и её актуальность не вызывает сомнений, цель поиска очевидна: найти полноценное изложение метода решения задачи данного типа (например, отчет о НИР, статью, учебник и т.д., достоверность которого не подлежит сомнениям). Во многих случаях (обычно, когда мы ищем уже известный человечеству метод) это можно сделать, не прибегая к «сложным» процедурам, использующим разнообразные, но *вспомогательные* средства: указатели, реферативно-библиографические БД и т.д. Достаточно «просто» пролистать разделы соответствующих учебников или монографий или, в крайнем случае – подшивку специальных журналов. Предложения библиотек и информационных служб использовать специальные справочно-поисковые средства, к тому же не содержащих полных текстов¹, кажутся многим современным пользователям абсурдными. Их использование становится неизбежным, когда собственные «подручные» ресурсы не позволяют найти решение, а характер ОД предполагает реальную ответственность (экономическую или юридическую) за принятие решения. Общеизвестными примерами являются задачи патентного поиска, позволяющего подтвердить приоритет изобретения, или научного поиска, доказывающего новизну решения.

Существенно другая роль вторичной информации на начальном и заключительном этапах ОД, в бизнес-планировании и в задачах управления качеством. Например, при определении направления деятельности, выборе решения при неполной информации,

¹ Необходимо отметить, что профессиональные БД вторичной научной информации обыкновенно позволяют почти автоматически выйти на полные тексты публикаций.

принятии решения о начале или завершении деятельности, оценке эффективности и применимости, оценке новизны и конкурентоспособность найденного решения. Очевидно, что искать документ, содержащий ответ, можно, если мы сознательно повторяем работы, выполненные кем-то ранее. То есть, информация такого рода – общее заключение (отдельное, «*дискретное*» решение), может быть только синтезирована на основе многоаспектного содержательного и статистического анализа «*непрерывного*» потока публикаций, отражающего не только разные подходы к решению, но и разные этапы жизненного цикла идеи.

Таким образом, можно сказать, что, в отличие от задач нахождения отдельной публикации о методе решения частной проблемы, своей конкретностью предопределяющих требование точности информационного представления, задачи ОД управленческого типа¹ предполагают широту и вариантность анализа предполагаемых решений, что предопределяет требование полноты и многоаспектности информационного обеспечения.

3. Обычно объектом информационного поиска является *предметное* содержание – данные, методы, инструкции и т.д., позволяющие решить или построить решение конкретной задачи ОД. При этом наиболее распространенной *коммуникативной* формой представления содержания является документ. Документ по своему статусу соответствует «завершенности» процесса ОД: излагаются, так или иначе, *проверенные* решения, *обоснованные* подходы, некоторым образом *принятые* гипотезы. Однако в некоторых случаях, когда исследование не завершено, или мы не знаем об этом, будет «естественным» обратиться к «источнику» – индивидуальному или коллективному автору, генерирующему новое знание. Отметим, что в информационной практике термин «источник» часто ассоциируется и с конкретными *средствами передачи* информации. С одной стороны, это – отдельные публикации (издания, специализирующиеся в данной предметной области), а с другой – организации (издательства, библиотечные коллекторы, книготорговцы),

¹ Выделение отдельных типов ОД здесь достаточно условно, поскольку при принятии «управленческого» решения мы обобщаем «частные» решения, а при поиске и выборе «частного» решения мы должны ориентироваться на общие направления и использовать общепринятые критерии.

обеспечивающие распространение публикаций, в том числе, по тематическому принципу.

Примечательно, что эта, в известной степени, искусственная типология не только отражает естественно сложившуюся технологию научных коммуникаций (отдельное решение, в абстрактной форме возникающее в сознании автора, оформляется в виде структурированного документа, который публикуется в издании, становясь, таким образом, доступным через систему доставки), но и, осознанно или нет, используется человеком в поисковой практике. Предметное содержание запроса пытаемся более или менее успешно выразить средствами естественного или искусственного языка: строим терминологическое выражение или соотносим предмет поиска с некоторой областью общепринятой классификации. Но столь же естественно мы используем и «ссылочные» технологии: прямое указание (чье-то мнение) о содержании публикации, или приставленные ссылки – косвенное упоминание содержательной связи, а, в крайнем случае – подряд просматриваем выпуски журналов, специализирующихся по данной тематике. Однако, как показывает анализ статистики запросов, поиск источников (трудов по имени индивидуального или коллективного автора, организаций, ведущих работу в этой области или выполняющих коммуникационные функции) в АИПС используется крайне редко.

4. С технологической точки зрения процесс поиска – это рутинный перебор документов¹, сосредоточенных в традиционных или электронных хранилищах более или менее полно представляющих интересующую нас тему. Отбор обыкновенно производится по содержанию документов. Однако здесь следует уточнить, что слово «содержание» в этом случае надо понимать условно. Содержание документа представляется в поисковой системе достаточно поверхностно – поисковым образом, перечисляющим основные понятия, который *система* и использует для «отбора» документа (а на самом деле – для формирования списка ссылок на эти докумен-

¹ Здесь и далее термин «документ» будет использоваться для обозначения собирательного понятия, связанного с такой формой представления информации, для которой характерна логическая завершенность (цельность содержания), а также физическая доступность и идентифицируемость (т.е., документ всегда имеет структуру, адрес, методы обработки и т.д.).

ты), и только на следующем шаге *человек*, обращаясь через сформированный системой список, получает собственно содержание документа и осмысливает возможность его использования.

Кроме того, достаточно очевидно, что по отношению к традиционным библиотечным *способы организации массивов и методы автоматизированного поиска* не отличаются принципиальной новизной. Поиск ведется либо путем последовательного просмотра ряда документов¹ до тех пор, пока не будет найдена нужная информация, либо с использованием указателей и каталогов, систематизирующих размещение документов по предметному, алфавитному или какому-либо другому принципу и, соответственно, облегчающих доступ к ним (просто сокращая объем перебора при просмотре).

Современным и достаточно простым для использования средством, обеспечивающим перемещение в пространстве Internet-документов, является Web-технология связывания документов гипертекстовыми ссылками, размещаемыми непосредственно в тексте. Такое очевидное, и, уже поэтому, эффективное средство навигации имеет, тем не менее, два существенных недостатка. С одной стороны, простановка ссылок – трудоемкая работа, выполняемая квалифицированными экспертами, и потому на самом деле ссылками обычно связано лишь очень ограниченное множество документов (а скорее, HTML-страниц). С другой стороны, связывание всегда проводится в контексте некоторой, но всегда *конкретной*, обычно, выбранной автором предметной области, что в некоторых поисковых ситуациях может привести к дезориентации или дезинформации пользователя, например, если тот же вопрос рассматривается в каком-либо другом аспекте или необходимо найти принципиально новое решение.

5. Важным фактором, влияющим на функциональные особенности реализаций информационных систем, является характер и организация хранения (доступа) информации. В этом смысле системы можно условно разделить на два класса:

- электронные каталоги и документальные ИС (как локальные базы данных);

¹ Что возможно только в случае очень маленького массива документов.

- поисковые машины (как системы поиска в распределенных массивах).

Однако следует отметить, что подобное деление скорее отражает не только единственно возможный на сегодняшний день компромисс практических потребностей и реальных возможностей промышленной реализации, но и в значительной степени обусловлено историей развития средств поиска. Первые – документальные ИС – берут начало от библиотечных информационных систем, ориентированных на традиционный каталожный поиск; вторые появились как вспомогательное средство поиска в сетевых средах, изначально предназначенных для оперативных электронных коммуникаций.

В основном именно электронные каталоги и документальные системы обеспечивают профессиональный поиск информации в локальных или распределенных базах данных. Наиболее известными примерами являются информационные ресурсы ВИНТИ и ИНИОН РАН, INIS, электронный каталог РГБ или библиотеки Конгресса США, которые обеспечивают, в основном, библиографический и тематический поиск.

Библиографический поиск обеспечивает *выявление* публикаций по их *выходным данным*, например, по именам авторов, датам публикаций и т.д. основополагающей предпосылкой здесь является фиксированная для конкретной базы данных модель представления информации, в соответствии с которой обеспечивается нормализованная (единообразная) запись элементов данных. Все это требует от пользователя далеко не очевидных знаний определенных соглашений. Кроме того, практика показывает, что подобные соглашения не всегда соблюдаются не только для различных баз данных, но иногда даже и в рамках одной базы данных.

Тематический поиск обеспечивает *отбор* документов по *семантическим признакам*, обобщенно представляющим его содержание. Здесь концептуальным положением является то, что содержание документа может быть представлено некоторой совокупностью понятий, характеризующих основной объект. Это позволяет достаточно эффективно использовать запросы в виде компактных комбинаций терминов – обычно двух-трех слов естественного или естественно-научного языка. Такое представление информации хорошо соответствует теоретико-множественным моделям поиска,

однако для многих пользователей создает ситуации непонимания поискового языка и самого процесса получения результата.

Отдельным направлением в развитии информационного поиска является *полнотекстовый поиск*, основная цель которого – обеспечить точный отбор за счет применения критериев, основанных на семантических категориях. Но здесь, не смотря на достаточно серьезные достижения в области анализа текста и появление промышленных полнотекстовых поисковых систем (в основном в сфере СМИ), ожидать скорого широкого внедрения систем полнотекстового поиска, в том числе и в область научной информации, не приходится уже хотя бы потому, что выявить и воспринять смысл (и тем более новые идеи) в научных публикациях может не всегда и не всякий естественный интеллект. И уж тем более сомнительно автоматически построить понятийную, точно отражающую смысл, структуру по тексту запроса из трех-четырех терминов.

Также важным, но часто не замечаемым фактором является реальная ограниченность полноты представления информации в конкретном массиве (точнее, источников информации, которые используются для формирования массива). Это особенно существенно при поиске в Internet: глобальная сеть сетей физически объединяет компьютеры практически всех крупнейших библиотек мира, однако вход пользователя в сеть не приводит автоматически к возможности использовать электронный каталог какой-либо из таких библиотек. То есть подключение к сети обеспечивает физическую доступность вычислительного комплекса хранящего ресурс, но доступность собственно *информационного ресурса* – обычно совокупности баз данных (документальных и фактографических массивов) и информационных технологий, часто ограничена технологическими, организационными, финансовыми или какими-либо другими условиями.

6. Наконец, следует отметить и некоторые особенности взаимоотношений человека и информационно-поисковых систем (ИПС), обусловленные «заторможенным развитием» последних: ИПС сохранили вопросно-ответную идеологию, свойственную уже ушедшим, традиционным системам пакетного информационного обслуживания, когда гарантом качества поиска был информационный работник. Задачей такого информационного посредника были не только понимание и интерпретация потребностей пользователя,

но и выбор ресурса и собственно поиск, что в итоге и обеспечивало эффективность (по крайней мере – профессиональность) поиска. В современных же условиях «информационного самообслуживания», пользователь, привыкший к интуитивному освоению программных средств, большинство из которых имеет существенно более простой интерфейс, часто неадекватно оценивает состояние и результаты поиска. Типичными, но «стратегически» фатальными, ошибками является принятие пользователем безапелляционного решения о «плохой» базе данных или о поисковой системе после получения неудовлетворительного или нулевого результата по первому же запросу (иногда даже не являющемуся правильным выражением поискового языка), или прекращение поиска, после получения известных публикаций (а не новой информации!).

1.3. Информационные коммуникации и свойства информации

Кругооборот информации, как и всякого продукта человеческой деятельности, подчиняется естественному циклу «создание–распространение–потребление».

Как отмечалось ранее, в традиционном цикле информационного обмена основной поток документальной информации идет по цепочке *автор–издательство–библиотека–читатель*, однако существуют и обходные пути, например: *автор–читатель*; *издательство–читатель* (подписка).

Новое знание (результат основной деятельности) воплощается обычно в форме сообщения – документа, реализующего преобразование смысла в текст. Такая «материализация» знаний субъекта обеспечивает унифицированную форму обобществления личного знания, а сравнительно низкая стоимость их тиражирования позволяет существенно расширить сферу потенциальных потребителей. Но, с другой стороны, низкая стоимость публикации (по сравнению со стоимостью получения самого результата) приводит к колоссальному и все возрастающему объему публикаций, а унифицированность форм представления обуславливает их внешнюю безликость. Кроме того, чтобы опубликованное сообщение стало стимулом для построения нового знания приемником, сообщение

должно быть не только воспринято (выделено среди других и помещено в среду последующей обработки), но также понято (выделен смысл) и вписано в систему наличного знания приемника (потенциально применено) или сохранено в долговременной памяти.

В свою очередь, формирование сообщения, явно или неявно связано с выбором или введением специальной терминологии, что часто сказывается не только на эффективности поиска, но и на адекватности передачи смысла.

Отсюда в частности следует, что «сигнальные» признаки, необходимые для обеспечения эффективности «узнавания» – первой фазы использования сообщения, могут быть сформированы путем введения явной систематизации. Это достаточно естественный путь – знания всегда системны, т.к. создаются в рамках некоторой системы понятий соответствующей отрасли знаний. И именно это обеспечивает приемлемые условия для ознакомления с более или менее представительным объемом публикаций в конкретной предметной области путем, отличным от полного перебора всех сообщений.

То есть процессы обработки и поиска информации не могут рассматриваться изолированно от процессов основной деятельности, поскольку обусловлены ими, а действенность информации обуславливается «коммуникационными» средствами ИС (например, лингвистическим обеспечением).

Основная и информационная деятельность

Процесс решения любой научной или практической задачи, где, так или иначе, возникает или используется информация, в общем случае включает следующие этапы [23].

1. Поиск сообщений. Создатель нового знания (потенциальный генератор информации) обращается к информационным ресурсам для получения информации, которая может быть использована им, например, в качестве концептуальной основы, экспериментальных, вспомогательных или опровергаемых данных и т. д. По отношению к среде он будет выступать в качестве потребителя информации, а информационная среда будет источником сообщений. При этом сообщения могут быть получены либо в виде услуги *информационных систем*, обеспечивающих поиск в различных

хранилищах, либо по другим каналам, например, путем личного общения с другими разработчиками, непосредственным обращением к результатам других исследований, в том числе еще не представленных в виде сообщений, и т. д.

2. Интерпретация сообщений. Вследствие уникальности конкретных условий решаемых задач язык полученного сообщения (в общем случае) может быть не тождественен «внутреннему языку» разработчика (создателя информации). Данный этап включает адаптацию сообщений, интерпретацию их содержания в терминах «внутреннего языка», а в итоге – извлечение из сообщений сведений, необходимых для решения поставленной задачи. Результат этапа – информационное обеспечение решаемой задачи, которое должно привести к повышению эффективности ее решения.

3. Решение задачи. На данном этапе, используя информационное обеспечение, а также собственные знания, прилагая определенные усилия, разработчик создает новую информацию, составляющую решение задачи. Эта информация зафиксирована на языке задачи, является достоянием разработчика или достаточно ограниченного множества лиц (организаций), связанных с конкретной разработкой и, как правило, для использования за пределами конкретной задачи будет требовать дополнительных затрат труда.

4. Создание сообщений. На данном этапе исследователь осуществляет интерпретацию полученного результата на «языке коммуникаций», т. е. подготавливает сообщение в «стандартной» форме, одной из тех, которые приняты на данном этапе развития предметной области вообще и информационных коммуникаций, в частности. Результатом этого этапа может быть статья, выступление на конференции, сообщение по электронной почте и т. д. Новая информация, оформленная как сообщение, уже представляет собой потенциальную общественную ценность для большого круга пользователей и решения других задач.

5. Распространение сообщений. На этом этапе создатели сообщений вступают в активное взаимодействие с системой научных коммуникаций, затрачивая определенные усилия (в основном организационного характера) по «вводу» сообщения в один (или несколько) из доступных каналов коммуникации (публикация, аудиторное выступление и т. д.). Эффективность данного этапа определяется как степенью усилий, предпринимаемых исследователем,

так и теми возможностями, которые ему предоставляет система коммуникации и, в частности, АИС.

Очевидно, что в общем случае перечисленные этапы реализуются сложным последовательно-параллельным образом, не обязательно все и в указанной последовательности. Кроме того, они не сопоставимы по затратам труда. Наиболее общее представление о взаимодействии потребителей-поставщиков информации иллюстрируется схемой на рис. 1.1.

Первый и пятый этапы являются этапами собственно информационной деятельности (ИД), поскольку их эффективность во многом определяется свойствами конкретных коммуникаций и информационных систем. Третий этап – собственно основная деятельность. Этапы второй и четвертый носят пограничный, диффузный характер и могут относиться как к основной, так и к информационной деятельности.

Показанные на рис. 1.1 «каналы» могут быть разделены на три типа:

- непосредственное рабочее взаимодействие разработчиков (связь 3-3) представляет собой постоянный обмен информацией в группе или коллективе в процессе совместной деятельности;
- непосредственное документальное взаимодействие (связь 4-2) заключается в оформлении результата и ограниченном контролируемом распространении (например – передача отчета или документации заказчику);
- опосредованное документальное взаимодействие (связь 5-1) состоит в опубликовании результата и его последующем неограниченном перемещении по каналам ИС. Информационные связи этого типа, осуществляющиеся через АИС, и являются объектом нашего рассмотрения.

На рис. 1.1, согласно терминологии [23], представлена схема информационного обмена. Реально же пользователь работает с источником информации по схеме информационного обслуживания, для которой характерна опосредованность, «разорванность», взаимодействия: сообщения «отчуждаются» от автора¹. Это приводит к

¹ Соответственно, источник информации ассоциируется уже с сообщением – носителем информации, а не с человеком или системой, которые являются источником в прямом смысле этого слова.

«рассеянию» информации, и поэтому для эффективного отыскания публикаций (по времени и необходимым ресурсам) необходимо создавать и использовать специальные справочно-поисковые средства.

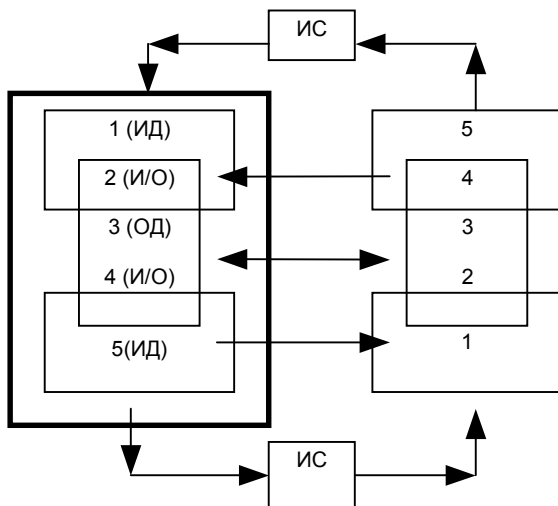


Рис. 1.1. Каналы взаимодействия потребителей-поставщиков информации

Концентрация и рассеяние информации

Рассеяние информации означает, что информация, которая была бы полезной для решения данной проблемы, может оказаться в документах, относящихся к совершенно другой предметной области. Следствием свойства рассеяния является дублирование информации – рассредоточение информации, полезной для некоторой предметной области по документам, изданиям и базам данных, относящимся к разным предметным областям.

Рассеяние информации обусловлено социально-экономическим характером научных коммуникаций, и в первую очередь экономическим и авторским правом. Действительно, автор в процессе обучения и работы не только относится к определенной научной школе, но и приобщается к достаточно ограниченному кругу изданий, которые, по его мнению, достаточно адекватно от-

ражают состояние предметной области. Очевидно, что вследствие ограниченности знаний, а также физической доступности многих изданий, этот список будет неполным, но, тем не менее, достаточным для более или менее успешной ОД. Аналогично, новую рукопись автор отправит в *одну* из редакций, руководствуясь разными факторами, например, авторитетностью, популярностью или даже нетребовательностью издания. В свою очередь, редакция, рассматривая рукопись, может счесть содержание не соответствующим тематике или уровню издания и отклонить её.

Процесс рассеяния имеет устойчивые закономерности. Журналы, которые посвящены непосредственно данной отрасли или предмету, образуют некоторое ядро, вокруг которого можно построить кольцевые зоны, содержащие столько же публикаций по данному вопросу, что и ядро. Этот закон С. Брэдфорд сформулировал следующим образом: «Если научные журналы расположить в порядке уменьшения числа помещенных в них статей по какому-либо заданному предмету, то в полученном списке можно выделить ядро журналов, посвященных непосредственно этому предмету, и несколько групп, каждая из которых содержит столько же статей, что и ядро. Тогда число журналов в ядре и в последующих группах будет относиться как $1 : n : n^2 \dots$ ». То есть, поскольку процесс генерации-публикации сообщений по своей природе является открытой динамической системой, тематическое распределение материалов по изданиям будет иметь ранговый характер.

Отметим, что свойство рассеяния информации собственно указывает не на ослабление, а на необходимость усиления междотраслевых связей. Именно поэтому при поиске информации следует сосредоточиваться не только на профильных для какой-либо отрасли науки или техники источниках, но и искать информацию в других, даже весьма отдаленных отраслях. По некоторым данным до 50% всей научной информации, получаемой современным исследователем, извлекается из смежных областей знания, причем именно она является наиболее значимой.

Здесь также следует отметить и не отраженную в явной форме специфику разделения информационных потоков по видам до-

кументов соответственно этапам жизненного цикла объекта ОД¹. Сложившаяся система информационных коммуникаций включает:

- «серую литературу» (отчеты о НИР и диссертации, впервые представляющие новые решения или гипотезы, еще не прошедшие широкую проверку практикой), представленную отдельными экземплярами;
- опубликованную научную информацию (статьи и монографии), отражающую в основном апробированные результаты, и изданную достаточно представительными тиражами;
- патенты – правовые документы, объективно отражающие область и уровень предлагаемых решений;
- экономическую, коммерческую и общественно-политическую информацию.

Основные ресурсы и инструменты информационной поддержки задач ОД разного типа приведены в табл. 1.1.

Концентрация информации достигается в процессе, который иногда называют *свертыванием* (т. е. обобщением, агрегированием).

Концентрация в сфере информационной деятельности проявляется, по крайней мере, в трех формах: документационной, фактографической и теоретико-концептуальной [20].

¹ Это означает, что реально цикл информационного обмена представляет собой спираль воспроизводства информации: результаты этапа фундаментальных исследований становятся основой на этапе прикладных исследований и разработок, которые в свою очередь, внедряются в производство и т.д. Причем каждый этап, естественно, включает все компоненты информационной деятельности.

Таблица 1.1

Основные цели, ресурсы и инструменты информационной поддержки

Тип цели поиска	Решение частной задачи			Бизнес-планирование, управление качеством
Тип объекта поиска	Документ	Источник	Лексика	Системное основание
Задачи	Выделение ядра/Периферии	Идентификация позиции в рамках научного направления		Анализ динамики
Операции	Выделение Существенного	Классификация/кластеризация	Выделение полей/терминологических структур	Оценка состояния и тенденций развития ПрО
Ресурсы	Тексты (ЭБ, ЭИ)	Реферативные БД	Рубрикаторы/Классификаторы	Онтологии Кодификаторы
Инструменты	Контекст-анализ	Мультиагентный поиск	Автоматическая классификация/Кластеризация	Построение терминологических структур
Результат	Автоматическое индексирование	Тематические подборки	Аспектные рубрикаторы, тезаурусы, онтологии ПрО	Частотные распределения и временные ряды
Итог	Построение документального представления частной задачи	Список основных поставщиков/генераторов информации		Формирование системного представления

Первая – документационная форма связана в основном с библиотечной и информационно-поисковой практикой, где в качестве единицы информации выступает *документ*. Простейшим из них является *библиографическая обработка документов*, в результате которой в обобщенном виде представляются основные, главным образом внешние *идентификационные* признаки документа (автор, заголовок, выходные данные и т. п.).

Более емким является *реферирование*, которое предполагает извлечение из документа основных положений содержания и их представление в виде реферата. Близким к этому виду концентрации информации является *классифицирование* и *индексирование*, в результате которого на некотором специальном, обычно искусственном, языке отражаются с большей или меньшей степенью глубины и полноты тематические или фактографические *признаки содержания* документов. Все эти виды концентрации информации предполагают неизбежные потери информации, поскольку в получаемом таким образом *вторичном документе* отражаются только основные аспекты содержания документов. Отметим также, что своеобразной формой концентрации информации является метainформация, например, системы классификации и кодирования, тезаурусы и онтологии предметных областей.

Таблица 1.2

Расcеяние/концентрация и доступность информации

	Авторские публикации	Отчет о НИР	Диссертация	Патент	Мат-л конф-ции	Статьи	Моно-графии
Тираж	1	3	3	300-1000	100-500	1000-10000	300-5000
Места хранения / доступа	Личная библиотека	ВНТИЦ	РГБ, ВНТИЦ	Патентная библиотека	РГБ, ГПНТБ, специализир. библи-ки.	Национальные, отраслевые, университетские, личные библиотеки	
Электронные средства доступа	Личный сайт	Портал ВНТИЦ	Портал РГБ, ВНТИЦ	Портал ФИПС	Сайт организации	ЭБ издательства, специализированных информационных центров	
Базы данных и эл. библиотеки	–	ВНТИЦ, INIS, CARI	ВНТИЦ	ФИПС	Международные, национальные, отраслевые информационные центры (INIS, ВИНТИ, АтомИнформ)		
Поисковые машины	Возможно	–	–	–	Возможно, ограничено	–	–

Вторая форма – фактографическая концентрация информации, ориентирована не на отдельный документ, а на совокупность

фактов или сообщений по определенной теме или проблеме. В качестве таких форм могут выступать реферативные обзоры, фактографические информационные картотеки и т. д.

Третья, высшая форма – теоретико-концептуальная концентрация информации – это такое преобразование (свертывание, агрегирование), которое позволяет представить научное знание на более высоком уровне обобщения. Примером являются аналитические формы представления математических зависимостей, законы, теоремы, позволяющие выводить различные следствия и т. д.

Характер рассеяния/концентрации и доступности в «традиционной» или электронной форме информации различных видов иллюстрирует пример, приведенный в табл. 1.2.

Отметим в заключение, что именно возможность работы с компактными хорошо структурированными вторичными документами (а не с их полнотекстовыми представлениями), предметно отражающими содержание первичного документа и обеспечивающим его идентифицируемость, как в целом, так и на уровне отдельных элементов данных, позволяет рационально реализовать поиск информации в сверхбольших документальных потоках.

1.4. Обобщенная схема информационного обслуживания

Схему взаимодействия потребителей-поставщиков информации (см. рис. 1.1) в общем случае можно преобразовать в схему информационного обслуживания, представленную на рис. 1.2.

Здесь информационные ресурсы, наряду с оригинальным авторским представлением материала, в большинстве своем характеризуются высокой систематизированностью (тематической профильностью источников и ядерностью тематических потоков), а также практически обязательным наличием метаинформации: поисковых образов документов и систем вторичной информации – рубрикаторов и тезаурусов, обеспечивающих единообразие представления и организации доступа к ресурсам.

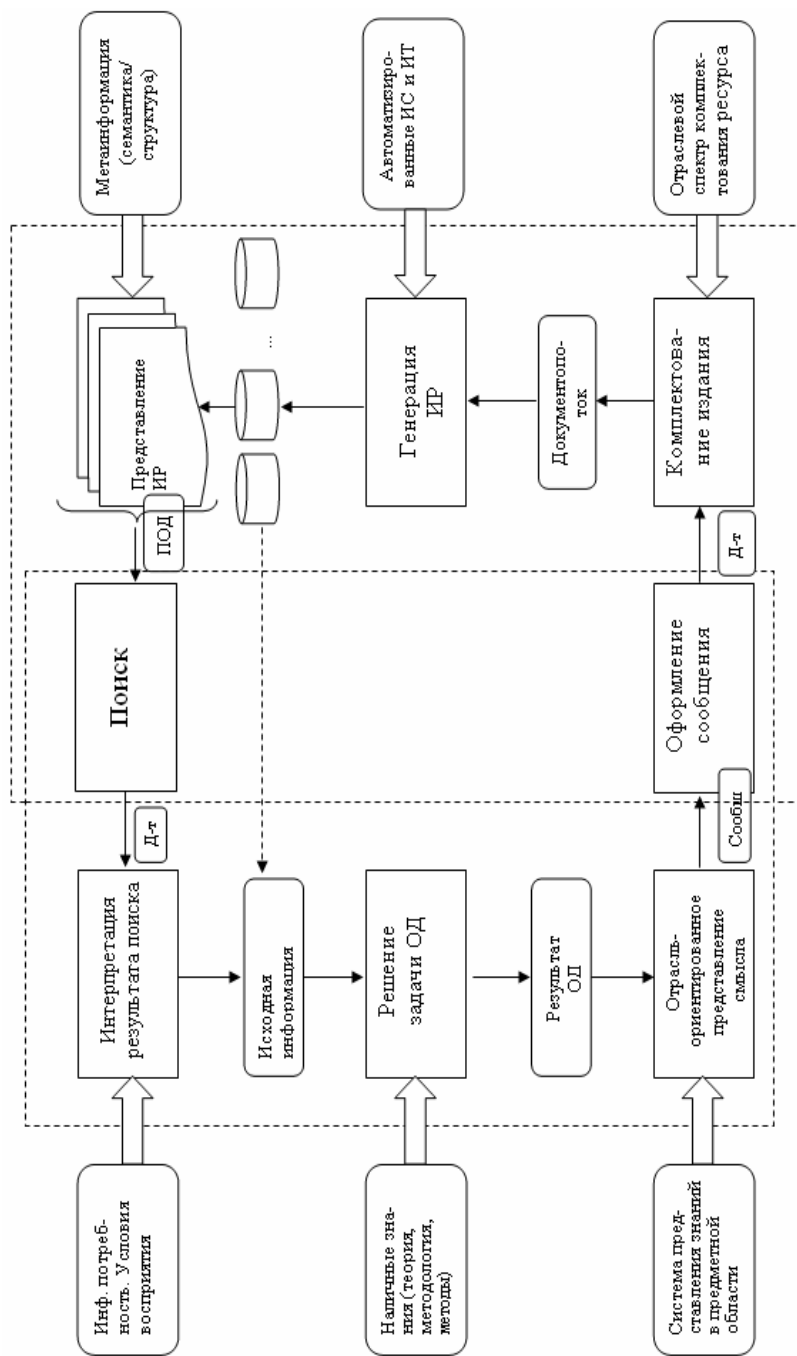


Рис. 1.2. Обобщенная схема информационного обслуживания

Операционными объектами собственно машинного поиска являются поисковый образ документа (ПОД) и поисковый образ запроса (ПОЗ), соответствие которых устанавливается поисковым механизмом АИПС на формальном уровне. Установление же истинного соответствия содержания документа информационной потребности предполагает соотнесение на смысловом уровне: пользователь как бы реконструирует возможное содержание по основным понятиям, представленным в ПОД, и далее полученный образ соотносит с реальной потребностью. При этом адекватность образа действительному содержанию документа определяется не только качеством индексирования, но и уровнем знания субъектом средств отражения – концептуальной схемы предметной области и возможностей информационно-поискового языка.

Применительно к конкретному информационному сообщению можно сказать, что каждый элемент ИД осуществляет семантические или форматные (но не синтетические) преобразования этого сообщения, внося ту или иную неопределенность. Например, неопределенность, вносимая структурно-форматными преобразованиями (выделение формальных поисковых признаков, форма представления и распространения сообщения и т. д.), может приводить к *ненахождению*, а неопределенность семантическая (адаптация содержания требованиям издания, например) – к *неузнаванию*.

Вследствие этого, информационно-поисковая деятельность должна представлять собой *не одноактную*, в общем случае, *итеративную* последовательность действий, обеспечивающих не только получение полезной информации, изменяющей состояние потребителя (точнее, в части решения задачи его ОД), но и данных, позволяющих объективно (и, желательно, количественно) оценить прагматические свойства найденной с помощью АИПС информации – полноту, достоверность, актуальность и т. д.

Это подчеркивает *активную* роль потребителя в «получении» информации: связь (типа 5-1 на рис. 1.1), выполняющая передачу информации, инициируется не системой, как это следовало бы из направления стрелки, а наоборот, *обращением* потребителя *посредством АИПС* к ресурсу, выбираемому им же, и именно он должен оценить результат взаимодействия и принять решение о его завершении или выборе другого ресурса.

Следует также отметить, что важной, но мало замечаемой, особенностью является принципиально разное восприятие пользователем и системой основных информационных объектов – документов и запросов. Человек рассматривает их как носители информации (смысл которой в общем случае может быть различным в зависимости от точки зрения пользователя), выделяя и преобразуя при этом отдельные фрагменты (часто не совпадающие с текстовым предложением, параграфом и т.п.) так, чтобы в сознании возникли устойчивые образы и понятия. Для АИПС те же объекты – это совокупности данных, из которых *механически* (не извлекая и не преобразуя смысла) выделены термины (слова, словосочетания, шифры, даты и т.д.), которые и сравниваются с терминами запроса.

Далее, человек считает документ полезным (наиболее соответствующим запросу), если тот несет новую, ранее не известную информацию, т.е. дает или позволяет найти ответ на некоторый *вопрос* («как?», «что?») при решении практической задачи. Система же считает наиболее соответствующим запросу документ, который содержит наибольшее количество терминов из запроса. То есть вполне вероятно, что пользователю в первую очередь будут выданы наиболее знакомые документы, возможно написанные им же, что вряд ли принесет ему новое знание.

Еще один важный момент связан с понятием *структура документа*. Для человека это понятие в большинстве случаев (по крайней мере, для текстовых документов) ассоциируется с удобством восприятия, т.е. *описание структуры* практически не используется, поскольку отдельные информационные поля документа *узнаются* обычно по косвенным признакам (угадываются). Для автоматизированных систем понятие структуры является неотъемлемым и изначально определяющим. Более того, для каждого структурного элемента (реквизита) документа обязательно определен свой формат, имя и, возможно, свой метод обработки. Например, способы записи дат или разные правила выделения терминов в разных текстовых полях (во многих системах знак пробела в поле ключевых слов не считается признаком разделения, позволяя таким образом выделять словосочетания).

1.5. Типология поисковых задач и форма выражения запроса

В зависимости от характера задачи основной деятельности пользователя по степени соотношения известного/неизвестного в предмете поиска можно выделить три типа поисковых задач.

К *задачам первого типа* относится поиск объекта, когда известно, что этот объект существует (например, поиск фактографии или трудов конкретного автора). Знания пользователя об искомом объекте полные, цель поиска – найти его документальное представление. Модель такого «атрибутивного» поиска может быть представлена как логическое выражение над именами понятий, задаваемыми терминами или их комбинациями.

Второй тип задач – подбор информации по некоторой теме, например, для обзора научной проблемы, обоснования или поиска метода решения научной или практической задачи. Пользователь, уже обладая знаниями, определяет место задачи (как вновь вводимое понятие в системе уже известных понятий), ищет документы, содержащие материал, с необходимой полнотой раскрывающий новую для него тему, или дающий возможность построения нового метода решения задачи. Поисковая модель в этом случае – это частично известные понятия, связи или комбинации. Тематический поиск реализуется как последовательность атрибутивных поисков, каждый из которых соответствует определенному (априорно заданному) аспекту представления объекта поиска.

Третий тип задач представляет собой проблемный поиск, который, по сути, является основной составляющей творческого процесса определения путей решения профессиональной задачи пользователя. Здесь изначально отсутствует четкость структуры знания: пользователь располагает отдельными фактами, возможно, не имеющими между собой доказанных связей. Проблемный поиск – это нахождение описаний объектов или их составляющих, актуально или потенциально существующих, и в совокупности, возможно, образующих целое, свойства которого возможно будут больше суммы свойств частей. То есть этим свойствам в явной форме могут не соответствовать «собственные» атрибуты, а новое свойство, например, может быть задано комбинацией уже извест-

ных атрибутов. В этом случае к неопределенности отображения объекта на предметную область ИС, свойственной тематическому поиску, добавляется неопределенность на уровне «субъект-объект ОД»: представление, которое субъект имеет об объекте поиска, может не соответствовать представлению конкретного источника. Логическая поисковая модель для этого случая – поиск «похожих» документов, содержание которых некоторым образом ассоциируется с задачей пользователя.

Собственно и информационная потребность имеет несколько форм, соответствующих разным стадиям процесса познания (когнитивным состояниям потребителя информации), для которых характерна разные формы проявления знания о незнании объекта поиска.

Реальная информационная потребность, отражая проблемную ситуацию пользователя в несистематизированной форме (она еще не вполне осознана), характерна для начальной стадии ОД.

В процессе понимания проблемной ситуации реальная ИП преобразуется в *осознанную ИП*, представленную в виде *вопроса или задачи* на привычном естественном или научном языке, и затем преобразуется в *поисковый запрос*. Для запроса характерно то, что вопросы типа «Как?» и «Почему?» должны быть преобразованы в вопрос типа «Есть ли?», поскольку именно такая форма представления потребности является наиболее адекватной теоретико-множественной модели поиска. Преобразование вопроса в запрос происходит в сознании человека и имеет качественный характер. Переход от реальной к осознанной ИП тем сложнее, чем менее определена задача ОД: для задач проблемного типа этот переход наиболее труден, так как пользователь не представляет, *какая именно* информация нужна для решения его задачи и не изменит ли она саму постановку задачи. Наиболее адекватной формой представления осознанной ИП как поискового запроса может быть семантическая сеть – граф понятий, характерных для объекта поиска.

Поисковый образ запроса – это *выраженная ИП* – представляется лингвистическими средствами конкретной АИПС, причем лексический состав ПОЗ уже в значительной степени будет зависеть от особенностей выбранного информационного ресурса. Формирование ПОЗ, в отличие от других форм ИП, производится в

среде АИПС, но эффективность процесса его составления определяется не только интерфейсными возможностями системы, но также и информационной грамотностью и профессионализмом пользователя.

Для человека идеальной коммуникативной формой представления знаний и потребностей является вербальная – в виде терминологического выражения. Принципиально важной особенностью вербального способа является изначальная контекстная определенность (хотя этот контекст, возможно, представлен только в сознании высказывающего). Т. е. отдельное высказывание, как грамматическая форма (предложение), в общем случае может порождать в сознании воспринимающего несколько смыслов, а исходный смысл высказывания будет воспринят только при условии одновременной передачи исходного контекста.

Запрос с точки зрения способа его представления – это так же, как и в случае документа, терминологическое выражение, представляющее гипотетический объект через описание свойств (атрибутов, связей), наличие которых как признаков (зачастую уже безотносительно характера атрибутов и связей) должна проверить ИС в документах БД. То есть ПОЗ должен быть построен по типу вопроса «Есть ли?».

Вербальной форме запроса свойственно то, что она предполагает построение завершенного, логически и синтаксически правильного выражения. Такой подход «по духу» отражает стремление к точности выражения, исключающей возможность многозначности ответа, и, соответственно, в минимальной степени учитывает свойство комбинативности. В условиях, когда семантическая неопределенность отсутствует (поисковая задача первого типа), вербальная форма, безусловно, предпочтительнее, как предпочтительнее аналитический способ задания математической функции по сравнению, например, с табличным.

С другой стороны, содержание потребности частично или полностью может быть представлено уже существующими документами (как сообщениями, содержащими частные решения проблемной ситуации или имеющими ними какую-либо семантическую общность). В этом случае можно говорить о **кластерной форме запроса**.

Эти формы являются альтернативными, но скорее взаимодействующими, чем взаимно исключаящими, что воплотилось в практике АИПС в виде двух уже привычных форм поискового запроса в диалоге «человек-система» – *запросно-ответной* и *гипертекстовой*.

С точки зрения свойства *комбинативности* и запрос, и документ являются моделями, представляющими средствами языка отдельные части и аспекты некоторого целостного фрагмента предметной области. Но при этом, цель создания ПОД – представить изначально уникальный смысл документа компактной композицией признаков (например, в случае дескрипторных ИПЯ – ключевыми словами), по возможности, не увеличивая комбинативность порождаемых ими возможных смыслов. Цель построения ПОЗ – сохраняя уникальность проблемной ситуации, увеличить комбинативность смыслов, порождаемых композицией поисковых признаков запроса, для того, чтобы максимально охватить аспекты представления объекта поиска.

Тот факт, что пользователь за новым знанием обращается в массив уже известного знания (хотя, возможно, и противоречивого¹), предопределяет очевидность того, что запрос представляется в виде *гипотетического документа*, описывающего реальный, создаваемый или воображаемый объект. То есть в этом контексте задача поиска может быть сформулирована следующим образом: *найти уже существующие документы, которые являются содержательным аналогом запрашиваемого гипотетического*.

1.6. Обобщенная схема поиска

Резюмируя ранее приведенное, можно сказать, что эффективность информационного поиска определяется следующими факторами:

- свойством концентрации-рассеяния информации, предопределяющим априорную неполноту любого отдельного ИР – источника информации практически по любой теме. Любой ресурс всегда ориентирован не только тематически (по отраслям знаний) и на

¹ Массив системы включает и документы, содержащие сведения, которые могут быть неполными, непроверенными, взаимно противоречивыми.

определенный вид информации (НТД, патенты, отчеты НИР и т. д.), но также имеет свои системы представления и средства доступа к информации;

- свойством эмерджентности информации, предполагающим множественность и комбинативность использования любого информационного сообщения;

- свойствами информационно-поисковой деятельности, зависящей как от характера задач ОД, так и от особенностей человека – его возможностей по восприятию и интерпретации найденных документов (информация может быть потенциально полезной, но актуально не воспринятой, например, по причине недостаточности знаний получившего её потребителя);

- свойствами вычислительной среды реализации АИПС, для которой характерна жесткость процедур и предопределенная ограниченность форм представления информации.

В целом процесс информационного поиска может быть представлен как итеративная цепочка операций, выполняемых в совокупной человеко-машинной среде (сознании человека и в вычислительной машине), последовательно снимающей неопределенности, обусловленные перечисленными ранее свойствами информации, и в итоге реализующей своеобразное преобразование информационной потребности в совокупность документов, содержание которых удовлетворяет эту потребность, т. е. информация найденных документов обеспечивает решение задачи ОД.

Укрупненный алгоритм итеративного процесса поиска информации представлен на рис. 1.3.

Такой подход позволяет рассматривать процесс поиска как *последовательное изменение состояний* (этапов) взаимодействующих подсистем (человека и автоматизированной информационно-поисковой системы), направленное на локализацию (снятие) неопределенностей следующих видов:

- 1) неопределенности соотношения «известного/неизвестного» в предмете поиска (свойственна реальной ИП);

- 2) неопределенности системы характеристических признаков для структуризации предмета поиска (свойственна осознанной ИП);

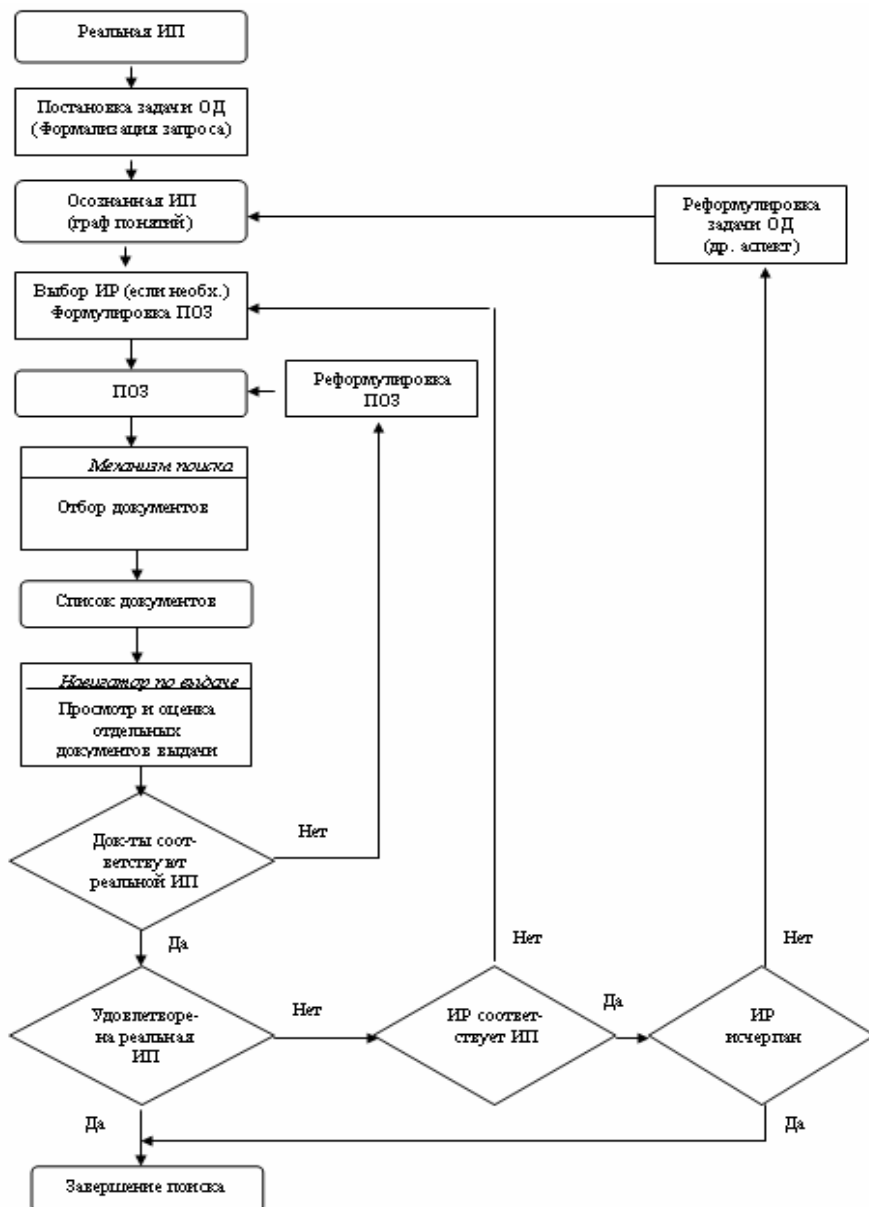


Рис. 1.3. Обобщенный алгоритм автоматизированного информационного поиска

3) лексической неопределенности, как фактора степени соответствия информационно-поискового языка естественно-научному языку предметной области (свойственна выраженной ИП);

4) неопределенности критериев сравнения поисковых образов (адекватность формальных мер близости, реализованных в конкретных АИПС);

5) неопределенности интерпретации ПОДов (субъективность и неполнота реконструирования пользователем смысла найденных документов);

6) неопределенность тематического соответствия и степени полноты представления проблематики в данном ИР.

Причем, первые четыре вида неопределенности имеют информационную природу (преобразование форм представления информации), пятая характеризует поисковый аппарат АИПС, а шестая отражает когнитивные особенности человека – приемника и генератора информации. Последняя существенна в том случае, когда в используемом ресурсе не была найдена информация, полностью обеспечивающая решение задачи ОД, и потребителю необходимо принимать одно из следующих решений:

1) продолжить поиск в этом ресурсе в надежде на то, что удастся так реформулировать запрос, что он выведет на нужный «пласт» информации;

2) перейти в другой ресурс (например, когда любая формулировка запроса дает отрицательный результат);

3) прекратить поиск и заняться непосредственно решением задачи (когда становится понятно, что легче открыть заново, чем найти описание открытия).

Не являясь практически измеримыми величинами, эти параметры, тем не менее, позволяют обозначить характер изменения состояния сторон.

Поскольку разные типы поисковых задач имеют разные типы и степени неопределенности, достаточно очевидно, что «траектория» поиска (циклы и число итераций) для каждого случая будет различным. И какой бы интеллектуальной система не была, роль человека будет определяющей. То есть в целом ситуация все еще отвечает тезису, что «...до тех пор, пока люди не научатся адекватно выражать на естественном языке свои информационные потреб-

ности, документальная ИПС должна быть организована таким образом, чтобы *человек* мог как бы *исследовать* поисковый массив, изменяя формулировку поискового предписания в зависимости от промежуточных результатов поиска» [18].

ГЛАВА 2. ТЕХНОЛОГИЧЕСКИЕ И ЛИНГВИСТИЧЕСКИЕ ОСНОВЫ ИНФОРМАЦИОННОГО ПОИСКА

2.1. Организация данных и критерии поиска

Как отмечалось ранее, процесс поиска всегда, так или иначе, сводится к процессу простого перебора – упорядоченного или случайного, полного или частичного. При этом достаточно очевидно, что степень соответствия («нужность») устанавливается не путем эмпирического «натурного» подбора – помещения очередного выбранного объекта непосредственно в конструкцию создаваемой системы, а путем сопоставления *образов* соотносимых объектов – описаний структуры и выполняемых функций, параметров, свойств и т. д.

Система производит отбор ПОД автоматически: механизм поиска включает в выдачу те документы, ПОЗ которых удовлетворяют формальному *критерию отбора*. Очевидно, что формальный (количественный) критерий соответствия может быть определен только в том случае, если соотносимые объекты имеют одинаковую природу (или приводятся к таковым), и, соответственно, сравниваемые атрибуты принадлежат одному пространству. То есть поскольку нельзя непосредственно сравнивать информационный образ с реальным объектом, для корректного соотнесения необходимо либо создать описание реального объекта, либо по образу (описанию) построить объект или его «действующий» макет – эквивалент объекта, удовлетворяющий требованиям решаемой задачи.

В отличие от логики сознания человека, где вопросы имеют форму «как», «почему», предполагающую развернутую форму от-

вета, машинная логика в основе своей может обрабатывать только вопросы типа «есть ли». В [15] приводится следующая типология простых (атомарных) запросов:

- 1). $A(E) = ?$ Каково значение атрибута A для объекта E ?
- 2). $A(?) = V$ Какие объекты имеют значение атрибута, равное V ?
- 3). $?(E) = V$ Какие атрибуты объекта E имеют значение, равное V ?
- 4). $?(E) = ?$ Какие значения атрибутов имеет объект E ?
- 5). $A(?) = ?$ Какие значения атрибут A имеет в наборе?
- 6). $?(?) = V$ Какие атрибуты объектов набора имеют значение, равное V ?

Для документальных систем, элементом данных, задающим смысл, является термин языка (слово, словосочетание) запрос сводится ко 2-му типу. А наиболее естественной формой критерия отбора будет являться предикат, построенный для выражения ПОЗ. Соответственно, документ считается *формально релевантным* (и, соответственно, включается в выдачу), если для данного документа предикат принимает значение «истина».

Простейшим вариантом критерия отбора является правило: документ считается *формально релевантным*, если количество общих слов, которые он имеет с запросом, не менее заданного порогового значения. Этот критерий может быть усилен путем взвешивания, в частности, двух-, трех- и многоуровневого ранжирования терминов по их важности. Сущность «весового» метода заключается в ранжировании пользователем терминов запроса по весовым коэффициентам и в последующем признании документа релевантным, если сумма весовых коэффициентов совпавших терминов больше заранее установленной величины.

Оценка семантической близости функцией «косинус» является своеобразной нормированной суммой «весовых» коэффициентов совпадающих терминов. Здесь запрос и документ представляются n -мерными векторами в пространстве терминов, где их i -е координаты принимают значения «единица» или «ноль» в зависимости от того, входит ли i -й термин в соответствующий поисковый образ. Документ считается *формально релевантным*, если мера принимает значение не меньше *порогового*.

В большинстве реальных ИПС критерий задается выражением алгебры логики, сформулированным пользователем на основании семантической структуры информационной потребности.

Неявным, но основным с точки зрения реализации алгоритма поиска, фактором здесь является порядок выборки. Выборка может проводиться в «естественном» порядке, соответствующем расположению объектов в массиве, или в «искусственном», соответствующем, например, некоторой классификации предметной области.

И в том и в другом случае имеем дело с *перебором* объектов, выбираемых для сравнения из хранилища. То есть рациональность построения процедуры поиска зависит от длины перебора, что в свою очередь определяется как характеристиками хранимых объектов (в первую очередь, размерами документов), так и характером запросов (в нашем примере – поиском по предмету или поиском по шифру хранения документов). Таким образом, оптимизация достигается сокращением перебора – длины *последовательно* проверяемого массива.

Заметим, что есть две классические технологии обработки запросов (так называемых *режимов информационного поиска*): режим *ретроспективного поиска* и режим *избирательного распределения информации*.

При ретроспективном поиске очередной ПОЗ сравнивается со всеми ПОД. В режиме избирательного распределения информации схема зеркально симметричная: ПОД каждого документа сравнивается со всеми поисковыми образами запросов. То есть, в первом случае запросы обрабатываются после создания массива ПОД, которые, накапливаясь, формируют *ретроспективную* БД, а во втором – массив ПОЗ создается до обработки документов (при этом хранится массив ПОЗ и необязательно – ПОД). Поэтому эти режимы иначе называют режимами обработки *разовых* и *постоянно действующих* запросов.

Соответственно, технологии (алгоритмы) поиска основываются на двух типах *организации массива* объектов поиска – *прямой* и *инвертированной*. Для рассмотрения взаимосвязи алгоритма поиска и организации массива здесь и далее используем знакомый

всем пример организации и поиска информации в традиционных библиотеках¹.

В случае прямой организации массива (хранилища) документы могут размещаться в последовательности, никак не связываемой с порядком какой-либо классификации или алфавита, в простейшем случае – в порядке их поступления. Но с точки зрения основного назначения АИПС – информационного обеспечения ОД, определяющим в понятии «*прямая организация*» является не характер размещения документов – единиц хранения, а размещение *содержания документов*, которое представлено изначальной «естественной» *последовательностью слов*, образующих, в том числе, и контекст их употребления. При прямой организации поиск в больших массивах будет требовать достаточно много времени, так как для сравнения с запросом надо последовательно выбирать *все*² документы из хранилища, по той простой причине, что, не обратившись к документу, мы не можем судить о его содержании.

В инвертированном массиве документы могут быть, например, разбиты на подмножества, которые, в свою очередь, упорядочены в соответствии с некоторой классификацией и, что особенно важно, обозначены идентификаторами, отражающими предметное содержание соответствующего класса (в пределе таким идентификатором может быть отдельный термин). Такое упорядочение документов в хранилище сопровождается построением вспомогательной структуры – *инвертированного справочника*, в котором с каждым *индексом* (идентификатором класса) связан список ссылок на документы, отнесенные к этому классу. Например, систематический каталог библиотеки имеет типично инвертированную организацию: карточки с шифром раздела представляют собой индекс, а

¹ Отметим, что выбор этого примера основывается не только на его «привычности» для человека, но и на том, что с методологической и системной точек зрения применяемые в библиотеках подходы, методы и технологии являются по существу универсальными и не зависящими от степени автоматизации.

² Конечно, перебор можно завершить, когда будет получен документ, отвечающий в рамках известного, и даже уже ставшего традиционным, подхода на все вопросы реальной ИП, однако, вполне возможно предположить, что мы при этом уже не дойдем до документов, опубликовавших новейшие достижения, опровергающие традиционный подход

инвертированный список – это расположенные за ней карточки с шифрами хранения соответствующих единиц. Для случая текстовых баз данных в качестве индекса может выступать термин из текста или термин из предопределенного словаря, выступающий в качестве смыслового эквивалента фрагмента текста, словосочетания или отдельного слова текста.

В контексте типологии простых запросов отметим, что запросы типа 1 выполняются поиском по «прямому» массиву: доступ к записи производится по первичному ключу. Запросы типа 2 выполняются поиском по инвертированному списку: доступ к записи производится по указателю, выбираемому из списка по значению вторичного ключа. Ответом в этих случаях будет *значение* атрибута или идентификатора. Запросы типа 3 имеют ответ – *имя* атрибута.

Запросы типа 2, 5, 6 относятся к нескольким атрибутам и в этом случае могут быть построены несколько индексов, облегчающих поиск по этим ключам.

Для документальных систем, которым свойственны в основном запросы 2-го типа, можно выделить три следующих типа архитектур доступа:

1. *Системы с вторичными индексами.* В этих системах последовательность расположения записей соответствует последовательности значений первичного ключа. Как правило, используется один первичный индекс и несколько вторичных.

2. *Системы частично инвертированных файлов.* В этих системах записи могут располагаться в произвольной последовательности. В отличие от систем первого типа первичный индекс отсутствует. Вторичные индексы применяются для прямой адресации записей, что существенно облегчает включение в файл новых записей, так как допускается их размещение в любом свободном участке файла.

3. *Системы полностью инвертированных файлов.* В этих системах предусмотрено наличие файлов, содержащих значения отдельных элементов данных, входящих в состав записей, – допускается раздельное хранение элементов данных записи. Значения элементов данных, составляющих конкретную запись или кортеж, в общем случае могут размещаться в памяти произвольно. Для ус-

корения процесса поиска в системе используют два набора индексов: *индекс экземпляров* (значений ключей) и *индекс данных* (инвертированный список). С помощью индекса экземпляров в файле можно найти элементы данных, имеющих заданное значение. С помощью индекса данных можно найти записи, связанные с заданными значениями элементов. Такая организация характерна для организации данных *документальных информационных систем*.

На рис. 2.1 приведена примерная схема организации данных для представления и поиска информации диалоговой системы поиска документов STAIRS (Storage and Information Retrieval System), разработанной фирмой IBM в 1970-х годах. Отметим, что такая структура не только хорошо иллюстрирует принципы организации данных в документальных системах, но и составляет основу большинства современных АИПС.



Рис. 2.1. Организация данных в документальной АИПС STAIRS

Физическая структура БД рассматриваемой системы включает в себя следующие четыре файла:

- файл частотного словаря, устанавливающий соответствие между словом, встречающимся в БД, его кодом и частотой, используется при текстовом поиске;

- инверсный (инвертированный, обратный) список, содержащий для каждого слова БД список документов, его содержащих, используется при текстовом поиске;

- текстовый файл, содержащий собственно документы, используется при выдаче (просмотре) документов;

- прямой, последовательный файл, содержащий «собранные» в одну строку фиксированной длины форматные поля и список кодов слов, находящихся в тексте данного документа. При необходимости в соответствующих местах находятся разделители сегментов и/или предложений. Файл используется при наличии в запросах конструкций *SENT*, *SEGM*, *CTX*, определяющих необходимость проверки взаимного расположения поисковых терминов.

На рис. 2.2 представлена структура индексных файлов *словарь слов*, в котором содержится перечень терминов, встречающихся в документах. Словарь содержит само слово, его характеристики и указатель на списки ссылок на документы, в которых встречается это слово.

Ввиду значительных размеров словаря, его организация предусматривает наличие специального индекса, представленного *матрицей пар знаков*. Каждой паре знаков поставлен в соответствие указатель на *блок словаря*, содержащий группу слов, начинающихся с этих знаков. Знаками могут быть буквы, цифры, а также специальные символы. Группы слов в словаре имеют переменную длину.

Некоторые слова в словаре могут иметь одинаковый смысл; такие слова связаны с помощью специального указателя «синоним» (на рисунке связи данного типа показаны штриховыми стрелками).

При использовании инвертированных форм представления информации на первом шаге проводится поиск в инвертированном справочнике и, если предмет запроса отождествлен с соответствующим классом, то на втором шаге для детального соотнесения содержания документа и запроса обращение будет производиться только к сравнительно небольшому числу документов – только к тем, которые отнесены к этому классу. Таким образом, за счет введения информационно-избыточной структуры и дополнительного шага поиска достигается существенный выигрыш во времени: суммарное время на поиск в инвертированном справочнике существенно меньше поиска в целом массиве документов, поскольку длина индекса, идентифицирующего содержание документа, обычно

Идентификация содержания с помощью индексов строится по принципам искусственных информационно-поисковых языков, где каждый индекс представляет то или иное множество характеристических признаков. Это позволяет сократить число просматриваемых документов: в соответствии с формулой композиции признаков (что хорошо реализуется выражением алгебры логики) производится слияние относящихся к разным индексам списков ссылок на документы, т. е. выбираются только те документы, которые описываются именно этим сочетанием.

Использование технологии индексирования, тем не менее, имеет принципиальные недостатки:

- 1) индексационная информация, относящаяся к документу, статична: индексы, приписанные к документу, будут всегда иметь смысл, заложенный при создании языка индексирования (например, классификации конкретного поколения);

- 2) нельзя без дополнительных затрат реализовать управление глубиной поиска, а также выполнить поиск с использованием критерия «частичного» соответствия.

И все же, автоматизация поиска информации основывается именно на технологии индексирования (как способа идентификации содержания и, соответственно, инвертированных форм представления информации) документов, поскольку документальные АИПС имеют следующие принципиально важные особенности построения и использования [27].

Во-первых, задачи в области документального поиска не сравнимы с другими задачами обработки текстов, такими, как, например, автоматический перевод. По существу документальные ИПС создаются только для того, чтобы указать потребителю те документы, которые, *скорее всего*, имеют отношение к данному интересующему его вопросу. Поэтому здесь можно ограничиваться довольно грубым раскрытием содержания документа, указывающим лишь основные моменты, вместо фразеологического анализа, необходимого, например, при переводе.

Во-вторых, поисковые системы создаются для обслуживания больших и часто разнородных групп потребителей. Поскольку последние могут иметь различные потребности и цели, поисковые запросы варьируются от вопросов обзорного или познавательного

характера до очень подробных аналитических запросов. При таких условиях слишком подробный анализ может оказаться излишне (или даже неприемлемо) специализированным для большинства пользователей.

В-третьих, в основе процесса оценки лежит некоторый критерий эффективности, обычно усредняемый по многим поисковым запросам. Это означает, что более предпочтительными оказываются такие методы, которые дают умеренно высокую общую эффективность, чем, может быть, более тонкие алгоритмы, которые могут превосходно обрабатывать одни запросы, но значительно хуже другие. Практически может оказаться, что для каждого вида запроса оптимальным будет некоторый специфический метод, но для среднего запроса наилучшими являются более простые методы индексирования.

Как очевидно следует из изложенного, высокое качество поиска, в первую очередь – полнота и точность, могут быть достигнуты только за счет метаинформационной и/или процедурной избыточности. Увеличение полноты достигается, например, следующими путями:

1) дополнительным индексированием документа (вплоть до индексирования несколькими методами каждого поля документа, включая полный текст);

2) расширением запроса за счет близкой по смыслу лексики, выбираемой пользователем или системой из дополнительных (метаинформационных) справочников, таких, как словари синонимов, тезаурусы;

3) использованием многостадийных итеративных процедур и/или нескольких механизмов поиска;

4) снижением точности запроса или порога выдачи, что позволяет, занижая требования к степени смыслового соответствия, увеличить вероятность попадания в выдачу истинно релевантных документов. Очевидно, что при этом в выдачу попадет во много раз больше нерелевантных документов и пользователь должен будет потратить больше времени на отбор истинно релевантных. Этот путь кажется малопривлекательным, однако, поскольку многие (если не большинство) промышленные ИР (см. примеры в разделе 4) достаточно ограниченно используют средства, перечисленные в первых трех пунктах, в некоторых случаях, это вариант является

единственно возможным для получения удовлетворительного результата.

Увеличение точности достигается, например, следующими путями:

1) использованием для индексирования словосочетаний, обычно дескрипторов ИПТ или словосочетаний, приведенных к нормализованной форме;

2) использованием при построении поискового образа документа и/или запроса статистики словоупотреблений и/или лингвистических процессоров, что позволяет «взвешивать» термины;

3) использованием сложных критериев отбора, дифференциально учитывающих роль и значимость терминов и терминологических конструкций;

4) использованием постобработки, упорядочивающей документы для просмотра, что позволяет сократить время пользователя.

Перечисленные средства не являются взаимоисключающими и широко исследуются и, в той или иной степени, применяются. Например, во многих поисковых машинах и системах (прежде всего в тех, поисковые интерфейсы которых в явной форме не ориентированы на алгебру логики) после отбора проводится постобработка, так или иначе упорядочивающая отобранные документы. В простейшем случае это может быть сортировка по какому-либо существенному для пользователя атрибуту (дате публикации, шифру классификации и т.д.) или степени соответствия (мере близости). В более сложных случаях система может рубрицировать выданные документы в соответствии с какой-то классификационной схемой, или кластеризовать их по степени взаимной близости. Существо постобработки с точки зрения сокращения объема документов, которые должен просматривать пользователь до момента фактического удовлетворения ИП, состоит в том, что достаточно большое множество всех найденных документов разделяется на сравнительно небольшие подклассы. И, поскольку классы обычно выделяются по тематическому признаку, пользователь, возможно, удовлетворится просмотром документов одного, двух классов, вполне обоснованно не обращаясь к остальным.

2.2. Функциональная обработка запросов и документов в АИПС

На рис. 2.3 представлена обобщенная схема обработки запросов и документов в АИПС.

С организационно-функциональной точки зрения в АИПС выделяются два контура: *обработки запросов* и *обработки документов*. В свою очередь, в контуре обработки документов могут выделяться (как отдельные подсистемы) контуры *первичной* и *вторичной* информации. Контур первичной информации выделяется в отдельную подсистему в том случае, если массив первичных документов размещается на иных типах носителей или использует отдельную систему управления данными, например, на микрофишах или специализированные хранилища CD ROM-носителей, не имеющих программных интерфейсов с АИПС.

С точки зрения функциональности в составе АИПС можно выделить три блока:

- *блок предобработки* – преобразование в машинную форму документов и запросов;
- *блок формирования базы данных АИПС* – загрузка ПОД и машинных форм документов (полных текстов) в базу данных
- *блок поиска* – отбор по поисковому образу запроса из множества ПОД, тех, которые удовлетворяют требованиям критерия смыслового соответствия;
- *блок постобработки* – упорядочение найденных документов.
- *блок выдачи* – форматирование и отображение материала найденных документов.

Как видно на рис. 2.3, изначальными являются процессы генерации информации и появление информационной потребности. Их возникновение происходит в сознании человека, однако выражение, так или иначе, связывается с конкретной предметной областью, её структурой и терминологией. При этом могут использоваться такие лингвистические средства, как тезаурусы предметных областей, язык представления онтологий (OWL) или язык представления знаний (KML). При этом для машинной формы материа-

лов, ориентированной на передачу, используются коммуникативные форматы, как например, ISO-2709 и ISO-8211, или XML, а для описания логической структуры ресурса, содержащего материалы, может использоваться язык описания ресурсов RDF.

Обработка поступающих в систему документов обычно включает:

- присвоение документу уникального идентификатора, необходимого для поиска, а также, возможно, для связывания ПОД с полным текстом документа, для чего может использоваться соответствующий кодификатор или, например, система идентификации цифровых объектов (Digital Object Identifier – DOI);
- преобразование во внутрисистемный формат, когда могут использоваться XML-схемы и язык определения документов DTD;
- индексирование и, возможно, реферирование – построение поискового образа (не обязательно автоматическое или автоматизированное) в рамках лингвистических средств АИПС, для чего используются словари, рубрикаторы, классификации, тезаурусы предметных областей;
- загрузку ПОД и, если в АИПС есть контур первичной информации, то и полного текста документа в базу данных. При этом используются языки определения и манипулирования данными соответствующей СУБД, а для оперативного взаимодействия с внешними ресурсами, например, XML-SQL.

При обработке запросов введенная пользователем формулировка преобразуется в соответствии с требованиями информационно-поискового языка (индексируется) и преобразуется во внутрисистемный формат в соответствии с правилами информационно-поискового языка конкретной АИПС. При этом используются словари системы, а для расширения (терминологического и тематического обогащения запроса) могут использоваться тезаурусы, онтологии, а также словари естественного языка.

На этапе отбора документов поисковый запрос, по тому или иному алгоритму, сопоставляется с поисковым образом документа и, если результат удовлетворяет критерию выдачи, который выступает в качестве *критерия смыслового соответствия*, то документ (точнее, его идентификатор в БД) включается в список результата поиска.

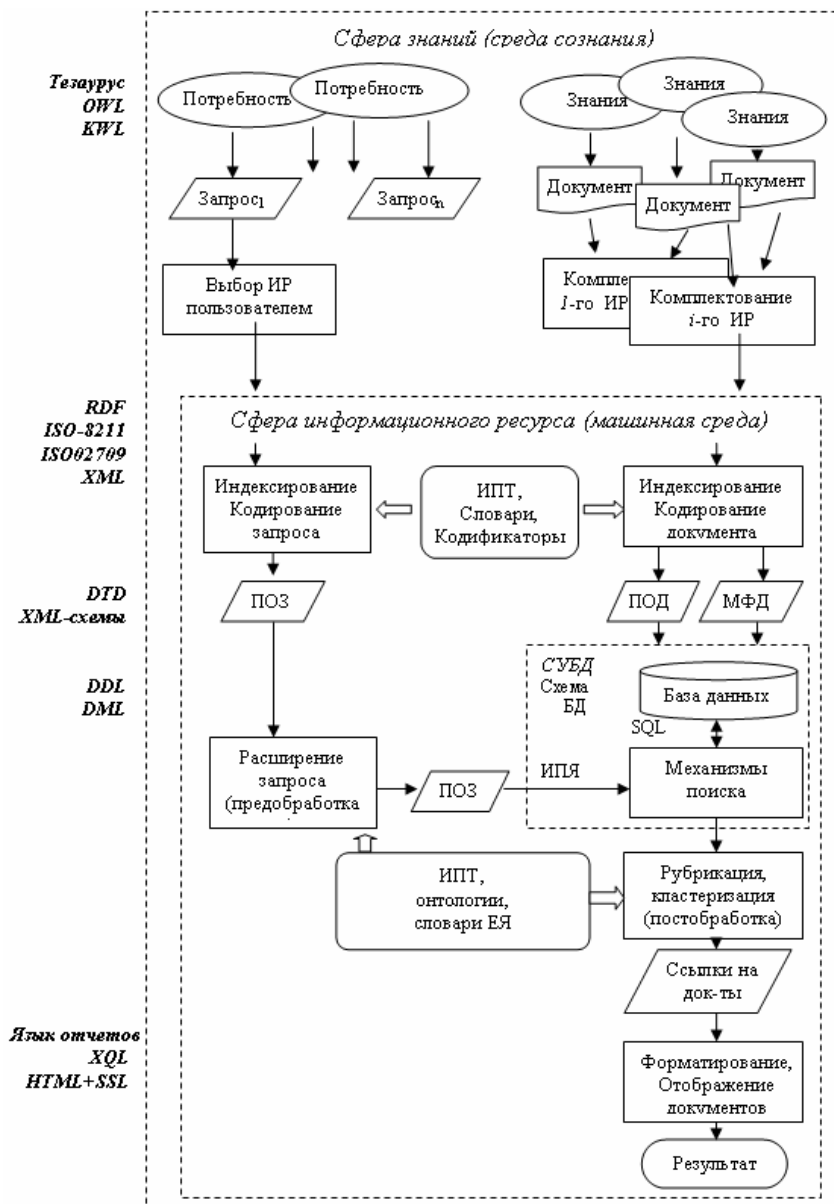


Рис. 2.3. Обобщенная схема обработки запросов и документов в АИПС

На этапе постобработки отобранные по ПОЗ документы могут группироваться (путем классификации или кластеризации) и ранжироваться, например, по степени соответствия запросу. При этом для обогащения ПОДа и уточнения возможных (осмысленных) сочетаний лексических единиц за счет устойчивых семантических связей используются словари, тезаурусы, онтологии предметной области, а также словари естественного языка.

На этапе выдачи документы из внутренней машинной формы преобразуются в форму удобную для восприятия человеком и, более или менее, адекватную его задачам. При этом используются языки отчетов, HTML+SSL и т.д.

Отметим, что в целом лингвистические средства, упомянутые выше, могут быть с той или иной точностью отнесены либо к группе, обеспечивающей форму представления информации, либо к группе, обеспечивающей представление содержания. К первой группе относятся коммуникативные форматы, схемы документов и баз данных, языки отчетов, HTML, XML, DTD, RDF. Ко второй – тезаурусы, классификации, рубрикаторы, кодификаторы, ИПЯ, языки онтологий. Более подробно основные из перечисленных компонентов ЛО будут рассмотрены далее.

2.3. Лингвистическое обеспечение и обработка информации в АИПС

Лингвистическое обеспечение – это совокупность языковых средств (в том числе и правила реферирования и индексирования!), позволяющих более или менее подготовленному пользователю взаимодействовать с машинной системой (по крайней мере, в части отыскания нужных пользователю документов).

Как уже отмечалось, принципиальное различие сред обработки информации (сознание человека – память ЭВМ) предопределяет, что общий для них язык общения будет далек от естественного, и его конкретные возможности будут определяться, прежде всего, спектром задач АИПС и требованиями к уровню эффективности. Он должен быть *безусловно приемлем* для машинной обработки, в то время как его приемлемость для человека может быть достаточно условной. Язык может обладать большими возможностями

ми, но быть сложным в освоении для человека (что ограничивает круг пользователей), и очевидно, что упрощение¹ языка будет снижать эффективность отбора.

Если язык запросов («утилитарная» поисковая часть ЛО) так или иначе известен всем, кто проводит поиск, то другая часть, обеспечивающая индексирование документов, не столь заметна пользователю (уже потому, что пользователь не выполняет индексирования, а синтаксис языка запросов не используется в ПОДе). ЛО имеет еще одно, «системное», предназначение – согласование точек зрения на предметную область как для различных поставщиков и потребителей информации, так и во времени: ИПЯ имеет искусственную природу и создается в конкретных условиях и в конкретное время (т.е. каждый его элемент имеет контекст, используемый в это время). Таким образом, если со временем контекст употребления термина изменился, то для адекватного восприятия смысла документа, содержащего этот термин, необходимо иметь возможность явно обратиться к «старому» контексту, который, следовательно, необходимо каким-то образом фиксировать.

Методы представления содержания документов

В основу методов представления смысла положена та или иная знаковая система, но при этом различают классификационный и описательный подходы.

Классификация как средство описания содержания документа представляет собой процесс соотнесения содержания документов с понятиями, зафиксированными в заранее составленных систематических схемах. Классификационные методы обеспечивают систематизацию объектов в соответствии с некоторой заданной классификационной схемой. Код, присвоенный отдельному классу (равно как и мнемоническое обозначение), обеспечивает его полную идентификацию в рамках конкретного классификатора.

¹ Удобным упрощением для пользователя является применение естественного языка, для которого разработаны достаточно хорошие анализаторы. Однако, в любом случае, это не будет идеальным решением, т.к. программа использует априорные сведения о предметной области и не может учитывать ситуативность пользовательской задачи.

Описательные методы идентификации используются, как правило, в тех случаях, когда необходимо идентифицировать конкретный объект или группу объектов путем описания произвольного набора его характеристик. Описательный метод предполагает наряду с указанием классификационных характеристик выделение дополнительных наборов свойств, углубляющих характеристику объекта и сужающих область поиска.

В основе любой классификации лежит принцип деления объектов рассматриваемой предметной области. Каждый объект (материальный или нематериальный) с точки зрения решаемых классификацией задач характеризуется фиксированным множеством свойств, совокупность значений которых может говорить об эквивалентности (или близости) данного объекта некоторому множеству объектов, обладающих определенной общностью.

При построении *иерархической классификации* деление на каждом уровне иерархии должно проводиться только по одному основанию. Но соблюдение установленных формально-логических правил классификации не устраняет ее главного недостатка – невозможность проведения группировки документов и информационного поиска по любому сочетанию характеристик: для построения иерархической классификации используется определенный ряд атрибутов (оснований деления), применяющихся только в одной последовательности. В иерархической классификации отдельные науки «разорваны» разветвлениями жесткого классификационного дерева. Таким образом, требуется периодическое изменение общей структуры той или иной реальной иерархической классификации, но на разработку новых таблиц классификаций уходит много времени и труда, и новый вариант устаревает раньше, чем удастся завершить работу по реклассифицированию документов.

Фасетная система классификации в отличие от иерархической позволяет выбирать признаки классификации независимо как друг от друга, так и от семантического содержания классифицируемого объекта. Признаки классификации называются *фасетами* (facet – рамка). При создании фасетной классификации деление на классы проводится на основе всех возможных комбинаций признаков. Каждый фасет содержит совокупность значений соответствующего классификационного признака. Недостатком фасетной системы классификации является сложность ее построения, так

как необходимо учитывать все многообразие классификационных признаков.

Описательный подход в большей степени поддерживается языками *дескрипторного типа*. Они реализуют *координатное индексирование*, которое заключается в формировании описания документа как совокупности дескрипторов, выбираемых из заранее созданных словарей понятий либо из текстов документов.

Метод координатного индексирования базируется на положении, что основное смысловое содержание документа и информационной потребности может быть с достаточной степенью точности и полноты выражено соответствующим списком так называемых *ключевых слов*, которые явно или в скрытом виде содержатся в тексте. Под ключевыми словами в данном случае понимаются наиболее существенные для этой цели слова и словосочетания, обладающие назывной (номинативной) функцией.

Контроль за ключевыми словами может иметь разные степени. При *отсутствии* контроля для координатного индексирования документа или информационного запроса ключевые слова выбираются непосредственно из текста документа без учета того, какие ключевые слова уже использовались ранее для индексирования таких же или близких по смыслу документов и информационных запросов. В этом случае не устраняется синонимия, полисемия и омонимия ключевых слов, а их грамматические формы даже не приводятся к нормальному виду. При *полном* контроле за словарным составом ИПЯ для индексирования документов и информационных запросов разрешено использовать лишь *дескрипторы*, т. е. такие ключевые слова, которые содержатся в некотором нормативном списке (например, в тезаурусе). В таком списке или словаре устранены синонимия, полисемия и омонимия ключевых слов, а также обозначены определенные парадигматические связи между ними.

Контролируемый словарь предполагает ведение некоторой лексической базы данных, добавление терминов в которую проводится администратором системы (то есть новые документы индексируются только теми терминами, которые есть в этой базе). Свободный словарь пополняется автоматически по мере появления новых терминов в новых документах.

В отличие от языка, построенного на основе классификации (УДК, рубрикаторы и др.), который позволяет потребителю легко найти свое место в информационной среде, как бы причислив себя к классу других потребителей, дескрипторный язык дает возможность «индивидуализироваться», отбирать документы по существенным только для него (потребителя) признакам. «Технологически», дескрипторный язык может выступать как дополнение к классификационному

Язык запросов документальной АИПС

Наиболее распространенным ИПЯ является язык, позволяющий составить логические выражения запроса из набора терминов. Термины могут связываться булевыми операторами AND, OR, NOT (И, ИЛИ, НЕ).

Например, запрос *((информационная AND система) OR ИПС) NOT СУБД* означает: найти все документы, которые содержат одновременно слова «информационная» и «система», и/или слово «ИПС», но не содержат слово «СУБД».

В случае, когда система позволяет создавать запросы на «естественном языке», фраза запроса обычно разбивается на слова, из этого списка удаляются запрещенные и общие слова, производится нормализация лексики, а затем все слова связываются либо логическим AND, либо OR. Таким образом, запрос: *Программы для Unix* будет преобразован в *Unix AND Программы*, что будет означать следующее: «Найти все документы, в которых слова «Unix» и «программы» встречаются одновременно».

Кроме обычного набора логических операторов AND, OR, NOT большинство систем позволяет использовать **контекстные** операторы NEAR, CTX, SENT, PAR, обеспечивающие уточнение запроса требованием взаимного расположения терминов в документе.

Поскольку все документы обычно состоят из **полей**, в запросе можно указать, в какой части (поле) документа пользователь хочет увидеть поисковый термин (в ссылке, заголовке и т.п.). Для отдельных систем можно также задать поле, по значению которого будет упорядочена выдача.

Информационно-поисковый запрос как операционный объект представляет собой совокупность семантически связанных¹ *предложений запроса*. Понятие «Запрос» надо отличать от понятия «Поисковый образ запроса». ПОЗ – это лингвистическая конструкция, задающая условие отбора документов, в то время как запрос отражает объединенную общей тематикой последовательность поисковых действий, направленных на получение полного и логически завершенного результата.

С одной стороны, запрос – это несколько отдельных условий, позволяющих «по частям» выразить ИП, а с другой – это совокупность результатов, каждый из которых не только отвечает некоторой части ИП, но и, возможно, получен своим методом поиска. Это позволяет избирательно использовать результаты отдельных предложений, объединять поисковые результаты, выделять общее множество релевантных документов и т. п.

Предложение запроса – это структурная единица *Запроса*. В нотации Бэкуса-Наура *Предложение запроса* имеет следующий синтаксис:

<Предложение запроса> ::= <Условие поиска> |
 <Предложение запроса><Логическая операция><Предложение запроса>|
 (<Предложение запроса>)
 <Логическая операция> ::= И | AND | ИЛИ | OR | , | НЕ | NOT | ^

Предложение запроса в общем случае состоит из произвольного числа *Условий поиска*, связанных логическими операциями И (AND, «пробел»), ИЛИ (OR, «,») и НЕ (NOT, «^»). Внутри предложения допускается использование скобок, задающих приоритеты выполнения условий поиска.

Условие поиска устанавливает критерии соответствия поисковых дескрипторов запроса некоторой области поиска, представляющей собой совокупность структурных единиц документа – полей:

<Условие поиска> ::=
 <Область поиска><Оператор критерия><Выражение условия> |
 <Результат поиска>

¹ В общем случае (и обычно) семантическая связь предложений между собой не поддерживается синтаксисом ИПЯ.

Область поиска внутри документа задается именем отдельного поля или логическим выражением, объединяющим имена нескольких полей.

Выражение условия – набор терминов (поисковых дескрипторов) или идентификаторов результатов поиска (см. п. 4.1.4), объединенных с помощью булевых или контекстных операторов в логическое выражение.

Оператор критерия задает условие включения или сравнения дескрипторов запроса и терминов, содержащихся в указанных полях документов.

В простейшем случае условие поиска состоит из имени поля, оператора вхождения и одного дескриптора, например:

KW: РОССИЯ

Область поиска задается именами структурных единиц документа – полей

$\langle \text{Область поиска} \rangle ::= \langle \text{Имя поля} \rangle |$

$(\langle \text{Область поиска} \rangle \langle \text{Логическая операция} \rangle \langle \text{Область поиска} \rangle)$

Из нотации видно, что допускается использование логических операций¹ при формировании области поиска. Например:

(AB OR TI): (РОССИЯ NOT СССР)

означает, что в результат поиска включаются все документы, в которых хотя бы в одном из заданных полей (или в обоих) встречается дескриптор РОССИЯ, но не встречается дескриптор СССР.

Если в условии поиска область поиска явно не задана, то поиск проводится в области, заданной «по умолчанию». Область поиска «по умолчанию» задается обычно либо средствами описания документа (схемой), либо параметрами интерфейсных форм построения запроса.

Оператор критерия. Для связи области поиска с выражением условия используются оператор вхождения или один из операторов сравнения.

¹ Не все ИПЯ допускают использование логических операций для указания области поиска.

Оператор вхождения (":") позволяет найти документы, которые в указанной области содержат термины в сочетании указанном выражением условия;

Операторы сравнения (= | **EQ** | < | **NE** | > | **GT** | >= | **GE** | < | **LT** | <= | **LE**) позволяют найти документы, значения указанного поля которых соотносятся со значением, указанным в правой части выражения условия согласно оператору. Например, условие **DT** < **2000** позволит отобрать документы, дата публикации которых (поле DT) имеет значение меньше «2000». Существенно знать тип поля, так как например, цифровые поля могут сравниваться как по правилам числовых, так и символьных данных.

Выражение условия. Синтаксис выражения условия в языке запросов следующий:

```
<Выражение условия> ::= <Дескриптор> |  
    <Выражение условия> <Операция> <Выражение условия> |  
    (<Выражение условия> <Операция> <Выражение условия>)  
<Операция> ::= <Логическая операция> | <Контекстная операция>  
<Контекстная операция> ::=  
    CTX|CTX[N]|+|NEAR|NEAR[N]|SENT|CON[N]
```

При использовании в запросе нескольких дескрипторов они должны быть связаны контекстными или логическими операторами, а для указания приоритета выполнения операций – помещены в круглые скобки.

Приведем описание основных логических операций, примеры их использования и графическую интерпретацию (результат операции – затемненная область):

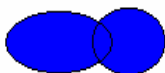
□ Логическая операция **OR (ИЛИ)**

Например:

KW:('ЯДЕРНЫЙ РЕАКТОР' OR 'ЯДЕРНЫЙ ЦИКЛ')

означает, что в результаты поиска включаются все документы, в которых в поле *KW* встречаются термины (словосочетания) «ЯДЕРНЫЙ РЕАКТОР» или «ЯДЕРНЫЙ ЦИКЛ», или оба вместе

ЯДЕРНЫЙ РЕАКТОР



ЯДЕРНЫЙ ЦИКЛ

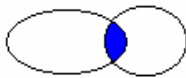
□ Логическая операция **AND (И)**

Например:

KW:('ЯДЕРНЫЙ РЕАКТОР' AND 'ЯДЕРНЫЙ ЦИКЛ')

означает, что в результаты поиска включаются только те документы, в которых в поле **KW** встречаются оба термина «ЯДЕРНЫЙ РЕАКТОР» и «ЯДЕРНЫЙ ЦИКЛ»

ЯДЕРНЫЙ РЕАКТОР



ЯДЕРНЫЙ ЦИКЛ

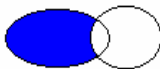
□ Логическая операция **NOT (НЕ)**

Например:

KW:(ЯДЕРНЫЙ РЕАКТОР' NOT 'ЯДЕРНЫЙ ЦИКЛ')

означает, что в результаты поиска включаются документы, в которых в поле **KW** встречается термин «ЯДЕРНЫЙ РЕАКТОР» и не встречается «ЯДЕРНЫЙ ЦИКЛ»

ЯДЕРНЫЙ РЕАКТОР



ЯДЕРНЫЙ ЦИКЛ

Контекстные операторы. К этой группе операторов относятся:

оператор расстояния (NEAR[N]);

оператор расстояния со строгим следованием (CTX[N]);

оператор предложения (SENT);

оператор пересечения полей (CON[N]).

Параметр N в операторах NEAR и CTX может принимать значения от 0 до 255 (по умолчанию N равно 0). Отсутствие параметра означает следование терминов в поле непосредственно друг за другом (идентично значению 0).

Оператор NEAR позволяет найти документы, в заданной области поиска которых в одном предложении присутствуют поисковые дескрипторы на расстоянии N слов друг от друга (в произвольном порядке). Выражение условия имеет вид:

<дескриптор1> NEAR[N] <дескриптор2>

Оператор CTX позволяет найти документы, в заданной области поиска которых в одном предложении присутствуют поисковые дескрипторы, расположенные в указанном порядке на расстоянии не более N слов друг от друга. Выражение условия имеет вид:

<дескриптор1> CTX[N] <дескриптор2>

Оператор SENT позволяет найти документы, в заданной области поиска которых поисковые дескрипторы находятся в одном предложении. Выражение условия имеет вид:

<дескриптор1> SENT <дескриптор2>

При задании поисковых дескрипторов допускается использование операторов (символов) маскирования, средств нормализации и ссылок на ранее полученные результаты поиска.

Маскирование реализует функции двух видов:

- маскирование (или замена) произвольного числа рядом стоящих символов дескриптора (оператор маскирования – символы «*» или «\$»);
- маскирование одного (непустого) символа дескриптора (оператор маскирования – символ «%»).

Символы маскирования могут использоваться вместо любого символа дескриптора, и их количество внутри дескриптора не ограничено. Параметризованные операторы маскирования $*(N)$, означают, что в дескрипторе на месте символа маскирования может стоять произвольная последовательность длиной не более N символов (где N от 0 до 255¹).

¹ Но, естественно, не более чем длина термина.

Использование ранее полученных результатов поиска. В качестве операнда условия поиска в предложении запроса может использоваться ранее полученный *результат поиска*:

<Результат поиска> ::= # <Идентификатор результата поиска>

Для включения в предложение поискового запроса результатов ранее проведенного поиска используются ссылки на номер предложения в текущем запросе.

Например, запрос может иметь вид:

#2 AND ((KW OR AB) : Россия)

где #2 – ссылка на результат второго предложения запроса.

Символ «#» является индикатором ссылки. За ним указывается номер одного из предыдущих предложений текущего запроса или имя сохраненного запроса, результат поиска по последнему предложению которого используется для уточнения в этом предложении.

Индексирование и реферирование

Ранее понятие индексирования рассматривалось в основном с точки зрения организации физического доступа к документу (точнее, к записи в БД). В этом разделе индексирование и реферирование рассматриваются как методы концентрации, сжатия информации. Это семантические аспекты представления информации, где характер преобразований определяется природой объектов, т.е. для них должны быть определены системы знаков, которые (или их коды) и будут использоваться для построения индексов.

Важно понимать, что целесообразность сжатия определяется вовсе не необходимостью экономии машинной памяти, а особенностями характера самого процесса поиска. Во-первых, как отмечалось ранее, собственно отбор сводится к простейшей (элементарной) процедуре сопоставления машинных кодов, представляющих запрос и документ. Во-вторых, характер взаимодействия потребителя с системой имеет «телеграфный» стиль: запрос очень похож на простое перечисление основных понятий, характеризующих основной предмет поиска. Это происходит не по причине лени пользователя. При поиске иначе и быть не может: подробное изложение существа проблемы не только займет все время пользовате-

ля, но и заставит сосредоточиться на *отдельном* решении, в то время как задачи поиска имеют обратную направленность – отыскать как можно больше вариантов и подходов к решению, среди которых пользователь сможет впоследствии (после акта поиска) выбрать наиболее подходящие.

В документальных системах в качестве таких смысловых сверток используют рефераты, аннотации и поисковые образы. При такой обработке полного текста основная задача состоит в возможно более полном извлечении из документа фактической информации. Характер обработки достаточно хорошо представлен в Методике индексирования [17], фрагмент которой приведен в Приложении 1.

Такая предобработка имеет качественный характер. По существу мы переходим от непрерывного текста (точнее непрерывного изложения содержания и, соответственно, последовательного и непрерывного процесса построения знания в сознании воспринимающего) к дискретной форме перечисления основных понятий, используемых источником и приемником информации для обозначения содержания в краткой неизбыточной перечислительной форме.

Реферирование предполагает извлечение из документа основных положений содержания и их представление в виде реферата.

Индексирование текста (в основном, автоматическое) предполагает выделение (или приписывание) слов и словосочетаний, обозначающих основные понятия, образующие содержание документа.

При индексировании часто используются списки запрещенных слов, которые не могут быть включены в ПОД (общие слова, предлоги, союзы и т.п.), а также иногда применяется нормализация лексики. Таким образом, даже то, что называется полнотекстовым индексированием, реально является выбором слов из текста документа и результатом сравнения с целым набором различных словарей, после чего термин попадает в поисковый образ документа, а потом и в индекс системы.

Например, для следующего реферативно-библиографического описания:

Trosow Samuel E. Управление библиотечным и информационным центром. Libr. Quart., 2000, № 70, 153-155

Рецензируемая книга давно стала стандартным учебником по курсу управления в рамках библиотечной и информационной науки, охват материала в котором расширялся с каждым очередным изданием. В предисловии отмечается, что значительные изменения в окружающих условиях, вызванные внутренними и внешними факторами, требуют более систематичного подхода к обзору функций в условиях организации. Технология, политическая, экономическая и социальная среда указываются как наиболее мощные силы для изменений. Помимо отдельных глав, посвященных перечисленным вопросам, авторы уделяют большое внимание теории организации, связи социологической теории и ее применений в организационном анализе, различным уровням анализа, на которых могут изучаться организационные явления.

По содержанию реферата будет построен поисковый образ с использованием нормализованной лексики, включающий термины: библиотеки, информационные центры, управление, функции руководителей, рецензии, США.

ПОД по заголовку документа будет построен на основе свободного индексирования и будет включать отдельные слова: управление, библиотечным, информационным, центром.

Механизмы отбора документальной информации

Механизмы отбора документов в результат поиска, зачастую отождествляемые с языком поиска, реализуются алгоритмами, в основе которых лежит какая-либо модификация вычисления *булева выражения*, соотносящего множество терминов запроса и множество терминов документов базы данных. К сожалению, системы редко информируют пользователя о применяемых механизмах и особенностях их использования.

Модификацией булевого поиска является *взвешенный булев поиск*. Запрос может формулироваться на ИПЯ, описанном выше, но выдача документов при этом будет ранжироваться в зависимости от степени близости запроса и документа.

Поиск «по сходству» (документы-аналоги, «Like this»). Наиболее простой моделью этого типа является линейная модель

индексирования и поиска, когда близость документа и запроса рассматривается, например, как угол между векторами запроса и документа в многомерном пространстве терминов индексирования. В соответствии со значением меры близости происходит ранжирование документов при выдаче. В простейшем случае документ считается соответствующим запросу, если он содержит не обязательно все, а только часть терминов запроса в любом сочетании.

Пороговые модели. Перечисленные выше алгоритмы предполагают, что на конечном этапе поиска выборка найденных документов ранжируется. Но совершенно очевидно, что это приводит к ранжированию практически всего массива документов. Ранжировать целиком такие массивы на практике невозможно, поэтому применяются пороговые модели, которые задают пороговые значения для документов, выдаваемых пользователю.

Поиск с коррекцией запроса по релевантности. Такой поиск является уже *интерактивным итеративным процессом*. После проведения первичного поиска с использованием какого-либо из вышеперечисленных методов пользователь отмечает в списке найденных документов истинно релевантные, т.е. соответствующие его информационной потребности, а не просто содержащие термины запроса. Некоторые системы имеют для этого специальное поле (область при документе), где пользователь может отметить документ как релевантный. При следующей итерации система уже сама расширяет запрос пользователя терминами из релевантных документов и снова выполняет поиск. Так продолжается до тех пор, пока пользователь не сочтет, что лучшего результата, чем он уже имеет, добиться не удастся. Здесь имеем два типа обратной связи. Для построения словников на основе лексики документов, определяемых пользователем как истинно релевантные, используется «внешняя» обратная связь. Для построения реформулированного запроса используется уже «внутренняя» обратная связь, позволяющая пользователю (непосредственно) или системе (ранжированием или кластеризацией по статистическим показателям) выделить семантически-значимые термины.

Этот простой и достаточно эффективный механизм, тем не менее, имеет особенность, требующую от пользователя сосредоточения только на одной цели (предмете) поиска. В том случае, если он при просмотре очередной итерации будет отмечать релевант-

ными документы, которые ему интересны в контексте другой темы, это создаст для системы ситуацию “погони за двумя зайцами”.

Приведенные механизмы можно отнести к эвристическим в том смысле, что в результате процесса поиска формируется одна или несколько документальных областей, в той или иной степени соответствующих потребности пользователя. Принципиальным моментом является то, что система таким образом готовит *альтернативы*, позволяя не только увеличить полноту поиска, но и выйти на новую лексику предметной области, неизвестную пользователю.

Постобработка поисковой выдачи

Основное назначение постобработки – снижение размерности пространства документов, которое необходимо «обработать» пользователю для получения уверенности в том, что полученная информация адекватна – полно и точно отражает состояние предметной области в аспекте задачи конкретного пользователя. Именно наличие конкретных особенностей содержания, формы представления или будущего использования информации позволяет предположить, что часть найденных формальными методами документов может заранее (до просмотра) считаться малозначимой, т.е. не обладающей *существенными признаками*. Соответственно, если множество найденных документов будет упорядочено в выдаче по убыванию значения этого признака, то пользователь достаточно просто, но *обоснованно*, может принять решение о завершении просмотра.

Такое упорядочение (ранжирование) документов достаточно просто реализуется, если в качестве *существенного признака* выступает, например, «свежесть» публикации: для этого система отсортирует документы по значению поля «Дата публикации». Существенно сложнее процедура в случае ранжирования по степени смыслового соответствия. Сложность предопределяется двумя факторами. Во-первых, системе достаточно трудно точно и однозначно установить действительный смысл реальной потребности пользователя по обычно очень короткому выражению запроса. Во-вторых, практически невозможно выбрать *адекватную* меру соответствия, поскольку она должна вычисляться по формальным *коли-*

качественным признакам, в то время как смысловое соответствие имеет *качественный* характер.

В технологиях информационного поиска используются две методики упорядочения. Первая – классификация, основана на сведении запроса к общепринятой (по крайней мере, в данной предметной области) системе классификации. Вторая – кластеризация, основана на предположении, что предметная область (и взгляды на нее) обладает свойством структурированности, и выданные документы могут быть разделены в соответствии с аспектами, один или несколько из которых, вероятно, будут соответствовать реальной потребности.

Классификация и кластеризация представляют собой две противоположные технологии: классификация заключается в автоматическом отнесении (определении тематики) документа к одному из классов, определенных на известном множестве признаков, в то время как задачей кластеризации является автоматическое построение классов семантически подобных документов. То есть, в случае классификации на первом этапе задается система признаков, определяющих фиксированное количество классов, а на втором – документы распределяются между этими классами. В случае кластеризации множество документов сразу разбивается на кластеры по степени близости признаков, которыми они обладают сами, т.е. происходит не столько выявление, обладает ли документ признаками, заранее объявленными в классификационной системе, сколько выявление признаков, которые могут быть объявлены (на втором этапе) в качестве классификационных.

Обзор методов автоматической классификации и кластеризации документов приведен в главе 5.

2.4. Поисковый интерфейс

Поисковый интерфейс (интерфейс пользователя) – это программа, обеспечивающая формирование запросов, просмотр найденных документов и управление поисковой сессией, включая, например, выбор ресурсов, установление параметров отбора, использование ранее полученных результатов и т.д.

Наибольшее разнообразие форм и возможностей (и проблем для пользователя) демонстрируют средства, которые пользователь

должен применять при подготовке запроса для его обработки системой. Тем не менее, здесь можно выделить два класса решений. Первые (в том числе и в действиях пользователя) – «вербальные» представлены средствами подготовки выражения поискового условия. Вторую группу составляют средства, позволяющие представить запрос в так называемой «кластерной» форме, например документом (поиск аналогов), множеством документов или терминов.

Средства представления запроса

Средства подготовки запроса в вербальной форме. Как отмечалось ранее, все ИПС, явно или неявно используют поисковые условия, основанные на алгебре логики. Однако пользователю не просто изучить искусственный язык. Это заставляет разработчиков применять сложные системы типа искусственного интеллекта или реализовывать меню-подобные схемы, позволяющие составлять выражение запроса.

По форме (структуре) диалога и, соответственно, по способу задания условия отбора средства такого рода можно разделить на две категории: *рубрикационного типа* и *структурно-логические*. Первые реализуются в виде иерархических, последовательно раскрывающихся списков, через которые обеспечивается доступ к тематически связанным группам документов. Раскрывая очередную рубрику и перемещаясь, таким образом, по тематической иерархии, пользователь уточняет предметную область и увеличивает степень соответствия выдаваемых документов своей информационной потребности. При таком решении недостатки предопределенности, а иногда и неточности соотнесения документов с отдельными рубриками компенсируются логичностью естественнонаучной классификационной схемы, заменяющей пользователю путеводитель.

Структурно-логические средства подготовки запроса – это меню-подобные схемы и процедуры, имитирующие «посредничество» и позволяющие в интерактивном режиме «сконструировать» булевоподобное выражение, удовлетворяющее требованиям ИПЯ системы.

Типичными примерами, реализующими структурно-логический подход, являются следующие:

- *конструктор запроса «по образцу»* реализует традиционный для библиографического поиска форматно-ориентированный интерфейс. Имеет жестко фиксированную модель поискового условия, предполагающую обязательное выполнение частных условий, относимых к полям, выбираемым из predetermined списков;

- *конструктор формирования запроса «по шагам»* характеризуется большей гибкостью. Здесь поисковые термины также выбираются из словаря, но могут связываться любыми отношениями. Причем, построенные таким образом лексические выражения, относимые к отдельным полям, в свою очередь могут связываться операторами, выбираемыми из списка. Такой конструктор позволяет формировать достаточно сложные предложения запроса последовательным наращиванием либо выражения условия (путем добавления очередного термина), либо всего предложения (путем добавления нового условия поиска). Необходимо отметить, что сложные предложения запроса требуют достаточно хорошей предварительной структуризации;

- *конструктор формирования логического выражения запроса* – это непосредственный клавиатурный набор выражения запроса с возможностью обращения в произвольном порядке к словарям, спискам имен полей и т. д.

Кластерные формы представления запроса. Кластерные формы имеют несколько другую технологию – для спецификации потребности используются документы, как образцы для поиска подобных.

Интерфейс поиска документов-аналогов: инициирование поиска осуществляется пользователем в момент просмотра отдельного документа, который им признан *истинно релевантным*. Функция поиска документов-аналогов использует наиболее информативные термины текущего документа, при этом пользователь в диалоге может указать поля, термины которых будут использованы системой для построения ПОЗ, логику и пороги критерия соответствия.

Интерфейс эвристического поиска достаточно прост, так как и значимые термины, и критерий соответствия система формирует автоматически после того, как пользователь из интерфейсной формы «*Протокол запросов*» иницирует для указанного им множества

истинно релевантных документов процедуру эвристического поиска.

Найденные документы упорядочиваются в соответствии со значимостью, определяемой статистически в контексте релевантности документов.

Интерфейс поиска по обратной связи реализуется последовательностью шагов:

- отбор и выделение множества релевантных документов;
- построение ранжированного словника релевантных документов и предоставление словника пользователю;
- выбор терминов из словника для формирования документального пространства;
- разбиение построенного информационного пространства на кластеры и предоставление пользователю возможности для просмотра документов каждого отдельного кластера.

Поэтому, помимо интерфейсов протокола и просмотра документов, этот вид поиска использует интерфейс работы со словниками, который позволяет пользователю выделять релевантные термины в списке, построенном и упорядоченном по статистической значимости системой.

Процедуры развития поиска. Такие процедуры обычно не имеют специализированных интерфейсов, и, как правило, реализуются средствами подготовки запроса и обработки результатов поиска на уровне множества документов или отдельного документа.

Интерфейсные средства обработки результатов и технология поиска

Интерфейсные средства обработки результатов поиска позволяют работать как с отдельными документами, так и с их коллекциями.

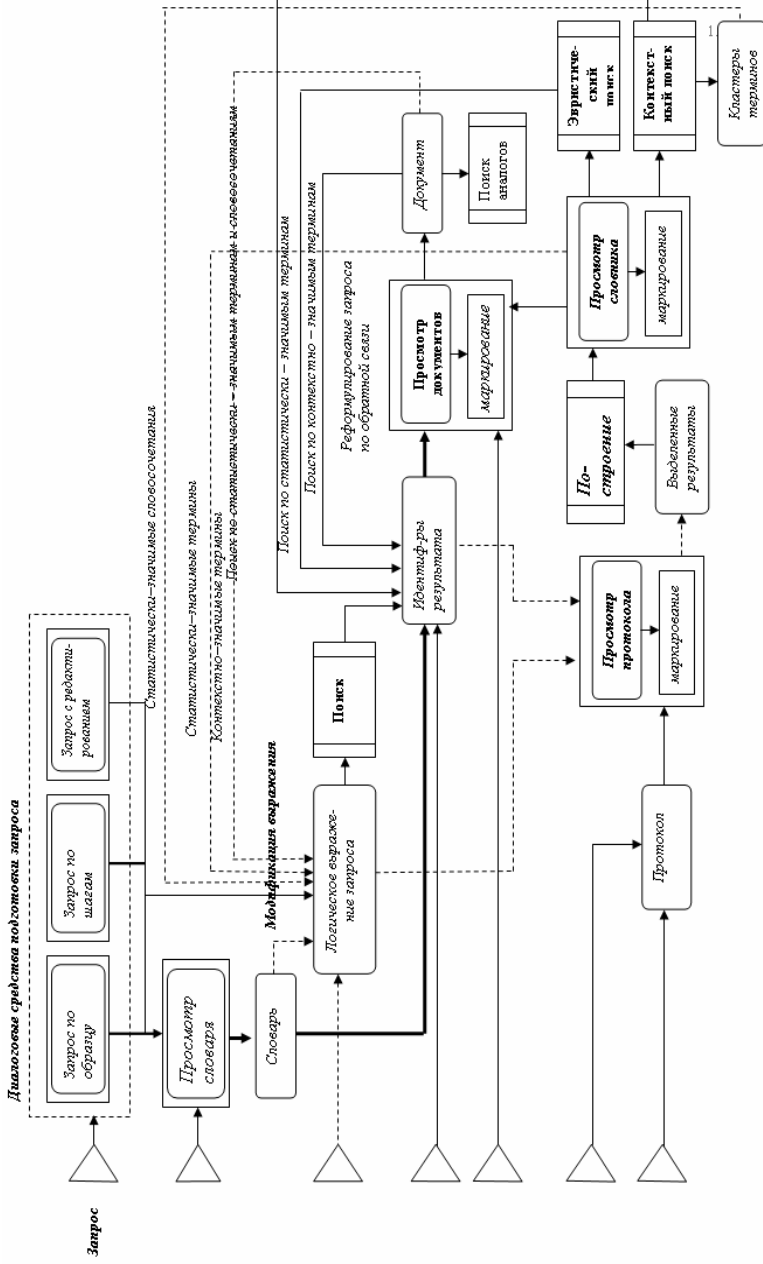


Рис. 2.4. Обобщенная технологическая схема и инструменты поиска.

Интерфейсная форма «Документ» помимо традиционных средств управления просмотром и выводом, может включать средства изменения формы представления материала, а также средства инициирования процедур поиска, использующих лексику документа. Интерфейс формы *«Документ»* позволяет пользователю отмечать релевантность документа теме запроса, что используется для итогового объединения в один файл всех отмеченных документов, а также для поиска по технологии обратной связи по релевантности.

В ряде случаев (обычно для БД вторичной информации) интерфейсная форма позволяет обращаться к текстам соответствующих первоисточников, размещенных в локальной среде, или во внешних ресурсах.

Интерфейсная форма «Протокол запросов» позволяет работать с результатами поисков, которые регистрируются в Протоколе отдельными предложениями в виде логических выражений или комментариев к результатам, полученным способом, отличным от поиска по логическому выражению.

Форма *«Протокол запросов»* используется в первую очередь для просмотра документов – результатов ранее выполненного поиска. Средства формы *«Протокол запросов»* также обеспечивают разделение результата на подмножества по уровню релевантности, сортировку документов по значениям указанных реквизитов, а также сохранение в файлах текстов запросов и результатов всех или выделенных пользователем предложений для последующей обработки. Выражения запроса могут быть переадресованы для поиска в других информационных ресурсах.

В заключение отметим, что интерфейсные средства собственно и являются для пользователя теми инструментами, которые он выбирает и использует для реализации конкретной поисковой траектории (навигации в базе данных).

Обобщенная технологическая схема поиска, приведенная на рис. 2.4, представляет основные группы интерфейсных средств на примере АИПС IRBIS и достаточно хорошо отражает последовательность и характер их использования. Эта схема наглядно демонстрирует, можно сказать, парадоксальную особенность поискового процесса: схема подобна лабиринту, имеющему много явно замет-

ных входов, но выход не указан – его должен определять пользователь индивидуально в зависимости от собственных целей, наличных ресурсов и способностей.

Контрольные вопросы к разделу

1. Охарактеризуйте соотношение понятий «информация», «данные», «знания».
2. Приведите примеры, которые отражают цикличность информационного обмена.
3. Определите свойства информации.
4. Охарактеризуйте свойство рассеяния информации.
5. Дайте определение понятия «информационная система», «информационная технология».
6. Определите понятие «информационная деятельность».
7. Определите назначение и организацию инвертированного списка.
8. Перечислите типы простых запросов.
9. Охарактеризуйте основные этапы процесса информационного поиска.
10. Перечислите основные и технологические объекты, используемые при поиске.
11. Определите назначение «обратной связи» в процессе информационного поиска.
12. Охарактеризуйте основные интерфейсные средства подготовки и модификации поисковых запросов.
13. Охарактеризуйте основные интерфейсные средства развития поисковых запросов.
14. Приведите типологию сценариев формирования выражения поискового запроса на ИПЯ.
15. Дайте характеристику интерфейсным средствам использования тезаурусных связей при модификации поисковых запросов.
16. Охарактеризуйте интерфейсные средства использования терминологических структур при подготовке запросов.
17. Перечислите информационные объекты, используемые для реализации технологии «обратной связи».
18. Охарактеризуйте влияние интерфейсных средств на адаптацию пользователя.

РАЗДЕЛ 2. ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ИС И АВТОМАТИЧЕСКАЯ ОБРАБОТКА ДОКУМЕНТАЛЬНОЙ ИНФОРМАЦИИ

Лингвистическое обеспечение рассматривается как совокупность языковых средств, обеспечивающих гибкость представления и обработки информации. К таким средствам относятся коммуникативные форматы, схемы документов и баз данных, языки отчетов, а также информационно-поисковые языки, тезаурусы, классификации, рубрикаторы, кодификаторы и т.п.

Основное назначение лингвистического обеспечения, рассматриваемое в рамках настоящего пособия, – предоставить поставщику и потребителю информации возможность строить в рамках информационного ресурса общее лексическое (и инструментальное) пространство, позволяющее с общих позиций, более или менее адекватно отображать как содержание ресурса, так и содержание информационной потребности.

ГЛАВА 3. КЛАССИФИКАЦИИ В ЗАДАЧАХ ИНФОРМАЦИОННОГО ПОИСКА

Классификация в том виде, в каком она используется в формальной логике, близка к процессу описания предмета, его спецификации. Под спецификацией в данном случае понимается точное, однозначное и непротиворечивое обозначение предмета без указания его относительного положения в классификационной схеме.

Рассмотрим далее содержательный аспект классификации как семантической системы, предназначенной для смысловой идентификации и поиска документов.

В основе любой *иерархической классификации* лежит представление о том, что вся совокупность накопленных человечеством знаний может быть разделена на взаимоисключающие классы и подклассы, где каждый класс делится лишь по одному основанию (признаку), порождая одну серию подклассов. Классификационное дерево, получающееся в результате такого деления, отличается большой жесткостью.

Фасетные классификации в этом смысле существенно отличаются от иерархических. Если при составлении иерархических классификаций в некотором систематическом порядке дается перечень всех предметных комплексов, которые уже описаны в литературе, то при построении фасетной классификации дается некоторое множество элементарных термов («строительных кирпичей»), из которых можно построить большое число тематических классов.

Развитие науки, как известно, характеризуется наличием двух противоположных тенденций: во-первых, дифференциацией, в результате которой каждая наука разделяется на все новые и новые ветви; во-вторых, взаимопроникновением не только смежных, но иногда очень далеких одна от другой наук, в результате чего появляются новые, ранее не существовавшие науки. Отсюда следует, что любая претендующая на научность и перспективность классификация должна учитывать особенности развития науки и иметь такую схему, которая бы позволяла адекватно отражать в классификации новые ветви уже сложившихся наук, новые науки и возникающие в результате дифференциации последних ветви новых наук.

Классификации обычно строятся на базе классификации наук с ее делением на отдельные отрасли, хотя имеется множество объектов, особенно в области естествознания, медицины и техники, изучение которых не является задачей какой-либо одной науки (например, одна и та же машина или аппарат может применяться в различных отраслях техники).

Рассматривая классификацию как систематическое распределение объектов множества по классам, возникающее в результате последовательного многоступенчатого деления, можно выделить следующие два вида классификаций:

- *естественные классификации* — классификации, в основание которых кладутся существенные для выделяемых классов признаки.

- *вспомогательные классификации* — классификации, в основание которых кладутся несущественные для выделяемых классов признаки.

В практических задачах информационного обслуживания наиболее широко и устойчиво используются следующие системы классификации:

- *библиотечно-библиографические*, специально предназначенные для систематизации книг и других документов;
- *патентные*, служащие для индексирования объектов промышленной собственности, заявленных или признанных изобретениями;
- *классификации наук*, призванные систематизировать научную информацию (Государственный рубрикатор НТИ и локальные (или отраслевые) рубрикаторы, построенные на его основе, используемые при формировании всех видов информационных изданий).

В то же время необходимо отметить, что использование в классификаторах и современной научно-технической терминологии в значительной степени ограничено, поскольку в качестве содержания классификационных индексов-рубрик выбраны достаточно общие родовые понятия, являющиеся, кроме того, по сути своей номинальными определениями индексируемых объектов, где в качестве терминов используются искусственные конструкции. Такая атерминологичность объясняется, прежде всего, тем, что выбор лексических единиц искусственного типа обеспечивает единообразную трактовку содержания рубрик, независимо от особенностей употребления специальной научно-технической терминологии для различных научных школ и языков.

Основными достоинствами классификаций являются следующие:

- весь поток научной информации индексируется в соответствии с классификациями;
- классификации отражают практически все направления в науке и технике, систематизируя объекты по основным существенным признакам;
- использование единой классификации не только облегчает поиск, но и обеспечивает платформу для единого понимания предмета рассмотрения.

Вышеперечисленное позволяет:

- повысить эффективность информационного обслуживания в традиционном и автоматизированном режимах;
- унифицировать структуру локальных рубрикаторов и обеспечить их совместимость;
- обеспечить совместимость тематического описания ИР;

- вести статистический анализ информационных массивов и потоков и т. д.

3.1. Библиотечно-библиографические классификации

Различают три основные функции библиотечно-библиографической классификации:

- *библиотечная* – организация фондов;
- *библиографическая* – организация знаний и информации о фондах;
- *когнитивная* – организация тематических описаний для поиска в фондах.

Библиотечная функция предназначена для ответа на один вопрос – где находится нужная книга.

Библиографическая функция обеспечивает систематический подход ко всему объему фондов. Необходимо, однако, отметить, что систематичность здесь определяется, в том числе, уровнем образования пользователей, их информационными потребностями, культурной подготовкой.

Когнитивная функция направлена на установление связей, ассоциаций, приближений и соответствий между элементами тематической области.

Из рассмотренных ранее свойств классификаций следует, что одна классификация не может полностью обеспечить все три перечисленные функции. В идеальной ситуации должно быть две классификации: одна для поиска самих документов, другая для выполнения библиографической и когнитивной функций.

В отличие от естественных классификаций, основной целью которых является установление и обозначение существенных связей между предметами, задача библиотечно-библиографической классификации совершенно иная: она должна обеспечить такую систематизацию множества документов, при которой документы (или их описания) по одному и тому же предмету будут *физически* собираться в одном, и только в одном, месте. Таким образом, для библиотечно-библиографической классификации исключительно большое значение имеет ясное обозначение относительного расположения классов в схеме.

В отличие от формально-логической классификации, для которой прежде всего важно образование классов предметов и установление *родо-видовых отношений* (иерархии) между этими классами, в библиотечно-библиографической классификации требуется установление *порядка подклассов* в общем ряду и определение их пространственного расположения относительно друг друга. Логическое отношение включения лишь определяет, что подклассы необходимо расположить таким образом, чтобы было показано их подчинение классу. Но формальная логика не дает универсального метода установления, в каком порядке следует располагать элементы деления. Например, если мы разделим класс «информационные системы» на подклассы «техническое обеспечение», «программное обеспечение», «лингвистическое обеспечение» и т. д., то порядок этих подклассов в ряду определяется содержательно, а не по формально-логическим правилам. Таким образом, библиотечно-библиографическая классификация отличается от формально-логической классификации своим *принципом упорядоченного размещения классов в пространстве*.

Существуют два подхода к построению библиотечно-библиографической классификации. *Реалистический подход* предполагает, что структура классификации должна отражать структуру взятой за основу теоретической классификации. *Прагматический подход* предполагает, что классификации изобретаются, а не открываются, и что годится любая организация тематики предметной области, соответствующая поставленной задаче [38].

В настоящее время сложились две группы библиотечно-библиографических классификаций:

- перечислительные классификации;
- аналитико-синтетические классификации.

К перечислительным библиотечно-библиографическим классификациям относят *иерархические* и *алфавитно-предметные классификации*; к аналитико-синтетическим – *фасетные* классификации.

Наиболее известными иерархическими классификациями на сегодняшний день являются Десятичная классификация Дьюи, Библиографическая классификация Бласса, Классификация Библиотеки Конгресса США, библиотечно-библиографическая клас-

сификация ББК, назначение, состав и использование которой рассмотрим ниже.

Иерархические библиотечно-библиографические классификации. ББК

Отечественная иерархическая библиотечно-библиографическая классификация (ББК) предназначена для организации библиотечных фондов, систематических каталогов и картотек. Ее основная задача – раскрыть содержание произведений печати, представить их в виде стройной научно обоснованной системы знаний и этим максимально облегчить читателю использование библиотечных фондов.

Методологической основой ББК является классификация наук и явлений действительности. Она выражается в определенной субординации наук, в соответствии с классификацией видов материи и форм ее движения, переходом от простого к сложному, от низшего к высшему.

В основу членения классификационных таблиц положены многообразные критерии: объект изучения, метод исследования, цель познания, структура объекта, его свойства, процессы, отношения, территория, исторический период и т.д. На более глубоких ступенях деления иногда применяется расположение понятий по алфавиту наименований (алфавит стран в пределах части света и др.). В то же время в ББК предусмотрена возможность характеризовать одно и то же понятие на основе разных признаков, что дает возможность более глубоко раскрыть содержание произведений печати.

Структура ББК. ББК имеет синтетическую структуру, позволяющую многоаспектно отражать содержание произведений печати. Единство между аналогичными по статусу, но различными по содержанию элементами универсальной классификации позволяют установить следующие аспекты обобщения ББК:

- основные таблицы,
- система типовых делений.

Таблица 3.1

Основные разделы ББК

Индекс для массовых библиотек		Разделы (таблицы) ББК	Индекс для научных библиотек
1		Общенаучное и междисциплинарное знание	А
2		Естественные науки	
	20	Естественные науки в целом	Б
	22	Физико-математические науки	В
	24	Химические науки	Г
	26	Науки о Земле (геодезические, геофизические, геологические и географические науки)	Д
	28	Биологические науки	Е
3		Техника. Технические науки	Ж/О
4		Сельское и лесное хозяйство. Сельскохозяйственные и лесохозяйственные науки	П
5		Здравоохранение. Медицинские науки	Р
6/8		Общественные и гуманитарные науки	
	60	Общественные науки в целом	С
	63	История. Исторические науки	Т
	65	Экономика. Экономические науки	У
	66	Политика. Политическая наука	Ф
	67	Право. Юридические науки	Х
	68	Военное дело. Военная наука	Ц
	70/79	Культура. Наука. Просвещение	Ч
	80/84	Филологические науки. Художественная литература	Ш
	85	Искусство	Щ
	86	Религия. Мистика. Свободомыслие	Э
	87	Философия	Ю
	88	Психология	Ю
9		Литература универсального содержания	Я

Основные таблицы. Перечень разделов первого уровня деления (состав основных таблиц ББК) представлен в табл. 3.1. Следующие (второй, третий, четвертый и т.д.) уровни классификации в основных таблицах образуются путем деления первого уровня (основного ряда) на подчиненные группы наук, отрасли деятельности, отдельные науки, проблемы, темы и т.п.. Первый уровень деления таблиц ББК для научных библиотек обозначен первыми буквами русского алфавита, а для массовых библиотек – арабскими цифрами, т.е. в таблицах ББК для массовых библиотек отдельные естественные и общественные науки, которые в таблицах для научных библиотек представлены на первом уровне деления, перенесены на второй уровень деления и подчинены обобщающим классам.

Науки, возникшие в результате взаимопроникновения наук и (или) лежащие на пересечении двух или трех наук, в ББК условно относят к одной из них (агробиология – к сельскому хозяйству, биогеохимия – к биологии), а от другой (или других) дается отсылка. Комплексная наука разделяется на части, каждая из которых относится к соответствующей «материнской» науке (микробиология к биологии, медицинская микробиология к медицине и т.п.), а от общей науки даются отсылки к ее отраслям и дисциплинам.

Система типовых делений. Типовые деления помогают выделить и единообразно разместить однотипную литературу в систематическом каталоге. Система типовых делений представлена таблицами основных типовых делений, используемых во всех отделах классификации, и таблицами специальных типовых делений, обслуживающих отдельные отрасли наук.

К таблицам основных типовых делений относятся:

- таблицы общих типовых делений;
- таблицы территориальных типовых делений;
- таблицы типовых делений социальных систем.

Общие типовые деления (ОТД) используются для дополнительного тематического (история науки, научные и культурные связи и т.п.) и формального (библиографические пособия, справочные издания, сборники и т.п.) деления. Например, во всех отделах могут встретиться книги по истории вопроса, справочники, учебники и т.п. Для каждого из этих видов изданий предусмотрены особые рубрики, которые включены в таблицу общих типовых де-

лений. Индексы ОТД присоединяются к основному индексу отрасли или темы непосредственно без всякого знака.

Территориальные типовые деления (ТТД) предназначены для единообразного подразделения материала по региональному (территориальному) признаку и отражают в основном современную политическую карту мира. Объектами классификации при этом являются территории и акватории земного шара. Основные деления ТТД выделены по физико-географическому признаку: весь мир (земной шар, Земля со всем существующим на ней), части света (суша) и Мировой океан. Исключения представляют деления "(2) Россия. СССР" и "(3) Зарубежные страны в целом". Россия (а раньше СССР) расположена в двух частях света (Европе и Азии) и не может быть представлена в целом ни в одной из них. Зарубежные страны в целом – это все страны, кроме тех, которые занимают обширные территории на всех континентах. Все страны, независимо от их величины и общественного устройства, расположены в пределах частей света в порядке алфавита их наименований. Дальнейшая детализация осуществляется по политико-административному и административно-территориальному признаку (для этого используются специальные типовые деления). В пределах стран все административные территории располагаются также по алфавиту их наименований.

Индексы ТТД присоединяются к индексу тех подразделений основных таблиц или общих типовых делений, в которых требуется выделить материал об отдельных странах, местностях, территориях.

Типовые деления социальных систем (ТДС) предназначены для единообразного подразделения и обозначения материала по признаку социальной системы.

Специальные типовые деления (СТД) используются для унификации построения классов по определенным признакам, типичным только для данной отрасли знания.

Способ кодирования, назначение и примеры применения типовых делений приведены в табл. 3.2.

Обозначения типовых таблиц отличаются от индексов основной таблицы и не имеют самостоятельного значения, а присоединяются к индексу основной таблицы непосредственно или же с помощью условных обозначений.

Использование системы типовых делений подчиняется следующим правилам:

- типовые деления не применяются, если для данного вопроса имеется специальный индекс в основной таблице или в таблицах специальных типовых делений: например: учебники по русскому языку для средней школы кодируются индексом 81.2Рус-922, а не 81.2Руся72; литература по истории образования и педагогической мысли кодируются индексом 74.03, а не 74г.

- ОТД могут комбинироваться с индексами таблицы территориальных типовых делений, например, "5г(2) История медицины и здравоохранения в России и СССР";

- СТД обычно помещаются непосредственно в индексах тех разделов, для детализации которых они предназначены. В то же время в ряде разделов основных таблиц иногда используются единичные специальные типовые деления, не выделенные в таблицы СТД. В этом случае под соответствующим индексом помещается методическое указание. Например, под индексом "82 Фольклор. Фольклористика" имеется указание: "В подразделениях 82 для выделения произведений устного народного творчества (текстов) используется типовое деление -6"

Каждое из типовых делений выражает определенный аспект рассмотрения явлений. В тех случаях, когда нужно отразить разные их аспекты, таблицы комбинируются между собой, образуя сложные индексы для выражения соответствующих понятий.

Связи в ББК. Библиотечно-библиографическая классификация неизбежно расчленяет явления, науки, проблемы. Это требует установления связей между разорванными понятиями.

В тех случаях, когда связи не могут быть выражены в структуре таблиц или когда предусмотренные структурой решения спорны, условны, односторонни, применяется система отсылок, ссылок и методических указаний. Из множества связей составители пытались отобрать и отразить с помощью ссылочно-отсылочного аппарата лишь те, которые имеют для библиотечной классификации практическое значение, т.е. позволяют получить более полную информацию по данной проблеме (табл. 3.3).

Кодирование. В ББК принята логически последовательная система кодирования, непосредственно связанная со структурой классификации. Система кодирования позволяет детализировать общее понятие путем присоединения к имеющимся индексам новых знаков справа, и, наоборот, при необходимости сокращать детализацию путем отбрасывания от индекса его конечных знаков.

В алфавит системы кодирования входят:

- арабские цифры;
- строчные и прописные буквы русского алфавита.
- точка («.»), двоеточие («:»), дефис («-»), круглые скобки («()»), косая черта («/»).

Таблица 3.2

Типовые деления ББК

Таблица	Кодирование	Назначение	Примеры
<i>Основные типовые деления</i>			
ОТД	Строчные буквы русского алфавита (первый уровень деления) Арабские цифры (второй, третий и последующие уровни деления)	Тематические и формальные деления	22.141я72 – Учебники для средней школы по алгебре. <i>Здесь:</i> 22.141 – Элементарная алгебра, я72 – Учебные издания для средней школы.
ТТД	Цифры и буквы русского алфавита, прописные и строчные, заключенные в круглые скобки	Деление по региональному признаку	<i>Страны мира:</i> (4) – Европа (4Авс) – Австрия (5Япо) – Япония 60.7(4) – Население Европы, 30г(2) – История техники в России и СССР <i>Крупные физико-географические регионы:</i> (53) – Западная (Передняя) Азия (531) – Кавказ

Таблица	Кодирование	Назначение	Примеры
ТДС	Арабские цифры с опознавательным знаком апостроф (')	Деление по признаку социальной системы	66.2(0)'6 – Политика и современное политическое положение развитых стран. <i>Здесь:</i> 66.2 – Политика и современное политическое положение; (0) – деление ТТД, обозначающее все страны; '6 – деление ТДС, обозначающее систему развитых стран.
<i>Специальные типовые деления</i>			
СТД	Последовательность арабских цифр и знак «-» (дефис) перед цифровой частью	Используются только в пределах одного или нескольких классов одной отрасли знания	(4Лат-5) – Районы Латвии, (5Гру-6) – Автономные республики Грузии, (4Укр-6Кры-2) – Города Крыма.

Сочетание всех знаков алфавита приведено в известную систему, в которой каждый знак имеет свое место, назначение и смысловую нагрузку.

Арабские цифры обозначают:

- основные классы ББК (2 – Естественные науки, 3 – Техника. Технические науки);
- второй, третьей и последующие уровни деления основных таблиц (24 – Химические науки);
- территориальные типовые деления неадминистративного характера ((4) – Европа, (9) – Мировой океан. Океаны и моря)
- государства, союзные республики – в сочетании с буквами русского алфавита ((5Япо) – Япония);

- второй и последующие уровни деления общих типовых делений (я2 –Справочные издания); специальные типовые деления (-2 – Грамматика).

Прописные буквы русского алфавита служат для обозначения отдельных понятий, предметов, территорий и т.п. при алфавитном расположении материала, используются в сочетании с цифрами и имеют мнемонический характер ("(4Авс) Австрия").

Строчные буквы русского алфавита применяются для обозначения первого уровня общих типовых делений (г – История науки, я2 – Справочные издания).

Точка применяется в качестве разделительного знака после первых двух цифр индекса, заменяющих прописную букву основных делений варианта ББК для научных библиотек, а также далее после каждой группы цифр из трех знаков, считая слева направо (22.151.1 – Неевклидовы геометрии, 74.900.6 – Организация жизни детей в семье).

Дефис является отличительным признаком специальных типовых делений («-4 Агротехника» – рубрика таблицы типовых делений).

Круглые скобки – отличительный признак индексов территориальных типовых делений ((5) – Азия).

Двоеточие применяется при комбинировании индексов рубрик одного основного класса с индексом рубрик других основных классов (91.9:85 – Библиографические пособия по искусству).

Косая черта применяется для обозначения сдвоенного индекса («6/8 Общественные и гуманитарные науки»). Такие рубрики не являются рабочими.

Апостроф является отличительным признаком индексов типовых делений социальных систем ('6 – Развитие страны).

Алфавитно-предметный указатель ББК. Алфавитно-предметный указатель является вспомогательным инструментом для индексатора. Он собирает в одном месте рассредоточенные в таблицах характеристики объекта. Это позволяет отразить все вопросы в двух разрезах: в предметном (указатель) и в систематическом (таблицы).

Таблица 3.3

Семантические связи в ББК

Связь	Форма представления в ББК	Назначение	Примеры
Отсылка	"см."	Показывает, что литература по данной теме (предмету, вопросу) собирается не в текущем классе, а в другом	36.996 Специальная кулинария Лечебное питание см. 53.51
Ссылка	"См. также"	Связывает темы (понятия), дополняющие друг друга	85.38 Художественное радиовещание и телевидение • См. также: 76.03 Радиовещание. Телевидение 85.335.42 Балетный театр • См. также: 85.335.41 Оперный театр
Методические указания		Представляют собой либо перечень тем, включенных в рубрику, либо указания о характере и типе литературы, собираемой в данной рубрике, указания о способе дальнейшего подразделения материала, о размежевании с другими рубриками.	28.06 Общая морфология. Общая гистология • Форма и размеры живых организмов, симметрия, асимметрия, изомерия, окраска и др. Гистогенез, гистохимия, гистофизиология и др. 22.68 Космогония Литература общего характера. Происхождение и эволюция планет, планетных систем и Солнечной системы см. под индексами космических тел и их систем. Например: происхождение Луны см. 22.654.1.

Понятия, содержащиеся в таблицах, в словнике указателя представлены в прямых и инвертированных формулировках. Наряду с понятиями, которые определяются одним словом или его синонимом, имеются такие, смысл которых передается несколькими словами, а иногда и целой фразой. В тех случаях, когда одно и то же понятие отражено в указателе синонимами, прямыми и инвертированными формулировками, для основных рубрик выбирается термин, имеющий наибольшее распространение, а от остальных к нему даются отсылки.

В указателе также даны рубрики методического характера. Эти рубрики разъясняют, где искать данный предмет или вопрос. Например:

История 63.3

библиографические пособия по И. 91.9:63 и см. в подразделениях 63

методика преподавания в общеобразовательной школе 74.266.3

отдельных наук или предметов см. в соответствующих подразделениях классификации

Аналитико-синтетические (фасетные) классификации. Универсальная десятичная классификация

К наиболее известным классификациям, которые основываются на использовании общих категорий и фасетов, следует отнести «Классификацию с двоеточием», «Универсальную десятичную классификацию» (УДК).

Первая фасетная «Классификация с двоеточием» (Colon Classification – CC) была опубликована ее разработчиком Ш. Р. Ранганатаном (Индия) в 1933 г. С тех пор она пересматривалась много раз. Уже в первом издании структура и система индексации классификации существенно отличались от распространенных тогда в мировой практике перечислительных классификационных систем. В изданиях 1939, 1950, 1952, 1957, 1960 гг. Ранганатан развил положенные в основу классификации принципы фасетного анализа и фасетного синтеза. По замыслу автора фасетные (или аналитико-синтетические) классификации должны обеспечить при систематизации одного и того же документа разными систематизаторами единое классификационное решение, стандартно оформ-

ленное в виде одного классификационного индекса, построенного в соответствии с фасетной формулой.

Ранганатан предложил пять основных фасетов, обозначаемых латинскими буквами: Р (Personality, Индивидуальность), М (Matter, Материя), Е (Energy, Энергия), S (Space, Место), Т (Time, Время). Фасеты (категории) в конкретных отраслевых классах отражают специфику той или иной отрасли. Каждое понятие представлено в тексте таблицы классификации только один раз и в случае необходимости отражается в синтезируемом классификационном индексе.

Совокупность первого уровня делений (главные классы классификации), по мнению Ранганатана, не имеет принципиального значения: время от времени могут возникать новые главные классы. Важно лишь строго выполнять условие: каждый предмет может относиться к одному и только к одному главному классу. Подразделения главных классов образуются, как правило, в соответствии с определенными для него фасетами (категориями).

Универсальная десятичная классификация (УДК) создавалась в основном еще до того, как был разработан фасетный принцип. Поэтому в УДК этот принцип получил лишь частичное воплощение, и она, по существу, является классификацией *полуфасетного* типа.

В качестве инструмента для систематизации материала была принята десятичная классификация Дьюи¹. В основу своей классификации М. Дьюи – сторонник и пропагандист метрической системы – поставил принцип деления всей совокупности знаний на десять основных классов, а отрасли знания и отдельные дисциплины, в свою очередь, также подразделялись и обозначались десятичным методом. Ведущим принципом явилась форма (десятичная индексация), ей подчинено содержание (классификация). Дьюи отрицал связь своей классификации с классификацией наук. Основной принцип классификации Дьюи – удобство и легкость: арабские

¹ М. Дьюи (1851–1931) – один из виднейших американских библиотечников. Помимо десятичной классификации в 1876 г. разработал стандартную каталожную карточку, вошедшую с 1901 г. в обиход сначала в Библиотеке Конгресса США, а затем и во всех библиотеках мира.

цифры понятны во всем мире, что обеспечивает независимость классификационной системы от языка.

УДК и классификация Дьюи едины в своей основе, совпадают для первых трех ступеней классификации (трехзначных индексов, т.е. для 1000 рубрик), но расходятся в дальнейших делениях, их детализации и методике составления индексов. УДК сохранила в своей основе иерархическую структуру. Вместе с тем, в УДК был внесен ряд дополнений и приемов, присущих фасетной классификации.

УДК, построенная по систематическому принципу, понимается как *система цифрового кодирования* не содержания, а *поисковых образов (признаков) документа*, которые формально отделяются от его содержания.

С 1963 г. классификация всех публикаций в области технических и естественных наук по УДК введена в качестве обязательной в научно-технических издательствах, редакциях научно-технических журналов, органах научно-технической информации.

Структура УДК. Все классы УДК сгруппированы в шесть фасетов (табл. 3.4), каждый из которых подразделяется по иерархическому принципу на несколько уровней. Для идентификации отдельного класса на каждом уровне используются десятичные цифры, тем самым, в индексах УДК каждая последующая цифра не меняет значения предыдущих, а лишь уточняет их, обозначая частное понятие. Индекс УДК представляет собой последовательность десятичных цифр, возможно разделенных на группы знаками «точка».

Фасет «Общий предмет» (основная таблица классификации) имеет десять основных подразделений, которые называются *главными классами*.

Основная таблица включает понятия и соответствующие им индексы, охватывающие классы от 0 до 9 с их дальнейшими делениями по содержанию. Главными классами фасета «Общий предмет» являются:

- 0 Общий отдел
- 1 Философия. Психология. Логика
- 3 Общественные науки
- 5 Математика. Естественные науки

- 6 Прикладные науки. Медицина. Техника
- 7 Искусство. Декоративно-прикладное искусство. Фотография. Музыка. Игры. Спорт
- 8 Языкознание. Филология. Художественная литература. Литературоведение
- 9 География. Биографии. История

Таблица 3.4

Фасеты УДК с их стандартным символическим обозначением

№ пп	Наименование фасета	Тип фасета	Нотация фасета
1	Общий предмет	Фасет содержания	индекс
2	Место	Фасет содержания	(индекс)
3	Народность	Фасет содержания	(=индекс)
4	Время	Фасет содержания	"индекс"
5	Язык документа	Фасет формы	= индекс
6	Форма документа	Фасет формы	(0индекс)

Остальные фасеты УДК получили название *вспомогательных таблиц*, предназначенных для классификации по дополнительным признакам. Вспомогательные индексы (определители) этих таблиц бывают двух видов: *общие*, которые могут соединяться с любыми индексами основной таблицы УДК, и *специальные*, присоединяемые только к понятиям текущего раздела.

Общие определители УДК. Общие определители отражают категории и признаки, применяемые по всей таблице (время, место, язык, формы документов и т.д.) и служат для стандартного обозначения этих категорий и признаков. Нотации общих определителей приведены в табл. 3.4.

Общие определители по их функциям можно разделить на две группы:

- определители для обозначения формальных признаков (определители языков и определители формы документа);

- определители для обозначения содержательных признаков (определители места, народности, времени).

Общие определители языков обозначают язык документа, содержание которого определяется основным индексом УДК:

= 111 Английский язык

= 112.2 Немецкий язык

= 161.1 Русский язык

Хотя теоретически язык какого-либо документа или информационного сообщения может быть указан всегда, практически это полезно делать только тогда, когда имеется потребность различить документы на разных языках, например, чтобы обеспечить возможность поиска по языку или упорядочивания выдачи по языку.

В индексе УДК определители языка обычно располагаются на последнем месте, например:

663.4(493)(075)=112.5 Пивоваренная промышленность

Бельгии. Учебник на фламандском языке

Определители языка могут располагаться в середине и даже в начале составного индекса, если есть потребность располагать документы по порядку языков, а не по их содержанию.

Для обозначения *переводов* применяется определитель =03 (переводные работы, переводы), например:

55=03.111.161.1 Документы по геологии, переведенные с английского языка на русский

Многоязычные документы могут быть обозначены кодом =00 или определителями отдельных языков, расположенных в порядке возрастания кода, например:

53(035)=00 Многоязычные справочники по физике

53(035)=111=112.2=133.1 Справочники по физике на английском, французском и немецком языках

Общие определители формы документов служат для систематизации документов по форме их публикации или представления. При этом форма документа должна четко отличаться от его содержания и касается только:

- его внешней (физической) стороны;
- формы его представления (рукопись, печатное издание и т.д.);

- способа представления информации об отдельном предмете, содержании документа (т. е. вида документа).

В сложных индексах общий определитель формы документа обычно занимает предпоследнее место, перед общим определителем языка.

Определители формы обычно используются в сочетании с индексами основной таблицы УДК, например:

54(035) Справочники по химии

54(051) Периодические издания, журналы по химии

555(075.8) Учебник по геологии для вузов

91(084.3) Географические карты

Общие определители места служат для обозначения географического положения (или другого пространственного аспекта) объекта, классифицируемого по УДК, путем присоединения определителя к индексу основной таблицы. Они применяются преимущественно для дальнейшего подразделения документов по географическому признаку.

Определители места относятся к самостоятельным общим определителям, т.е. их можно использовать как самостоятельные индексы УДК с тем же значением.

При изменении политического статуса территорий (например, при образовании новых государств) Общие определители места обычно не менялись. Поэтому получилось, что в ряде случаев одним определителем места объединены территории, принадлежащие разным странам, или, наоборот, части одной страны отнесены к разным определителям места.

Если какой-либо определитель места не совпадает по значению с совокупностью территорий, обозначенных всеми десятичными подразделениями этого определителя, то в квадратных скобках дается соответствующее пояснение.

Примеры использования определителей места:

621.311(410) Электростанции Великобритании

631.4(44) Почвы Франции

69(571.53/.62) Строительство в зоне БАМ

Общие определители народов (этнические определители) обозначают национальный или этнический аспект объекта, представленного основным индексом УДК, например:

398(=81/=82) Фольклор Северной Америки (североамериканских индейцев)

Эти определители образуются, как правило, из общих определителей языка путем помещения их в круглые скобки. Они служат для обозначения народов, наций, народностей и этнических групп в отличие от языков, на которых эти народы говорят. Этнические определители могут также отражать культурно-языковые группы населения, например:

(=111) Англоязычное население (в отличие от англичан – населения Англии)

(1.2) Народы (население, жители) определенных физико-географических районов

(1.23) Горцы

Политическая национальность (гражданство национальных государств) отражается главным образом определителями (=1.4)/(=1.9), которые образованы из общих определителей места, но для некоторых аспектов могут оказаться более подходящими сами определители места как таковые.

Общие определители времени служат для отражения понятия, связанного с датой, периодом или другими аспектами времени.

Определители времени обычно добавляются после основного индекса.

Примеры использования определителей места:

53"196" Физика 1960-х гг.

54"196" Химия 1960-х гг.

Специальные определители УДК. *Специальные определители* чаще всего *используются (присутствуют) в основной таблице* и приводятся непосредственно в разделе, в котором применяются. В УДК предусмотрены три вида нотаций специальных определителей:

- «дефис» (знак «-», за которым следует индекс, используется только в горном деле);
- «точка-ноль» (последовательность знаков «.01», за которой следует индекс);
- «апостроф» (знак «'», за которым следует индекс).

Специальные определители «дефис» и «точка-ноль» выполняют двоякую роль:

во-первых, специальный определитель в соединении с общим индексом раздела обозначает общее понятие рассматриваемой отрасли знания, если данного понятия нет среди основных индексов раздела, например:

622-1/-9 Характеристики и детали машин,

где

622 Горное дело. Горные предприятия.

Добыча нерудных ископаемых

622.01 Теоретические, геофизические и технические вопросы горного дела

во-вторых, тот же определитель как подвижная часть индекса в соединении с любым детальным индексом данного раздела выполняет аналитическую функцию, уточняя конкретное понятие, выраженное этим индексом, например:

550.34.01 Теория, моделирование, исследования и разработка,

где **550.34 Сейсмология. Землетрясения в целом**

Определитель «апостроф» выполняет синтетическую функцию и служит для комплексного обозначения отдельных составляющих элементов, свойств и других характеристик, например:

553.411'435 Золотоносные медно-колчеданные месторождения,

где

553 Экономическая геология.

Месторождения полезных ископаемых

553.411 Золотые месторождения. Месторождения золотоносных конгломератов и комплексные месторождения

Правила построения индексов и кодирования. В алфавит системы кодирования индексов входят:

- арабские цифры;
- знак «точка» («.») как разделитель групп арабских цифр для удобства читаемости индекса;
- знаки нотаций фасетов (табл. 3.4);
- знаки «дефис» («-») и «апостроф» («'») как нотации специальных определителей;
- знаки построения сложных индексов (табл. 3.5).

Таблица 3.5

Связи в УДК

Знак	Назначение	Примеры применения
+ присоединение	Применяется, если содержание документа не может быть выражено одним индексом	622+669 Горное дело и металлургия (7+8) Северная и Южная Америка
/ распространение	Применяется, если содержание документа можно выразить с помощью нескольких индексов, следующих друг за другом в десятичном ряду	592/599 Систематическая зоология (вместо 582+593+...+599) 669.2/.8 Металлургия цветных металлов (вместо 669.2+669.3+...+669.8)
: двосточие	Применяется для выражения общих отношений, отношений соподчинения и обратимых отношений	341.63(44:450) Международный арбитраж между Францией и Италией 341.63(450:44) Международный арбитраж между Италией и Францией
[...] квадратные скобки	Применяются в качестве алгебраического обозначения группы из двух или более индексов, связанных между собой знаками «(+)» или «(:)», с целью отражения какого-либо понятия, если эти индексы подразделяются дальше при помощи двосточия или определителей (общих или специальных)	31:[622+669](485) Статистика горного дела и металлургии в Швеции 004.3:[621.771.016.3:669.14] Применение ЭВМ при холодной прокатке стали

Знак	Назначение	Примеры применения
:: двойное двоеточие	Применяется для закрепления определенной последовательности двух или более элементов в составном индексе	575::576.3 Цитогенетика 77.044::355 Военные фото- съемки

В УДК принята определенная последовательность присоединения индексов общих и специальных определителей к основному индексу. Таким образом, в УДК применяется определенная фасетная формула, хотя эта формула имеет меньшую жесткость, чем в «чистых» фасетных классификациях.

Кроме общей фасетной формулы в УДК применяется также субформула для построения сложных предметных индексов. Если предметный индекс УДК состоит из нескольких простых индексов, соединяемых друг с другом знаками присоединения, распространения, объединения или отношения, то эти простые индексы рекомендуются всегда располагать в следующей последовательности:

1. Предмет в целом
2. Виды предмета
3. Части предмета
4. Материалы
5. Свойства
6. Процессы
7. Действия
8. Агенты

Основными недостатками правил построения индексов в УДК являются:

1) низкая специфичность правил построения (небольшое число символов используется для выражения значительно большего числа различных связей между характеристиками);

2) правила построения не позволяют выражать в линейном виде многомерные связи между характеристиками, что приводит к появлению многозначных выражений (индексов);

3) неоднозначность правил построения (не нарушая этих правил, два систематизатора могут по-разному выразить одно и то же отношение между характеристиками).

В Приложении 2 приведен фрагмент таблицы одного из основных делений фасета «Общий предмет» УДК «004 Информационные технологии. Вычислительная техника. Теория, технология и применения вычислительных машин и систем».

3.2. Международная патентная классификация

Наиболее известными, используемыми для работы с объектами патентования изобретений, являются Международная патентная классификация (МПК) и Национальная классификация изобретений США (НКИ).

Международная патентная классификация¹ (МПК) обеспечивает достаточно полное индексирование предмета изобретения с помощью ограниченного числа рубрик за счет ориентации последних на важные с точки зрения патентоведения аспекты, такие как характеристика вещества или устройства, его функции, применение, получение. Классификация представляет собой специальную, имеющую технико-прикладной характер, линейную систему иерархического типа, предназначенную для ручного индексирования. Согласно основным принципам применения Международной патентной классификации, указывающим, что изобретение, подлежащее классификации, не может рассматриваться как чистая идея в отрыве от ее технического воплощения в устройстве, способе или веществе, выделяют следующие предметы (объекты) изобретения:

- 1) вещество или материал;
- 2) устройство, прибор, конструкция;
- 3) процесс, способ, метод;

Изобретение рассматривается как некоторый объект, взятый относительно среды, т.е. формальное описание изобретения включает описание объекта и некоторый предикатор (аспект), отра-

¹ Ранее МПК называлась Международная классификация изобретений (МКИ)

жающий его отличительную особенность. В качестве предикаторов наиболее часто используют:

- для вещества или материала: материал как таковой, его применение или получение;
- для устройства прибора или конструкции: применение, построение, функциональное назначение, принцип действия;
- для процессов: назначение, использование.

Правила классифицирования. Основным назначением МПК является облегчение поиска аналогичных технических решений. В связи с этим МПК разработана и должна использоваться таким образом, чтобы одинаковые технические решения классифицировались одинаково, и, следовательно, могли быть найдены в одном и том же месте системы классификации. *Методика классифицирования* опирается на выявление технической сущности изобретения. Техническую сущность изобретения может представлять собой способ, продукт или устройство (или способ их использования или применения), причем эти термины используются в МПК в самом широком смысле. Например:

способ – полимеризация, ферментация, разделение, формование, транспортирование, обработка текстиля, передача и преобразование энергии, строительство, методы работы на машинах, способы работы машин, обработка и передача информации;

продукт – химическое соединение, состав, ткань;

устройство – установка для проведения химических или физических процессов, инструмент, орудие, машина, устройство для выполнения технологических операций.

Структура МПК. МПК охватывает все области знаний, которые могут подлежать защите охранными документами и имеет иерархическую 4-уровневую структуру.

1. *Раздел.* Обозначается латинской заглавной буквой от А до Н и снабжен *заголовком*, укрупнено отражающим его содержание.

МПК включает следующие восемь разделов:

А – удовлетворение жизненных потребностей человека;

В – различные технологические процессы, транспортирование;

С – химия, металлургия;

D – текстиль, бумага;
E – строительство, горное дело;
F – механика, освещение, отопление, двигатели и насосы, оружие, боеприпасы, взрывные работы;
G – физика;
H – электричество.

2. *Класс*. Каждый раздел делится на классы, обозначаемые двузначным числом. Содержание класса отражает *заголовок класса*. Некоторые классы снабжаются кратким перечнем относящейся к ним тематики – *указателем класса*.

3. *Подкласс*. Каждый класс содержит один или более подклассов, обозначаемых заглавной буквой латинского алфавита. Содержание подкласса определяет *заголовок подкласса*. Некоторые подклассы снабжаются кратким перечнем относящейся к ним тематики – *указателем содержания подкласса*.

4. *Группа, подгруппа*. Каждый подкласс разбит на подразделения, которые в дальнейшем именуются «дробными рубриками». Среди дробных рубрик различают основные группы и подгруппы.

Дробная рубрика обозначается двумя числами, разделенными наклонной чертой. Первое (максимум трехзначное, обычно нечетное) число индексирует *основную группу*, второе (минимум 2 цифры, обычно четное) – *подгруппу*. Для основной группы код подгруппы имеет значение 00. Каждую третью или четвертую цифру после наклонной черты следует понимать как дальнейшее десятичное деление предыдущей цифры. Отсюда следует, что подгруппа с индексом 5/417 должна стоять после подгруппы 5/41, но перед подгруппой 4/42.

Текст основной группы определяет область, которая считается целесообразной для проведения поиска. Текст и индексы основных групп выделены жирным шрифтом.

Текст подгруппы понимается всегда в пределах объема ее основной группы и определяет тематическую область, в которой считается целесообразным проведение поиска. Перед текстом подгруппы ставится одна или более точек, которые определяют степень ее подчиненности, т.е. указывают на то, что подгруппа является рубрикой, подчиненной ближайшей вышестоящей рубрике, напечатанной с меньшим сдвигом, т.е. имеющей на одну точку меньше.

Полный классификационный индекс МПК состоит из комбинации символов (табл. 3.6), используемых для обозначения раздела, класса, подкласса и основной группы или подгруппы.

Таблица 3.6

Уровни деления МПК

Уровень	Наименование	Индекс	Примеры
1	Раздел	Латинская заглавная буква от А до Н	Н – Электричество
2	Класс	Двузначное число	Н 04 Техника электрической связи Н 05 Специальные области техники, не отнесенные к другим классам.
3	Подкласс	Заглавная буква латинского алфавита	Н 04 J Многоканальные системы связи Н 05 K Печатные схемы, корпуса или конструктивные элементы электрических приборов; изготовление блоков электронной аппаратуры
4	Группа	Одно-, двух- или трехзначное число, наклонная черта и два нуля	Н05 K 1/00 Печатные схемы A 63 F 13/00 Аспекты игр с использованием двух- и более размерного дисплея, работающего на электронных средствах, например, телевизионного экрана, и показывающего изображения, относящиеся к игре A 63 F 9/00 Различные игры

Уровень	Наименование	Индекс	Примеры
5	Подгруппа	Одно-, двух- или трехзначное число основной группы, наклонная черта и, по крайней мере, две цифры, кроме 00. Каждая следующая цифра обозначает дальнейшее десятичное деление предыдущей цифры	A 63 F 13/10 Управление ходом игры, например началом, развитием, концом A 63 F 9/28 Игры с цепной реакцией опрокидывания частей

В том случае, когда изобретение приближенно соответствует группе, но не подпадает ни под одну из входящих в нее подгрупп, производится приближенная классификация: чтобы отметить неточный характер индекса МПК, в конец его ставится знак "X", например, F 01 B 10/00X.

Дополнительные средства выражения семантики МПК. К дополнительным средствам относятся:

- подзаголовки;
- составные заголовки или тексты дробных рубрик;
- отсылки;
- примечания;
- аннулированные группы или подгруппы.

В табл. 3.7 приведены характеристики и примеры применения дополнительных семантических средств.

Форма этих дополнительных средства выражения семантики не стандартизована, поэтому при интерпретации любого заголовка или выражения часто бывает необходимо рассмотреть связанные с ними подклассы или группы.

В тех случаях, когда содержание одной рубрики находится в отношении подчинения или подчиненности к другой, а также, если одна является уточнением другой, используется аппарат отсылок

(это указание границ компетентности рубрики, указание о преимуществе, пояснение о близости обозначаемых). Формы выражения связи приведены в табл. 3.8.

Таблица 3.7

Семантические средства МПК

Наименование	Использование	Пример
<i>Подзаголовок</i>	Для обозначения нескольких следующих друг за другом групп, имеющих родственную тематику, и относится ко всем группам, расположенным после него до следующего подзаголовка	
<i>Составной заголовок или текст дробных рубрик</i>	Там, где желательно в одном месте объединить различные виды технических объектов, которые неудобно охватить одной фразой, и состоят из двух или более частей, разделяемых точкой с запятой	А 63 Н 27/127 Игрушечные летательные аппараты с вертикальным взлетом и посадкой; Летящие волчки
<i>Отсылка</i>	Указывает на то, что тематика охватывается другой рубрикой или рубриками МПК и включается в заголовок класса, подкласса, текст группы, подгруппы, подзаголовки или примечание в виде заключенного в скобки индекса другой рубрики (или рубрик) МПК	Е 02 F 3/00 Экскаваторы и другие землеройные машины (для специальных целей 5/00; машины и оборудования для горного дела Е 21 С; для строительства туннелей Е 01 G)
<i>Примечание</i>	Разъясняет значение употребляемых терминов и выражений в некоторых частях системы классификации; определяет объем рубрик; указывает, каким образом следует классифицировать ту или иную тематику	В 01 J 31/00 При классификации по этой рубрике присутствие воды не учитывается

Наименование	Использование	Пример
<i>Аннулированная группа или подгруппа</i>	При индексах аннулированных групп или подгрупп (т.е. групп или подгрупп, существовавших в структуре предыдущей редакции классификации и не существующих в текущей редакции) указывает индексы, относящиеся к рубрикам, содержащим тематику аннулированных групп или подгрупп	"(Рубрика аннулирована. Содержание перенесено...)"

Заголовки разделов, подразделов и классов лишь приблизительно отражают объем последних и, как правило, не дают точного представления о тематике, охватываемой этими разделами и классами. Вообще, заголовки разделов или подразделов очень свободно определяют содержание тематики, отнесенной к данному разделу или подразделу, а заголовок класса содержит лишь общее указание на объекты, отнесенные к подклассам этого класса. В противоположность заголовкам разделов и классов, заголовки подклассов с учетом всех относящихся к ним отсылок и примечаний с максимальной точностью определяют сущность тематики, охватываемой подклассом.

Таблица 3.8

Связи в МПК

Связь	Назначение	Пример
<i>«Относящийся к», «предусмотренный, охватываемый (какой-либо рубрикой)», «перекрываемый (какой-либо рубрикой)»</i>	Используются в заголовках, текстах групп и подгрупп при наличии особенностей, позволяющих классифицирование тематики в указанных после выражения рубриках	В 41 В 21/00 Общие конструктивные элементы фотонаборных машин, отнесенных к группам 17/00 и 19/00

Связь	Назначение	Пример
<i>"Не предусмотренный, не отраженный в других группах"</i>	Используются в заголовке класса, подкласса или в тексте группы, подгруппы, если тематика, о которой идет речь, не относится ни к одной из других групп или подгрупп данного или какого-либо иного индекса.	В 23 К 28/00 Способы сварки или резки, не отраженные в предыдущих группах, например электролитическая сварка
<i>"Другие...", "прочие...", "различные..." и "...не отнесенные к другим группам или подгруппам",</i>	Охватывают тематику, обладающую особенностями, не позволяющими отнести ее ни к одной другой группе или подгруппе данного подкласса	А 63 Н 21/00 Прочие игрушечные железные дороги А 63 F 7/36 Конструктивные детали, не отнесенные к группам 7/24-7/34, например, рамы, игральные диски, направляющие желобки
<i>"Отнесенные более чем к одной из предшествующих групп"</i>	Охватывают тематику, включающую большое количество альтернатив (или особенностей, не охватываемых в целом ни одной из предшествующих групп), каждая из которых, взятая отдельно, охватывается одной из предшествующих групп	В 23 Q 41/00 Комбинирование или агрегирование станков, отнесенных к классам В 21 – В 32
<i>"То есть"</i>	Две фразы, соединенные этим выражением, должны рассматриваться как эквивалентные, т.е. одна фраза является толкованием понятия, названного в другой фразе	В 22 D 11/00 Непрерывное литье металлов, т.е. отливка изделий неограниченной длины

Связь	Назначение	Пример
<i>"Например"</i>	Поясняет смысл предшествующей фразы	A 41 D 17/00 Гетры, краги, гамаши; щитки для защиты ног (например при сварке)
<i>"А и В"</i>	Означает наличие <i>А</i> и <i>В</i> (логическое И)	B 28 C 1/00 Способы и устройства для получения или обработки глины
<i>"А или В"</i>	Означает наличие <i>А</i> или <i>В</i> , или <i>А</i> и <i>В</i> вместе (логическое ИЛИ)	B 29 H 19/00 Регенерация и/или переработка использованной резины
<i>"или А, или В"</i>	Означает наличие либо <i>А</i> , либо <i>В</i> , но не <i>А</i> и <i>В</i> вместе.	B 23 K 15/00 Сварка или резка электронным лучем
<i>"Вообще"</i>	Указывает на то, что объект рассматривается с точки зрения, не зависящей от применения в какой-либо конкретной области техники	F 02 B Поршневые двигатели внутреннего сгорания. Двигатели внутреннего сгорания вообще
<i>"Специально предназначенный для"</i>	Означает, что рассматриваемый объект особым образом сконструирован для определенного использования или определенной цели	A 45 D 33/00 Коробки, флаконы и т.п. принадлежности, специально предназначенные для туалетной или косметической пудры
<i>"И т.д.", "и т.п."</i>	Применяются для того, чтобы подчеркнуть, что рассматриваемая рубрика не ограничена только той тематикой, которая упомянута в ее формулировке	A 01 D 46/00 Устройство для сбора фруктов, овощей, хмеля и т.д. A 01 J 19/00 Ручное устройство для формирования брусков масла и т.п.

3.3. Отраслевые классификационные системы

Государственный рубрикатор научно-технической информации (ГРНТИ)

ГРНТИ (прежнее наименование – Рубрикатор ГАСНТИ) представляет собой универсальную трехуровневую иерархическую классификацию областей знания, принятую для систематизации всего потока научно-технической информации.

Рубрикатор имеет многоцелевое назначение, отвечает потребностям всех информационных органов в силу универсальности охвата тематики, обеспечивает многофункциональное использование.

Рубрикам (в качестве справочной информации) поставлены в соответствие индексы УДК и коды Номенклатуры специальностей научных работников ВАК. Это обеспечивает взаимосвязь между классификационными системами, а также возможность поиска документов в информационных массивах и базах данных, систематизированных по УДК и номенклатуре специальностей ВАК. На основе Рубрикатора построена система локальных (отраслевых, тематических, проблемных) рубрикаторов в органах научно-технической информации.¹

Локальный рубрикатор строится как выборка произвольного множества рубрик Государственного рубрикатора НТИ с возможным дальнейшим развитием на глубину ниже третьего уровня. Глубина классификации локального рубрикатора не ограничивается и определяется информационной потребностью органа НТИ. При этом разные рубрики могут быть развиты на разную глубину.

¹ Рубрикатор является частью общесистемных средств лингвистического обеспечения, в состав которых также входят: комплекс базисных тезаурусов, номенклатура грамматических средств ИПЯ, правила представления данных в коммуникативных форматах, методики индексирования для обмена информационными материалами в режиме сети.

Состав и структура Рубрикатора. Рубрикатор имеет три уровня иерархии, при этом весь универсум знаний условно разделен на четыре подкласса:

- «Общественные науки» (значение кода первого уровня от 00 до 26);
- «Естественные и точные науки» (значение кода первого уровня от 27 до 43);
- «Технические и прикладные науки. Отрасли экономики» (значение кода первого уровня от 44 до 81);
- «Межотраслевые и комплексные проблемы» (значение кода первого уровня от 82 до 90).

Каждая рубрика состоит из кода (нотации) и наименования (описание класса), а также может иметь при себе ссылки и примечания.

На каждом уровне Рубрикатора возможно деление на 100 подклассов. Коды рубрик состоят из цепочки пар арабских цифр, разделенных точкой. В конце кода точка не ставится.

Наименование рубрики представляет собой текст, отражающий ее смысловое содержание, например:

37.23.31 Моделирование климата. Прогноз климата.

Рубрикатор является иерархической классификацией и построен так, что классы одного уровня, как правило, не пересекаются, а понятия одного уровня находятся в состоянии подчинения к понятию более высокого уровня, например:

73 Транспорт

73.29 Железнодорожный транспорт

73.31 Автомобильный транспорт

73.34 Водный транспорт

73.37 Воздушный транспорт

73.39 Трубопроводный транспорт

Признаком, по которому проведено деление класса «**73** Транспорт», является признак отнесения к видам технических средств перевозки. Перечисленные виды транспорта не пересекаются, а в совокупности составляют класс «Транспорт» — обобщенное понятие более высокого уровня иерархии.

В Рубрикаторе использовано сочетание иерархии с фасетным принципом, который проявляется в наличии совокупности рубрик,

повторяющейся в разных классах в виде группы «Общие вопросы», а также в применении типовых классификационных делений в разных разделах Рубрикатора. Например, фасет «Общие вопросы» имеет одинаковую структуру в разных классах:

59 Приборостроение	73 Транспорт
59.01 <i>Общие вопросы</i>	73.01 <i>Общие вопросы</i>
59.01.17 <i>Международное сотрудничество</i>	73.01.17 <i>Международное сотрудничество</i>
59.01.29 <i>Информационная деятельность</i>	73.01.29 <i>Информационная деятельность</i>
59.01.79 <i>Кадры</i>	73.01.79 <i>Кадры</i>

Для исключения дублирования разделы Рубрикатора включают тематическое «ядро» предметной области, оснащенное ссылками на связанные с ним смежные вопросы, помещенные в других тематически близких разделах Рубрикатора.

Уровень рубрики, соответствующей определенной области знания, отражает не ее значимость, а только степень обобщения при логической группировке понятий. В пределах одного уровня располагаются примерно равнозначные по объему понятия. Иерархическое строение Рубрикатора, реализованное при упорядочении понятий, отражено в нотации: количество двухразрядных цифровых элементов кода рубрики соответствует иерархическому уровню класса, например:

70 Водное хозяйство	(1 уровень)
70.21 Оросительные системы	(2 уровень)
70.21.31 Виды орошения	(3 уровень)

Рубрикатор обеспечивает отнесение каждого документа к минимально необходимому числу рубрик, поскольку между рубриками, имеющими логико-семантические связи, существуют разграничения аспектов, отраженных в наименованиях рубрик, примечаниях и ссылках.

Средства выражения семантики. Наряду с иерархической классификационной структурой в Рубрикаторе с помощью *аппарата ссылок и примечаний* отражаются полииерархические связи, т.е. подчинение одного понятия двум или более классам, размещенным в разных местах иерархии. При этом могут указываться аспекты, уточняющие признаки деления понятий, например:

60.31 Репрография

См. также 13.20.31 Техническое оснащение библиотек

Аппарат ссылок и примечаний при рубрике способствует уточнению и разграничению содержания рубрик, а также адекватному отнесению документов к релевантным рубрикам при индексировании.

Виды ссылок, используемые при составлении рубрикатора, приведены в табл. 3.9.

Табл. 3.9

Средства выражения семантики в МПК

Связь	Назначение	Примеры
<i>Смотри</i> ("см.")	Содержит наименование понятия или темы, не находящей отражения в данной рубрике, и одновременно указывает ее место в рубрикаторе	19.21 Массовая коммуникация. Социология массовой коммуникации см. 04.51.54 Социология информации и коммуникации
<i>Отсылка от</i> ("Отс. от")	Является обратной к ссылке "см." и ставится при рубрике, на которую имеется ссылка "см."	66.15.17 Валка леса. Выборочные рубки и рубки ухода см. 68.47.29 Лесопользование 68.47.29 Лесопользование Отс. от 66.15.17 Валка леса
<i>Смотри также</i> ("См. также")	Отражает частичное пересечение содержания между рубриками, симметрична и присутствует в обеих связанных рубриках	38.36.17 Кристаллография минералов См. также 31.15.17 Кристаллохимия и кристаллография 31.15.17 Кристаллохимия и кристаллография См. также 38.35.17 Кристаллография минералов

Связь	Назначение	Примеры
Эквивалентно ("Экв.")	Используется для указания случаев тождественного наименования двух рубрик в разных разделах рубрикатора и также является симметричной	15.81.43 Военная психология Экв. 78.21.15 78.21.15 Военная психология Экв. 15.81.43
"Примечание"	Конкретизирует тематическое содержание рубрики, уточняет порядок расположения материала на более глубоком уровне иерархии, заменяет группу однотипных ссылок либо представляет собой комбинацию перечисленных случаев	34.33 Зоология Примечание. В рубрике отражаются вопросы систематики, фаунистики, морфологии, физиологии и экологии (по видам животных)

Связь с другими классификациями. Каждой рубрике в качестве справочной информации приписаны индексы УДК и Номенклатуры специальностей ВАК. Соблюдение структурной и терминологической эквивалентности при этом являлось не обязательным. Сопоставление классов классификационных систем осуществлялось на понятийном уровне, с учетом аспекта рассмотрения понятия.

При сопоставлении с УДК в первую очередь используется простой индекс, а при его отсутствии — комбинированный (сложный или составной):

10.07.53 Теория государства УДК 321.01

50.05.09 Языки программирования УДК 004.43; 811.93

61.39.51 Люминесцентные органические красители.

Оптические отбеливатели

УДК 667.28:661.143; 667.28:535.683

Для организации связи с Номенклатурой специальностей ВАК рубрикам приписывается один, либо несколько кодов специальностей по Номенклатуре специальностей ВАК, разделенных

точкой с запятой. Если при этом специальность ВАК заведомо не исчерпывает тематику рубрики, перед кодом специальности ставится знак плюс:

27.15 Теория чисел ВАК 01.01.06

28.23.13 Инженерия знаний. Представление знаний
ВАК 05.13.01; 05.13.17; 05.25.05

29.17.27 Жидкие металлы и полупроводники
ВАК +01.04.07; +01.04.10

В Приложении 3 приведен фрагмент классификации наук ГРНТИ в области физики.

Рубрикатор INIS/ETDE

Рубрикатор содержит перечень предметных рубрик и описание их тематического охвата и является важным инструментом, который используется для предметной классификации научно-технических публикаций при их обработке для ввода в INIS, а также при автоматизированном поиске информации. Вместе с другими публикациями, входящими в Справочную серию INIS, рубрикатор определяет правила и стандарты, а также дает практические рекомендации по формированию и использованию БД INIS.

Рубрикатор, входящий в Справочную серию INIS, имеет две цели:

- определить тематический охват Международной системы ядерной информации (INIS)
- определить схему предметной классификации INIS.

За время существования INIS используемый в этой системе рубрикатор, построенный по многоуровневой иерархической схеме, постоянно совершенствовался и дополнялся. Последняя седьмая версия Рубрикатора, опубликованная в декабре 1991 года, включала 337 предметных рубрик и имела 5-уровневую схему иерархических связей между ними. Эта версия соответствовала схеме предметной классификации по физике, принятой Международным советом по научной и технической информации (the International Council for Scientific and Technical Information –ICSTI). Практически в это же время, в связи с тем, что ряд ведущих стран-участниц INIS стали членами ETDE (the Energy Technology Data Exchange),

встал вопрос о необходимости создании единой схемы классификации для этих двух систем. Основная сложность заключалась в том, что в схеме классификации ETDE практически отсутствовали иерархические связи между рубриками. После многочисленных дискуссий участники 27-ого Консультативного совещания представителей по связи INIS (Вена, Австрия, 25-27 мая 1999 г.) приняли решение об упрощении схемы предметной классификации INIS и создании объединенного рубрикатора INIS/ ETDE. На сессии 5-го Объединенного технического комитета INIS/ETDE (Ноксвилл, США, 28-29 октября 1999 г.) был одобрен вариант рубрикатора, предложенный специально созданной для этой цели рабочей группой. Принципиально новая редакция рубрикатора INIS/ETDE была введена в действие в INIS с января 2000 г (INIS Circular Letter No. 161, 1999-12-01) и используется, начиная с INIS Atomindex 31/01.

Состав и структура Рубрикатора. Упрощенная схема предметной классификации объединенного рубрикатора INIS/ETDE включает 45 широких рубрик, из которых только 42 соответствуют тематическому охвату INIS. Рубрики из общей схемы классификации **S32** (ЭНЕРГОСБЕРЕЖЕНИЕ, ПОТРЕБЛЕНИЕ И ИСПОЛЬЗОВАНИЕ ЭНЕРГИИ), **S33** (УСОВЕРШЕНСТВОВАННЫЕ ДВИГАТЕЛЬНЫЕ СИСТЕМЫ), и **S47** (ПРОЧИЕ КОНТРОЛЬНО-ИЗМЕРИТЕЛЬНЫЕ ПРИБОРЫ) не входят в тематический охват INIS.

Иерархические связи между рубриками полностью отсутствуют. Каждая предметная рубрика имеет буквенно-цифровой код, состоящий из буквы S и двух цифр. В определении рубрики дан подробный перечень входящих в ее тематический охват научных и/или технологических направлений, а также пояснения по возможным особенностям и ограничениям, связанным с ее применением, например:

ETDE	ИНИС
S02 НЕФТЬ Рубрика охватывает все тематические направления, связанные с нефтью, такие как запасы, геология и разведка месторождений; бурение и добыча; переработка; основные и побочные продукты переработки; свойства и состав; сжигание; транспортировка, погрузочно-разгрузочные операции и хранение; обращение с отходами; проблемы защиты окружающей среды; охрана здоровья людей и безопасность; законодательство и нормативные документы; экономические, производственные и коммерческие аспекты.	S02 НЕФТЬ <i>Проблемы защиты окружающей среды</i> в связи с различными операциями, реализуемыми в процессе бурения, разработки месторождений, переработки, транспортировки и использования нефти и нефтепродуктов для производства энергии, а также разливами нефти. Рубрика включает исследования условий жизни животных и роста растений; экологические аспекты находящихся в зоне влияния промышленных объектов освоенных человеком природных ресурсов; загрязнение земли, воздуха, поверхностных водоемов и грунтовых вод; восстановление растительного покрова территорий и общие эффекты. <i>Экономические аспекты</i>, включая проблемы спроса и предложения, потребления, продажи и ограничений на продажу, ценообразования; тенденции рынка; прогнозирование; финансирование исследований и разработок; людские ресурсы; налогообложение и налоговые льготы; влияние энергетической политики и цен на энергоносители на индивидуальное хозяйство, экономику регионов и стран; влияние погодных факторов на спрос и предложение; экономику электростанций, использующих нефть в качестве топлива; экономические аспекты аварийных ситуаций.

Полный перечень предметных рубрик INIS/ETDE приведен в табл. 3.10.

Таблица 3.10

Перечень предметных рубрик INIS/ETDE в порядке цифровых кодов

Код	Наименование рубрики
S01	УГОЛЬ, ЛИГНИТ И ТОРФ
S02	НЕФТЬ
S03	ПРИРОДНЫЙ ГАЗ
S04	ГОРЮЧИЕ СЛАНЦЫ И НЕФТЕНОСНЫЕ ПЕСКИ
S07	ИЗОТОПЫ И ИСТОЧНИКИ ИЗЛУЧЕНИЙ
S08	ВОДОРОД
S09	ТОПЛИВО НА ОСНОВЕ БИОМАССЫ
S10	СИНТЕТИЧЕСКОЕ ТОПЛИВО
S11	ЯДЕРНЫЙ ТОПЛИВНЫЙ ЦИКЛ И ТОПЛИВНЫЕ МАТЕРИАЛЫ
S12	ОБРАЩЕНИЕ С РАДИОАКТИВНЫМИ И НЕРАДИОАКТИВНЫМИ ОТХОДАМИ ЯДЕРНЫХ УСТАНОВОК
S13	ГИДРОЭНЕРГЕТИКА
S14	ЭНЕРГИЯ СОЛНЦА
S15	ГЕОТЕРМАЛЬНАЯ ЭНЕРГИЯ
S16	ЭНЕРГИЯ ПРИЛИВОВ И ВОЛН
S17	ЭНЕРГИЯ ВЕТРА
S20	ЭЛЕКТРОСТАНЦИИ НА ИСКОПАЕМОМ ТОПЛИВЕ
S21	ЯДЕРНЫЕ РЕАКТОРЫ КОНКРЕТНЫХ ТИПОВ И СВЯЗАННЫЕ С НИМИ УСТАНОВКИ
S22	ОБЩИЕ ПРОБЛЕМЫ ЯДЕРНЫХ РЕАКТОРОВ
S24	ПЕРЕДАЧА И РАСПРЕДЕЛЕНИЕ ЭНЕРГИИ
S25	НАКОПЛЕНИЕ ЭНЕРГИИ
S29	ЭНЕРГЕТИЧЕСКОЕ ПЛАНИРОВАНИЕ, ПОЛИТИКА И ЭКОНОМИКА
S30	ПРЯМОЕ ПРЕОБРАЗОВАНИЕ ЭНЕРГИИ
S32	<i>ЭНЕРГОСБЕРЕЖЕНИЕ, ПОТРЕБЛЕНИЕ И ИСПОЛЬЗОВАНИЕ ЭНЕРГИИ (только для ETDE)</i>
S33	<i>УСОВЕРШЕНСТВОВАННЫЕ ДВИГАТЕЛЬНЫЕ СИСТЕМЫ (только для ETDE)</i>
S36	МАТЕРИАЛОВЕДЕНИЕ

Код	Наименование рубрики
S37	НЕОРГАНИЧЕСКАЯ, ОРГАНИЧЕСКАЯ, ФИЗИЧЕСКАЯ И АНАЛИТИЧЕСКАЯ ХИМИЯ
S38	РАДИАЦИОННАЯ ХИМИЯ, РАДИОХИМИЯ И ЯДЕРНАЯ ХИМИЯ
S42	ТЕХНИКА
S43	УСКОРИТЕЛИ ЗАРЯЖЕННЫХ ЧАСТИЦ
S46	КОНТРОЛЬНО-ИЗМЕРИТЕЛЬНЫЕ ПРИБОРЫ, ИСПОЛЪЗУЕМЫЕ В ЯДЕРНОЙ НАУКЕ И ТЕХНОЛОГИИ
S47	<i>ПРОЧИЕ КОНТРОЛЬНО-ИЗМЕРИТЕЛЬНЫЕ ПРИБОРЫ (только для ETDE)</i>
S54	ЭКОЛОГИЯ И ЗАЩИТА ОКРУЖАЮЩЕЙ СРЕДЫ
S58	НАУКИ О ЗЕМЛЕ
S60	ПРИКЛАДНЫЕ ЕСТЕСТВЕННЫЕ НАУКИ
S61	РАДИАЦИОННАЯ ЗАЩИТА И ДОЗИМЕТРИЯ
S62	РАДИОЛОГИЯ И ЯДЕРНАЯ МЕДИЦИНА
S63	ВЛИЯНИЕ РАДИАЦИОННЫХ, ТЕПЛОВЫХ И ДРУГИХ ЗАГРЯЗНЕНИЙ ОКРУЖАЮЩЕЙ СРЕДЫ НА ЖИВЫЕ ОРГАНИЗМЫ И БИОЛОГИЧЕСКИЕ МАТЕРИАЛЫ
S70	ФИЗИКА ПЛАЗМЫ И ТЕХНОЛОГИЯ ТЕРМОЯДЕРНЫХ УСТАНОВОК
S71	КЛАССИЧЕСКАЯ И КВАНТОВАЯ МЕХАНИКА, ОБЩАЯ ФИЗИКА
S72	ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И ПОЛЕЙ
S73	ЯДЕРНАЯ И РАДИАЦИОННАЯ ФИЗИКА
S74	АТОМНАЯ И МОЛЕКУЛЯРНАЯ ФИЗИКА
S75	ФИЗИКА КОНДЕНСИРОВАННЫХ СРЕД, СВЕРХПРОВОДИМОСТЬ И СВЕРХТЕКУЧЕСТЬ
S98	ЯДЕРНОЕ РАЗОРУЖЕНИЕ, ГАРАНТИИ И ФИЗИЧЕСКАЯ ЗАЩИТА
S99	ОБЩИЕ ВОПРОСЫ И ПРОЧИЕ ДОКУМЕНТЫ

Средства выражения семантики. Большинство рубрик снабжено набором перекрестных ссылок с указанием рубрик, которые при определенных условиях должны использоваться вместо данной конкретной рубрики или могут быть использованы для рубрицирования публикаций наряду с ней (например, «см.» – смотри, «исп.» – используй).

Таблица 3.11

Средства выражения семантики в Рубрикаторе INIS/ETDE

Связь	Назначение	Примеры
<i>Смотри</i> ("см.")	Содержит наименования рубрик, которые могут быть использованы для рубрицирования публикаций вместе с заглавной рубрикой	S07 ИЗОТОПЫ И ИСТОЧНИКИ ИЗЛУЧЕНИЙ Для: ... неизотопных источников излучений см. S71 или рубрику, соответствующую области применения
<i>Смотри также</i> ("См. также")	Отражает частичное пересечение содержания между рубриками	S22 ОБЩИЕ ПРОБЛЕМЫ ЯДЕРНЫХ РЕАКТОРОВ Для: ... топливных элементов и сборок см. также S21
<i>Используй</i> ("исп.")	Содержит наименования рубрик, которые при определенных условиях должны использоваться вместо данной конкретной рубрики	S07 ИЗОТОПЫ И ИСТОЧНИКИ ИЗЛУЧЕНИЙ Для: химических свойств тяжелой воды исп. S37 физических свойств тяжелой воды исп. S36 фундаментальных исследований по физической химии обмена и разделения стабильных изотопов исп. S37
"Примечание"	Конкретизирует тематическое содержание рубрики и ее использование при индексировании	S70 ФИЗИКА ПЛАЗМЫ И ТЕХНОЛОГИЯ ТЕРМО-ЯДЕРНЫХ УСТАНОВОК <i>Физика плазмы.</i> <i>Примечание: рубрика включает только исследования плазмы, имеющие отношение к ядерному синтезу.</i> S21 ЯДЕРНЫЕ РЕАКТОРЫ КОНКРЕТНЫХ ТИПОВ И СВЯЗАННЫЕ С НИМИ УСТАНОВКИ <i>Примечание: эта рубрика должна присваиваться релевантным документам, в которых указаны конкретные типы реакторов.</i>

Перекрестные ссылки на другие предметные рубрики призваны помочь правильно выбрать рубрику для конкретной публикации. Они, по существу, указывают научные и/или технологические проблемы, которые не входят в рассматриваемую рубрику.

Виды ссылок, используемые при составлении рубрикатора, приведены в табл. 3.11.

При индексировании публикаций определением предметных рубрик занимаются предметные специалисты, в задачу которых входит выделение основного предмета исследований для каждой конкретной публикации и отбор только тех публикаций, в которых содержится информация по проблемам, входящим в тематический охват. Основной задачей является определение первой рубрики (с точки зрения важности для ядерной науки и технологии). Первая рубрика должна отвечать основной теме исследований, результаты которых рассматриваются в данной конкретной публикации. Если в публикации приведена информация, относящаяся к другим тематическим направлениям, правила INIS позволяют приписать ей одну или несколько второстепенных предметных рубрик. Теоретически количество второстепенных рубрик не ограничено, однако обычно не превышает одной или двух.

ГЛАВА 4. ТЕРМИНОЛОГИЧЕСКИЕ СТРУКТУРЫ

В большинстве информационных систем помимо ИПЯ на этапах индексирования и поиска документов применяются различные средства, имеющие лингвистическую природу, например, тематические рубрикаторы, тезаурусы, словари как информативных, так и неинформативных лексических единиц, словари синонимов, словари словосочетаний и т.п.

Организационная типология терминологических структур, приведенная на рис. 4.1, тесно связана с типологией по семантическому признаку. С точки зрения семантики словоупотребления терминологические структуры могут быть разделены на семантически упорядоченные и семантически неупорядоченные. При этом семантически неупорядоченные терминологические структуры всегда имеют линейную организацию, а семантически упорядоченные – иерархическую или сетевую организацию.

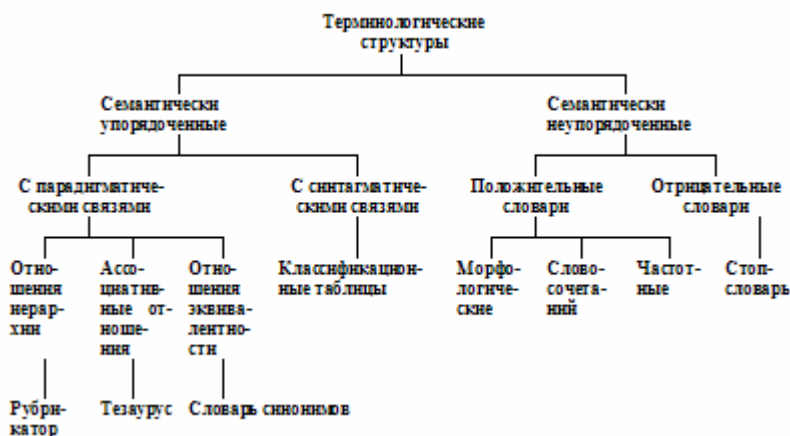


Рис. 4.1. Типология терминологических структур по семантическому признаку

Семантически упорядоченные терминологические структуры отражают оба типа связей, которые могут существовать между отдельными терминами – парадигматические и синтагматические. Парадигматические связи характеризуют различные виды отноше-

ний – отношения иерархии, ассоциативные отношения и отношения эквивалентности. Синтагматические связи показывают логические отношения между понятиями (рис. 4.1).

4.1. Линейные терминологические структуры

К линейным терминологическим структурам относятся линейные словари различного назначения, обычно упорядоченные по лексикографическому принципу. С точки зрения своего участия в процессах индексирования документов и запросов такие словари делятся на *положительные* и *отрицательные*.

Положительные словари объединяют лексику, которую можно использовать в процессе индексирования. Отрицательные словари содержат лексику, запрещенную для использования при индексировании.

Морфологические словари. Морфологические словари содержат основы слов и аффиксы (суффиксы, окончания). Такие словари могут быть использованы, с одной стороны, для нормализации поисковых образов документов, а с другой, для нормализации лексики поисковых запросов.

Грамматический строй естественных языков нередко расходится со структурой логического мышления, и поэтому при поиске информации желательно иметь возможность исключить (полностью или частично) влияние аффиксов словоизменительного формата. Для этого можно предусмотреть наращивание документов всеми потенциально возможными словоформами, которые можно составлять, например, на базе основ слов, первоначально содержащихся в документах. Наличие в паре «документ – запрос» словоформ, совпадающих с точностью до общности их корней, в результате такого наращивания может привести к появлению в документе словоформы, полностью совпадающей со словоформой, имеющейся в запросе. Такое наращивание снимало бы различие употреблений словоформ в документах и запросах.

Другой технологический вариант, позволяющий снимать различие употреблений словоформ, состоит в использовании кодирования слов.

Сущность метода автоматического кодирования слов с помощью наперед заданных (составленных экспертами-лингвистами) словарей аффиксов заключается в автоматической проверке на наличие в словах естественных языков элементов, вошедших в такие словари, и возможном их отсечении.

От качества составления таких словарей в значительной мере зависит качество автоматического кодирования слов естественных языков, а, следовательно, и функциональная эффективность ИПС в целом. Ошибки могут быть следствием такого алгоритма, когда после включения очередной морфемы в словарь, она удаляется из всех слов естественно-языкового употребления в базе данных (даже если является для отдельных слов частью корня).

Словарь словосочетаний. Такой словарь используется для определения наиболее часто встречающихся устойчивых комбинаций слов. Словарь словосочетаний повышает эффективность анализа содержания, выделяя для идентификации содержания однозначные словосочетания вместо множества в общем случае неоднозначных слов (например, пара отдельных терминов «программа» и «язык» является менее определенной, чем словосочетание «язык программирования»).

Лингвистической особенностью словаря является то, что термины – одиночные слова зачастую не выражают никакого смысла, являясь только составной частью словосочетания.

Основываясь на том, что наиболее информативными терминами являются термины-словосочетания, наиболее правомерно использовать именно их для составления поискового запроса.

Например, количественный анализ ключевых слов, которые были использованы при индексировании документов БД ВИНТИ РАН «Информатика» за период с 1986–2000 гг., по следующим рубрикам Рубрикатора ВИНТИ РАН: 201.23.15 «Информационно-поисковые языки», 201.23.01 «Информационный поиск, общие вопросы», 201.01.04 «Информатизация общества, информационная политика», показал лидирующую роль словосочетаний при индексировании (см. табл. 4.1).

Частотный словарь. Частотный словарь – перечень дескрипторов и ключевых слов. Термины располагаются в алфавитном

порядке, либо в порядке убывания (возрастания) частоты использования их в информационном массиве.

Частотная характеристика термина показывает количество документов информационного массива, в которых термин встретился хотя бы один раз. Частота встречаемости ориентирует пользователя в лексике информационного массива с точки зрения включения какого-либо термина в поисковый запрос.

Рассмотрим, например, фрагмент частотного словаря ретроспективной реферативной БД «Информатика» (1986–2002 гг):

51	ИНФОРМАЦИОННАЯ ГРАМОТНОСТЬ
1	ИНФОРМАЦИОННАЯ ГРАНИЦА ВСЕЛЕННОЙ
1	ИНФОРМАЦИОННАЯ ДЕМОКРАТИЯ
1563	ИНФОРМАЦИОННАЯ ДЕЯТЕЛЬНОСТЬ
2	ИНФОРМАЦИОННАЯ ДИАГНОСТИКА
2	ИНФОРМАЦИОННАЯ ДОКТРИНА
1	ИНФОРМАЦИОННАЯ ДОКУМЕНТАЦИЯ

Из приведенного примера следует нецелесообразность использования в поисковых запросах ключевого слова *информационная деятельность* без соответствующих уточнений о видах такой деятельности (например, *информационная деятельность* и *лингвистика*).

Таблица 4.1

Количественный анализ ключевых слов

Рубрика	Количество				% словосочетаний
	документов	терминов	слов	словосочетаний	
201.23.15	868	1645	455	1190	72,34
201.23.01	1040	1867	445	1422	76,16
201.01.04	1475	2016	453	1563	77,52

Словарь отрицаний. Словарь отрицаний («стоп-слов») содержит термины, которые признаны не информативными для данной предметной области. Использование их запрещается для индексирования содержания документов. Например, термины «исследование», «вопросы», «требования», «проблемы» и др. являются

политематическими и удаляются из поисковых образов документов и запросов.

Словарь стоп-слов может использоваться как при построении частотных словарей, так и при разборе выражения информационной потребности на ИПЯ. Запрещенные термины не заносятся в словарь. Таким образом, неинформативные термины автоматически исключаются из поискового процесса.

4.2. Иерархические терминологические структуры

Иерархическая организация терминов или понятий, подобная схеме библиотечной классификации, позволяет для данного входа словаря найти более широкое понятие, перемещаясь вверх по схеме, или более узкое, перемещаясь вниз по схеме. Иерархические терминологические структуры отражают отношения «род-вид» или «часть-целое».

Указатель иерархических отношений терминов может быть создан только после решения проблемы классификации понятий. Он разрабатывается на основе классификационных схем понятий, поскольку фактически в иерархическом указателе находят отражение иерархические цепочки подчинения понятий, зафиксированные в классификационных схемах. В качестве основных входов этого указателя выбираются те заглавные термины, которые не имеют родовых дескрипторов (т. е. стоят на верхней ступени иерархического дерева) [1].

Иерархические классификационные структуры. К таким структурам относятся различные рубрикаторы и классификаторы, фиксирующие подчинение терминов в определенной предметной области (например, рассмотренный ранее рубрикатор ГРНТИ, рубрикатор ВИНИТИ и др.)

На рис. 4.2 приведен фрагмент Рубрикатора ВИНИТИ для заглавной рубрики «201 Информатика». Рубрикатор ВИНИТИ является локальным (отраслевым) по отношению к Государственному рубрикатору НТИ и отличается большей детализацией рубрик с развитием на глубину до восьмого уровня.

201.01	Общие вопросы информатики
201.01.04	Информатизация общества.
	Информационная политика
201.01.07	Теория и методология информатики
201.01.07.03	Предмет и объекты исследования информатики
201.01.07.03.03	Информация как знание
201.01.07.03.07	Научная коммуникация
201.01.07.03.11	Информационная деятельность
201.01.07.06	Методы информатики
201.01.07.06.03	Наукометрический анализ
201.01.07.06.05	Семиотический подход
201.01.07.06.07	Кластерный анализ
201.01.07.06.09	Теория информации
201.01.07.06.11	Эвристические методы
201.01.07.06.13	Системный подход
201.01.07.06.15	Моделирование
201.01.07.06.99	Другие методы
201.01.07.08	Проблемы информатики
201.01.07.08.03	Взаимосвязи данных, информации и знания
201.01.07.08.05	Представление знаний
201.01.07.08.07	Разработка классификаций

Рис. 4.2. Фрагмент Рубрикатора ВИНИТИ

Словарь синонимов. Словарь синонимов, который для каждого входа словаря определяет одну или более синонимичных категорий, также с точки зрения своей структуры может быть отнесен к иерархической организации терминов. Такие словари широко используются при индексировании, а также позволяют искать не только по запрошенному слову, но и по его синонимам.

Ниже приведен фрагмент словаря синонимов для области «Информатика».

...

ЭФФЕКТИВНОСТЬ ПОИСКА

информационная эффективность

техническая эффективность

эффективность информационного поиска

ЮНИСИСТ

Всемирная система научной и технической информации
ЮРИДИЧЕСКАЯ ДЕЯТЕЛЬНОСТЬ

юридическая практика

ЮРИСПРУДЕНЦИЯ

право

правоведение

юридические аспекты

ЯДЕРНАЯ ФИЗИКА

ядерная энергия

ЯЗЫКИ

подъязыки

языковые системы

ЯЗЫКИ МАНИПУЛИРОВАНИЯ ДАННЫМИ

ЯМД

процедурные языки манипулирования данными

ЯЗЫКИ ОПИСАНИЯ ДАННЫХ

ЯОД

ЯЗЫКИ-ПОСРЕДНИКИ

ЯП

ЯЗЫКИ ПРОГРАММИРОВАНИЯ

машинно-зависимые языки

машинно-независимые языки

машинно-ориентированные языки

машинные языки

проблемно-ориентированные языки

процедурно-ориентированные языки

...

Иерархическая организация терминов. В основу построения тематических словников двухуровневой иерархической структуры, которые в дальнейшем могут рассматриваться как исходные для формирования проблемно-ориентированных мини-тезаурусов, положено ранжирование терминов. При использовании таких структур происходит существенное сокращение необходимого терминологического пространства без ущерба для полноты тематического охвата.

При формировании мини-тезаурусов используется лексика представительной подборки документов из рассматриваемой предметной области. Для выбора терминов первого уровня (заглавных) используются формальные оценки, позволяющие вы-

явить так называемые «ядерные» для темы термины. Второй (и при необходимости – последующие) уровень составляют термины, присоединенные к заглавным терминам по принципу включения лексических единиц заглавного термина.

Приведем фрагмент иерархического словника для предметной области «Информационно-поисковые языки»:

АВТОМАТИЗАЦИЯ

АВТОМАТИЗАЦИЯ ВЕДЕНИЯ

АВТОМАТИЗАЦИЯ ПОДГОТОВКИ

АВТОМАТИЗИРОВАННОЕ ВЕДЕНИЕ

АВТОМАТИЗИРОВАННЫЕ БИБЛИОТЕЧНЫЕ СИСТЕМЫ

АВТОМАТИЗИРОВАННЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ

АВТОМАТИЗИРОВАННЫЕ ИПС

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ

ДОКУМЕНТОВ

АВТОМАТИЧЕСКОЕ ИНДЕКСИРОВАНИЕ

АВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ

ПОЛУАВТОМАТИЧЕСКОЕ ПОСТРОЕНИЕ

АВТОМАТИЧЕСКОЕ СОСТАВЛЕНИЕ

АКТУАЛИЗАЦИЯ

АЛГОРИТМЫ

АНАЛИЗ

АНАЛИЗ ДАННЫХ

АНАЛИЗ ДОМЕНОВ

АНАЛИЗ СОДЕРЖАНИЯ

АНАЛИЗ ТЕКСТА

ИСКРИМИНАЦИОННЫЙ АНАЛИЗ

ДИСПЕРСИОННЫЙ АНАЛИЗ

ИНФОРМАЦИОННЫЙ АНАЛИЗ

КАТЕГОРИАЛЬНЫЙ АНАЛИЗ

КЛАСТЕРНЫЙ АНАЛИЗ

ОЦЕНКА И АНАЛИЗ ХАРАКТЕРИСТИК

ПРОБЛЕМЫ АНАЛИЗА

СЕМАНТИКО-СИНТАКСИЧЕСКИЙ АНАЛИЗ

СЕМАНТИЧЕСКИЙ АНАЛИЗ

СИНТАКСИЧЕСКИЙ АНАЛИЗАТОР

СИСТЕМНЫЙ АНАЛИЗ

СРАВНИТЕЛЬНЫЙ АНАЛИЗ

СТАТИСТИЧЕСКИЙ АНАЛИЗ

4.3. Терминологические структуры с сетевой организацией

Наиболее употребительными формами информационного представления знаний в сетевой структуре являются тезаурусы и онтологии.

Тезаурусы

Основными представителями сетевых терминологических структур являются *тезаурусы*.

Весь окружающий мир можно рассматривать как множество, состоящее из двух элементов: предметов и их отношений. Этот реально существующий мир отражается в сознании человека в форме взаимосвязанных понятий, т. е. в такой форме мышления, при которой в сознании фиксируются только существенные связи и признаки предмета.

Все понятия естественного языка, служащие для описания окружающего мира, входят во всеобщий тезаурус, отражающий весь универсум знаний. Такой тезаурус представляет собой список понятий, выраженных на естественном языке, с обозначением отношения между ними.

Всеобщий тезаурус можно подразделить на частные тезаурусы путем выделения совокупности однородных понятий по их иерархическому уровню или путем выделения понятий, которыми можно описать какую-либо специфическую часть мира. Таким образом, на основе всеобщего тезауруса можно составить бесконечное множество тезаурусов по различным областям знаний, по отдельным проблемам и задачам.

Термин *тезаурус* достаточно древнего происхождения. Впервые его применил в значении, близком сегодняшнему, еще в тринадцатом столетии Брунто Латини в заголовке своего труда – систематизированной энциклопедии «Книга о сокровище». Особенно большую известность получил тезаурус, составленный в

1852 г. англичанином Роджетом «для облегчения выражения мыслей и помощи при написании сочинений» [3].

Тезаурус может быть представлен как семантическая сеть, в которой понятия связаны регулярными и устойчивыми семантическими отношениями – иерархическими (например, род-вид, целое-часть), ассоциативными, а также отношениями эквивалентности. При этом отдельное понятие определенной области знаний в тезаурусе представлено словом или словосочетанием, соотносящимся с другими словами и словосочетаниями и образующим вместе с ними замкнутую систему [3].

Иерархические отношения в тезаурусе представляют собой классификацию, основанную на словах естественного языка, а не на абстрактных категориях, поэтому нарушается правильная структура дерева – один и тот же термин может иметь несколько «родителей» – вышестоящих терминов на предыдущем уровне.

Например, в Тезаурусе по информатике [11] словосочетание *Автоматизированная обработка информации* имеет два вышестоящих родителя: *Автоматизированная обработка* и *Обработка информации*, а слово *Буквы* – целых три родителя: *алфавиты*, *символы*, *буквенно-цифровая информация*.

Тезаурус, отображая возможные семантические связи терминов, представленных в БД, является идеальным лексическим инструментом информационно-поисковых систем, с помощью которого можно найти необходимую лексику для составления запросов или их модификации с целью достижения наилучших показателей эффективности поиска.

Информационно-поисковые тезаурусы. Информационно-поисковые тезаурусы позволяют решить проблему соотнесения:

- авторской терминологии (понятиями и словами естественного языка, которые автор использует для обозначения этих понятий);
- терминологии системы (понятиями и терминами, которые используются для выражения этих понятий при вводе документов в ИПС);
- терминологии потребителя (понятиями и терминами, которые потребитель использует для представления этих понятий при формировании запросов).

Таким образом, тезаурус выступает как средство уменьшения семантического расстояния между выражением тематики документа и поисковым запросом.

Как сказано выше, являясь лексическим инструментом информационно-поисковых систем, тезаурус состоит из контролируемого, но изменяемого словаря терминов, между которыми указаны смысловые связи. Такой словарь исчерпывающим образом покрывает некоторую специфическую область знаний и представляет собой перечень *лексических единиц*, упорядоченных по систематическому и алфавитному принципам. Кроме этого между лексическими единицами заданы смысловые отношения как иерархического (родо-видового), так и неиерархического типа (ассоциативного).

Лексическая единица (ЛЕ) информационно-поискового тезауруса – выбранное для использования в тезаурусе слово, словосочетание или лексически значимый компонент сложного слова естественного языка.

Дескриптор – лексическая единица, предназначенная для использования в поисковых образах документов и/или запросов.

Аскриптор (недескриптор) – лексическая единица, которая в поисковых образах документов (запросов) подлежит замене на дескриптор при поиске или обработке информации.

Соответственно, информационно-поисковые тезаурусы подразделяют на два типа:

- 1) тезаурусы, выделяющие среди своих лексических единиц дескрипторы и аскрипторы;
- 2) тезаурусы, все лексические единицы которых являются дескрипторами.

Тезаурус по атомной науке и технике. Рассмотрим основные определения и виды связей, обозначаемых в тезаурусах, на примере информационно-поискового «Тезауруса по атомной науке и технике» в САРИ.

Лексические единицы тезауруса поделены на дескрипторы (графически выделенные написанием прописными буквами) и *ключевые слова* – недескрипторы (написание строчными буквами).

Ключевые слова – термины естественного языка (слова или словосочетания), служащие для точного обозначения понятий оп-

ределенной предметной области (предметов, явлений, свойств, отношений, процессов и т.д.). Термины, являющиеся абсолютными или условными синонимами (в рамках данной предметной области), объединяются в классы условной эквивалентности.

Один из терминов класса условной эквивалентности выбирается в качестве дескриптора. Он обозначает данный класс и выражает основное значение всех слов и словосочетаний, входящих в него.

Дескрипторы – нормализованные термины естественного языка. Каждое ключевое слово, не являющееся дескриптором, но входящее в тот или иной класс условной эквивалентности, имеет отсылку к соответствующему дескриптору.

ЛЕ нормализованы следующим образом:

- имена существительные, обозначающие исчисляемые предметы, представлены в форме именительного падежа множественного числа;
- существительные, обозначающие неисчисляемые объекты, представлены в форме именительного падежа единственного числа;
- для всех словосочетаний-дескрипторов, включая словосочетания с именем собственным, используется естественный (прямой) порядок слов.

Тезаурус содержит более 65 000 лексических единиц. Каждый дескриптор снабжен словарной статьей, отражающей его парадигматические отношения.

В тезаурусе приняты три вида парадигматических отношений:

- тождество (синонимия);
- подчинение (иерархические родо-видовые отношения);
- сходство (ассоциативные отношения).

Под родо-видовыми отношениями понимаются иерархические отношения между понятиями, обозначающими классы предметов, такие, что родовое понятие отражает существенные признаки всех видовых понятий, а видовое понятие содержит все признаки родового понятия, а также отражает конкретные свойства предмета или явления, выраженного этим видовым понятием.

Ассоциативные отношения – смысловые отношения дескрипторов типа часть-целое, причина-следствие, производитель-объект и т.п.

Словарная статья дескриптора состоит из собственно дескриптора (заглавного дескриптора) и списка дескрипторов и ключевых слов, связанных с заглавным дескриптором парадигматическими отношениями. В словарных статьях используются перекрестные ссылки, т.е. термины, содержащиеся в словарной статье дескриптора, должны иметь этот дескриптор в своих словарных статьях.

В *дескрипторной статье* лексические единицы располагаются в следующем порядке:

- заглавный дескриптор;
- ключевые слова, условно синонимичные заглавному дескриптору;
- вышестоящие дескрипторы (иерархически подчиняющие понятия);
- нижестоящие дескрипторы (иерархически подчиненные понятия);
- дескрипторы, связанные с заглавным дескриптором одним из ассоциативных отношений.

При записи дескрипторной статьи применяются условные обозначения.

Ключевое слово, не используемое в качестве дескриптора, снабжается символом «-» (знак минус).

Отсылка от ключевого слова к дескриптору обозначается сокращением «исп.» (англоязычный вариант – USE), например:

температура абсолютного нуля

USE температура абсолютного нуля

Обозначение «см.» отсылает от ключевого слова к одному из перечисленных за ним дескрипторов (англоязычный вариант – SEE):

теория биденхарна-роуза

SEE угловая корреляция

SEE угловое распределение

Если в дескрипторной статье необходимо задать ключевое слово, вместо которого *должен* использоваться заглавный дескрип-

тор, применяется условное обозначение «**исп. вместо**» (англоязычный вариант – **UF**). Отсылка «**см. вместо**» (англоязычный вариант – **SF**) задает ключевое слово, вместо которого *может* использоваться заглавный дескриптор, например:

ТЕМПЕРАТУРА АБСОЛЮТНОГО НУЛЯ

UF температура абсолютного нуля

RT диапазон температуры

RT криогенная техника

ЯДЕРНОЕ ОРУЖИЕ

UF атомное оружие

UF атомные бомбы

UF термоядерное оружие

UF ядерное нападение

SF проект тамблер

RT гражданская оборона

RT договор паротонга

RT договор тлателолко

RT испытательный полигон шт. невада

RT локальные выпадения

RT манхэттенский проект

RT нагасаки

RT национальная оборона

RT политика нераспространения ядерного оружия

RT проект кастл

RT проект пламббоб

RT проект редвинг

RT проект типот

RT противоракетная оборона

RT радиоактивные выпадения

RT снаряды

RT убежища

RT хиросима

RT ядерная зима

RT ядерное разоружение

RT ядерные взрывы

Дескриптор, являющийся в иерархическом отношении вышестоящим (родовым) по отношению к заглавному дескриптору, снабжается в словарной статье условным обозначением «**РТ**» (англоязычный вариант – **ВТ**). Если дескриптор является родовым для

большой группы дескрипторов, перечень которых в словарной статье не приводится, перед обозначением помещается символ «*» – «*РТ» («*ВТ»):

АЦЕТАЛЬДЕГИД

UF уксусный альдегид

UF этаналь

UF этилальдегид

*ВТ1 альдегиды

РТ ацеталь

РТ хлораль

Дескрипторы, нижестоящие по отношению к заглавному дескриптору (видовые), вводятся в дескрипторную статью со ссылкой «ВТ» (англоязычный вариант – NT). Этой ссылкой обозначаются как видовые дескрипторы, так и дескрипторы, обозначающие компонент (часть) понятия, выраженного заглавным дескриптором:

ПОГЛОЩЕНИЕ

UF торможение (в веществе)

ВТ1 сорбция

NT1 k-поглощение

NT1 всасывание в кишечнике

NT1 поглощение корнями

NT1 поглощение полярной шапкой

NT1 поглощение через кожу

NT1 поглощение энергии

NT1 резонансное поглощение

NT1 самопоглощение

РТ абсорбционная спектроскопия

РТ абсорбционный цикл охлаждения

РТ гетерогенные эффекты

РТ замедление

РТ излучения

РТ пробег

РТ пропускание

РТ самоэкранирование

РТ слой половинного поглощения

РТ спектры поглощения

РТ тормозная способность

РТ точечные ядра

РТ экранирование

Цифры 1, 2, 3, стоящие после обозначений РТ (ВТ) и ВТ (NT), указывают на соответствующий уровень подчинения:

УСКОРИТЕЛИ

NT1 когерентные ускорители

NT1 коллективные ускорители

NT2 плазменные бетатроны

NT2 ускорители с фронтом ионизации

NT2 ускорители электронных колец

NT1 линейные ускорители

NT2 квадрупольные лин. ускорители

NT2 линейные ускорители тяжелых ионов

NT3 сверхпроводящий линейный ускоритель atlas

NT3 суперхайлак

NT2 линейный уск-ль тяжелых ионов с перестройкой частоты

NT2 линейный ускоритель fmit

NT2 линейный ускоритель jaeri

NT2 линейный ускоритель kek

NT2 линейный ускоритель lampf

NT2 линейный ускоритель mea

NT2 линейный ускоритель mit bates

NT2 линейный ускоритель nrl

NT2 линейный ускоритель orela

NT2 линейный ускоритель на 200 мэв бнл

NT2 линейный ускоритель орсэ

NT2 линейный ускоритель протонов beijing

NT2 линейный ускоритель сакле

NT2 линейный ускоритель свирк

NT2 линейный ускоритель фрасскатти

NT2 линейный ускоритель хфти

NT2 линейный ускоритель церн

NT2 станф. лин. ускор. на энергию 1,2 гэв

NT2 станфордский лу на 20 гэв

NT2 универсальный линейный ускоритель unilac

NT2 ускорители на волнах биения

NT2 ускорители с использованием спут-ого поля

NT2 усовершенствованный испытательный ускоритель нллл

NT2 установка на основе ускорителя с непрерывным электр. пучком

NT2 электрон-позитронный коллайдер beijing

NT1 мезонные фабрики

NT2 линейный ускоритель lampf

NT2 синхротрон lampf ii
 NT2 установки для медицин-го обсл-ния с использованием пионов
 NT1 рельсовые ускорители
 NT1 ускорители для термоядерного синтеза
 NT1 ускорители тяжелых ионов
 NT2 коллайдер релятивистских тяжелых ионов бл
 NT2 линейные ускорители тяжелых ионов
 NT3 сверхпроводящий линейный ускоритель atlas
 NT3 суперхайлак
 NT2 линейный уск-ль тяжелых ионов с перестройкой частоты
 NT2 миланский сверхпроводящий циклотрон
 NT2 мюнхенский свсе циклотрон
 NT2 сверхпроводящий циклотрон crnl
 NT2 синхротрон sis
 NT2 техасский сверхпроводящий циклотрон
 NT2 универсальный линейный ускоритель unilac
 NT2 ускоритель hhif
 NT2 ускоритель vicksi
 NT2 ускоритель нуматрон
 NT2 циклотрон ganil
 NT2 циклотрон hirfl
 NT2 циклотрон kvi
 NT2 циклотрон u-120 в кракове
 NT2 циклотрон в варшаве
 NT2 циклотрон в калькутте
 NT2 циклотрон института физичеких и химических исследо-
 ваний (ифх)
 NT2 циклотрон иссл-го центра ядер-й физики (унив-т в осака)
 NT2 циклотрон нуц, юар
 NT2 циклотрон токийского ин-та ядерных иссл-ий
 NT2 циклотрон тохоку
 NT2 циклотрон у-400 оияи
 NT2 циклотрон циклон
 NT1 циклические ускорители
 NT2 бевалак
 NT2 бетатроны
 NT2 синхротроны
 NT3 беватрон
 NT3 кембриджский ускоритель электронов сша
 NT3 корнеллский синхротрон на 10 гэв

NT3 космотрон
 NT3 накопительное кольцо cosy
 NT3 накопительное кольцо esca
 NT3 накопительные кольца lep
 NT3 принстон-пенсильванский синхротрон
 NT3 сверхпроводящие суперколлайдеры
 NT3 серпуховской тэватрон
 NT3 синхротрон ags бнл
 NT3 синхротрон ipns-i
 NT3 синхротрон lampf ii
 NT3 синхротрон luby
 NT3 синхротрон mura
 NT3 синхротрон ps в черне
 NT3 синхротрон sis
 NT3 синхротрон sps
 NT3 синхротрон zgs
 NT3 синхротрон боннского университета
 NT3 синхротрон в пахре
 NT3 синхротрон ефти
 NT3 синхротрон итэф
 NT3 синхротрон ифвэ
 NT3 синхротрон кек
 NT3 синхротрон лаборатории во фрасскати
 NT3 синхротрон оияи
 NT3 синхротрон токийского ун-та
 NT3 синхротрон томского университета
 NT3 синхротрон фиан
 NT3 тэватрон лаборатории им. ферми
 NT3 ускоритель desy
 NT3 ускоритель nimrod
 NT3 ускоритель nina
 NT3 ускоритель saturne
 NT3 ускоритель saturne ii
 NT3 ускорительный комплекс pal
 NT2 синхроциклотроны
 NT3 синхротрон iko
 NT3 синхроциклотрон uppsala
 NT3 синхроциклотрон в беркли
 NT3 синхроциклотрон в орсэ
 NT3 синхроциклотрон в черне
 NT3 синхроциклотрон гарвардского университет

NT3 синхроциклотрон мак-гилла
 NT3 синхроциклотрон оияи
 NT3 синхроциклотрон харуэлл
 NT2 циклотроны
 NT3 изохронные циклотроны
 NT4 изохронный циклотрон орнл
 NT4 казахстанский циклотрон
 NT4 киевский циклотрон
 NT4 миланский сверхпроводящий циклотрон
 NT4 мюнхенский компактный циклотрон
 NT4 мюнхенский свсе циклотрон
 NT4 принстон-пенсильванский циклотрон
 NT4 сверхпроводящий циклотрон csnl
 NT4 техасский сверхпроводящий циклотрон
 NT4 циклотрон, юлих
 NT4 циклотрон aabo
 NT4 циклотрон aic-144 в кракове
 NT4 циклотрон alice
 NT4 циклотрон ganil
 NT4 циклотрон hirfl
 NT4 циклотрон iu
 NT4 циклотрон kvi
 NT4 циклотрон nirs
 NT4 циклотрон nrl
 NT4 циклотрон sara (система уск-лей в роун-альпе)
 NT4 циклотрон sin
 NT4 циклотрон triumf
 NT4 циклотрон агро-механического колледжа
 NT4 циклотрон бнл
 NT4 циклотрон в варшаве
 NT4 циклотрон в гренобле
 NT4 циклотрон в дебрецине
 NT4 циклотрон в карлсруэ
 NT4 циклотрон в орсэ
 NT4 циклотрон в осло
 NT4 циклотрон в эйндрховене
 NT4 циклотрон гайзи
 NT4 циклотрон института физичеких и химических исследо-
 ваний (ифх)
 NT4 циклотрон иссл-го центра ядер-й физики (унив-т в осака)
 NT4 циклотрон ияи (шанхай)

NT4 циклотрон нуц, юар
 NT4 циклотрон токийского ин-та ядерных иссл-ий
 NT4 циклотрон тохоку
 NT4 циклотрон циклон
 NT4 циклотроны msu
 NT4 циклотроны uclrl
 NT5 88-дюймовый циклотрон в ллб
 NT4 циклотроны оияи
 NT5 циклотрон у-400 оияи
 NT3 микротроны
 NT4 рейстрек-микротроны
 NT3 сверхпроводящие циклотроны
 NT4 миланский сверхпроводящий циклотрон
 NT4 техасский сверхпроводящий циклотрон
 NT3 циклотрон u-120 в кракове
 NT3 циклотрон ин-та нильса бора
 NT3 циклотроны с переменной энергией
 NT4 циклотрон в калькутте
 NT4 циклотрон в чандигаре
 NT3 циклотроны с разделенными орбитами
 NT1 электростатические ускорители
 NT2 динамитроны
 NT2 тандемные электростатические ускорители
 NT3 тандемный ускоритель вивитрон
 NT3 тандемный ускоритель ияи японии
 NT3 тандемный ускоритель мр
 NT3 тандемный ускоритель орсе
 NT2 ускорители ван-де-граафа
 NT3 тандемный ускоритель вивитрон
 NT3 тандемный ускоритель ияи японии
 NT3 тандемный ускоритель мр
 NT3 тандемный ускоритель орсе
 NT2 ускорители кокрофта-уолтона
 NT2 ускорители-пеллетроны
 NT3 ускоритель-пеллетрон 5u
 RT бустеры
 RT вакуумные системы
 RT динамика пучка
 RT драйверы ударных реакций синтеза
 RT мишенные камеры
 RT накопительные кольца

RT поглотители пучка
 RT производство изотопов
 RT сепараторы пучков
 RT ускорение
 RT ускорители-размножители
 RT ускорительный комплекс

Дескрипторы, связанные с заглавным дескриптором другими видами отношений, включены в дескрипторную статью со ссылкой «АТ», обозначающей ассоциацию (англоязычный вариант – RT).

АБСОРБЦИОННАЯ СПЕКТРОСКОПИЯ

UF атомная абсорбционная спектроскопия
 UF колориметрия
 SF спектрохимия
 BT1 спектроскопия
 RT инфракрасные спектры
 RT лазерная спектроскопия
 RT методы двойного резонанса
 RT поглощение
 RT спектры крайнего ультрафиолетового излучения
 RT спектры поглощения
 RT структурный химический анализ
 RT ультрафиолетовые спектры
 RT фотоакустические спектрометры

В табл. 4.2 приведены возможные связи между дескрипторами и ключевыми словами в тезаурусе INIS с указаниями типа отношения.

Таблица 4.2

Связи в тезаурусе

Обозначение		Название	Тип отношения
Рус.	Англ.		
см.	SEE	смотри	синонимия
см. вместо	SF	смотри вместо	синонимия
исп.	USE	используй	синонимия
исп. вместо	UF	используй вместо	синонимия
РТ	BT	вышестоящий	иерархия
BT	NT	нижестоящий	иерархия
АТ	RT	ассоциативный	ассоциация

Характеристика словарного состава Тезауруса. Лексика тезауруса может быть представлена множеством терминов

$$V = \{ M_1 M_2 M_3 M_4 M_5 M_6 \},$$

которое включает следующие подмножества [14]:

M_1 – ненормализованные термины (недескрипторы), которые являются условными или истинными синонимами дескрипторов.

M_2 – дескрипторы без связей, т.е. понятия, для которых фиксация в тезаурусе родо-видовых отношений была признана нецелесообразной. К данному подмножеству относятся также *категории* – широкие понятия, которые не входят ни в какие более широкие понятия, и, кроме того, *единичные понятия*, отражающие признаки какого-либо одного предмета и не имеющие видового развития.

M_3 – родоначальные дескрипторы. К родоначальным дескрипторам относятся дескрипторы понятий, которые не имеют в тезаурусе родового развития, но имеют подчиненные видовые понятия. Эти дескрипторы являются корнями иерархических деревьев.

M_4 – видовые дескрипторы первого уровня.

M_5 – видовые дескрипторы второго уровня.

M_6 – видовые дескрипторы третьего и низшего уровней.

Иерархические структуры в тезаурусе могут иметь несколько уровней, и на каждом из них возможно отражение выделенных при анализе понятий, что позволяет описывать содержание предмета по тематике тезауруса с различной степенью специфичности. Однако для большинства родо-видовых семейств в тезаурусе характерно развитие до первого и второго видовых уровней.

На I-м уровне специфичности находятся дескрипторы без связей и дескрипторы родоначальных понятий. Эта величина характеризует предметную широту тезауруса – представление в нем различных понятий.

II-й уровень специфичности, основу тезауруса, составляют дескрипторы без связей, дескрипторы родоначальных понятий и дескрипторы видовых понятий первого уровня.

Специфичность описания документа или запроса можно увеличить последовательным расширением используемой лексики до видовых понятий II (III уровень специфичности) и низших уровней (IV уровень специфичности). Объем словарного состава при этом увеличивается до 90% и 100% соответственно.

Способность языка индексирования отражать понятия в точном соответствии с тем объемом, с каким они выделены при анализе предметного содержания, характеризуется его специфичностью. Для тезауруса с развитой иерархией понятий специфичность может быть приближенно оценена степенью развития иерархических связей.

Полнота индексирования определяется возможностью перевода на ИПЯ всех понятий, выявленных при анализе содержания предмета в заданной тематической области, и зависит, таким образом, от наличия в языке дескрипторов, представляющих эти понятия, т. е. от предметной широты тезауруса.

Развитие иерархических связей входного словаря, а также предметная широта тезауруса характеризуют его «семантическую силу», определяемую как «способность точно и полно выражать смысл любого сообщения».

Иерархические деревья понятий в тезаурусе отличаются как по числу иерархических уровней, так и по количеству видовых понятий на различных уровнях. *Степень развития иерархической структуры* тезауруса можно оценить отношением числа всех дескрипторов к числу разных понятий:

$$I = \frac{M_2 + M_3 + M_4 + M_5 + M_6}{M_2 + M_3}.$$

Степень развития входного словаря, а следовательно, и возможности тезауруса по переводу содержания текстов по его тематике с естественного языка на ИПЯ (индексирование) можно оценить с помощью коэффициента синонимии:

$$S_n = \frac{M_1}{V}.$$

Обзор правил и методов построения тезауруса. Построение тезауруса включает в себя следующие этапы.

1. Определение тематического охвата информационно-поискового тезауруса путем анализа информационной потребности абонентов (потребителей).

2. Сбор массива лексических единиц в том случае, если отсутствуют тезаурусы по заданной тематике. Первоначальный сбор лексики осуществляется выделением лексических единиц из представленной коллекции документов и/или запросов. В полученный массив лексических единиц дополнительно должны быть включены соответствующие тематике лексические единицы, выделенные в соответствии с методикой построения данного тезауруса из указанных в ней источников: рубрикатора ГРНТИ; энциклопедических и терминологических словарей, справочников; терминологических стандартов; классификаторов технико-экономической информации; таблиц УДК, МКИ и других систем классификации.

3. Формирование словника тезауруса. В словник могут быть включены следующие типы лексических единиц:

- одиночные слова (существительные, прилагательные, глаголы, наречия, причастия);
- именные словосочетания;
- лексически значимые компоненты сложных слов;
- аббревиатуры; при наличии аббревиатур в словник должны включаться также соответствующие им полные формы (если только она не отсутствует).
- сокращения слов и словосочетаний.

Одиночные существительные следует заменять формой именительного падежа. Формы прилагательных и причастий следует приводить к именительному падежу. Прилагательные и причастия в единственном числе приводятся к форме мужского рода. Глаголы рекомендуется заменять отглагольными существительными. Глаголы, включенные в словник, приводят к форме инфинитива.

4. Построение словарных статей. При построении словарных статей информационно-поискового тезауруса лексическим единицам приписываются ссылки, устраняется неоднозначности ЛЕ, устанавливаются отношения эквивалентности, выбирается дескриптор, представляющий класс эквивалентности при индексиро-

вании (для ИПТ, различающих дескрипторы и аскрипторы), устанавливаются иерархические и ассоциативные отношения между дескрипторами.

5. Построение лексико-семантического указателя. Лексико-семантический указатель является упорядоченной последовательностью словарных статей ИПТ и формируется путем расположения их в алфавитном порядке заглавных ЛЕ.

6. Создание алгоритма автоматизированного построения ИПТ, например, проведение частотного анализа, коррекции статей, алфавитной сортировки словника, проверки взаимности и непротиворечивости ссылок, составление указателей, распечатка в требуемых форматах [29].

Основные проблемы, с которыми приходится сталкиваться при построении тезауруса, можно разделить на три класса [35]:

- относящиеся к форме терминов;
- относящиеся к организации связей между терминами;
- относящиеся к занесению их в тезаурус, то есть описанию связей между дескрипторами и аскрипторами.

Проблема организации связей между терминами, по существу, имеет две компоненты: те, что относятся к иерархической структуре, и те, которые относятся к неиерархическим связям между терминами (ассоциативные отношения).

Проблемы, относящиеся к форме терминов, включают в себя принятие решения о том, использовать единственное или множественное число, какие классы терминов (прилагательные, существительные, глаголы) могут служить в качестве дескрипторов и уровень, до которого контролируемый словарь может содержать сложные (составные) словосочетания.

Проблемы, попадающие под название «ввод терминов», включают в себя синонимию, отношение «частное-целое», отношение «один-ко-многим», омонимию, а также необходимость сокращенного написания (аббревиатуры).

Эффективность поиска, безусловно, повышается при использовании на этапе формирования запроса отраслевых словарей и тезаурусов. Следует, однако, отметить, что составление таких

средств вручную занимает несколько лет, причем за это время многое меняется и в проблематике, и в лексике отрасли.

Онтология как средство формализованного представления информации

Одним из перспективных направлений в области формализации знаний, которое дает возможность использования накопленных знаний для компьютерной обработки, являются онтологии.

Онтология – это набор определений (на формальном языке) фрагмента декларативных знаний, ориентированный на совместное многократное использование различными пользователями. В онтологии вводятся термины, типы и соотношения (аксиомы), описывающие фрагмент знания [2]. Таким образом, онтология определяет общий словарь для специалистов, которым нужно совместно использовать информацию в предметной области.

Понятие онтология активно применяется в информатике и искусственном интеллекте. Этот термин пришел из философии, где обозначал часть метафизики – учение обо всем сущем, о его наиболее общих философских категориях, таких как бытие, субстанция, причина, действие, явление. При этом онтология как наука претендовала на полное объяснение причин всех явлений [4; 33].

Онтологии представляют понятия в виде, пригодном для машинной обработки. Нередко онтологии используются в качестве посредника между пользователем и информационной системой: с их помощью можно формализовать договоренности о терминологии.

По степени зависимости от конкретной задачи или предметной области обычно различают следующие онтологии.

Онтологии верхнего уровня – описывают наиболее общие, независимые от конкретной проблемы или области концепты (пространство, время, материя, объект, событие, действие и т. д.), которые могут быть унифицированы для больших сообществ пользователей.

Онтологии, ориентированные на предметную область – стандартные онтологии, которые могут использоваться экспертами для совместной информационной деятельности в своей области.

Онтологии, ориентированные на задачу – онтологии, используемые при выполнении конкретной задачи. Они отражают специфику решения задачи, но могут также содержать некоторые общие термины.

Прикладные онтологии – описывают понятия, которые могут относиться как к онтологии задач, так и к онтологии ПрО. Примером может служить онтология автомобилей, строительных материалов, вычислительной техники. Прикладная онтология обобщает понятия, используемые в некоторых задачах ПрО, абстрагируясь от самих задач (так, онтология автомобилей независима от любых особенностей конкретных марок машин).

Язык описания онтологий. Ключевым моментом в проектировании онтологий является выбор языка спецификации онтологий (Ontology specification language). Цель таких языков – существенно повысить выразительные возможности концептуального моделирования слабо структурированных данных.

К традиционным языкам спецификации онтологий относятся: Ontolingua; CycL; языки, основанные на дескриптивных логиках (LOOM [34]); языки, основанные на фреймах (OKBC, OCML, Flogic [4]).

Более поздние языки ориентированы на Web-стандарты (XOL, SHOE, UPML). Специально для обмена онтологиями через Web были созданы RDF(S), DAML, OIL, OWL.

Отношения в онтологиях. При создании онтологии предметной области выделяются отношения, которые необходимо использовать для задания связей между объектами. Далее приводится краткий обзор отношений в существующих онтологиях. Набор этих отношений призван обеспечить эффективную работу в информационно-поисковых приложениях.

Отношения иерархии делятся на три подвида.

1. Родовидовое или внутрикатегориальное отношение *род ↔ вид*.

Основные конструкции предложений для поиска родового понятия В и видового понятия А:

А относится к В;

родом A является B ;
 A принадлежит классу/семейству B .
 B имеет следующие виды: $A_1, A_2, \dots, A_N$;
к видам B относятся $A_1, A_2, \dots, A_N$.

2. Отношение *признак* $\leftrightarrow$ *значение признака* или отношение типизация-конкретизация. Здесь признак или тип – это категория, а значение – это имя конкретной категориальной формы.

Основные конструкции предложений для поиска с использованием отношения, если A – понятие-категория, B – имя категории:
 A имеет имя B ;
 A именуется B ;
 A называется B .

3. Отношение *инвариант* $\leftrightarrow$ *вариант*. В этом отношении инвариант – это неизменная структурная единица языка, а вариант – разновидность структурной единицы языка.

Основные конструкции предложений для поиска с использованием отношения для понятия инварианта A и варианта B :
к вариантам A относятся $B_1, B_2, \dots, B_N$;
инвариантом B является A .

Отношения агрегации делятся на четыре подвиды.

1. Отношение *целое* $\leftrightarrow$ *часть (компонент)*. Целое в отношении – то, что перестает существовать, если отнять любую часть, а часть – то, без чего целого не существует.

Для поиска используются следующие конструкции (целое A – часть B):

B входит в состав A ;
 B составляет часть A ;
 B является частью;
 B является элементом A ;
 A состоит из $B_1, B_2, \dots, B_N$;
 A включает в себя B как часть;
 A включает в свой состав B как часть;
 A имеет своими частями/элементами $B_1, B_2, \dots, B_N$.

2. Отношение *объект* $\leftrightarrow$ *пространство реализации объекта*.
Пространство реализации – это то пространство, где объект A проявляет свои свойства и функции.

Выражения для поиска:

А происходит в В;

А находится в В;

А входит в В.

3. Отношение *объект ↔ свойство/признак*. Свойство – это то, что присуще объекту, что отличает его от других. Свойства делятся на существенные и несущественные.

Основные выражения для поиска следующие:

А обладает свойством В,

А имеет В существенным признаком;

А характеризуется наличием В;

для А характерно В.

4. Отношение *уровень ↔ единица уровня*. Выделяют такие составляющие, как уровень, единица уровня, тип структуры.

Для поиска используются следующие выражения:

А принадлежит уровню В;

А – единица уровня В;

А рассматривается на уровне В.

Семиотические отношения предназначены для выражения соответствия между понятиями знаковых систем. Посредством таких отношений можно создавать иерархии понятий знаковых систем, при этом каждый уровень иерархии будет соответствовать одному метаязыку. Виды семиотических отношений:

1. Отношение *термин ↔ способ выражения*. Отражает фундаментальные свойства языка как системы знаков, имеющей план выражения и план содержания. Как знаковая система, предназначенная для коммуникации, язык обладает совокупностью средств выражения самых разнообразных значений. Отсюда следует, что в языке должна быть терминология, относящаяся к формальной стороне языка, и терминология, называющая значения и функции. В данном отношении осуществляется связь двух языковых объектов. Диагностирующие выражения:

А обычно выражает В;

А используется для выражения В;

А является В.

2. Отношение *термин ↔ способ представления*. Фиксирует связь языкового объекта и его представления в модельном языке лингвистики (метаязыке), где ненаблюдаемый теоретический язы-

ковый объект может получить наглядный аналог во фрагменте или элементе некоторой модельной записи. Диагностирующие выражения:

А обычно представляется посредством В;

А используется для представления В;

А используется с помощью В.

3. Отношение *термин одного метаязыка ↔ термин другого метаязыка*. Позволяет установить соответствие между двумя знаковыми системами как в одной предметной области, имеющей разные уровни представления знаний, так и между смежными предметными областями.

Диагностирующие выражения:

А представлено в виде/как В;

А представлено с помощью В;

способом представления А является В.

Функциональные отношения – отношения, принадлежащие сфере процессуальности, представляют собой многоместный предикат «Операция», в котором выделяют следующие аргументы (места):

А: субъект операции;

В: инструмент или способ/метод/алгоритм выполнения операции;

С: начальный объект или исходные данные;

Д: конечный объект или результат/выходные данные;

Е: событие, активизирующее операцию.

Субъектом операции может выступать человек или устройство. Основное выражение для выявления субъекта операции:

А совершает В.

Начальным объектом (исходными данными) называется объект, над которым совершается операция. Выражения для выявления начального объекта:

А осуществляется над В;

А применяется к В;

А меняет В.

Конечный объект или результат/выходные данные – это объект, являющийся конечным результатом выполнения операции. Выражения для выявления конечного объекта:

в результате А получается В;

в результате А образуется В;
А приводит к В;
А превращает В₁ в В₂.

В [30] предлагается следующая типология отношений (связей):

- *таксономические связи*, выражающие отношение «является видом» или отношение «общее/специфичное»;
- *композиционные связи*, выражающие отношение «является частью» или «агрегация объектов друг из друга»;
- *топологические связи*, отражающие, как различные составляющие (компоненты) системы связаны друг с другом через определенные связи, или показывающие «пути» физических взаимодействий между составляющими.

Каждый из таких типов связей представлен в конкретных онтологиях множеством вариантов.

К наиболее типичным таксономическим отношениям относятся отношения, ставящие две сущности в зависимость «частное-общее», и обозначаемые «является потомком», «является подклассом», «является представителем».

К композиционным связям в искусственном интеллекте и концептуальном моделировании относят различные типы отношений «часть-целое»: «компонент-объект», «вещество-ингредиент», «член из набора», «порция массы или большого количества», «место в области», «фаза, шаг-процесс».

Топологическими связями являются следующие связи:

- связь объекта и свойства: сущность «Имеет атрибут» свойство;
- связи сущностей с процессами и связи сущностей с сущностями посредством процессов: «Сущность поддерживает процесс», «Сущность является инструментом процесса», «Сущность участвует в процессе», «Сущность выполняет процесс», «Сущность управляет процессом» и т.п.;
- причинно-следственные связи: «Процесс влияет на прохождение другого процесса» (например, препятствует, оберегает от), «Процесс влияет на сущность», «Процесс вызывает другой процесс» и др.;

- временные связи (или связи сущностей, протекающих во времени): «Процесс связан по времени с другим Процессом», «Временной интервал является частью другого», «Временной интервал связан с другим интервалом», «Процесс является частью другого Процесса»;
- пространственные связи: «Сущность частично располагается в другой сущности», «Сущность наполняет другую сущность», «Сущность является носителем (содержит внутри себя) процесса»;
- родственные биологические связи: «Организм имеет родителем Организм», «Организм имеет потомком Организм», «Организм имеет братом/сестрой Организм».
- «умственные» связи: «Человек владеет чем-то», «Человек является гражданином чего-то», «Сущность является родиной Организма».
- целевые связи: «Сущность имеет целью абстракцию», «Сущность является результатом, является продуктом Процесса» и т.д.

На рис. 4.3 приведен фрагмент онтологии по естественным наукам и технологиям для приложений в сфере информационного поиска [8].

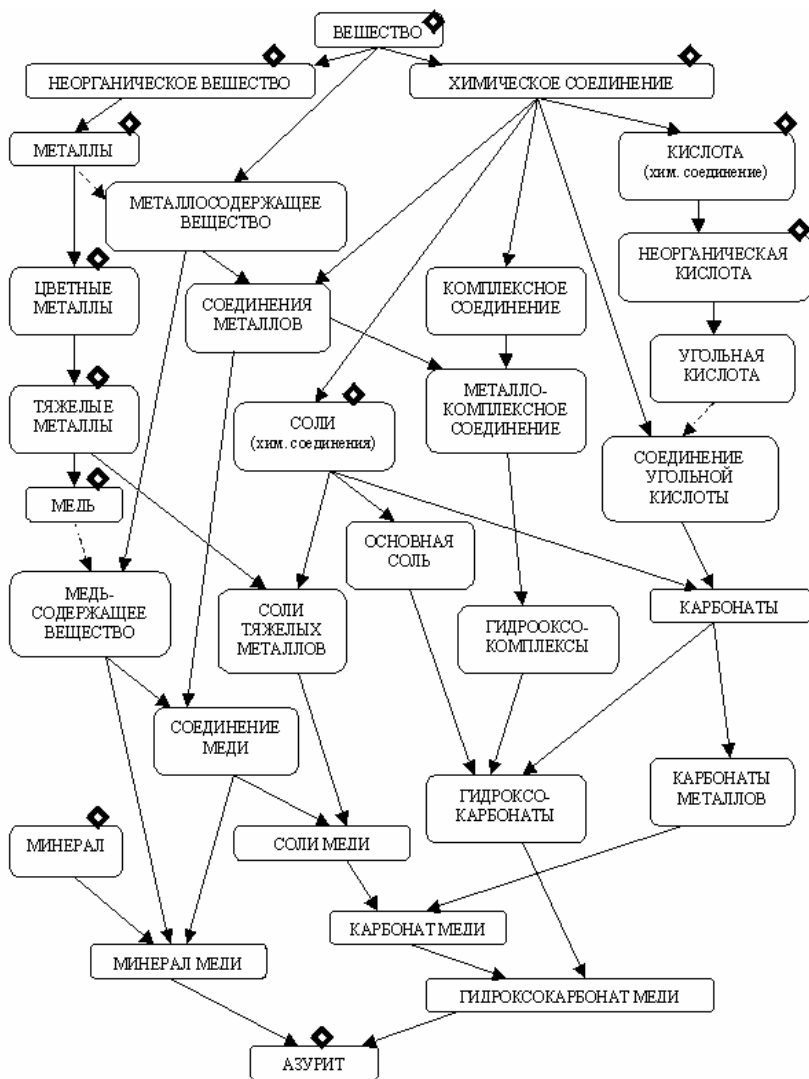


Рис. 4.3. Фрагмент онтологии по естественным наукам и технологиям

ГЛАВА 5. МЕТОДЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ И ЛЕКСИЧЕСКОГО АНАЛИЗА ДОКУМЕНТАЛЬНОЙ ИНФОРМАЦИИ

В основе решения таких задач анализа и синтеза, как индексирование, классификация и кластеризация, лежит понятие семантического соответствия между текстами, которое в свою очередь сводится к определению «силы» семантической связи между соответствующими терминами. Поэтому методы автоматической обработки документальной информации рассматривают текст как ограниченный набор терминов (отдельных слов и словосочетаний), образующих в контексте документа семантическую систему, представляющую его смысл. В каждой предметной области формируется своя, в той или иной степени устойчивая и динамичная, терминологическая система. В информационном аспекте (в задачах обработки информационных потоков) такую систему могут составлять следующие объекты:

- терминологические таблицы, организованные в виде семантических сетей, каждому узлу которых сопоставлен некоторый набор документов;
- тезаурусы и онтологии – сетевые словарные структуры, отражающие парадигматические и синтагматические связи терминов в конкретной предметной области;
- классификации, тематические рубрикаторы и др.

В целом можно сказать, что именно терминологические системы позволяют не только идентифицировать информационные объекты хранения и поиска, но и, оценивая динамику изменения терминологической системы, наблюдать и прогнозировать развитие (по крайней мере, изменение) предметной области. Количественной мерой при этом будет изменение частот употребления лексики, характерной для этой предметной области.

5.1. Методы выделения информативных терминов

Первой задачей при построении терминологической системы является задача выделения из текстов предметной области отдельных слов или словосочетаний, которые для данной предметной области обозначают ключевые понятия. Для выявления терминов –

кандидатов для включения в терминологическую систему – применяются модели «взвешивания» и ранжирования. Вычисление веса отдельного термина в рамках информационного массива проводится в соответствии со следующими основными методиками:

TF (Term Frequency) – в основе методики лежит частота встречаемости термина в документе (или в массиве документов).

TF*IDF (Term Frequency*Inverse Document Frequency) – в соответствии с этой методикой вес слова зависит от частоты его появления в конкретном предложении и в документе в целом.

TL*TF (Term Length*Term Frequency) – методика базируется на том, что слова, которые появляются часто, стремятся быть краткими. Такие слова не описывают основную тему документа, т.е. являются стоп-словами. Наоборот, слова, которые появляются редко, стремятся быть длинными.

Рассмотрим некоторые конкретные применения перечисленных методик.

Мера инверсной частоты термина (TF-методика). Расчет весовых коэффициентов может быть основан на так называемой *мере инверсной частоты* термина [31; 37], которая задается следующим выражением:

$$w_i = \log \frac{N_0}{F_i}, \quad 0 \leq w_i \leq \log N_0,$$

где N_0 – общее число документов информационного массива;

F_i – частота i -го термина (количество документов массива, содержащих данный термин).

Такая мера разбивает все терминологическое пространство на $[\log N_0]$ кластеров со значениями весов $0 \leq w_i < 1$, $1 \leq w_i < 2, \dots, [\log N_0] - 1 \leq w_i < [\log N_0]$. При этом термины, у которых $F_i = 1$, попадают в последний кластер – кластер самых «тематически точных» терминов, а в первом кластере находятся термины, наиболее общие для информационного массива.

Мера Чена (TL*TF-методика). В [32] рассматривается модифицированная мера инверсной частоты, дополненная показателем длины термина:

$$w_i^{ch} = \log L_i + \log \frac{N_0}{F_i},$$

где L_i – количество слов в термине.

Рассмотрим этот весовой показатель с точки зрения структурных свойств терминологической системы:

$$w(di) = \log(L_i) + (\log N - \log(F_i))$$

Первое слагаемое отражает степень точности отождествления термина с определенным понятием (успешность угадывания смысла термина, образованного из l слов, по крайней мере, в l раз выше), второе – «специфичность», как величину обратную «общности» в рамках множества возможных значений коллекции из N документов.

В пространстве терминов образуется еще один кластер со значениями весов $[\log N_0] \leq w_i < [\log N_0] + 1$, а термины, у которых $F_i = 1$, распределены уже по нескольким кластерам, причем в кластер с номером $[\log N_0] + 1$ попадают самые длинные словосочетания. Это отражает тот факт, что словосочетания являются более специфичными (менее общими) терминами для информационного массива.

TF*IDF-методика. Методика расчета весов заключается в следующем:

Пусть документ d состоит из m предложений и может быть представлен в виде набора предложений:

$$d = (s_1, s_2, \dots, s_m).$$

Каждое предложение идентифицируется взвешенным вектором слов, присутствующих в документе

$$s_i = (w_{i1}, w_{i2}, \dots, w_{in}),$$

где n – количество слов в документе d .

Вес w_{ij} слова j зависит от частоты его появления в конкретном предложении i и во всем наборе предложений (в документе):

$$w_{ij} = f_{ij} \log_2 \left(\frac{m}{m_j} \right), \quad i = 1, \dots, m;$$

$$j = 1, \dots, n.$$

где m_j – количество предложений, в которых присутствует слово j .

Функция f_{ij} частоты появления слова j в предложении i вычисляется следующим образом:

$$f_{ij} = \frac{n_{ij}}{n \cdot \text{len}(s_i)},$$

где n_{ij} – количество появления слова j в предложении i ;
 $\text{len}(s_i)$ – длина (количество слов) предложения s_i .

Во избежание смещения, вызванного длиной предложения, функция f_{ij} нормализована относительно длины предложения.

Неинформативными являются термины с малой частотой встречаемости (часто среди них встречаются слова с ошибками) и большой частотой встречаемости (кандидаты на новые стоп-слова) в наборе.

Выделение словаря информативных слов проводится на основе граничных значений частот. После определение интервала термины, попадающие в этот интервал, составляют множество информативных терминов (ключевых слов), полученное после этапа статистической обработки.

Приведем еще один способ вычисления веса термина по методике TF*IDF, который не требует информации об отдельных предложениях (интегральная значимость):

$$tf * idf = \log_2 \left(\frac{N}{n_t} \right) \times \left(\frac{1}{2} + \frac{1}{2} \times \frac{m_{dt}}{m_d} \right),$$

где N – количество документов,

n_t – количество документов, в который входит термин t ,

m_{dt} – число вхождений термина t в документ d ,

$m_d = \max_d m_{dt}$.

5.2. Статистические меры близости терминов

В общем случае семантические связи отражают качественно различные, как правило, не сводимые друг к другу связи между денотатами. Поэтому, если два термина связаны с некоторым третьим двумя различными видами семантической связи, то сопоставление их с целью определить, какой из них «ближе по значению» к третьему (то есть попытка сравнить семантические связи между ними и третьим термином), строго говоря, неправомерно. Семантическая связь (в изложенном выше понимании) характеризуется не одним, а несколькими параметрами, и свести эти параметры к одному (семантическому расстоянию) нельзя.

Однако необходимо учитывать, что наличие определенного вида связи между терминами иногда свидетельствует о наличии между этими же терминами некоторого другого вида связи. Например, для языка RX-кодов [26] наличие семантической связи между двумя терминами означает, что в их состав входит одинаковая характеристика. Если эта характеристика не является термином словаря, то вместо нее можно поставить соответствующий семантический код (т. е. развернуть сравниваемые термины). В результате у двух терминов окажутся одинаковыми еще две характеристики (другими словами, если совпадают вершины деревьев, то совпадают и подчиненные им поддеревья). Тем самым, между данной парой терминов окажется возможным установить по новым характеристикам еще некоторый вид (или виды) семантической связи (в частном случае по новым характеристикам устанавливается тот же вид связи).

Таким образом, количественно семантическое расстояние может быть определено как функция от числа семантических составляющих, присущих одновременно обоим словам. Считается, что чем больший удельный вес общих составляющих, тем меньше семантическое расстояние между словами. При полном совпадении семантических составляющих семантическое расстояние равно нулю (значения обоих слов тождественны — синонимия). Если, напротив, ни одна семантическая составляющая не совпадает, семантическое расстояние считается бесконечно большим (значения слов не имеют ничего общего).

Рассмотрим некоторые подходы, рассматривающие семантическое расстояние между терминами как функцию вычисления меры их близости.

Степень семантической связи. Основная трудность при определении семантического расстояния заключается в том, что отсутствуют точные методы как адекватного описания значений слов в терминах семантических составляющих, так и учета характера связей. Для языка RX-кодов [26] в пределах каждого отдельного вида семантической связи можно выделить различные степени этой связи. Степень семантической связи можно определить с помощью следующей формулы, в которой учитываются основные из перечисленных выше параметров:

$$D = \frac{\rho_i + \rho_0}{2},$$

где D — степень семантической связи;

ρ_i — удельный вес совпавшей характеристики в термине запроса;

ρ_0 — удельный вес совпавшей характеристики в термине объекта.

Удельный вес характеристики в термине определяется по формуле:

$$\rho = \frac{v}{\sum_{k=1}^m v_k},$$

где v — абсолютный вес характеристики в термине, $v=s/2^p$;

m — число конечных вершин в дереве термина (одной из которых является совпавшая вершина);

p — разряд характеристики;

s — категория характеристики (0 — общеупотребительное понятие, 2 — понятие смежной области, 3 — специальное понятие данной области).

Семантическое расстояние пары терминов, основанное на семантических множителях. Другой метод ([24]), учитывающий

семантическую значимость терминов, которые представлены в пространстве смысловых категорий (семантических составляющих), определяет семантическое расстояние пары терминов как

$$\delta = 1 - \frac{n_{IJ}}{n_I + n_J},$$

где n_{IJ} – число общих семантических множителей для терминов d_i, d_j ;

$(n_I + n_J)$ – общее число семантических множителей терминов d_i, d_j , а в качестве семантических множителей используются рубрики тезауруса, в которые входит термин.

Эта мера позволяет устанавливать силу семантической связи между словами, но не ее характер. Действительно, может оказаться, что семантическое расстояние между словами «стул» и «кресло» равно расстоянию между словами «стул» и «мебель». Однако ясно, что характер семантической связи между этими двумя парами слов различный: в первом случае слова выражают смежные понятия, во втором – вид и род.

Меры сходства. Рассмотренные выше меры семантической связи основаны на использовании характеристик, которые свойственны языкам со сложной грамматикой и практически отсутствуют в ИПЯ современных систем. При этом следует отметить первые реализации автоматического извлечения некоторой апостериорной семантической информации (установления ассоциативных связей между терминами) с помощью мер сходства (коэффициентов ассоциаций), широко используемых в методах автоматической классификации и кластерного анализа [36]. Наиболее известными мерами являются:

$$\Delta_1 = \frac{n_{IJ}}{n_{IJ} + n_{Jj} + n_{iJ}} \quad (\text{Jaccard});$$

$$\Delta_2 = \frac{n_{IJ} + n_{ij}}{d} \quad (\text{Soral});$$

$$\Delta_3 = \frac{n_{IJ}}{d} \quad (\text{Rao});$$

$$\Delta_4 = \frac{2n_{IJ}}{2n_{IJ} + 2(n_{Ij} + n_{iJ})} \quad (\text{Dice});$$

$$\Delta_5 = \frac{n_{IJ} + n_{ij}}{d + n_{iJ} + n_{Ij}} \quad (\text{Tanimoto});$$

где n_{IJ} – число совместных появлений d_i и d_j в документах массива;

n_{ij} – число совместных непоявлений d_i и d_j ;

n_{Ij} – число документов, содержащих d_i и не содержащих d_j ;

n_{iJ} – число документов, содержащих d_j и не содержащих d_i ;

$n_I = n_{IJ} + n_{Ij}$ – общее число появлений d_i ;

$d = |D|$ – размер словаря.

Дистрибутивно-статистический анализ. В основу дистрибутивно-статистического анализа [16; 25] положена гипотеза о том, что сила связи пары слов определяется величиной отклонения эмпирических данных о совместной их встречаемости от теоретических (математическое ожидание) в предположении о независимости появления этих слов в тексте. Хорошей иллюстрацией является работа [19], в которой метод дистрибутивно-статистического анализа используется при построении тезауруса для установления семантических связей, причем сила парадигматических связей между d_i и d_j (понимаемых как обобщение синтагматических) определяется как

$$P_{ij} = S_{ij} + \sum_k \lg(S_{ik} m_{ik} S_{jk} m_{jk}) ,$$

где P_{ij} – сила парадигматической связи между d_i и d_j ;

S_{ij} – сила синтагматической связи;

m_{ij} – математическое ожидание совместной встречаемости d_i и d_j .

Данный подход позволил успешно выделить устойчивые семантически регулярные словосочетания, однако используемая классификация выявленных связей достаточно условна. Парадигматические связи, понимаемые как обобщение синтагматических, в тексте обычно явно не присутствуют, а выделяются априорно в понятийной структуре языка при его разработке.

Другим показателем является *мера парадигматической близости терминов* В.Джулиано:

$$A_{kl} = \frac{n * \sum_{i=1}^D \frac{f_{ki} * f_{li}}{f_l}}{f_k * f_l},$$

где n – объем выборки (число документов),

f_{ij} – число документов, в которых вместе встретились i -й и j -й термины,

f_k – число документов, в которых встретился k -й термин.

Поскольку семантические связи внутри каждого конкретного текста определяются автором в значительной степени субъективно, то для уверенности в правильности установления ассоциативных отношений необходимо исследовать значительные массивы документов. Причем на оригинальных текстах дистрибутивно-статистический анализ позволяет установить лишь синтагматические отношения, поскольку слова, связанные парадигматическими отношениями, встречаются вместе крайне редко (например, в определениях). Поэтому в работах [21; 22] для определения меры близости этих слов предлагается использовать толковые и/или энциклопедические словари, что требует дополнительных затрат при автоматизации соответствующих процедур.

Несколько более детальный подход к выявлению устойчивых структурных связей с оцениванием связи пары терминов в цепочке, в иерархической и замещаемой структурах приведен в [6].

Несимметричный коэффициент подобия. Приведенные оценки меры семантической близости строятся только на данных о частотах употребления, никак не учитывая места этих терминов в понятийной структуре языка. Более того, почти все коэффициенты ассоциаций симметричны, что не всегда выполняется в языковых системах, по крайней мере, для родо-видовых отношений.

Предложенный в [28] несимметричный коэффициент подобия

$$C_{ij} = \frac{\sum_k \min(V_k^i, V_k^j)}{\sum_k V_k^i}$$

использует K -мерные вектора свойств V^i , V^j , представляющие информативность и специфичность терминов d_i и d_j соответственно с точки зрения удобств их использования при индексировании. Такой коэффициент, по всей видимости, лишенный указанных ранее недостатков, использует, однако, субъективно-задаваемые векторы свойств.

5.3. Методы классификации и кластеризации документов

Если рассмотренные выше методы выделения в тексте информативных терминов ассоциируются с задачами анализа, то методы классификации и кластеризации связываются с задачами синтеза – построения системы классов документов или терминов. Хотя необходимо отметить, что методы классификации широко используются на этапе предобработки для определения тематики документа и приписывания ему соответствующего шифра.

Приведем обзор методов, осуществляющих кластеризацию или классификацию документов.

Технология построения иерархической структуры папок.

Суть технологии заключается в автоматическом распределении множества объектов – текстовых документов по папкам. Папки имеют иерархическую структуру, что дает возможность всё более и более сужать область просмотра документов. Подсчёт коэффициентов близости для всего объёма документов выполняется по мере поступления новых документов в коллекцию.

Например, в поисковом сервере NorthernLight иерархия папок содержит около 20000 элементов. Однако, поскольку набор папок построен вручную и определен заранее, технология обладает следующими недостатками:

- иерархия папок имеет общий характер, поэтому не может быть использована, например, для поиска в специализированных базах данных;

- структуру папок необходимо периодически обновлять.

Методы факторного анализа. В основе методов лежат принципы выявления латентной структуры изучаемых явлений или объектов.

Пространство терминов рассматривается как пространство элементарных признаков (факторов). Документы, содержащие семантически близкие термины, группируются в определённых местах пространства терминов. Фактор – это некая семантическая сущность, которая может не иметь определённого названия. После определения числа и параметров факторов происходит отображение множества документов на пространство факторов.

Задачей факторного анализа является выделение главных факторов из пространства элементарных. В большинстве случаев задача нахождения главных факторов решается с помощью Метода Главных Компонентов [9]. Количество главных факторов выбирают существенно меньшим, чем количество элементарных, редуцируя таким образом пространство, на которое происходит отображение документов. Определение количества главных факторов и конфигурации их в пространстве являются основными проблемами факторного анализа.

Метод факторного анализа позволяет успешно преодолевать проблемы синонимии (когда одно и то же понятие может быть описано разными словами – синонимами) и омонимии (когда одно слово может иметь разные смыслы в разных контекстах), основываясь только на статистической информации о множестве документов/терминов.

Метод использует статистическую информацию о множестве документов/терминов и в общем случае не нуждается в предварительной настройке на специфический набор документов (не требует обучения). Однако некоторые реализации метода могут быть как обучаемыми, так и необучаемыми. Практически обучение заключается в фиксировании предопределённого количества центроидов кластеров (центроид кластера – вектор, который вычисляется как среднее арифметическое векторов всех документов кластера), а классификация заключается в отнесении документа к кластеру, расстояние до центроида которого наименьшее.

Основными недостатками метода можно назвать следующие:

- большое количество вычислений при реализации алгоритма;
- отсутствие подходящего наименования для вычисленных факторов;
- полученные кластеры не могут пересекаться.

Метод суффиксных деревьев. Суффиксные деревья изначально были разработаны для быстрого поиска подстрок. Суффиксное дерево – это дерево, содержащее все суффиксы рассматриваемой строки. Ветви дерева обозначаются буквами или буквосочетаниями, которые являются частями суффиксов строки. Суффикс, соответствующий определённой вершине, можно получить путем объединения всех букв, которые находятся на ребрах дерева, начиная от корневой вершины и заканчивая заданной.

Суффиксное дерево кластеризации строится из слов и фраз входных документов, т.е. единиц, находящейся на ребре дерева, является слово или словосочетание. Построение дерева осуществляется поэтапно. Сначала документы подвергаются предварительной обработке – отчистка от пунктуации, приведение слов в начальную форму (с помощью, например, алгоритмов морфологического анализа) и т.д. Затем для набора документов строится дерево. Каждой вершине такого дерева соответствует фраза. Ее можно получить, объединив все слова/словосочетания, находящиеся на ребрах на пути от корня дерева к данной вершине дерева. В тех вершинах дерева, которые имеют потомков, находятся ссылки на документы, в которых встречается фраза, соответствующая вершине. Множества документов, на которые указывают эти ссылки, образуют базовые кластеры.

К достоинствам метода можно отнести высокую скорость работы (по времени и занимаемой памяти дерево строится пропорционально количеству документов), наглядность представления результатов (общие фрагменты текстов и фраз выступают в качестве названия кластеров). Алгоритм не нуждается в обучении и задании порога срабатывания, а также допускает пересечение кластеров.

Недостатки метода:

- не используется уже имеющаяся информация о значения близости документов или весах слов/словосочетаний;

- не выявляется семантика документов, присутствующая не только на текстовом уровне;
- не решаются проблемы синонимии и омонимии.

Методы кластерного анализа. Общая суть этих методов заключается в выполнении следующих шагов.

1. Вычисление значений близости между документами и формирование матрицы близости.
2. Отнесение каждого документа в свой отдельный кластер.
3. Объединение в один кластер наиболее близких пар документов.
4. Обновление матрицы близости путем удаления колонок и строк для документов, которые были объединены, и перерасчет матрицы близости.
5. Переход на шаг 3 до тех пор, пока не выполнится критерий останова.

Достоинством методов является использование матрицы близости документов, а также тот факт, что алгоритмы методов не нуждаются в обучении;

Основные недостатки методов:

- необходимо задавать порог – максимальное количество документов в кластере;
- для получения хороших результатов кластеризации значения близости между парами документов должны быть упорядочены;
- кластеры не пересекаются.

Метод стабилизации центроидов кластеров. Основной характеристикой кластера является его центроид и вся работа алгоритма направлена на стабилизирование или, в лучшем случае, полное прекращение изменения центроида кластера. Алгоритм метода состоит из следующих шагов.

1. Выбираются начальные центроиды для множества документов. Существует несколько методик выбора центроидов. По одной из них, например, из множества документов случайным образом выбираются k документов, где k равно требуемому числу кластеров и на первом шаге эти документы становятся центроидами кластеров.

2. Все документы множества распределяются по кластерам. Документ попадает только в один кластер – тот, у которого метрика близости центроида и документа имеет наибольшее значение.

3. Пересчитываются центроиды кластеров, исходя из нового множества документов в каждом кластере. Если центроид кластера изменился, то цикл вычислений повторяется с пункта 2. Иначе, если центроид стабилизировался в некоторой окрестности или полностью, процесс кластеризации завершается.

Метод характеризуется линейной скоростью работы алгоритма, использует значения весовых коэффициентов терминов, не нуждается в обучении.

Недостатки метода:

- требует задания количества кластеров, как минимум на начальных этапах – до использования априорной информации;
- в том случае, когда центроиды кластеров выбираются случайным образом, результаты, получаемые над одной и той же выборкой документов, могут отличаться;
- кластеры не пересекаются.

Нейронная сеть Кохонена. Метод разработан для отображения многомерных данных на двумерную плоскость. Сети Кохонена называют также топологическими картами.

Для обучения сети алгоритм Кохонена использует информацию о предыдущем шаге обучения, поэтому нельзя чётко говорить о скорости обучения такой сети.

Параметры кластеров при кластеризации с помощью сети Кохонена формируются в результате специального алгоритма обучения сети, применяемого к множеству документов. Если сеть не может выделить объект в определённый кластер, то имеет место факт зарождения нового кластера данных.

Метод прост в обучении и обеспечивает наглядность представления результатов кластеризации. В качестве входных данных для сети используются расстояния между документами, т.е. значения матрицы близости. Полученные кластеры могут пересекаться.

Недостатки метода:

- необходимость обучения сети для каждого нового множества документов;

- процесс обучения требует фиксирования числа кластеров и набора обучающих данных.

Методы, основанные на гипертекстовых ссылках. Одним из наиболее часто используемых форматов для представления документов в Internet является HTML, который позволяет создавать документы со ссылками. Использование гиперссылок с одних Web-страниц на другие позволяет реализовать достаточно эффективную навигацию. Но, кроме этого, анализ взаимосвязей документов при индексировании их поисковыми машинами позволяет обеспечить относительно высокую точность поиска. Это достигается за счет использования алгоритма ранжирования документов в соответствии с индексом их «цитирования», который можно оценить числом ссылок, ведущих к данному документу из других.

Наиболее известным применением и развитием индекса цитирования является PageRank, который рекурсивно определяет важность Web-страницы на основе информации о страницах, ссылающихся на неё. Вероятностная модель подсчета ранга Web-страницы PageRank рассматривает следующий процесс: пользователь открывает случайную Web-страницу, с которой переходит по случайно выбранной гиперссылке на другую страницу, где он снова активизирует случайную гиперссылку и так далее. Иногда пользователь переходит на случайную Web-страницу не по ссылке, а набрав URL вручную. В этом случае вероятность того, что «скользящий» пользователь перейдет на некоторую определенную Web-страницу и будет ее рангом, который тем выше, чем большее число страниц ссылается на нее.

Алгоритм, учитывающий связь с запросом (локальная «популярность») – HITS (hyperlink induced topic search) выбирает подграф из гипертекстовой сети документов, отвечающих запросу. Из этого подграфа выделяются два вида узлов: «первоисточники» – страницы, на которые ведут ссылки с достаточно большого числа других страниц, и страницы-«посредники», которые содержат в основном ссылки на страницы, соответствующие запросу. Подмножество наиболее «популярных» страниц строится путем расширения множества найденных по запросу страниц за счет добавления страниц, связанных с ними через заданное число ссылок.

При этом для каждого документа вычисляется его значимость как первоисточника и как посредника.

Другим путем является категоризация – классификация найденных документов в соответствии с некоторой схемой предметных категорий. Соответственно, если пользователь указывает интересующую его категорию, документы могут быть ранжированы в контексте именно этого подмножества.

Однако эти модели не учитывают динамику развития WWW, так как в момент индексирования анализируется граф фиксированной структуры. Во-вторых, взвешивание и категоризация хорошо работают для сложившихся предметных областей (и, соответственно, с предопределенной системой категорий) и не учитывают индивидуальностей конкретного пользователя. Поэтому, например, пионерская статья или запрос, открывающие новое направление, скорее всего будут иметь низкий вес и не попадут в выдачу первых страниц.

5.4. Автоматическое реферирование и аннотирование

Программы автоматизированной обработки текстов, реализующие индексирование, реферирование и другие формы информационного анализа и синтеза вторичных документов, имеют историю, сопоставимую с историей развития вычислительных машин и автоматизированных систем управления. Именно на основании вторичных документов, существенно меньших по объему по отношению к исходному тексту, пользователи могут составить более или менее обоснованное представление о содержании первичного документа, затратив на это значительно меньше усилий в сравнении с его полным прочтением. Но существенно более полезной такая форма представления содержания оказывается для автоматизированного поиска. Содержание практически любого документа сводится к форме, в которой основные объекты, методы и результаты представляются компактными не избыточными выражениями, к тому же легко сводимыми к перечислению, в том числе нормализованных, терминов и отношений¹. Это, как отмечалось ранее, по-

¹ Характер преобразований хорошо представлен в Методике индексирования, соответствующий фрагмент которой представлен в Приложении 1.

зволяет, с одной стороны, реализовать эффективные алгоритмы поиска, а с другой – за счет избыточности выдачи дает уверенность пользователю в том, что он, обращаясь к системе с запросом из трех–четырёх слов, может достигать достаточной полноты выдачи даже в случае, когда информационная потребность представляет узкую тематику.

Задачи автоматического анализа и синтеза неоднословных текстов связаны с решением следующих проблем:

1. Предложения текста могут содержать элементы, значение которых определяется элементами других предложений.

2. Для определения общего смысла текста существенен не только смысл отдельных предложений, но и отношения между ними.

3. Текст имеет такие неотъемлемые характеристики, как тема и композиционная структура, что существенно при определении смысла не только отдельных фрагментов, но и всего текста в целом.

На сегодняшний день сложились два направления анализа первичных документов — квазиреферирование и синтез краткого содержания (квазианнотирование).

Квазиреферирование основано на экстрагировании фрагментов документов, т.е. выделении наиболее информативных фраз.

Квазианнотирование основывается на выделении из текстов с помощью методов искусственного интеллекта наиболее существенных семантических единиц и порождении нового текста, перечисляющего основные объекты (цели, средства, положения, использованные или разработанные методы, результаты и т.д.) первичного документа.

Алгоритмы квазиреферирования. Квазиреферирование сводится к извлечению из документов некоторых минимальных фрагментов текста, в максимальной степени связанных с основными положениями документа. Создание квазиреферата, таким образом, представляет собой соединение выбранных фрагментов.

Выделяют три главных направления, применяемые для квазиреферирования:

- индикаторные методы, основанные на оценке элементов

текста исходя из наличия в них специальных слов и словосочетаний – так называемых маркеров важности («в заключение», «было отмечено, что...» и пр.), характеризующих их смысловую значимость;

- позиционные методы, опирающиеся на предположение о том, что информативность элемента текста находится в зависимости от его позиции в документе;

- статистические методы, основанные на оценке информативности различных фрагментов текста по частотным показателям, таким, как расположение этого блока в оригинале, частота появления терминов в тексте, частота использования терминов в ключевых предложениях и т.д.

Позиционные методы опираются на предположение, что в разных структурных составляющих документа вероятность присутствия терминов, соответствующих основным составляющим смысла, различна. Согласно исследованиям [10] наиболее вероятно ключевые термины встречаются в Заглавиях (0,33), заголовках (0,22), первом предложении (0,18), последнем предложении (0,04).

Определение веса фрагментов исходного текста в статистических методах выполняется по алгоритмам, разработанным еще в 1960–70-е годы и ставшим уже традиционными. Общий вес текстового блока на этом этапе вычисляется по формуле:

$$W = L + K + S,$$

где L определяется расположением блока в исходном тексте и зависит от того, где появляется данный фрагмент — в начале, в середине или в конце, а также используется ли он в ключевых разделах текста, например, в заключении. K учитывает использование блока в «резюмирующих» конструкциях типа «в заключение», «в данной статье», «результатом является» и т.п.. Статистический вес текстового блока S вычисляется как нормированная по длине этого блока сумма весов входящих в него терминов – слов и словосочетаний. Блоки с наивысшими весовыми коэффициентами и будут включены в текст квазиреферата.

Алгоритмы квазианнотирования. Большинство алгоритмов квазианнотирования включает три основных этапа: анализ текста, определение весомых фрагментов (предложений или целых абзацев) и формирование вывода.

Первый этап начинается с выделения из исходного текста лексических единиц (слов или словосочетаний) и их взвешивания. Сначала выполняется выделение из исходного текста всех лексических единиц и построение из них словарного массива. При этом каждой лексической единице присваивается коэффициент, зависящий от ее расположения в исходном тексте. Затем выполняется их нормализация. Для выполнения последующего семантического анализа каждой словоформе ставятся в соответствие значения грамматических категорий (род, падеж, число). На этом этапе с помощью стоп-словаря выполняется удаление из словарного массива слов, не несущих явной смысловой нагрузки. Для этого каждой из лексических единиц присваивается весовой коэффициент, который определяется как результат учета нескольких составляющих: частоты появления, тематического словаря (определяемого, например, тематикой запроса пользователя) и «плюс-словаря», включающего наиболее важную лексику общего назначения.

Последний этап формирования массива лексических единиц заключается в выборе некоторого ограниченного количества самых весомых терминов. Полученный массив лексических единиц, кроме задачи автореферирования, в дальнейшем может быть использован и при различных лингвистических исследованиях текста.

Автореферирование на основе семантических методов. Подход, опирающийся на методы искусственного интеллекта, исходит из предположения, что если удастся определить семантику текста, то аннотация будет более качественной. Используемые при этом базы знаний должны постоянно поддерживаться в актуальном состоянии и сопровождаться экспертами. Для подготовки рефератов при таком подходе требуются мощные информационные ресурсы, обеспечивающих обработку текстов на естественных языках, в том числе базы грамматических правил и словари для синтаксического разбора естественно-языковых конструкций. Для реализации этого метода нужны многочисленные необходимые для анализа и определения наиболее важной информации справочники, которые отражают понятия, ориентированные на предметные области.

В отличие от частотно-лингвистических методов, обеспечивающих квазиреферирование, подход, основанный на базах знаний,

опирается на автоматизированный качественный контент-анализ, состоящий, как правило, из трех основных этапов. Первый – сведение исходной текстовой информации к заданному числу фрагментов – смысловых единиц. На втором этапе производится поиск регулярных связей между ними, после чего наступает третий этап – формирование выводов и обобщений. Такая структурная аннотация представляет содержание текста в виде совокупности концептуально связанных единиц значения.

Семантические методы формирования рефератов-изложений предполагают два основных подхода: метод синтаксического разбора предложений и методы, опирающиеся на синтез семантической структуры текста.

В первом случае используются деревья разбора текста. Вторым подход основывается на системах искусственного интеллекта, в которых на этапе анализа также выполняется синтаксический разбор текста, но синтаксические деревья не порождаются. В этом случае формируются семантические структуры, которые накапливаются в базе знаний и затем объединяются путем слияния графов. В результате преобразования формируется концептуальная структура текста – аннотация, т.е. семантически значимые «выжимки» из исходного текста.

Контрольные вопросы к разделу

1. Охарактеризуйте сходства и отличия перечислительных и аналитико-синтетических классификаций.
2. Охарактеризуйте свойства и приведите примеры перечислительных классификаций.
3. Охарактеризуйте свойства и приведите примеры аналитико-синтетических классификаций.
4. Охарактеризуйте систему типовых делений ББК.
5. Перечислите основные фасеты УДК.
6. Приведите примеры семантических связей в УДК.
7. Приведите примеры отраслевых классификационных систем.
8. Приведите примеры средств выражения семантики в МПК.
9. Перечислите уровни деления МПК.

10. Приведите типологию терминологических структур.
11. Приведите примеры использования линейных терминологических структур при индексировании и поиске.
12. Приведите примеры иерархических терминологических структур.
13. Дайте определение понятия «тезаурус».
14. Охарактеризуйте назначение и структуру информационно-поискового тезауруса.
15. Приведите примеры тезаурусов.
16. Охарактеризуйте типологию отношений в тезаурусе.
17. Охарактеризуйте типологию отношений в онтологии.
18. Перечислите и охарактеризуйте основные методики вычисления веса термина.
19. Приведите примеры вычисления веса термина в соответствии с TL*TF-методикой.
20. Охарактеризуйте типы методов вычисления семантического расстояния между терминами.
21. Назовите основное различие задач классификации и кластеризации.
22. Приведите примеры методов классификации.
23. Приведите примеры методов кластеризации.
24. Охарактеризуйте методы автоматического реферирования и аннотирования.

РАЗДЕЛ 3. РАСПРЕДЕЛЕННАЯ ОБРАБОТКА ИНФОРМАЦИИ

Повышение эффективности и оперативности обслуживания большого числа клиентских запросов, помимо простого увеличения ресурсов и вычислительной мощности машины, может быть достигнуто двумя путями:

- снижением суммарного расхода памяти и вычислительных ресурсов за счет буферизации (кэширования) и совместного использования (разделяемые ресурсы) наиболее часто запрашиваемых данных и процедур;
- распараллеливанием процесса обработки – использованием разных процессоров для параллельной обработки изолированных подзапросов и/или для одновременного обращения к частям базы данных, размещенным на отдельных физических носителях.

Исходя из введенного в первом разделе положения, что информация – это данные и контекст их обработки, будем рассматривать архитектуры, реализующие совместную обработку информации, выделяя вопросы, относящиеся к собственно обработке данных, и вопросы, связанные с контекстом – функциональностью обработки метainформации.

ГЛАВА 6. РАСПРЕДЕЛЕННЫЕ ФАЙЛОВЫЕ СИСТЕМЫ

Ключевым компонентом любой распределенной обработки информации является файловая система, обеспечивающая хранение программ и данных, а так же предоставление доступа к ним по мере необходимости. Файловая система поддерживается одной или более машинами, называемыми файл-серверами. Файл-серверы перехватывают запросы на чтение или запись файлов, поступающие от других машин (не серверов), которые называются клиентами. Каждый посланный запрос проверяется и выполняется, а ответ отсылается обратно. Файл-серверы обычно содержат иерархические файловые системы, каждая из которых имеет корневой каталог и каталоги более низких уровней. Рабочая станция может под-

соединять и монтировать эти файловые системы к своим локальным файловым системам. При этом монтируемые файловые системы остаются на серверах.

Важно понимать различие между файловым сервисом и файловым сервером. *Файловый сервис* – это описание функций, которые файловая система предлагает своим пользователям. Это описание включает имеющиеся примитивы, их параметры и функции, которые они выполняют. С точки зрения пользователей файловый сервис определяет то, с чем пользователи могут работать, но ничего не говорит о том, как все это реализовано. В сущности, файловый сервис определяет интерфейс файловой системы с клиентами.

Файловый сервер – это процесс, который выполняется на отдельной машине и реализует файловый сервис. В системе может быть один файловый сервер или несколько, но в хорошо организованной распределенной системе пользователи не знают, как реализована файловая система, они не знают количество файловых серверов, их месторасположение и функции. Они только знают, что если процедура определена в файловом сервисе, то требуемая работа выполняется, и им возвращаются требуемые результаты. Более того, пользователи даже не должны знать, что файловый сервис является распределенным. В идеале он должен выглядеть так же, как и в централизованной файловой системе.

Так как обычно файловый сервер – это просто пользовательский процесс (или иногда процесс ядра), выполняющийся на некоторой машине, в системе может быть несколько файловых серверов, каждый из которых предлагает различный файловый сервис. Например, в распределенной системе могут быть два сервера, и любой пользовательский процесс пользуется подходящим сервисом.

Файловый сервис в распределенных файловых системах (впрочем, как и в централизованных) имеет две функционально различные части: собственно файловый сервис и сервис каталогов. Первый имеет дело с операциями над отдельными файлами, такими, как чтение, запись или добавление, а второй – с созданием каталогов и управлением ими, добавлением и удалением файлов из каталогов и т.п.

Существенным фактором, определяющим архитектуру и функциональность файловой системы, является способ распреде-

ления серверной и клиентской частей между машинами. В некоторых системах (например, NFS) нет разницы между клиентом и сервером, на всех машинах работает одно и то же базовое программное обеспечение, так что любая машина, которая хочет предложить файловый сервис, свободно может это сделать. Для этого ей достаточно экспортировать имена выбранных каталогов, чтобы другие машины могли иметь к ним доступ.

В других системах файловый сервер – это только пользовательская программа, так что система может быть сконфигурирована как клиент, как сервер или как клиент и сервер одновременно. Третьим, крайним случаем, является система, в которой клиенты и серверы – это принципиально различные машины, как в терминах аппаратуры, так и в терминах программного обеспечения. Серверы могут даже работать под управлением другой операционной системы.

Вторым важным вопросом реализации файловой системы является структуризация сервиса файлов и каталогов. Один подход заключается в комбинировании этих двух сервисов на одном сервере. При другом подходе эти сервисы разделяются. В последнем случае при открытии файла требуется обращение к серверу каталогов, который отображает символьное имя в двоичное, а затем обращение к файловому серверу с двоичным именем для действительного чтения или записи файла.

Аргументом в пользу разделения сервисов является тот факт, что они на самом деле слабо связаны, поэтому их раздельная реализация более гибкая. Например, можно реализовать сервер каталогов MS-DOS и сервер каталогов UNIX, которые будут использовать один и тот же файловый сервер для физического хранения файлов. Разделение этих функций также упрощает программное обеспечение. Недостатком является то, что использование двух серверов увеличивает интенсивность сетевого обмена.

Постоянный поиск имен, особенно при использовании нескольких серверов каталогов, может приводить к большим накладным расходам. В некоторых системах делается попытка улучшить производительность за счет кэширования имен. При открытии файла кэш проверяется на наличие в нем нужного имени. Если оно там есть, то этап поиска, выполняемый сервером каталогов, пропускается, и двоичный адрес извлекается из кэша.

Последний рассматриваемый здесь структурный вопрос связан с хранением на серверах информации о состоянии клиентов. Существует две конкурирующие точки зрения.

Первая состоит в том, что сервер не должен хранить такую информацию (stateless-сервер). Другими словами, когда клиент посылает запрос на сервер, сервер его выполняет, отправляет ответ, а затем удаляет из своих внутренних таблиц всю информацию о запросе. Между запросами на сервере не хранится никакой текущей информации о состоянии клиента. Другая точка зрения состоит в том, что сервер должен хранить такую информацию (statefull-сервер).

Рассмотрим эту проблему на примере файлового сервера, имеющего команды ОТКРЫТЬ, ПРОЧИТАТЬ, ЗАПИСАТЬ и ЗАКРЫТЬ файл. Открывая файлы, statefull-сервер должен запоминать, какие файлы открыл каждый пользователь. Обычно при открытии файла пользователю дается дескриптор файла или другое число, которое используется при последующих вызовах для его идентификации. При поступлении вызова, сервер использует дескриптор файла для определения, какой файл нужен. Таблица, отображающая дескрипторы файлов на сами файлы, является информацией о состоянии клиентов.

Для сервера типа stateless каждый запрос должен содержать исчерпывающую информацию (полное имя файла, смещение в файле и т.п.), необходимую серверу для выполнения требуемой операции. Очевидно, что эта информация увеличивает длину сообщения.

Однако при отказе statefull-сервера теряются все его таблицы, и после перезагрузки не известно, какие файлы открыл каждый пользователь. Последовательные попытки провести операции чтения или записи с открытыми файлами будут безуспешными. Stateless-серверы в этом плане являются более отказоустойчивыми, и это аргумент в их пользу.

6.1. Файловый сервис

Для любой файловой системы, независимо от того распределенная она или централизованная, самым главным является вопрос, что такое файл? Во многих ОС, пользователи представляют файл

как последовательность записей. Каждая запись – это байтовый агрегат постоянного или переменного размера. Записи можно читать или записывать последовательно или позиционировать файл на запись с указанным номером.

Большинство современных распределенных файловых систем поддерживает определение файла как последовательности байтов. Из файла можно прочесть указанное число байтов, либо начиная с его начала, либо предварительно произведя его позиционирование на байт с указанным номером. Аналогично можно записать указанное число байтов в конец файла, либо предварительно выполнив позиционирование файла.

Важным аспектом файловой модели является возможность модификации файла после его создания. Обычно файлы могут модифицироваться, но в некоторых распределенных системах единственными операциями с файлами являются *создать* и *прочитать*. Такие файлы называются неизменяемыми. Для неизменяемых файлов намного легче осуществить кэширование файла и его репликацию (тиражирование), так как исключаются все проблемы, связанные с обновлением всех копий файла при его изменении.

Файловый сервис может быть разделен на два типа в зависимости от того, поддерживает ли он модель загрузки-выгрузки или модель удаленного доступа. В модели загрузки-выгрузки пользователю предлагаются средства чтения или записи файла целиком. Эта модель предполагает следующую схему обработки файла: чтение файла с сервера на машину клиента, обработка файла на машине клиента и запись обновленного файла на сервер. Преимуществом этой модели является ее концептуальная простота. Кроме того, передача файла целиком очень эффективна. Главным недостатком этой модели являются высокие требования к дискам клиентов. Кроме того, неэффективно перемещать весь файл, если нужна его маленькая часть.

Другой тип файлового сервиса соответствует модели удаленного доступа, которая предполагает поддержку большого количества операций над файлами: открытие и закрытие файлов, чтение и запись частей файла, позиционирование в файле, проверка и изменение атрибутов файла и так далее. В то время как в модели загрузки-выгрузки файловый сервер обеспечивал только хранение и перемещение файлов, в данном случае вся файловая система вы-

полняется на серверах, а не на клиентских машинах. Преимуществом такого подхода являются низкие требования к дисковому пространству на клиентских машинах, а также исключение необходимости передачи целого файла, когда нужна только его часть.

6.2. Сервис каталогов

Природа сервиса каталогов не зависит от типа используемой модели файлового сервиса. В распределенных системах используются те же принципы организации каталогов, что и в централизованных, в том числе многоуровневая организация каталогов.

Многоуровневое именование файлов обеспечивается за счет поддержания во внешней памяти дополнительных файлов со специальной структурой – каталогов. Каждый каталог содержит имена каталогов и/или файлов, содержащихся в данном каталоге. Таким образом, полное имя файла состоит из списка имен каталогов плюс имя файла в каталоге, непосредственно указывающем на данный файл. Разница между способами именования файлов в разных файловых системах состоит в том, с чего начинается эта цепочка имен.

Известны два базовых варианта:

- поддержание изолированных файловых систем. Во многих системах управления файлами требуется, чтобы каждый архив файлов (полное дерево справочников) целиком располагался на одном дисковом пакете (или логическом диске, разделе физического дискового пакета, представляемом с помощью средств операционной системы как отдельный диск). В этом случае полное имя файла начинается с имени дискового устройства, на котором установлен соответствующий диск. Такой способ именования используется в файловых системах фирмы DEC, очень близко к этому находятся и файловые системы персональных компьютеров;

- полностью централизованная файловая система. Впервые была реализована в операционной системе Multics. В файловой системе Multics пользователи представляли всю совокупность каталогов и файлов как единое дерево. Полное имя файла начиналось с имени корневого каталога, и пользователь не обязан был заботиться об установке на дисковое устройство каких-либо конкретных дисков. Сама система, выполняя поиск файла по его имени, запрашивала оператора об установке необходимых дисков.

Централизованные файловые системы во многом удобнее изолированных: система управления файлами принимает на себя больше рутинной работы. Но в таких системах возникают существенные проблемы, если, например, кому-то требуется перенести поддерево файловой системы на другую вычислительную установку.

Компромиссное решение применено в файловых системах ОС UNIX. На базовом уровне в этих файловых системах поддерживаются изолированные архивы файлов. Один из этих архивов объявляется корневой файловой системой. После запуска системы можно смонтировать корневую файловую систему и ряд изолированных файловых систем в одну общую файловую систему. Технически это производится с помощью создания в корневой файловой системе специальных пустых каталогов. Специальная команда (системный вызов) **mount** ОС позволяет подключить к одному из этих пустых каталогов корневой каталог указанного архива файлов. После монтирования общей файловой системы именование файлов производится так же, как если бы она с самого начала была централизованной.

Большинство распределенных систем использует какую-либо форму двухуровневого именования: на одном уровне файлы имеют символические имена, такие как `text.doc`, предназначенные для использования людьми, а на другом – внутренние, двоичные имена, для использования самой системой. Каталоги обеспечивают отображение между двумя этими уровнями имен. Отличием распределенных систем от централизованных является возможность соответствия одному символическому имени нескольких двоичных имен. Обычно это используется для представления оригинального файла и его архивных копий. Имея несколько двоичных имен, можно при недоступности одной из копий файла получить доступ к другой. Этот метод обеспечивает отказоустойчивость за счет избыточности.

6.3. Семантика разделения файлов

Когда два или более пользователей разделяют один файл, необходимо точно определить семантику чтения и записи, чтобы избежать проблем. В централизованных системах, разрешающих раз-

деление файлов, таких как UNIX, обычно определяется, что, когда операция ЧТЕНИЕ следует за операцией ЗАПИСЬ, то читается только что обновленный файл. Аналогично, когда операция чтения следует за двумя операциями записи, то читается файл, измененный последней операцией записи. Тем самым система придерживается абсолютного временного упорядочивания всех операций, и всегда возвращает самое последнее значение. Будем называть эту модель семантикой UNIX'a. В централизованной системе (и даже на мультипроцессоре с разделяемой памятью) ее легко и понять, и реализовать.

Семантика UNIX может быть обеспечена и в распределенных системах, но только, если в ней имеется лишь один файловый сервер, и клиенты не кэшируют файлы. Для этого все операции чтения и записи направляются на файловый сервер, который обрабатывает их строго последовательно. На практике, однако, производительность распределенной системы, в которой все запросы к файлам идут на один сервер, часто становится неудовлетворительной. Эту проблему иногда устраняют путем разрешения клиентам обрабатывать локальные копии часто используемых файлов в своих личных кэшах. Если клиент сделает локальную копию файла в своем локальном кэше и начнет ее модифицировать, а вскоре после этого другой клиент прочитает этот файл с сервера, то он получит неверную копию файла. Одним из способов устранения этого недостатка является немедленный возврат всех изменений в кэшированном файле на сервер. Такой подход хотя и концептуально прост, но не эффективен.

Другим решением является введение так называемой сессионной семантики, в соответствии с которой изменения в открытом файле сначала видны только процессу, который модифицирует файл, и только после закрытия файла эти изменения могут видеть другие процессы. При использовании сессионной семантики возникает проблема одновременного использования одного и того же файла двумя или более клиентами. Одним из решений этой проблемы является принятие правила, в соответствии с которым окончательным является тот вариант, который был закрыт последним. Менее эффективным, но гораздо более простым в реализации, является вариант, при котором окончательным результирующим файлом на сервере может оказаться любой из этих файлов.

Следующий подход к разделению файлов заключается в том, чтобы сделать все файлы неизменяемыми. Тогда файл нельзя открыть для записи, а можно выполнять только операции СОЗДАТЬ и ЧИТАТЬ. Тогда для изменения файла остается только возможность создать полностью новый файл и поместить его в каталог под именем старого файла. Следовательно, хотя файл и нельзя модифицировать, его можно заменить (автоматически) новым файлом. Другими словами, хотя файлы и нельзя обновлять, но каталоги обновлять можно. Таким образом, проблема, связанная с одновременным использованием файла, просто исчезнет.

Четвертый способ работы с разделяемыми файлами в распределенных системах – это использование механизма неделимых транзакций, которые представляют собой неделимую последовательность действий над файлами (чтения, удаления, вставки, модификации) такую, что:

- либо результаты всех операторов, входящих в транзакцию, отображаются в файловой системе;
- либо воздействие всех этих операторов полностью отсутствует.

Для достижения свойства неделимости транзакций в распределенных системах используется специальный протокол, называемый протоколом двухфазной фиксации транзакций. Хотя он и не является единственным протоколом такого рода, но он наиболее широко используется.

Суть этого протокола состоит в следующем. Один из процессов выполняет функции координатора (см. рис. 6.1). Координатор начинает транзакцию, делая запись об этом в своем журнале регистрации, затем он посылает всем подчиненным процессам, также выполняющим эту транзакцию, сообщение «подготовиться к фиксации». Когда подчиненные процессы получают это сообщение, то они проверяют, готовы ли они к фиксации, делают запись в своем журнале и посылают координатору сообщение-ответ «готов к фиксации». После этого подчиненные процессы остаются в состоянии готовности и ждут от координатора команду фиксации. Если хотя бы один из подчиненных процессов не откликнулся, то координатор «откатывает» подчиненные транзакции, включая и те, которые подготовились к фиксации.

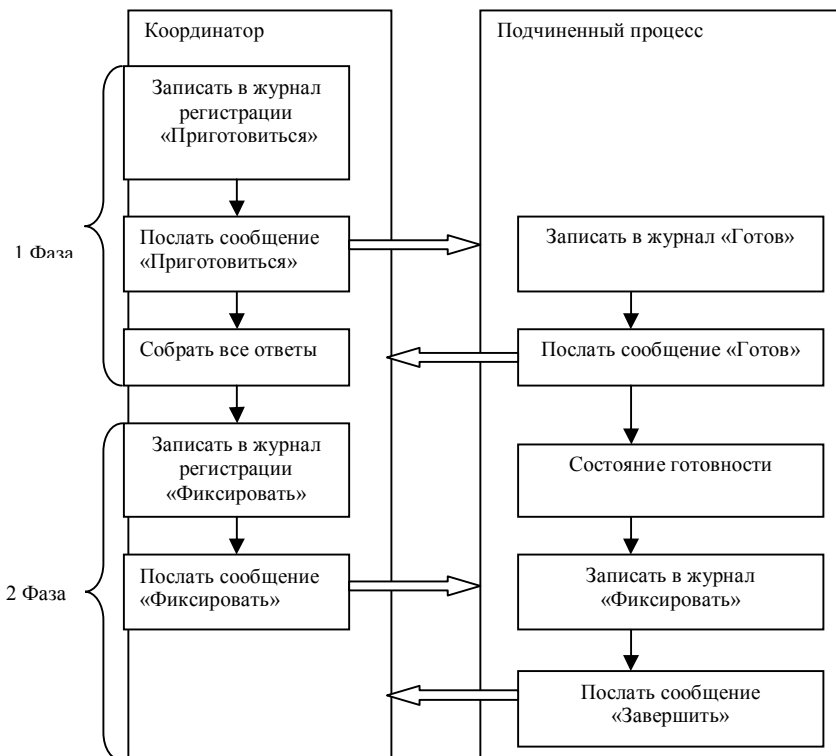


Рис. 6.1. Двухфазный протокол фиксации транзакции

Выполнение второй фазы заключается в том, что координатор посылает команду «фиксировать» (*commit*) всем подчиненным процессам. Выполняя эту команду, последние фиксируют изменения и завершают подчиненные транзакции. В результате гарантируется одновременное синхронное завершение (удачное или неудачное) распределенной транзакции.

ГЛАВА 7. РАСПРЕДЕЛЕННЫЕ ВЫЧИСЛЕНИЯ

Под распределенными вычислениями понимают такие, которые производятся на оборудовании, состоящем из более, чем одного процессора или блока оперативной памяти. Кроме того, к распределенным вычислениям относят параллельную обработку данных, а также одновременное выполнение слабосвязанных между собой программ.

Обычно при распределенных вычислениях программы разбиваются на части, которые исполняются на множестве компьютеров, объединенных сетью.

Распределенные вычисления – это вид параллельных вычислений. Однако, следует заметить, что в параллельных вычислениях обычно используются разделенные на части программы, которые обрабатываются на множестве процессоров одного компьютера.

Программы, разработанные для распределенных вычислений, отличаются от программ для параллельных вычислений тем, что их части между собой связаны слабее, предназначены для вычислений на различном оборудовании и приспособлены к межсетевому обмену данными. Кроме того, они должны быть устойчивы к проблемам связи и сбоям вычислительной техники.

При разработке программного обеспечения распределенных вычислений чрезвычайно важно учитывать следующие факторы: возможность организации межпрограммного взаимодействия через разнородные компьютерные сети и возможность транспортировки самих программ по сети для их исполнения на удаленных компьютерах и взаимодействия между собой.

Главной целью распределенных вычислений является предоставление пользователям имеющихся в наличии распределенных вычислительных мощностей. При этом для пользователей использование этих мощностей должно быть прозрачным, открытым и легко масштабируемым.

При этом под открытостью понимается способность вычислительной системы взаимодействовать с другой системой в стиле и по стандартам Открытых Систем. Часто под такими стандартами понимают стандарты Web Services, хотя это и сужает понятие открытости.

Прозрачность подразумевает свободное перемещение кода и его исполнение на любой системе, а также взаимодействие с любой другой системой, исполняющей другие части кода.

Масштабируемость подразумевает возможность использования любых дополнительных ресурсов по мере надобности в процессе вычислений.

К основным проблемам распределенных вычислений следует отнести: надежность, сложность диагностики отказов, невозможность распараллеливания многих вычислительных задач.

7.1. Архитектура распределенных вычислений

Для организации систем распределенных вычислений используются самые разные модели межпрограммного обмена данными и транспортировки программного кода. Остановимся на нескольких базовых моделях.

Модель «клиент-сервер». В этой модели клиент отправляет запросы на сервер, который на эти запросы отвечает. Взаимодействие инициирует клиент. Сервер постоянно находится в ожидании возможных запросов. И клиент, и сервер достаточно сложные программы, которые взаимодействуют друг с другом посредством протокола.

Модель с посредником между клиентом и сервером. В этой модели основная нагрузка по взаимодействию с сервером ложится на посредника, который собирает запросы от «легких» клиентов. В протоколе взаимодействия клиента и посредника, как правило, используется множество упрощений, по сравнению со схемой «клиент-сервер». Это призвано существенно упростить сложность разработки приложений-клиентов.

Модель с серверами приложений. В этой модели между клиентом и системой управления информационными ресурсами, которая выполняет функции универсального сервера, появляется цепочка серверов-посредников, которые реализуют различные функции по агрегированию запросов и ответов от групп клиентов и различных серверных подсистем. Такая модель характерна для

больших корпоративных информационных систем и систем управления.

Модель кластеров. Кластерная модель применяется при организации параллельных вычислений на группах компьютеров работающих совместно, которые называются кластерами. Такие комплексы имеют специальную архитектуру, оптимизированную для решения задач, допускающих высокую степень параллелизма отдельных операций.

Модель «точка-точка». Данная модель предполагает отсутствие некоего единого управляющего центра и единого центрального информационного ресурса. Все компьютеры в такой модели равноправны и могут выступать как в качестве клиентов, так и в качестве серверов.

Модель единого пространства. В этой модели вся вычислительная архитектура предоставляется конечному потребителю в виде единого вычислительного пространства. Все операции по распараллеливанию вычислений, использованию распределенного адресного пространства, обмену данными между частями программы скрыты от пользователя. Ресурсы наращиваются по мере необходимости в стиле, как если бы для вычислений использовался только один компьютер.

7.2. Виды параллелизма в распределенных вычислениях

Несмотря на развитие в последнее время различных способов виртуализации представления о реальной инфраструктуре вычислений, следует признать, что во многом распределенные вычисления ориентируются на архитектуру современных вычислительных комплексов. С другой стороны, и сами эти комплексы долгое время создавались под определенные вычислительные задачи. В настоящее время выделяют несколько системных архитектур, ориентированных на распределенные вычисления.

Многопроцессорные системы. Как правило, это компьютеры, которые имеют более одного процессора. Операционная систе-

ма при этом способна распределять задачи или потоки (нити) задач по различным процессорам, добиваясь одновременного использования всех процессоров.

Многоядерные системы. Данные архитектуры специально предназначены для исполнения легковесных процессов или потоков, на которые разбивается задача при ее исполнении. При этом предполагается, что различные задачи могут использовать одни и те же потоки, но с различными данными.

Мультикомпьютерные системы. Данная архитектура предполагает согласованное использование нескольких компьютеров для решения одной задачи или комплекса взаимосвязанных задач. При этом связность компонентов задачи гораздо ниже, нежели при решении задач с использованием кластеров. Как правило, мультикомпьютерные архитектуры применяются в случаях, когда существуют ограничения по энергопотреблению и площади размещения комплекса. Типовыми задачами являются задачи определения местоположения и медицинская 3D-диагностика.

Вычислительные кластеры. Кластеры – это многомашинные вычислительные комплексы, компьютеры которых объединены локальными высокоскоростными линиями связи. Довольно часто кластерные вычисления отличаются от прочих распределенных вычислений. Кластеры решают групповые задачи, требующие высокой согласованности вычислений, что, в свою очередь, требует компактного совместного размещения компьютеров, объединенных в кластер. При «традиционных» распределенных вычислениях таких тесных связей между компьютерами не требуется. Для решения задач распределенных вычислений часто применяются компьютеры, размещенные на площадках существенно удаленных друг от друга.

GRID-вычисления (GRID-Computing). Один из видов распределенных вычислений, когда принято оперировать понятием виртуального суперкомпьютера, который, на самом деле состоит из множества слабосвязанных вычислительных машин. При этом виде вычислений, как правило, используются вычислительные ресурсы,

которые в данный момент времени простаивают. Это могут быть суперкомпьютеры или даже множество персональных компьютеров. Обычно в качестве связи используют Internet.

Как правило, речь идет о решении очень больших задач, для которых не может быть применен ни один из суперкомпьютеров. Чаще всего это научные задачи, например, расшифровка генома человека или обработка результатов экспериментов, проведенных на супер-коллайдере. Для данного вида вычислений характерна слабая связь между крупными блоками задач, что позволяет использовать для решения этих задач вычислительные ресурсы, связанные между собой через Internet.

Grid-архитектуру от кластерной отличает слабая связность компонентов системы, разнородность оборудования и системного программного обеспечения, а также географическая удаленность компьютеров друг от друга. Grid-вычисления принято рассматривать как самостоятельный вид вычислений, а не только как под вид распределенных вычислений.

В настоящее время существует несколько определений Grid-вычислений. Например, в CERN определяют GRID как сервис для совместного использования вычислительных мощностей и памяти для хранения данных с использованием возможностей Internet.

Довольно точно суть Grid определили в IBM: «Это возможность, используя набор стандартов и протоколов, предоставить доступ к приложениям и данным, вычислительным мощностям, памяти и множеству других вычислительных ресурсов через Internet».

Основные преимущества Grid-вычислений, которые называют специалисты, поддерживающие эту идею, – это возможность за низкую стоимость получить высокопроизводительную масштабируемую систему.

Однако существуют и значительные ограничения по типам задач, которые могут быть решены на основе Grid-архитектур. Во-первых, из-за разнородности компьютеров, которые объединяются в Grid, разработка программного обеспечения превращается в сложную и нетривиальную задачу. Стоимость такой программы может существенно превышать стоимость обычного ПО. Во-вторых, слабая связность компьютеров и ненадежность коммуникаций заставляет уделять значительное внимание протоколам об-

мена информацией и согласованию расчетов в ненадежной вычислительной среде, что также повышает стоимость ПО. В-третьих, существует проблема стоимости каналов обмена информацией и самих вычислительных мощностей. В-четвертых, разрешение исполнения «чужих» программ на своих вычислительных мощностях чревато проблемами с безопасностью.

Internet-вычисления (Cloud computing). Этот тип вычислений предполагает использование Internet-технологий для организации распределенных вычислений. В основе Internet-вычислений лежит концепция «программное обеспечение как сервис». На основе этой модели развиваются многие современные технологии, такие, например, как Web 2.0.

Internet-вычисления – это, в некотором смысле, развитие идеи Grid-вычислений. Во-первых, в Internet-вычислениях была успешно решена проблема биллинга. Это значит, что была найдена жизнеспособная коммерческая схема использования распределенных вычислительных мощностей. Во-вторых, во многом благодаря использованию простаивающих мощностей Internet-компаний, например, Google, Amazon, Microsoft, удалось создать достаточно надежную распределенную децентрализованную инфраструктуру вычислений. В-третьих, Internet-вычисления опираются не на набор библиотек для сбора программ, которые могли бы исполняться на разнородной архитектуре, а на программные интерфейсы приложений (API), которые позволяют представлять вычислительную среду в качестве виртуального компьютера и разрабатывать ПО для этого компьютера, не заботясь о конкретной архитектуре вычислительных систем, задействованных в Internet-вычислениях.

При Internet-вычислениях пользователь оперирует понятиями аренды вычислительных мощностей, памяти и каналов обмена данными. Тарифицируется время счета, объем хранения данных и объем передаваемых по сети данных. При этом пользователь платит только за те объемы, которые реально использует, а не за возможность использования ресурсов.

Вычисления-услуги. Еще один вид распределенных вычислений – это пакетные вычисления (Utility computing) или вычисления-услуги. В некотором смысле это подмножество Internet-

вычислений. Данный тип вычислений строится по принципу общественных услуг (водоснабжение, электроснабжение, телефонные сети и т.п.).

Фактически речь идет об аренде вычислительных мощностей. Сама аренда вычислительных мощностей тема не новая. Арендовали компьютеры еще в эпоху IBM-360/370. В 1990-х в Internet появилась услуга хостинга, т.е. аренды вычислительных мощностей отдельных дата-центров. При хостинге можно арендовать часть вычислительных мощностей одного компьютера (виртуальный хостинг), компьютеры (dedication) или даже группы компьютеров или дисковые массивы.

7.3. Примеры проектов распределенных вычислений

Наиболее модным и одновременно «испытанным» с точки зрения биллинга направлением распределенных вычислений принято считать аренду вычислительных мощностей. Наиболее яркими коммерческими проектами данного направления считаются: сервисы Amazon (EC2), Google (Google Apps), Microsoft (S+S). Остановимся на них.

Проекты Amazon. Amazon предлагает несколько сервисов, которые одновременно можно отнести как к Internet-вычислениям, так и к вычислениям-услугам.

Во-первых, это аренда мест хранения (Amazon Simple Storage Service). Данная услуга предоставляется в виде API. Пользователь может сам написать программу, использующую данный интерфейс и использовать распределенную систему хранения как внешнее устройство хранения. Распределенная система представляется пользователю в виде, похожем на обычный внешний носитель. Плата берется за объем, который пользователь реально использует для хранения своих данных, а также за объем данных переданных/полученных при обращении к системе хранения.

Во-вторых, это аренда места в базе данных (Amazon Simple DB). В данном случае API предоставляет доступ пользователю к своим данным в виде SQL-запросов. SQL в данном случае довольно урезанный, если сопоставлять его со стандартом языка. Главным образом, это касается вопросов управления транзакциями. Тем не

менее, создавать таблицы, загружать в них данные, осуществлять довольно сложный поиск в этой системе можно. Как и в случае с распределенным хранилищем, плата при использовании виртуальной СУБД берется за объем хранения и за объемы данных передаваемых по сети.

В-третьих, это, собственно вычисления (Elastic Compute Cloud). В данном случае речь идет об API, позволяющем использовать распределенную вычислительную среду, а также обмениваться данными между программами, исполняющимися в данной среде. Оплачивают здесь время использования процессоров и объемы передаваемой между ними информации.

Собственно, у Amazon существует и отдельный сервис, реализованный в стиле Unix-конвейеров, который позволяет создавать распределенные системы, компоненты, которых могут использовать данный сервис для организации обмена данными. Канал работает по принципу стека FIFO. Причем, данные в стеке хранятся в течение недели.

На основе упомянутых выше API существует множество сервисов, которые ориентированы на конечного пользователя. Так, можно установить на персональном компьютере под управлением Windows ПО, которое будет изображать систему хранения Amazon в виде сетевой папки неограниченного размера. Другой пример использования – система электронной почты «a la Yandex» полностью реализованная на основе сервисов Amazon. Если кому-то вздумается организовать такой сервис, то не нужно ничего писать и даже покупать серверы. Можно просто завести эккаунт Amazon и начать регистрировать клиентов.

В качестве основы для пользовательского интерфейса в Amazon используется Xen virtualization software и различные клоны Unix.

В данном месте мы вплотную подошли к концепции «Software as a Service» (SaaS). Наиболее активно ее продвигают Google и Microsoft.

Сервисы Google. Google предоставляет два вида API для разработки распределенных систем вычислений. Во-первых, это API доступа к приложениям Google (Google Apps). В данном случае речь идет об очень широком наборе программ, которыми может

воспользоваться пользователь в своей повседневной работе. Во-вторых, это API управления эккаунтами пользователей.

Интерфейс доступа к приложениям Google в первую очередь ориентирован на конечных пользователей. В состав этих приложений входят: почта Google (gmail), набор офисных приложений (Writer, Calc, Note, DB), приложения для коллективной работы (Talk), приложения для работы с фото и видео файлами, например, широко известный сервис Picasso, и многое другое. В общей сложности это составляет почти два десятка приложений.

Все эти приложения бесплатные. Пользователю нужно только создать эккаунт в системе Google и сконфигурировать свой профиль для работы со всеми этими программами. Пользователь ничем не ограничен: ни объемами памяти, ни временем обработки данных, ни объемом передаваемых данных. Впрочем, собственно распределенных вычислений выполняется немного, так как большинство приложений загружаются на компьютер пользователя и там выполняются. Из распределенных платформ главным образом используется файловая система Google.

В отличие от Amazon в Google довольно сложно реализовать систему с эккаунтами, отличными от эккаунтов Google. Фактически любой, кто использует Google Apps, должен создавать эккаунт пользователя в системе Google. Сторонние приложения, таким образом, либо расширяют список приложений Google, либо позволяют организовывать агентские цепочки по использованию Google-приложений.

Тем не менее, API управления эккаунтами в Google существует. Прикладное ПО может создавать пользователей, назначать им определенные ресурсы, разрешать использовать то или иное API и ряд других функций.

Довольно часто архитектуру сервисов Google и Amazon сравнивают между собой. Наиболее типичной иллюстрацией этого сравнения можно считать стек приложений, представленный на рис. 7.1. Этот рисунок показывает степень доступности различных инструментов для конечного пользователя в стеках API обоих провайдеров услуг.

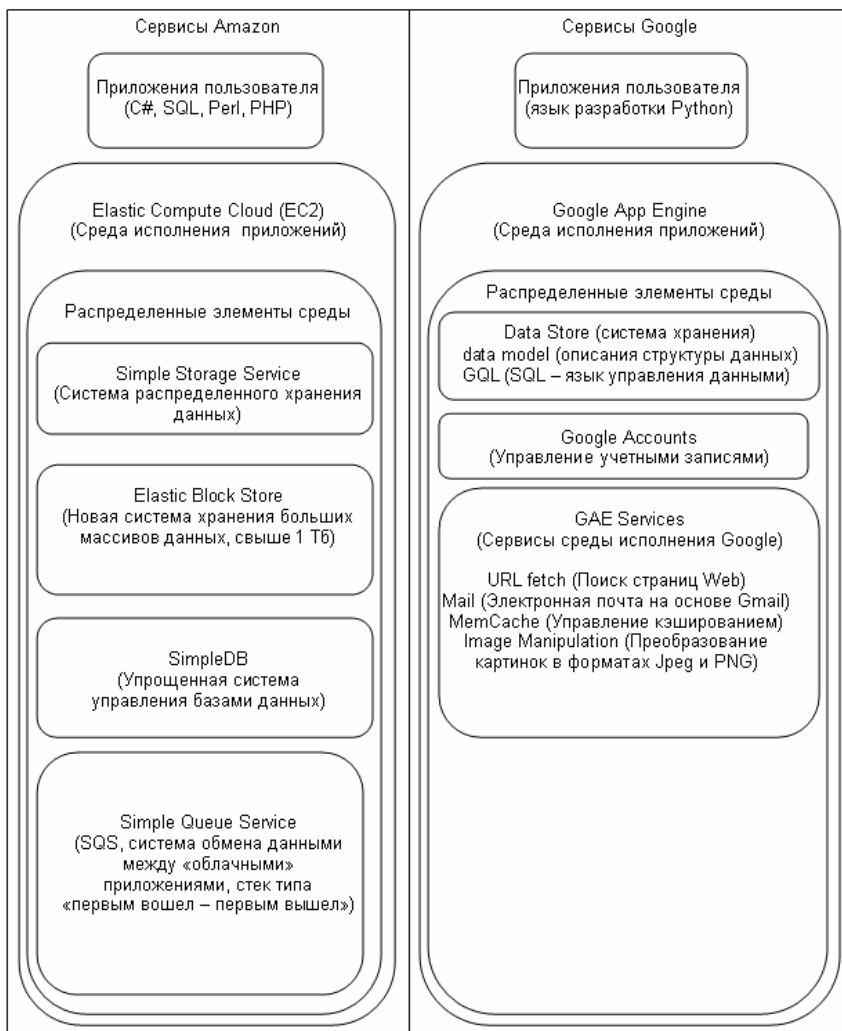


Рис. 7.1. Стеки приложений вычислений Amazon и Google.

Сердце системы Amazon – EC2, которое опирается на сервисы хранения, вычислений, обработки очередей и т.п.

Сердце системы Google – GAE, которое скрывает от конечного пользователя API уровня доступа к памяти, вычислениям и обмену данными с операционной средой.

Считается, что стек Google более детализирован и позволяет более гибко развивать собственные приложения конечному пользователю, однако, это иллюзия. Архитектура файловой системы и доступа к вычислительным мощностям скрыта в обоих случаях.

Сервисы Microsoft. Концепция Microsoft по организации распределенных вычислений претерпела существенные изменения. В первую очередь, это касается офисных приложений от Microsoft. Первоначально дело не шло дальше построения системы обновлений ПО посредством Internet. Сами программы и коды их активации стало возможным получать по Сети. Таким образом, речь не шла о собственно распределенных вычислениях. Была решена задача распространения программного кода по сети с последующим учетом того, на каких вычислительных установках он работает.

Вслед за решением этой задачи Microsoft реализовала распределенное хранение информации. Теперь пользователь мог работать с массивами данных, размещенными в распределенном файловом хранилище Microsoft и ее партнеров по данному проекту. Вместе с данными там могут размещаться и приложения. Таким образом, на вычислительной установке не требуется хранить и ПО, и данные. Можно просто загрузить программу по Сети, отредактировать, опять-таки загруженный по Сети файл, и сохранить его в виртуальной файловой системе для дальнейшего использования. Данная концепция, схематично представленная на рис. 7.2, получила название SaaS, т.е. «программное обеспечение, как сервис».

В данное время Microsoft продвигает несколько видоизмененную концепцию «Software + Services» («ПО+сервисы»). В рамках данной концепции уже нет необходимости загружать ПО на локальную вычислительную установку. Можно для обработки своих данных арендовать удаленные программы, взаимодействуя с ними через пользовательский интерфейс, главным образом, он ориентирован на обмен XML файлами между удаленно исполняемой программой и браузером пользователя. Такой подход наиболее близок к распределенным вычислениям типа Cloud computing. Архитектура вычислений полностью скрыта от пользователя за интерфейсом доступа к программным средствам.

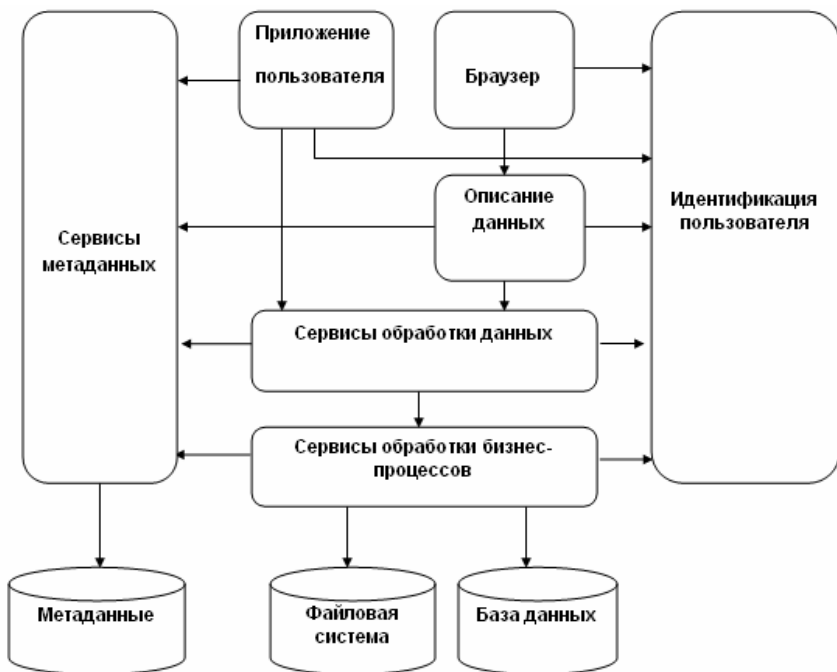


Рис. 7.2 Архитектура S+S от Microsoft.

От коммерческих реализаций распределенных вычислительных архитектур перейдем к разработкам, ориентированным на научные приложения. Наиболее известные проекты такого типа сосредоточены в области Grid Computing.

Примеры проектов Grid-вычислений. Впервые само понятие Grid-вычислений появилось в работе Яна Фостера(Ian Foster) и Карла Кессельмана(Carl Kesselman) «The Grid: Blueprint for a new computing infrastructure». На основе усилий добровольцев в 1997 г. появился проект distributed.net, а чуть позже, в 1999 г., SETI@home. В данных проектах была сделана попытка объединить вычислительные мощности персональных компьютеров по всему миру для решения вычислительно емких задач. Результатом этой работы должен был стать Globus Toolkit, который позволил бы не только организовать распределенные расчеты, но также и распределенное

хранение, распределенное управление и другие, необходимые для распределенных вычислений, компоненты.

В настоящее время существует множество реализованных Grid-проектов. Самый старый проект SETI@home (<http://setiathome.ssl.berkeley.edu/totals/html>) насчитывает около 3 миллионов компьютеров, суммарная производительность которых составляет порядка 23,37 терафлопа. Другой аналогичный проект Folding@home достиг вычислительной мощности в 1502 терафлопа на 270000 машинах в марте 2008 года.

Другой успешный пример Grid-системы – это NAS (NASA Advanced Supercomputing facility). Данная система используется для моделирования генетических алгоритмов. Программы в рамках этого проекта используют около 350 рабочих станций SUN и SGI.

Следует заметить, что Grid-вычисления довольно популярны и в других медицинских проектах. Например, до апреля 2007 г. Проект исследования рака применял Grid MP, который использовал персональные компьютеры добровольцев, подсоединенные к Internet. По данным 2005 г. в этом проекте участвовали 3100000 машин.

Используют Grid и в Европейском Союзе. Это так называемый European DataGrid – самый большой Grid проект в мире. Он так же, как и LCH Computing Grid (LCG), был разработан для обработки данных ядерно-физических экспериментов на большом адронном коллайдере в CERN. Вычислительная мощность этой системы около 10 петабайт в год.

На рис. 7.3 представлена архитектура gLite – «легкого» Grid, который используется в проекте обработки данных большого адронного коллайдера.

На этой схеме в правой стороне рисунка изображены статусы задания по мере его выполнения различными компонентами Grid-окружения, а слева собственно сами компоненты. Буквы на диаграмме обозначают состояния задания при его исполнении

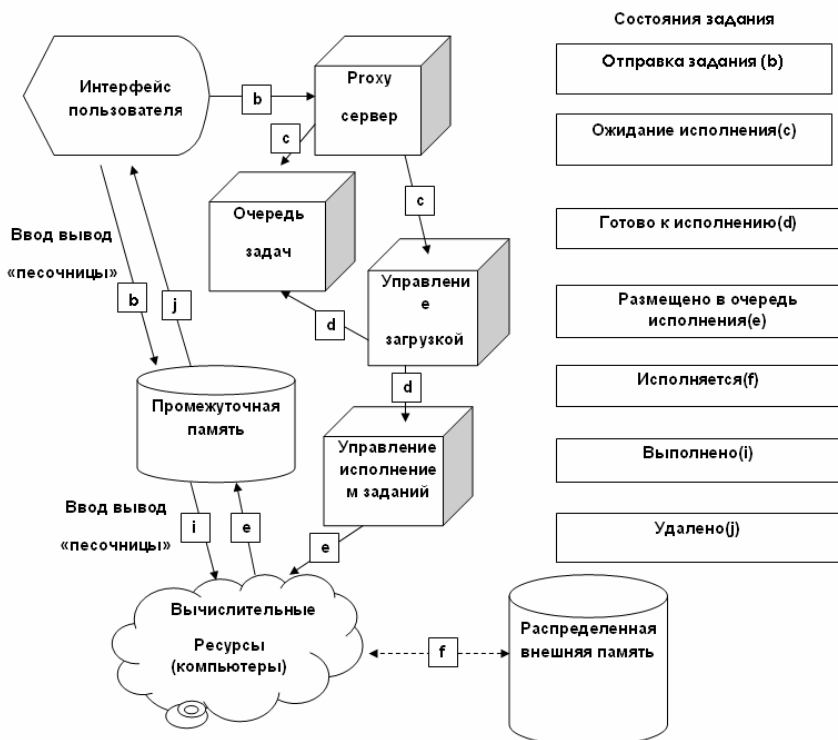


Рис. 7.3. Схема обработки данных в WLCG/EGEE Grid

Примеры других проектов Internet-вычислений. Кроме Grid-вычислений и Internet-вычислений существуют и другие распределенные вычислительные архитектуры, ориентированные на решение различных специфических задач. Одной из них является криптография.

Проект HashClas@home исследовал алгоритмы генерации кодов MD5 и SHA1, широко используемые в современных системах защиты каналов связи. Отдельно алгоритму SHA1 посвящен другой проект – SHA-1. Речь идет об устойчивости алгоритмов к различного рода исходным данным, используемым при генерации ключей.

Другой интересный научный проект – прогноз погоды Climateprediction.net. Задача проекта – организовать расчет предсказания климата в 21-м веке. Но больше всего проектов сосредоточено в области физики и биологии. Это и исследования по искусственному интеллекту, и моделирование деятельности мозга, и исследования магнетизма материалов в нанотехнологиях, и моделирование больших наносистем, и предсказание землетрясений, и многое другое.

7.4. Андеграунд распределенных вычислений

Возможность использования распределенных вычислений в современных вычислительных средах, особенно в Internet, породило не только массу полезных применений, но не меньшее количество угроз. Современные распределенные атаки на различные сайты и сети во многом похожи на распределенные вычисления. Здесь можно найти различные виды таких вычислений.

Рассмотрим, например, различные способы организации распределенных атак. Как правило, для организации такой атаки требуется большое количество компьютеров, с которых атаку проводят, так как в этом случае трудно обнаружить центр управления атакой и трудно заблокировать компьютеры, с которых атакуют¹.

При организации таких атак на машинах устанавливается стандартное ПО, которое выполняет подавление активности атакуемого сайта. Для такого подавления требуется согласованные действия множества машин (несколько тысяч). Как правило, управлением этими машинами производится из одного управляющего центра.

Участие в такого рода мероприятиях чревато тем, что на машине добровольца остается ПО, которое выполняет некоторые функции, о которых владелец компьютера может даже и не подозре-

¹ За последнее время мы стали свидетелями двух крупных атак, организованных добровольцами на сайты государственных органов. Во-первых, это атака 2007 года на эстонские сайты в связи с переносом памятника, и совсем свежие события августа 2008 года, когда была организована синхронная атака на российские и грузинские правительственные сайты, вызванная югоосетинским кризисом.

ревать. То есть компьютер в дальнейшем может быть использован третьими лицами в коммерческих целях без согласия его владельца.

Сети компьютеров, которые управляются третьей стороной без согласия их владельцев, принято называть бот-сетями. Традиционный способ создания бот-сетей – это заражение их программами-троянами. Заразить компьютер можно через почтовую рассылку, через уязвимости сетевого программного обеспечения, через уязвимости веб-сервисов и т.п.

В настоящее время бот-сети – это коммерческий клиент-ориентированный продукт. Наиболее известной из коммерческих сетей принято считать RBN (Russian Business Network). ФБР считает эту сеть «худшей из худших». Считается, что сервисы этой сети расположены на неизвестном российском пуленепробиваемом хостинге.

Суть работы данной сети заключается в том, что сервис состоит из нескольких подразделений, решающих узкоспециализированные задачи, ориентированные на конечную бизнес-цель, создание коммерческой бот-сети и продажу ее услуг.

Есть группа разработчиков, которая совершенствует методы внедрения программ-троянов и сами программы-трояны, есть группа разработчиков, которая совершенствует методы управления «затронутыми» компьютерами, группа разработчиков биллинга (система сбора платежей за предоставляемую услугу), группа разработчиков пользовательского интерфейса.

Фактически речь идет о распределенной вычислительной платформе, на основе которой предоставляются услуги по нелегальному использованию распределенных вычислительных мощностей, приватной информации и банковской информации.

Биллинг в такой сети очень напоминает принципы биллинга для Internet-вычислений (Cloud computing). Пользователь платит только за тот набор услуг, который реально использует. Например, только за то время, в течение которого заранее выбранное им количество компьютеров используются для организации DDos-атаки, или только за конкретное количество экаунтов электронной почты, которые можно использовать для спам-рассылки, или за точно определенное количество номеров кредиток, которые могут быть использованы системой для оплаты услуг своего клиента. В по-

следнем случае, система заботится о том, чтобы списание денег было не очень заметно для действительных владельцев карт и банков, эти карты выпустивших.

ГЛАВА 8. АРХИТЕКТУРА РАСПРЕДЕЛЕННЫХ РЕСУРСОВ

8.1. Типология и структура распределенных ИР

Рассматривая в предыдущих главах технологии и средства информационного поиска, мы не акцентировали внимание на организационно-физических аспектах ИР и доступа к ним, тем самым показывая, что принципы информационного поиска и архитектура ИС имеет универсальный характер.

Реальные АИПС по своим возможностям различаются достаточно существенно (по крайней мере, для пользователей). Это связано не столько с решениями тех или иных разработчиков, сколько с многочисленными факторами, в той или иной степени, ограничивающими или даже исключаящими некоторые функциональные возможности. Такими факторами являются, например, назначение и область применения, масштабность и интенсивность использования ИР, характер распределения информационных компонент (для сетевых систем), тип информации, источники и режим пополнения, предполагаемый профессиональный уровень потребителей и т.д.

По архитектуре, определяющей физическую доступность ресурсов, информационно-поисковые системы (в том числе и Internet-машины) могут быть разделены на следующие классы:

- локальные системы – локализованные данные и их обработка;
- частично-распределенные – локальная обработка распределенных данных;
- полностью распределенные системы, где реализуются принципы распределенных вычислений и распределенного хранения данных.

Локальные системы обеспечивают доступ удаленных пользователей к ресурсам, сосредоточенным на поисковом сервере. Эти

системы в большинстве случаев функционально эквивалентны локальным системам, например, на CD-ROM-носителях.

Ко второму типу относятся системы, использующие данные, которые находятся на Web-серверах, в качестве распределенных *первичных* ИР; вторичные и индексные данные сосредоточены на поисковом сервере, осуществляющем обслуживание пользователей. Это такие системы, как AltraVista, Google, Yandex и пр.

К третьему типу относятся системы, в которых процесс поиска реализуется на совокупности серверов, распределенных по сети, которые при обработке запроса опрашивают друг друга, причем исходные и промежуточные данные поиска также имеют распределенный характер.

По тематическому и видовому спектрам ИР могут быть однородными (иметь четко выраженную тематику и работать с документами определенного типа и состава) и гетерогенными (политематическими и не имеющими требований к составу и форме документов).

По способу формирования ИР подразделяются на те, которые используют предопределенные источники, например публикации издательств, рецензирующих материалы, и те, которые используют все свободно доступные материалы. Примерами здесь, соответственно, являются базы данных научной информации и поисковые машины Internet, индексирующие открытые HTML-страницы.

Если рассматривать электронные библиотеки в этом контексте, то можно заметить, что в качестве компонентов здесь выступают электронные каталоги (библиографические и реферативные базы данных), полнотекстовые массивы (электронные журналы, фактографические базы данных, хранилища электронных копий источников в том или ином виде и т.д.), справочно-нормативные файлы (рубрикаторы, тезаурусы, авторские, предметные, географические и другие указатели), возможно связанные между собой ссылками, указателями хранения или условиями поиска, хотя уже по своей сути эти компоненты всегда были и будут связаны, по крайней мере, на концептуальном уровне. Например, записи электронных каталогов содержат указания местоположения книг, а справочно-нормативные файлы традиционно используются в качестве «точек входа» в библиографические и реферативные базы дан-

ных. С появлением технических возможностей создания полнотекстовых баз данных справочно-поисковый аппарат и собственно массив информации технологически становятся единым целым, а на первый план выходит задача организации такой взаимосвязи, чтобы переход по ссылке от компонентов одного уровня к компонентам другого воспринимался пользователем как простейший одномоментный процесс, подобно перевороту страницы книги.

С организационной же точки зрения ЭБ помимо справочно-поискового сервиса обычно включает службу электронной доставки, обеспечивающую нахождение (или изготовление) и предоставление копий документов.

С точки зрения характера и формы представления информации (и, соответственно, логики организации поиска) архитектура ИР включает три уровня: уровень собственно документов (полных текстов), уровень поисковых образов и метаинформационный уровень. Характер и логика взаимосвязей информационных элементов отдельных уровней отражен схемой на рис. 8.1, где в скобках даны примеры, характерные для автоматизированных или традиционных информационных систем.

Взаимосвязь между компонентами разных уровней может быть реализована как для компонентов в целом, так и для их элементов. Иллюстрацией служит, например, такая связь, как «библиографическая запись электронного каталога – запись полнотекстовой базы данных» или «библиографическая запись - оцифрованная копия источника (изображение)». К другому типу (на уровне элементов) могут относиться такие связи, как «пристайная ссылка – библиографическая запись» или «фрагмент библиографической записи – запись нормативной базы данных» и т.д. Как говорится, все в этом мире взаимосвязано. Однако вербально или алгоритмически выразить эту взаимосвязь иногда весьма сложно.

Рассматривая эту схему как «технологию» поиска, можно видеть, что ссылки вполне узнаваемы и представляют собой традиционные правила и приемы отыскания информации в условиях «бумажной» библиотеки, когда поиск начинается с классификационной схемы или указателя, далее через библиографические карточки к первоисточнику, где, используя пристайные ссылки и указатели, снова продолжается с метаинформационного уровня.

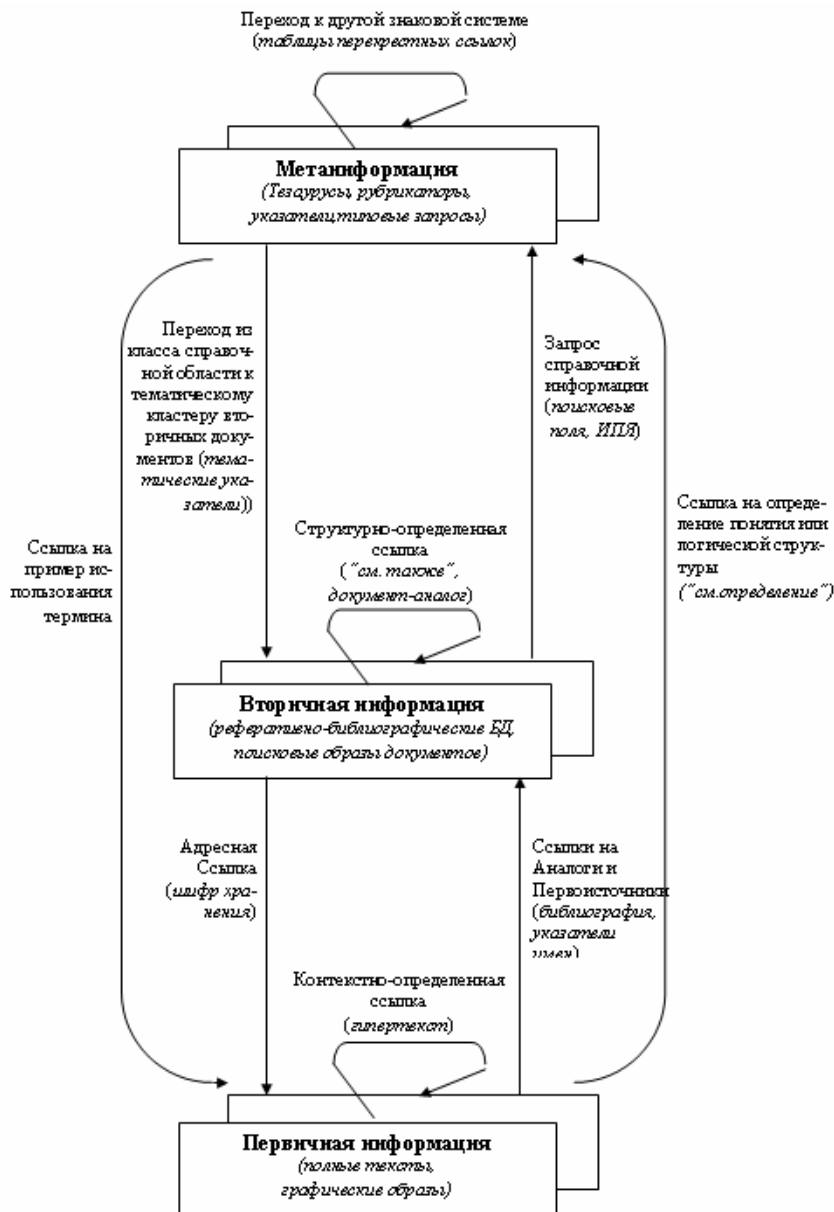


Рис. 8.1. Характер взаимосвязей информационных элементов

С широким внедрением телекоммуникационных сетей и некоторой стандартизации представления данных в Internet задача взаимосвязи становится еще более сложной. Ее решение путем создания статичных связей практически невозможно, даже если бы все компоненты имели свои уникальные идентификаторы и незыблемое место в информационном пространстве (чего зачастую невозможно добиться даже для локальных массивов). Что уж говорить о, скажем, информационных объектах, появляющихся на многочисленных сайтах Internet. Любое изменение местоположения информационного объекта влечет за собой возникновение "ложных" связей в распределенных электронных библиотеках. И число этих связей с течением времени возрастает. Поэтому на смену статичным связям приходят динамические, связи. Их особенностью является то, что они генерируются программно, по предопределенным алгоритмам во время обращения к объекту. Таким образом, можно связать информационные объекты, но при условии достаточной определенности (специфицированности) элементов.

Связи внутри могут быть построены на таких идентификаторах, как давно применяемые ISBN и ISSN или сравнительно недавно возникшие DOI (Digital Object Identifier). В тех случаях, когда такие идентификаторы отсутствуют (а таких случаев большинство), одним из решений может быть генерация динамических связей. В качестве основы для построения идентификаторов здесь могут выступать либо уникальные элементы записи, либо их свертки.

Кроме того, в качестве идентификаторов, используемых для установки активных связей, могут служить части элементов библиографического описания, организованные, например, в виде поисковых индексов. Таким способом можно связывать, скажем, компоненты справочно-нормативных файлов и массивов библиографических записей. Конечно, это возможно лишь при достаточно строгой структуризации компонентов баз данных и применении алгоритмов свертки, допускающих минимальный процент дублирования.

Отметим в заключение, что именно эти различия архитектуры, стратегии комплектования и организации доступа в итоге определяют не только функциональные возможности средств поиска конкретного ИР, но и круг потенциальных пользователей, а так же результативность решаемых ими задач.

8.2. Распределенная обработка в поисковых машинах Internet

Особенностью современных поисковых систем является фактическая разработка собственных операционных сред. Типичным примером является среда, на базе которой функционируют серверы Google.

Фактической базой системы является ОС FreeBSD. Однако способ управления вычислениями и файловая система в Google – собственная разработка компании.

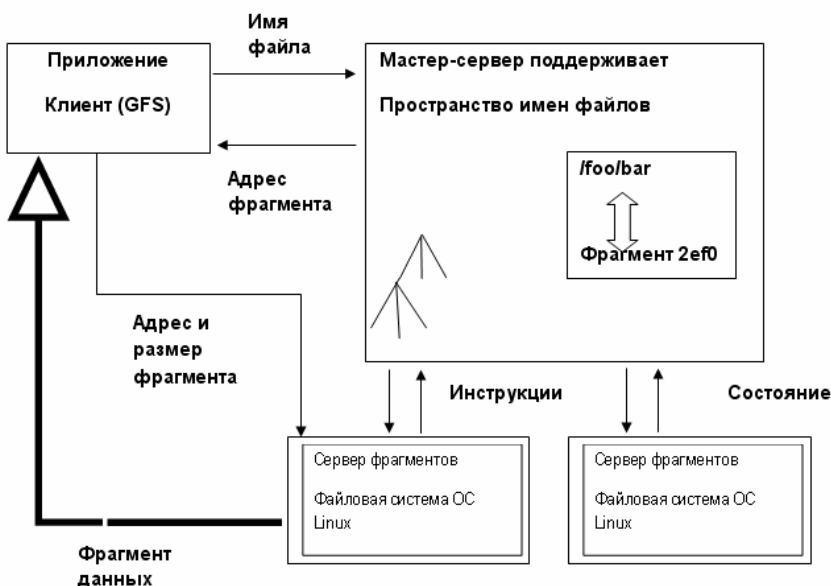


Рис. 8.2. Архитектура файловой системы Google (GFS)

Особенность файловой системы Google, представленной на рис. 8.2, – ориентация на очень большие, мало изменяющиеся файлы, которые должны храниться на множестве ненадежных вычислительных систем.

GFS состоит из множества кластеров. Каждый кластер состоит из отдельного Master-компьютера (мастера) и множества подчиненных ему серверов (chunkservers). Все файлы в системе состоят и

кусочков фиксированной длины – 64 Мб. Каждый кусочек помечен 64-битовым идентификатором.

GFS – это файловая система с журналом. Master, собственно, и ведет журнал изменений и обращений. Он же и восстанавливает систему при сбоях.

Система постоянно реплицируется. Кроме того, постоянно ведется контроль целостности данных и их непротиворечивости.

В GFS хранятся и индексы, и сами страницы, которые индексируются роботами Google. Собственно, GFS – это только часть технологии Google, предназначенной для работы с собранными данными. Кроме GFS, для поиска и записи данных в Google применяют технологии MapReduce (распараллеливание процедуры поиска и агрегирования результатов поиска) и BigTable (распределенное хранение больших объемов структурированных данных).

Технологии Google предназначены для параллельной обработки запросов. При этом процедура записи в GFS и поиска данных в GFS симметричны, что дает возможность использовать MapReduce и BigTable как при поиске, так и при записи данных.

8.3. Организация доступа к распределенным документальным ИР

Электронные (или цифровые) ИР, по сути своей, призваны хранить разнородную информацию и предоставлять пользователям удаленный доступ к ней по сети. Но при наличии в настоящее время огромного числа стандартов представления и передачи электронных документов, неременным требованием становится возможность представлять информацию в некой единой, удобной пользователю, форме, чтобы обеспечить её эффективное использование.

На самом деле большинство этих требований реализуется не на уровне конкретного приложения, а на уровне протоколов обмена данными.

В концепции WWW база данных – это набор текстовых файлов, написанных на языке HTML, который определяет форму представления информации (разметка) и структуру связей этих файлов (гипертекстовые ссылки). Такой подход предполагает наличие еще

одной компоненты технологии – интерпретатора языка. В World Wide Web функции интерпретатора разделены между сервером гипертекстовой базы данных и интерфейсом пользователя. Сервер, кроме обеспечения доступа к документам и реализации гипертекстовых ссылок, осуществляет также предобработку документов, в то время как программа просмотра (браузер) на машине пользователя проводит интерпретацию конструкций языка, связанных с представлением получаемой от сервера информации, и формирует изображение.

На сегодняшний день в области автоматизации доступа к распределенным ИР, кроме HTTP, достаточно широкое применение нашел протокол Z39.50.

Оба протокола, несмотря на то, что они используются для одних целей, различаются по назначению. Z39.50 изначально создавался как протокол поиска в библиографических базах данных и поэтому обладает большими возможностями по поиску, извлечению и представлению информации.

Список метаданных для доступа к ИР определяется в первую очередь используемым протоколом доступа. Для ресурсов, работающих по HTTP, единственным обязательным параметром является адрес документа/ресурса в сети (локальной или глобальной), т.е. URL или IP-адрес. Для Z-ресурсов минимальный набор метаданных должен включать адрес сервера, номер порта, имя поисковой базы и явное задание используемых поисковых полей. Параметры и структура поискового запроса определяются информационно-поисковым языком и типом запроса.

Также для Z-ресурсов обязательным является явное задание формата выдачи результата, в то время как при работе с HTTP он может быть опущен и результатом будет полный текст документа (для полнотекстовых баз) и реферат или библиографическая карточка для реферативных и библиографических ресурсов соответственно. При этом, информационно-поисковые системы (в том числе и работающие по Z-протоколу) обычно хранят и выражение запроса, и список идентификаторов документов, который на следующих шагах может объединяться/пересекаться со списками, полученным при поиске документов по новым выражениям. В то же время, возможность сохранения списка идентификаторов найденных доку-

ментов в поисковых машинах World Wide Web практически отсутствует, что связано с особенностью протоколов взаимодействия программы-клиента и сервера системы, которые не поддерживают сеансовый режим работы.

Протокол z39.50. Особенностью протокола Z39.50 является именно возможность сохранения состояний системы, что позволяет производить «навигацию во времени» т.е. в любой момент можно вернуться в определенную точку поиска, произведенного ранее. Наличие памяти в протоколе позволяет также использовать результаты поиска, полученные ранее, в составлении дальнейших запросов. Например, возможно составление запроса типа: (Result1) AND NOT (Result2) и др.

В качестве дополнительных возможностей Z-протокол позволяет выполнять параллельные операции поиска, уведомлять пользователя о состоянии сервера, сортировать данные на сервере, получать информацию о подключенных базах, наборах атрибутов, синтаксисах записей и т.д. Для описания баз данных внутри протокола был создан соответствующий абстрактный формат описания.

Следует заметить, что ни в одной сети не существует программных комплексов, реализующих полностью стандарт Z39.50. Современные системы в той или иной степени реализуют его подмножества, опуская подчас важные элементы, которые связаны как с обработкой запросов, так и с генерацией отчетов на эти запросы.

Z39.50 описывает прикладной уровень взаимодействия распределенных информационно-поисковых систем. Протокол определяет сам механизм информационного обмена в процессе обработки поисковых запросов и протокол обмена данными в системах, которые осуществляют поиск. Область применения протокола – библиотечные системы и системы научно-технической информации. Стандарт не определяет протоколы взаимодействия с физическими устройствами или их виртуальными аналогами, например, терминалами.

Термин *база данных* в спецификации Z39.50 означает набор файлов, каждый из которых имеет свое уникальное имя. Группа файлов внутри базы данных может также иметь свое собственное имя и образовывать отдельную базу данных. Такой подход аналогичен понятию схемы и подсхемы для систем управления базами

данных. Единицей хранения информации, которая может быть найдена при обращении к базе данных, является запись файла. Все записи одного файла должны иметь одинаковую структуру (т.е. состоять из одного и того же набора элементов и точек доступа). Точка доступа – это уникальный или неуникальный ключ (поисковое поле), который может быть указан самостоятельно или в совокупности с другими ключами в поисковом критерии. Ключ может быть элементом данных, состоять из нескольких элементов или быть частью элемента.

Запрос выполняется по всей базе данных. При этом проверяется совпадение точек доступа записей и элементов запроса. Набор записей, удовлетворяющих запросу, является ответом системы на запрос. Ответ может быть использован для последующих поисков в базе данных, в том числе, в виде ссылок на результаты предыдущих запросов.

В общем случае при выполнении запроса не требуется физического доступа к записям баз данных, так как в отчет о проведенном поиске включаются только идентификаторы записей, удовлетворяющих запросу, но не сами записи.

В рамках работы распределенной информационно-поисковой системы рассматриваются два типа прикладных задач:

- задача, инициирующая взаимодействие и посылающая запросы на обслуживание;
- задача, отвечающая за ответы на запросы первой задачи.

Первая задача называется «источник» (origin), а вторая – «мишень» (target). Взаимодействие источника и мишени осуществляется путем установки соединения. Соединение может быть инициализировано только источником и может быть разорвано либо другим источником, либо мишенью, либо по внешним причинам (например, физический разрыв связи). В процессе взаимодействия источник и мишень не могут поменяться ролями. Таким образом, протокол Z39.50 описывает *интерактивную сессию* между источником запросов и мишенью, которая эти запросы обслуживает, т.е. реализует типичное взаимодействие по схеме «клиент-сервер».

Согласно Z39.50, в рамках распределенной информационно-поисковой системы существуют семь основных видов информационного обмена:

- инициализация сессии;

- поиск информации по запросу;
- представление результатов поиска;
- удаление результатов поиска;
- контроль доступа к информационному ресурсу;
- контроль прав доступа к информационному ресурсу;
- завершение сессии.

Как было показано выше, с точки зрения состава и характера информационных ресурсов можно выделить: электронные каталоги, библиографические и реферативные базы данных, полнотекстовые массивы (электронные журналы, фактографические базы данных, хранилища электронных копий статей и других публикаций), справочно-нормативные файлы (рубрикаторы, тезаурусы, авторские, предметные, географические и другие метаинформационные продукты). Записи этих массивов могут быть взаимно связаны между собой ссылками, указателями хранения или условиями отыскания. И даже если такие ссылки в явном виде не существуют, то уже по своей сути эти компоненты всегда были и будут связаны, по крайней мере, на концептуальном уровне.

8.4. Обобщенная схема доступа к ресурсам ЭБ

Рассмотрим обобщенную информационную систему, обеспечивающую унифицированный доступ к разнородным распределенным документальным ресурсам, образующим в общем случае электронную библиотеку. На схеме, представленной на рис. 8.3, можно выделить три блока:

- клиентское приложение;
- информационный сервер;
- сервер управления доступом пользователей к информационным ресурсам.

Клиентское приложение обеспечивает взаимодействие конечного пользователя с серверной частью приложения. Ее основное назначение – формулирование поискового запроса и передача его локальной или внешней (распределенной через web- или Z39.50-шлюз) поисковой системе, а по окончании поиска отображение результата в удобной для пользователя форме. В качестве клиентского приложения может выступать как Web-браузер или Z-

клиент, так и поисковая система, имеющая собственные локальные БД и обеспечивающая взаимодействие с внешними ИР через шлюзы, согласующие требования к представлению запросов и выдач.

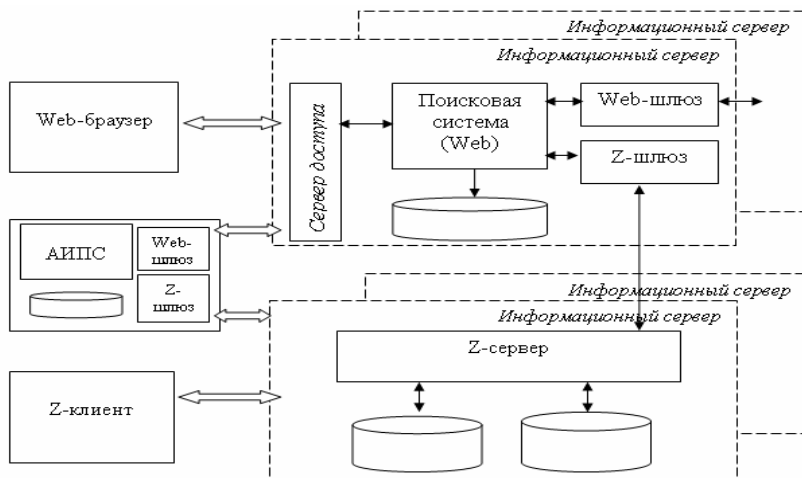


Рис. 8.3. Организация доступа к распределенным ИР

Серверная часть обеспечивает через сервер связи (Web-сервер или Z39.50-сервер) взаимодействие клиента и собственно информационного сервера. Эта работа скрыта от конечного пользователя.

Информационный сервер – это хранилище данных, представляющих коллекции разнородных логически или физически взаимосвязанных документов, доступ к которым обеспечивается средствами информационно-поисковой системы. При этом поисковые (реферативные или библиографические) и метаинформационные компоненты обычно реализованы в виде структурированных баз данных, в то время как компоненты уровня первичной информации могут быть представлены и полнотекстовыми БД, и коллекциями файлов, каждый из которых хранит, обычно, образ в том или ином графическом или текстовом формате.

Сервер управления доступом к информационным ресурсам реализует ограничения доступа к компонентам ИР. Обычно он представляет собой набор специализированных баз данных, как,

например, каталог ИР, хранящий сведения о локальных и ассоциированных ресурсах, каталог пользователей, хранящий учетные записи и список доступных конкретному пользователю информационных ресурсов, БД учетно-статистической информации об использовании ресурсов и т.д.

8.5. Доступ к полным текстам документов

Для пользователей наибольшую ценность представляют не библиографические описания, а сами документы. При этом требуется получить их с минимальными временными и финансовыми затратами. Современные поисковые порталы позволяют искать библиографическое описание и аннотации статей из периодических изданий, таких как электронные версии журналов или журналы, выходящие только в электронной форме. В большинстве случаев полные тексты статей бесплатно недоступны.

Источниками полнотекстовой информации могут являться:

- коммерческие или общественные издательства (например, Elsevier или American Chemical Society);
- коллекции (например, Bio OnLine);
- полнотекстовые БД (например, EBSCO);
- провайдеры услуг, обеспечивающие доступ к электронным версиям статей и электронную доставку документации (например, eLibrary);
- OnLine издатели (например, Highwire Press);
- независимые журналы (например, D-Lib);
- библиотечные каталоги (например, каталог библиотеки Конгресса США);
- специализированные службы электронной доставки документов (например, CISTI);
- общедоступные ресурсы поисковых машин Internet (например, Google).

Полнотекстовый документ, описание которого было получено при проведении первоначального поиска, может находиться в целом ряде источников, как платных, так и открытых. Кроме того, этот же документ может присутствовать в свободном доступе на веб-сайте автора или университета, в котором он работает. Для

решения задачи оптимизации формирования ссылок на полнотекстовые электронные документы был создан стандарт OpenURL. В основе HTTP, которая является самой распространенной Internet-технологией на сегодняшний день, лежит использование гиперссылок, содержащих адрес ресурса в формате URL. На веб-страницах могут применяться как статические, так и динамические URL. Статические URL применяются, когда адрес ресурса известен и неизменен, а динамические URL – при программном формировании ссылок, на основе сведений выбранных из БД, а также случаев передачи параметров, необходимых для последующей обработки.

Стандарт OpenURL. В основе стандарта OpenURL лежит формирование динамических ссылок, содержащих в качестве параметров гиперссылки трех блоков данных:

- базового OpenURL – содержит адрес используемого определителя ссылок;
- идентификатора sid (уникальный идентификатор, формируемый службой, создавшей ссылку OpenURL);
- метаданных, описывающих ресурс.

Например, для поиска статьи со следующим библиографическим описанием:

Автор: L.Feng

Название статьи: “The methodology for XML documents”.

Источник: “ACM Transactions on Informational Systems”.

ISSN: 1046-8188.

Дата публикации: 2002

Том: 20

Выпуск: 3

Страницы: 390-421

гиперссылка в формате OpenURL будет иметь следующий вид:

[http://www.lib.web.edu/openresolver/&sid=uweb:openurl&genre=article&article=The methodology for XML documents&title= ACM Transactions on Informational Systems&issn=1046-8188&date=2002&volume=20&issue=3 &spage=390&epage=421 &aulast=Feng&auinit=L](http://www.lib.web.edu/openresolver/&sid=uweb:openurl&genre=article&article=The%20methodology%20for%20XML%20documents&title=ACM%20Transactions%20on%20Informational%20Systems&issn=1046-8188&date=2002&volume=20&issue=3&spage=390&epage=421&aulast=Feng&auinit=L)

Первая часть гиперссылки
(<http://www.lib.web.edu/openresolver>)

содержит адрес определителя ссылок, т.е. адрес сервера, на котором работает программы, позволяющая найти гиперссылки полнотекстовых документов на основе метаданных, содержащих библиографические описания.

Определитель ссылок (Link Resolver) – распознающий сервер, является функциональным узлом, формирующим переход от описания на уровне библиографических элементов к полному тексту статьи. Хотя, в некоторых случаях пользователю будут доступны не сами статьи, а их аннотации, что связано с лицензионными ограничениями. В настоящее время предлагаются следующие системы распознавания ссылок: SFX (*ExLibris*), LinkFinderPlus (*Endeavor*), 1Cate (*Openly Informatics*). Article Linker (*Serials Solutions*), WebBridge (*Innovative Interfaces*), LinkSource (*EBSCO*), Sirsi Resolver (*Sirsi Corp.*).

Возможен вариант настройки, при котором для получения сведений о доступности полного текста документа определитель ссылок будет обращаться по протоколу z39.50 к соответствующей полнотекстовой базе данных (если сервер ресурса это поддерживает).

Версия стандарта OpenUrl 1.0 получила официальное обозначение z39.88-2004 (стандарт США). В этой версии предусматриваются возможности для поиска документов по библиографическим и контекстным описаниям, в частности, расширяется список типов поддерживаемых документов (в предыдущей версии предусматривался поиск только по статьям). Кроме того, есть передача сведений об источнике цитирования, если ссылка представляет собой цитату, найденную в другом полнотекстовом документе. Одним из параметров может быть тип запрошенного сервиса, что позволяет осуществлять непосредственную передачу данных различным службам, например, службе электронной доставки документов. Большое внимание также уделено идентификации пользователя с целью защиты данных, разграничения прав доступа к информации и поддержки биллинговых (расчетных) систем при работе с платными ресурсами. С этой целью в версии стандарта OpenUrl 1.0 предусмотрена передача сведений о пользователе.

8.6. Интероперабельность в распределенных ИР

Как уже отмечалось, эффективность использования накопленной человечеством информации в значительной степени сегодня связывается с обеспечением интероперабельности информационных ресурсов, т.к. отсутствие унифицированности и стандартизации способно свести к нулю все преимущества работы с разнообразными ресурсами по различным протоколам.

Интероперабельность имеет целью добиться однообразного представления разных коллекций, хотя, безусловно, нет необходимости устранять все различия. Несмотря на то, что коллекция карт отличается от собрания музыкальных записей, пользователь должен иметь возможность легко работать в обоих, в т.ч. выполнять поиск и быть защищенным от «особенностей» компьютерных систем или менеджмента этих коллекций.

Выделяют три основных уровня интероперабельности информационных ресурсов.

- *Техническая.* Предполагает регламентированность принципов, стандартов для общих механизмов передачи данных, стандартизация метаданных с использованием общего языка

- *Синтаксическая.* Приложения должны иметь возможность читать данные и получать представления, с которыми они могут работать. Уровень синтаксической интероперабельности достаточно высок, когда можно без труда построить синтаксический анализатор и API-интерфейсы, необходимые для манипулирования данными.

- *Семантическая.* Одно из наиболее важных требований к формату обмена состоит в том, что данные должны быть понятными. Если синтаксическая интероперабельность неразрывна с синтаксическим разбором данных, то семантическая связана с установлением соответствия между терминами, используемыми в данных, что требует анализа содержимого.

Интероперабельность и стандартизация взаимосвязаны. К сожалению, формальный процесс создания международных стандартов зачастую противоречит тому, что требуется для реальной интероперабельности ЭБ. И не только потому, что официальный процесс стандартизации слишком медлителен для быстроменяющегося мира ЭБ; получающиеся стандарты еще и исключительно

сложны. Многие интернациональные стандарты никогда не испытывались в реальной жизни. На практике только те стандарты работают, которые часто используются. Иногда стандарт принимается де-факто за счет того, что ведущая исследовательская группа его использует.

Приведем основные аспекты интероперабельности ЭБ.

Наименование и идентификация. Необходимо несколько способов для идентификации материалов ЭБ. Каждый компьютер в Internet имеет IP адрес и доменное имя. Универсальный указатель ресурсов (URL) веб расширяет эти имена до индивидуальных файлов. Однако этих возможностей недостаточно. Библиотечные объекты нуждаются в идентификаторах, которые обозначают именно материал, а не его расположение в данный момент времени. Такие локально-независимые (location-independent) идентификаторы иногда называют Универсальными Именами Ресурсов (Uniform Resource Names, URN).

Форматы. Во всех известных ЭБ материалы хранятся в цифровой форме. Web де-факто создал стандарты для новых форматов, к примеру, HTML для простого текста или GIF и JPEG для графики. Однако на этих форматах согласие разработчиков заканчивается. С точки зрения интероперабельности, особую трудность представляет текст. Формат ASCII, в 1980-е годы ставший практическим стандартом кодировки для компьютеров, имеет ограниченное число символов кроме английских. В настоящее время развивается кодировка Unicode – расширенный набор символов, способный поддерживать очень широкий набор шрифтов. Но он все еще не поддерживается многими компьютерными системами. Широко поддерживаемый формат SGML используется в некоторых ЭБ как язык разметки и для текста и для метаданных, является столь гибким и сложным, что достичь на его базе интероперабельности весьма нелегко. XML, упрощенная версия SGML, стала популярна в последнее время и возможно ему удастся занять нишу между простым HTML и исчерпывающим SGML.

Метаданные. Метаданные важны во многих аспектах ЭБ, но особенно для достижения операбельности. Как уже отмечалось,

метаданные часто делят на три категории: описательные (используются для библиографических целей, поиска и обработки), структурные (которые связывают различные объекты или части объектов между собой) и административные (для управления коллекцией и доступом к ней). Для интероперабельности необходим обмен частью этих метаданных между компьютерами. Это требует отдельного соглашения об именах метаданных, форматах кодировки и их семантике. Очевидный пример важности семантики: если в одной коллекции поле «Дата» используется для даты создания, а в другой — для даты включения в коллекцию, то это поле представляет относительную ценность. Некоторые предложения о системах метаданных приведены в Приложении 4.

Распределенный поиск. Пользователи часто хотят найти информацию из нескольких независимых коллекций одновременно. Они могут быть организованы логично и единообразно, но может существовать различие в описательных метаданных, используемых для поиска. Традиционный подход к поиску информации в различных коллекциях одновременно состоит в том, чтобы во всех них был стандартный набор метаданных и поддерживался один и тот же протокол. Все большее число специалистов осознает нереалистичность таких требований. Должны быть возможность поиска в различных коллекциях, даже если материал в них организован по-разному.

Сетевые протоколы. Перемещение материалов с одного компьютера на другой требует интероперабельности на сетевом уровне. Практически универсальный набор Internet-протоколов в значительной степени решает эту проблему, но не всегда. К примеру Internet-протоколы не слишком подходят для передачи непрерывных потоков данных, как видео- или аудиоматериалы, которые должны прибывать в качестве целостного потока в предсказуемые интервалы времени.

Поисковые протоколы. В процессе выполнения операций, компьютер посылает запрос другому с целью получить некоторые данные. Это послание должно быть передано по определенному протоколу, который может быть как простой (вроде HTTP), так и

сложный (например, z39.50). Идеальный протокол должен поддерживать идентификацию обоих компьютеров (с целью обеспечения безопасности), высокоуровневые запросы для определения доступных ресурсов, варианты поиска и возможности обработки, методы хранения и модификации промежуточных результатов, а также интерфейс для различных многочисленных форматов и процедур (наиболее близок к достижению этих целей, Z39.50, но он уже малоприменим из-за собственной сложности – и при этом все еще не отвечает всем критериям).

Идентификация и безопасность. Ряд наиболее сложных проблем интероперабельности между ЭБ включает идентификацию. Здесь есть три категории проблем. Первая – это идентификация пользователей. ЭБ вынуждены присваивать каждому пользователю "пользовательский ID" и пароль. Вторая категория – идентификация компьютеров. Системы, содержащие ценную информацию, особенно финансовую или конфиденциальную, должны знать, с каким компьютером они соединяются. Можно полагаться на IP-адреса, но это крайне ненадежно. Третья категория – идентификация материалов. Люди должны быть уверены, что они получают аутентичную версию, которая не была изменена (случайно или намеренно).

Семантическая интероперабельность – это широкий термин, который обозначает необходимость для компьютеров иметь одинаковую семантическую интерпретацию передаваемой информации. Семантическая интероперабельность связана с возможностью для пользователей находить одинаковые электронные объекты, распределенных по разнородным коллекциям.

Веб предоставляет некий базовый уровень интероперабельности, но именно та простота технологии, которая привела к его широкому распространению, принесла и ряд проблем: возможностей URLs недостаточно для длинных терминов; HTML позволяет представлять ограниченное число типов информации; MIME, который позволяет определять тип каждого объекта, достаточно хорош в настоящее время, однако библиотечная информация намного шире, чем возможности MIME.

Контрольные вопросы к разделу

1. Перечислите особенности организации распределенной файловой системы.
2. Сформулируйте главную проблему разделения файлов.
3. Приведите способы решения конфликтов при разделении (sharing) файлов.
4. Перечислите основные виды параллелизма при распределенной обработке данных.
5. Обоснуйте целесообразность разделения «клиентских» и «серверных» функций.
6. Обоснуйте назначение механизма неделимых транзакций.
7. Проведите сравнительный анализ распределения функций для различных базовых архитектур распределенной обработки.
8. Определите основные принципы и примерные структурные схемы распределенной обработки.
9. Приведите отличие модели кластеров от модели единого пространства
10. Чем отличаются мультимпьютерные системы от кластеров
11. Проведите сравнительный анализ видов параллелизма.
12. Перечислите сходства и различия сервисов Amazon и Google.
13. Перечислите компоненты SaaS (S+S) Microsoft.
14. Приведите примеры Internet-вычислений.
15. Назовите основные классы распределенных информационных ресурсов.
16. Определите алгоритм поиска с использованием информационных объектов разного уровня (первичная, вторичная, метainформация).
17. Назовите основные виды обмена информацией согласно Z39.50.
18. Перечислите основные преимущества и недостатки протокола Z39.50.
19. Перечислите основные факторы, определяющие интероперабельность ИР.

РАЗДЕЛ 4. ДОКУМЕНТАЛЬНЫЕ РЕСУРСЫ НАУЧНОЙ ИНФОРМАЦИИ

В этом разделе будут описаны некоторые, наиболее представительные, информационные ресурсы. Описания сгруппированы так, что их характер отражает не только форму организации, но и такие важные характеристики, как источники комплектования массива, тематический охват, вид информации (научная, патентная, отчеты о НИР и т.д.), форма представления (первичная, вторичная), которые в целом обуславливают использование ИР.

В данном изложении в девятой главе представлены отдельные ресурсы, а по существу, организации, основной задачей которых является сбор, обработка, хранение и обеспечение доступа к конкретным видам литературы и документации (первичной информации – книгам, сборникам, стандартам, выпускам журналов и т.д.). В десятой главе приведены примеры организаций (ресурсов), основной задачей которых является формирование справочно-информационных фондов и баз данных на основе собираемых сведений о первичной информации. В одиннадцатой главе – примеры организаций, основной задачей которых является выпуск первичной информации (или непосредственно с ними связанные). Двенадцатая глава содержит примеры отдельных специализированных баз данных и ресурсов, имеющих узкую направленность и функциональность. В примерах приведены Internet-версии ресурсов, но большинство продуцентов ИР предоставляют те же БД и в «локальной» версии на CD или DVD-носителях.

При этом надо понимать, что такое деление достаточно условно: ресурсы всех типов, так или иначе, выполняют все технологические функции; в каждом классе ресурсы характеризуются и, в большинстве случаев, организационно выделяются по тематическому (химия, биология, медицина, и т.д.) или видовому (патенты, отчеты о НИР, научно-техническая информация, стандарты и т.д.) признаку и, конечно, каждый ресурс имеет свою специфику.

В пособии не рассматривается достаточно серьезный класс служб, сервисов и информационных порталов, обеспечивающих доступ к ресурсам организаций перечисленных выше типов. Такие службы обеспечивают сбор и систематизацию информации о

внешних ресурсах, организуют доступ к отдельному или нескольким ИР, проводят обслуживание и обучение пользователей. Примерами служб такого класса являются STN International (Scientific and Technical Information Network), Международный центр научной и технической информации (МЦНТИ). Близкое по функциональности назначение имеют информационные порталы университетов или исследовательских центров.

Представленные материалы не являются руководствами по использованию конкретного ресурса. Свою задачу авторы видели в том, чтобы, показывая разнообразие интерфейсов и функциональных возможностей, привести читателя к пониманию того, что все системы, так или иначе, укладываются в достаточно простые схемы и технологии, рассмотренные в первых главах. Именно поэтому (и вследствие ограниченности объема пособия) в каждом классе систем достаточно подробно рассматривается какая-либо одна.

ГЛАВА 9. ЭЛЕКТРОННЫЕ КАТАЛОГИ И БИБЛИОТЕКИ

9.1. Электронные ресурсы

Библиотеки по естественным наукам РАН

Библиотека по естественным наукам РАН (БЕН РАН) ведет работы по предоставлению своих ресурсов пользователям всемирной компьютерной сети с 1996 г.. Значительное внимание в Библиотеке уделяется развитию системы электронных каталогов, доступных любому пользователю сети Internet.

БЕН РАН поддерживает два вида каталогов – *каталог журналов* и *каталог книг и продолжающихся изданий*.

Доступ к каталогам открыт для любого пользователя, однако доступ к полным текстам предоставляется только пользователям, зарегистрированным в соответствии с правилами, устанавливаемыми каждым издательством. Зарегистрированные пользователи могут заказать нужные им материалы или их копии из фондов ЦБС БЕН РАН через Internet.

На сервере БЕН РАН (<http://www.benran.ru>) реализована система информирования о новых поступлениях. Еженедельно загружаются два вида указателей: *Поступления в Центральную библиотеку* и *Поступления в библиотеки институтов РАН*.

Внешний вид домашней страницы библиотечного комплекса представлен на рис. 9.1.

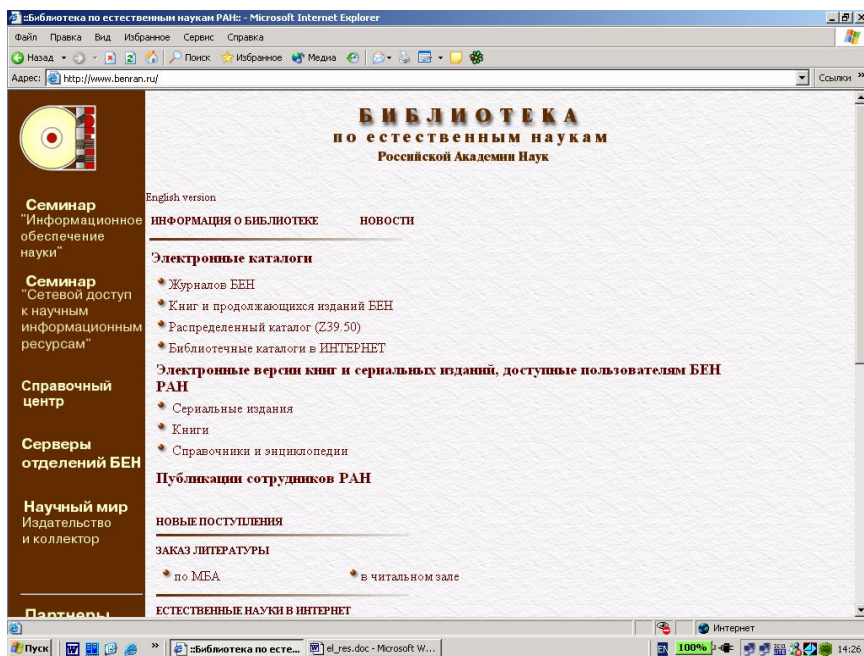


Рис. 9.1. Домашняя страница сайта БЕН РАН

Электронный каталог журналов БЕН РАН

Каталог журналов содержит ссылки на оглавления журналов, поступивших в ЦБС БЕН РАН, а также ссылки на их электронные версии (если таковые существуют). Актуальность этих ссылок постоянно контролируется. В настоящее время в электронном каталоге отражено около 7.5 тысяч названий отечественных и зарубежных журналов. Сведения о журналах, имеющихсх в фондах центра-

лизованной библиотечной системы (ЦБС БЕН РАН), представлены с 1994 г. (для журналов России и ближнего зарубежья) и с 1990 г. издания (для всех остальных).

Каталог предусматривает возможность просмотра журналов в алфавитном или тематическом порядке, а также поиска по фрагментам описания (рис. 9.2).

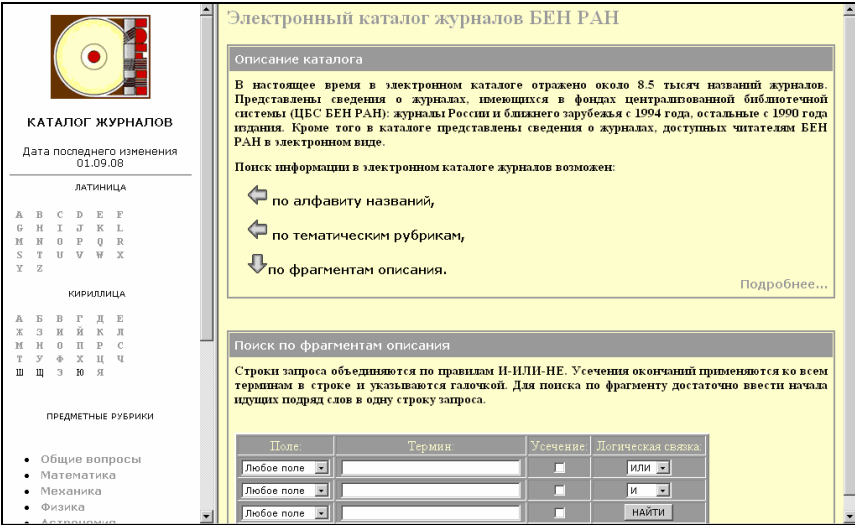


Рис.9.2. Электронный каталог журналов

Для просмотра в *алфавитном порядке* в левой части экрана выбирается алфавит (латиница или кириллица) и буква алфавита. После выбора нужной буквы на экран выводится список журналов, упорядоченный по алфавиту названий (рис. 9.3).

Получить сведения о журнале можно также и при помощи *тематической рубрикации*, выбрав сначала рубрику в левой части экрана, а затем – выбрав подрубрику и далее название журнала в правой части экрана (рис. 9.4, 9.5).

Если около названия журнала в списке журналов стоит символ , это означает, что для данного журнала на сайте издательства доступны полные тексты за несколько последних лет.

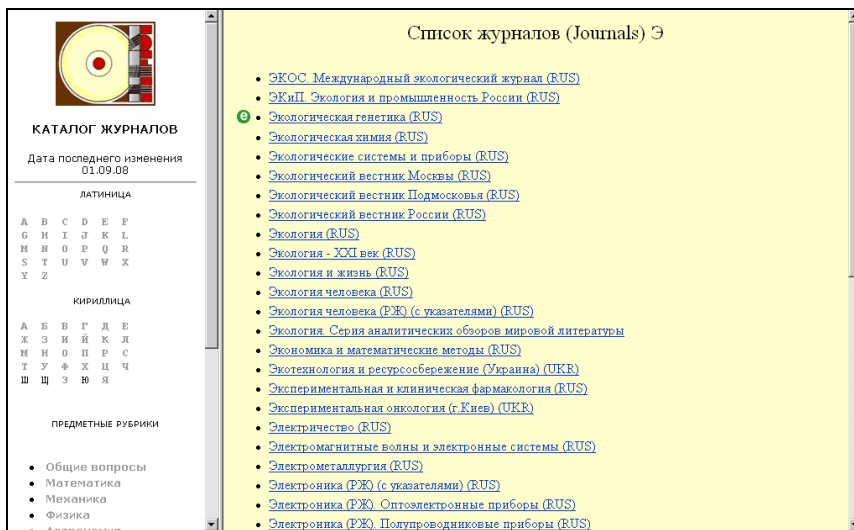


Рис. 9.3. Просмотр каталога журналов в алфавитном порядке

Выбрав название журнала, пользователь получает в правой части экрана страницу, содержащую описание журнала, (название журнала, код страны, ISSN, шифр хранения БЕН) и список упорядоченных по годам томов и номеров, имеющихся в фондах ЦБС БЕН РАН (рис. 9.6).

Для расширения возможностей при работе с каталогом журналов БЕН РАН на страницах многих журналов помещены от од-

ной до трех ссылок: на сайт издательства (), на сайт «Научной

электронной библиотеки» (НЭБ) РФФИ () и на сайт системы

Electronic Journals Service (EJS) ().

По ссылке на сайт издательства можно получить полные тексты или дополнительную информацию о журнале, а также (в большинстве случаев) ознакомиться с оглавлениями выпусков журнала и рефератами статей.

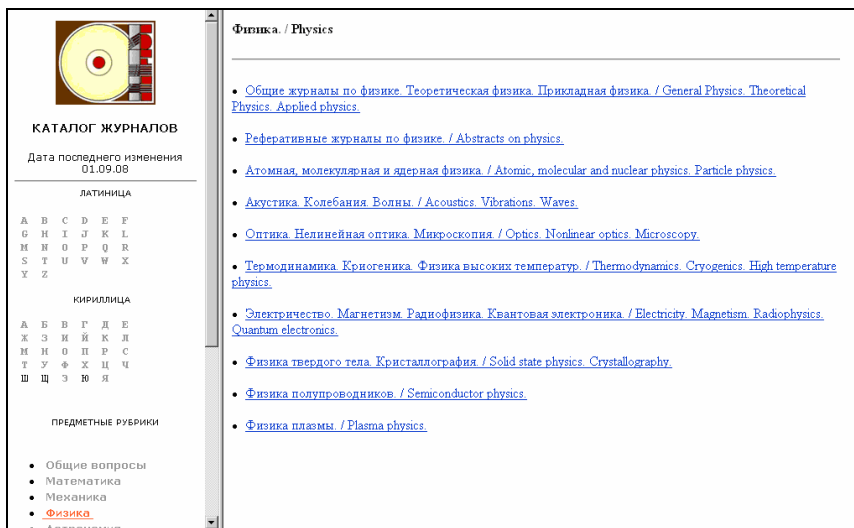


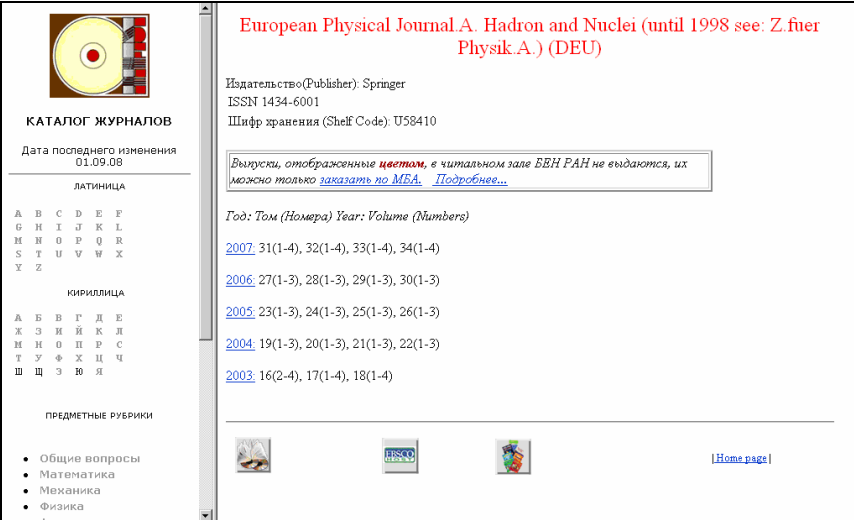
Рис. 9.4. Выбор подрубрики



Рис. 9.5. Выбор названия журнала

Ссылка на сайт НЭБ позволяет пользователям, зарегистрированным в НЭБ, выйти на страницу соответствующего журнала в НЭБ, ознакомиться с оглавлениями выпусков и получить полные тексты статей.

Система EJS, созданная компанией EBSCO и доступная российским пользователям в рамках проекта Международного информационного центра, позволяет получать доступ к оглавлениям журналов, рефератам статей из большинства этих журналов и к полным текстам статей из тех журналов, к которым БЕН РАН имела право доступа в рамках консорциумов и по подписке. На странице каждого журнала, отраженного в EJS, указаны годы, для которых можно получить дополнительную информацию по журналу.



КАТАЛОГ ЖУРНАЛОВ

Дата последнего изменения
01.09.08

ЛАТИНИЦА

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z				

КИРИЛЛИЦА

A	B	B	Г	Д	Е
Ж	З	И	Й	К	Л
М	Н	О	П	Р	С
Т	У	Ф	Х	Ц	Ч
Ш	Щ	Э	Ю	Я	

ПРЕДМЕТНЫЕ РУБРИКИ

- Общие вопросы
- Математика
- Механика
- Физика

European Physical Journal.A. Hadron and Nuclei (until 1998 see: Z.fuer Physik.A.) (DEU)

Издательство(Publisher): Springer
ISSN 1434-6001
Шифр хранения (Shelf Code): U58410

Выпуски, отображенные **цветом**, в читальном зале БЕН РАН не выдаются, их можно только **заказать по МБА**. [Подробнее...](#)

Год: Том (Номера) Year: Volume (Numbers)

[2007](#): 31(1-4), 32(1-4), 33(1-4), 34(1-4)
[2006](#): 27(1-3), 28(1-3), 29(1-3), 30(1-3)
[2005](#): 23(1-3), 24(1-3), 25(1-3), 26(1-3)
[2004](#): 19(1-3), 20(1-3), 21(1-3), 22(1-3)
[2003](#): 16(2-4), 17(1-4), 18(1-4)

[Home page](#)

Рис. 9.6. Просмотр описания журнала

Для поиска журнала *по фрагментам описания* пользователь должен заполнить бланк запроса (см. рис. 9.2), руководствуясь следующими правилами:

- термины, введенные в одной строке, объединяются логической связкой «И»

- строки запроса объединяются выбранными в правой колонке логическими связками И-ИЛИ-НЕ.

- усечения окончаний (помечаются флажком) применяются ко всем терминам, введенным в строке.

Для поиска по заглавию, например, достаточно ввести начала идущих подряд слов из заглавия в одну строку запроса и включить флажок усечения.

Электронный каталог книг и продолжающихся изданий БЕН РАН

Каталог содержит библиографические описания книг и продолжающихся изданий (сборники трудов конференций, трудов организаций и научных обществ и т.п.), поступающих в централизованную библиотечную систему БЕН РАН. В каталоге отражены поступления отечественной литературы, начиная с 1993 г., и зарубежной, начиная с 1995 г..

Результаты поиска в каталоге выдаются в виде стандартных библиографических описаний.

Запрос на поиск (рис. 9.7) может содержать от одной до трех строк терминов, соответствующих выбранным поисковым полям. Строки соединяются логическим И, ИЛИ, НЕ.

В каждую строку может быть введено через пробел несколько терминов, связанных логическим И или ИЛИ.

Для правого усечения терминов используется символ «*» (например, на запрос *реактор** будут выданы все документы, содержащие термины *реактор, реактора, реакторный* т.п.).

При работе с каталогом возможен просмотр, выбор и автоматический ввод в запрос терминов из словарей, соответствующих различным поисковым полям. Для просмотра словаря нужно в строку запроса ввести поисковый термин или его фрагмент, выбрать соответствующее поле и нажать кнопку «Словарь». На экран будет выдан фрагмент словаря из 20 терминов, начиная с введенного в строку запроса (рис. 9. 8). Отмеченные в словаре термины функцией «Возврат на поиск» могут быть помещены в строку запроса.

Поиск в каталоге БЕН

Дата обновления: 19.08.2008

Термины

тори* ядерн* цикл*

И

Слова из библиографического описания

Словарь

ИЛИ

тори* ядерн* реактор*

И

Слова из библиографического описания

Словарь

ИЛИ

ИЛИ

Слова из библиографического описания

Словарь

Порция выдаваемых документов

20

Поиск

Сброс

Тематический поиск

Уважаемые господа!
 Чтобы поиск по каталогу обеспечивал максимальную полноту и точность, ознакомьтесь, пожалуйста, с особенностями поиска по различным полям в описании каталога.

Описание каталога

Выберите раздел

[Home Page](#)

Нал / Об

319012 / 479

Рис. 9.7. Бланк поискового запроса

Выберите термины

ТОРИЕВОГО

ТОРИИ

ТОРИЙ

ТОРИКАВИЧ

ТОРИКОВ

ТОРИЧЕСКИЕ

ТОРИЧЕСКИХ

ТОРИШИМА

ТОРИЯ

ТОРКИНГТОН

ТОРЛЕЦКИЙ

ТОРЛОПОВА

ТОРЛОПОВА

ТОРМАСОВ

ТОРМОЖЕНИИ

ТОРМОЖЕНИЯ

ТОРМОЗА

ТОРМОЗНОЕ

ТОРН

ТОРНАДО

Возврат на поиск

Просмотр словаря

Вперед

Назад

Введите новый термин

Словарь

Рис. 9.8. Словарь поискового поля

235

Поиск осуществляется по следующим поисковым полям.

- *Слова из библиографического описания* – включает все слова, введенные в библиографическое описание издания, за исключением служебных и неинформативных слов (предлоги, союзы, артикли и т.п.), перечень которых определяется словарем «стоп-слов», используемым при формировании каталога.

- *Автор* – содержит не только сведения о собственно авторе издания, но и о любом лице, несущем интеллектуальную ответственность за издание (редактор, составитель и т.п.).

- *Название* – содержит полностью заглавие издания (при поиске может быть введен начальный фрагмент, последнее слово которого может иметь правое усечение).

- *Коллективный автор (организация, конференция...)* – содержит наименования организаций, коллективов, конференций и т.п.

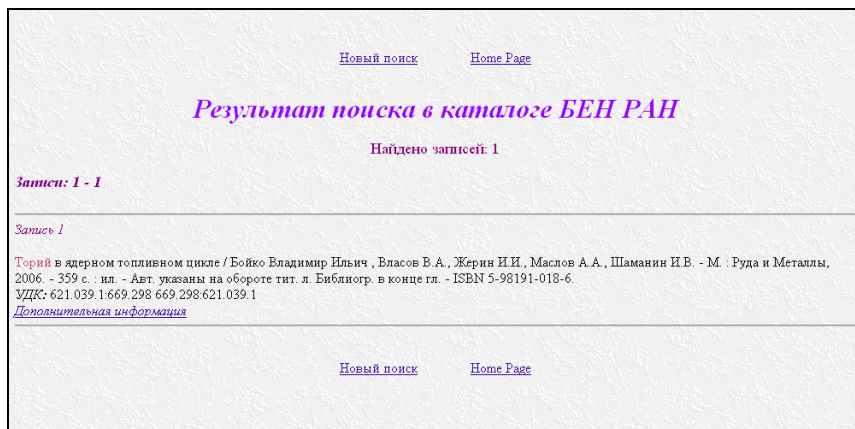


Рис. 9.9. Результат поиска по запросу, представленному на рис.9.7

Пример вывода страницы результатов поиска представлен на рис. 9.9.

Тематический поиск. Режим «Тематический поиск» позволяет искать библиографические описания по рубрикам, присвоенным изданиям в соответствии с системой Универсальной десятич-

ной классификации (УДК). При нажатии кнопки на экране раскрывается список наименований рубрик верхнего уровня, позволяющий осуществлять по гипертекстовым отсылкам переход на рубрики более низких уровней (рис. 9.10).

Для осуществления поиска необходимо отметить нужные рубрики и активизировать функцию «Поиск». Функция «Уточнение запроса» позволяет перейти на поисковую страницу, в первую строку бланка запроса которой будут внесены индексы выбранных рубрик. Во вторую и третью строки можно добавить термины, уточняющие запрос, а затем выполнить запрос, нажав на кнопку «Поиск».

Отметьте нужные рубрики

- ☐ [Астрономия](#)
- ☐ [Биология](#)
- ☐ [Ботаника](#)
- ☐ [Информатика, Кибернетика](#)
- ☐ [Математика](#)
- ☐ [Медицина](#)
- ☐ [Науки о Земле](#)
- ☐ [Науки об окружающей среде, Природоведение, Энвайронментология](#)
- ☐ [Науки об окружающей среде и антропогенных воздействиях на нее, Антропогенные воздействия на окружающую среду в целом, Зависимость окружающей среды от человека](#)
- ☐ [Общественные науки](#)
- ☐ [Палеонтология](#)
- ☐ [Природа, Охрана природы](#)
- ☐ [Психология](#)
- ☐ [Сельское хозяйство, Лесное хозяйство, Охота,](#)
- ☐ [Технические науки](#)
- ☐ [Физика, Механика](#)
- ☐ [Химия](#)

Поиск Уточнение запроса

[Новый поиск](#) [Home Page](#)

Рис. 9.10. Верхний уровень УДК

Распределенный каталог (Z39.50)

Распределенный каталог на базе библиотечного протокола Z39.50 разработан в БЕН в рамках проектов LibWeb и LibWeb 2 (на базе сервера ZooPark).

Кроме электронных ресурсов самой библиотеки по естественным наукам (по Z39.50 доступны каталог книг и продолжающихся изданий, каталог журналов, база данных “Наука России”)

пользователю предоставляется доступ к ресурсам других участников проекта LibWeb, таким как каталоги ГПНТБ, библиотеки МГУ, ЦБП, базе “Агрос” и базе журналов ЦНСХБ, а также любому другому внешнему ресурсу данного проекта.



Шлюз WWW - Z39.50

Выбор источников данных.

Каталоги и базы данных БЕН РАН.

☐ Каталог книг и продолжающихся изданий БЕН РАН.

☐ Каталог журналов БЕН РАН.

☐ База данных "Наука России".

Каталоги внешних библиотек.

☐ Каталог ГПНТБ России.

☐ Каталог библиотеки МГУ.

☐ Каталог ЦБП.

☐ База "Агрос" ЦНСХБ РАСХН.

☐ База журналов ЦНСХБ РАСХН.

1. Отметьте каталоги или базы данных, в которых будет осуществляться поиск.

2. Добавьте, при необходимости, известные Вам сервер и базу данных, отсутствующие в списке.

3. Нажмите на кнопку "Сформулировать запрос".

Введите информацию об источнике данных, если его нет в списке.

	- Сервер Z39.50
	- Порт
	- Имя БД

Сформулировать запрос

Текущая версия шлюза Z39.50 реализована в БЕН РАН с использованием программного обеспечения, разработанного ОИГГИМ СО РАН в рамках проекта LibWeb.
Контакты: varakin_v@mail.ru, var_vp@mail.ru.

© LNS RAS, 2003 г..

Рис. 9.11. Распределенный каталог Z39.50

Перед началом работы пользователю предлагается выбрать список ресурсов для поиска (рис. 9.11). Форма поиска организована в соответствии с текущей версией стандарта z39.50. От пользователя требуется указать поисковое поле, стандартный формат выдачи результата, количество записей на одной странице (рис. 9.12).

В качестве результата поиска возвращается наименование поискового ресурса и количество формально релевантных запросу документов (рис. 9.12) с возможностью дальнейшего просмотра отобранных библиографических карточек (рис. 9.13).

Поиск в отмеченных источниках данных

	Поле:	Термин:	Логическая связь:	Усечение:
	Заглавие	горий	ИЛИ	☒
	Ключевое слово	горий ядерный ре	ИЛИ	☒
	Заглавие			☐
	Формат:	Тип списка выводимых полей:	Начальный номер записи:	Количество:
	text	Полный список	1	1

LNB Выбор источников данных **LNB**

Результат поиска по запросу:

БЕН РАН: Каталог книг и продолжающихся изданий	CATALOG_LNS	1	Просмотр
БЕН РАН: Каталог журналов	serms	0	Просмотр
БЕН РАН: База данных "Наука России"	SCIENCE_LNS	0	Просмотр
ГПНТБ России: Каталог	katb	20	Просмотр
www.lib.ras.ru/210	LMSU	Error 1016	Init/AC: Blocked network address

Всего по запросу найдено: **21** записей

Рис. 9.12. Бланк запроса и поисковый результат
Распределенного каталога

Публикации сотрудников РАН (Информационно-аналитическая система «Наука России»)

Разработка информационно-аналитической системы «Наука России» ведется с 1994 г. при поддержке РФФИ.

В настоящее время для поиска доступны следующие базы данных:

- публикации библиотеки по естественным наукам РАН;
- публикации физического института РАН;
- публикации института органической химии РАН;
- публикации института астрономии РАН;
- патенты Пушинского научного центра РАН;
- публикации института общей и неорганической химии РАН;
- публикации института элементоорганических соединений РАН;
- публикации института теоретической и экспериментальной биофизики РАН;

- публикации сотрудников биологических институтов: ИПЭЭ, ИНБИ, ИНПА, ПИН.

Работа с системой начинается с выбора базы данных, в которой предполагается вести поиск (рис. 9.14).

Выбрав одну из предложенных баз данных и нажав на кнопку «Ввод», пользователь переходит на страницу выбора ресурса (*Персоналии*, *Публикации* или оба вместе – рис. 9.15).

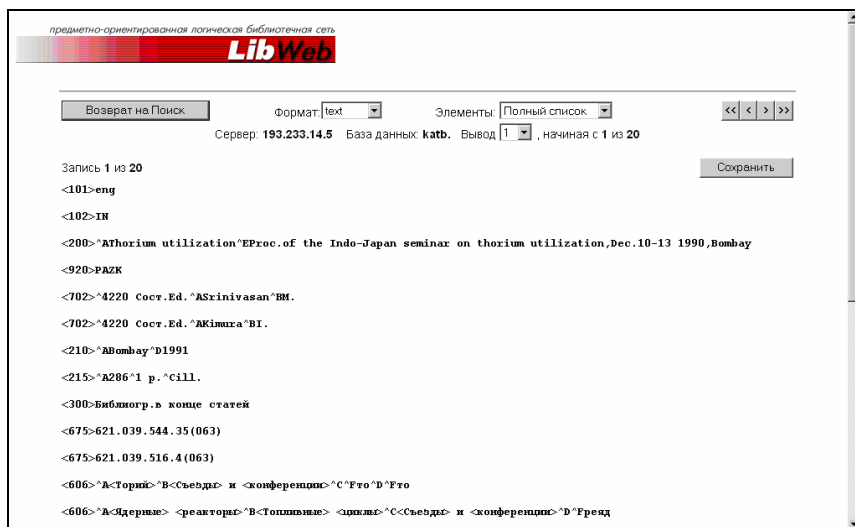


Рис. 9.13. Просмотр каталога ГПНТБ

Ресурс *Персоналии* обеспечивает возможность поиска и просмотра записей, содержащих информацию о персоналиях с последующим просмотром публикаций, авторами которых они являются.

Ресурс *Публикации* позволяет формулировать запрос по полям, отражающим сведения о публикациях (журналах, сборниках, в которых опубликованы работы) и просматривать найденные записи.

Выбор обоих ресурсов дает возможность формулировать комбинированные запросы.

Проект поддерживается РФФИ

Система "Наука России"

Выберите базу данных

Публикации Физического института РАН

ВВОД СБРОС

[Home Page](#) [Help](#)

Not found
318021 488

Рис. 9.14. Начало работы с системой «Наука России».

После выбора одного или обоих ресурсов (кнопка «Ввод»), пользователь переходит на страницу формирования запроса (рис. 9.16).

Бланк запроса на поиск позволяет вводить от одной до трёх строк терминов, соответствующих выбранным поисковым полям. В каждую строку может быть введено несколько терминов, связанных логическим оператором И (символ «#») или ИЛИ (символ «|»). Термины могут вводиться с правым усечением, признаком которого служит символ «*».

Система "Наука России"

База данных "Публикации Физического института РАН "

Ресурсы для формирования запроса

☐ Персоналии
☒ Публикации

ВВОД СБРОС

[Выход](#) [Home Page](#) [Help](#)

Рис. 9.15. Выбор ресурсов для формирования запроса

"Публикации" и "Персоналии"

Значения полей

тори*

Выбор поисковых полей

СП. НАЗВАНИЯ*

И

СП. НАЗВАНИЯ*

И

СП. НАЗВАНИЯ*

Порция выдаваемых документов

[Выход](#)
[Home Page](#)
[Help](#)

Рис. 9.16. Бланк запроса в системе «Наука России»

Строки, отнесенные к разным полям, могут соединяться логическим И, ИЛИ, НЕ. Набор поисковых полей определяется структурой выбранного ресурса и выводится в соответствующие окна строк запроса.

[Выход](#)
[Новый поиск](#)
[Home Page](#)
[Help](#)

Публикации

Найдено записей: 3

Записи: 1 - 3

Запись 1

Автор : Фок М.В.
 Шифр : I-12905
 Источник : I-12905 5-е Совещание по люминесценции (Кристаллофосфоры) 25-30 июня 1956. Тез. докл.
 Название : О послесвечении иона Би-+++ в фосфорах на основе окиси тория
 Год издания : 1956
 Страницы : С.76-77
 Язык : rus

Запись 2

Автор : Фок М.В.
 Название : О послесвечении иона Би-3+ в фосфорах на основе окиси тория
 Источник : Известия Академии наук СССР. Сер. физ.
 Год издания : 1957
 Том, (номер) : Т.21, N 4
 Страницы : С.505-506

Рис. 9.17. Результаты поиска

При вводе запроса можно установить ограничение выдачи документов на одной странице (по умолчанию порция выдаваемых на одной странице документов равна 20-и).

После ввода запроса и нажатия кнопки «Ввод» происходит поиск и на экран выводится информация о количестве найденных документов и собственно документы установленными порциями (рис. 9.17).

При работе с ресурсом *Персоналии* можно перейти по кнопке «Просмотр публикаций» к списку публикаций авторов.

9.2. Электронные ресурсы Государственной публичной научно-технической библиотеки России (ГПНТБ)

Работы по использованию сети Internet для обслуживания как читателей, пришедших в библиотеку, так и удаленных пользователей, в ГПНТБ активно ведутся с 1994 г..

В настоящее время на стартовой странице сайта библиотеки – <http://www.gpntb.ru> (рис. 9.18) – размещены ссылки на электронные ресурсы – *Имидж-каталог* и *Электронный каталог*.

Имидж-каталог

Электронный ресурс обеспечивает доступ к ретроспективным базам данных каталожных карточек изданий 1984–1990 г.г. В каталоге представлены следующие БД ретрофонда ГПНТБ:

- «Отечественная книга»;
- «Зарубежная книга»;
- «Отечественная периодика»;
- «Зарубежная периодика».

Поисковый запрос формируется как последовательность терминов, в общем случае соединяемых логическим оператором «ИЛИ» (рис. 9.19). *Дополнительные параметры* позволяют задавать требуемые расстояния между терминами запроса (для определения формальной релевантности каталожной карточки) и ограничение постраничной выдачи.

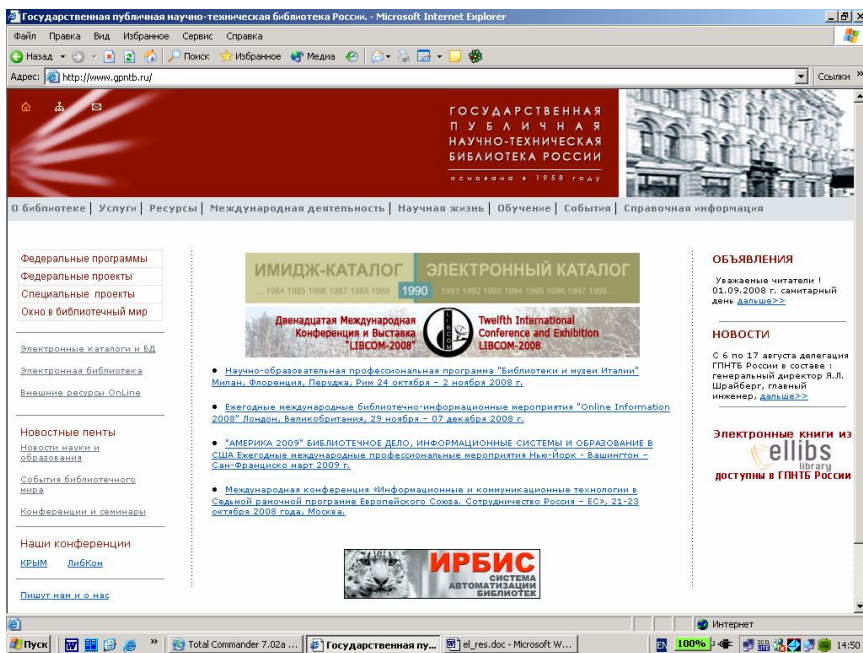


Рис. 9.18 . Стартовая страница сайта ГПНТБ

Кнопка «Искать» инициирует поисковый процесс и выдачу первой порции отобранных документов. На странице размещаются материалы документов выбранной при вводе поискового запроса БД, однако поиск проводится по всем доступным базам, а к результатам по другим БД можно перейти по соответствующим ссылкам (рис. 9.20).

Активизировав ссылку *Показать каталожные ящики*, можно перейти в режим просмотра содержимого каталожных ящиков (рис. 9.21).

Ссылка *Поиск по тематическому навигатору* позволяет использовать для получения результата иерархический тематический рубрикатор (рис. 9.22). Отсылки от размещенных на странице рубрик ведут к рубрикам, нижестоящим по иерархии. Найти рубрику по терминам из ее заглавия можно с помощью бланка запроса, размещенного справа от списка рубрик.

ИРБИС64 ГПНТБ России Поиск в полнотекстовых базах данных

ИРБИС64 ПОЛНОТЕКСТОВЫЕ БАЗЫ ДАННЫХ

База данных:

Запрос:

Дополнительные параметры Показать каталожные ящики Поиск по тематическому навигатору

Расстояние между словами: Выдавать по документов

слов в рядом
 1 слово
 2 слова
 3 слова
 5 слов
 6 слов
 7 слов
 8 слов
 9 слов
 10 слов

Copyright © 2006 ИРБИС64 Полнотекстовые базы данных. Все права защищены.

Рис. 9.19 . Запрос на поиск в Имидж-каталоге

Запрос:

☐ в найденном

Дополнительные параметры Показать каталожные ящики Поиск по тематическому навигатору

Расстояние между словами: Выдавать по документов

Найдено в БД: [Зарубежная книга](#) [Отечественная периодика](#) [Зарубежная периодика](#)

Запрос: торий ядерный реактор
 Общее количество найденных документов : 5
 1 ... 5

3.

Ар8764-22302

Цуриков Д.Ф.

Использование усовершенствованных
 быстрых реакторов для вовлечения горю-
 чих ресурсов в ядерную энергетику: авто-
 реф. дис. ... канд. техн. наук: 05.14.03. - М.,
 1987. - 19 с. - В надзат. : Ил.-т. атом. энергии
 им. И.В. Курчатова. Библиогр. : с. 18-19 (13
 назв.). Для служ. пользования. Экз. №...
 64.039.576.4 (043)

☐ Показать оборот каталожной карточки

Рис. 9.20. Результаты поиска в Имидж-каталоге

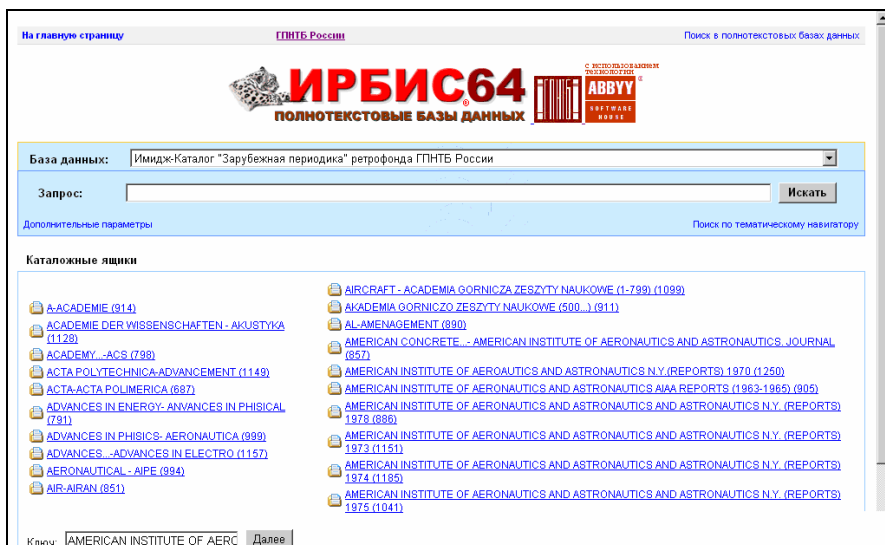


Рис. 9.21. Режим просмотра каталожных ящиков



Рис. 9.22. Поиск по тематическому рубрикатору

Бланк запроса предусматривает возможность ввода нескольких терминов, соединяемых одним из трех логических операторов – «И», «ИЛИ», «Фраза целиком». Окончания слов при этом можно учитывать или не учитывать, включив соответствующий флаг.

Результат поиска содержит список отобранных рубрик, отсылки от которых позволяют получить рубрикационное дерево (рис. 9.23). Для дальнейшего поиска и отображения каталожных карточек необходимо отметить нужные рубрики и активизировать кнопку «Поиск», расположенную внизу списка рубрик.

На главную страницу ГПНТБ России Поиск в полнотекстовых базах данных

ИРБИС64 ПОЛНОТЕКСТОВЫЕ БАЗЫ ДАННЫХ

База данных: Имидж-Каталог "Отечественная книга" ретрофонда ГПНТБ России

Тематический рубрикатор

Вышестоящий: [АТОМНЫЕ ОБЪЕКТЫ](#)

Рубрика: ☒ ЯДЕРНЫЕ РЕАКТОРЫ

Нижестоящие: ☒ РЕАКТОРЫ РАЗНОЖИТЕЛИ

Смотри также: ☐ РАДИОАКТИВНЫЕ ВЕЩЕСТВА

Поиск по отмеченным в текущей базе данных:

Поиск рубрик

Логика:

Окончания слов: ☒ не учитывать / ☐ учитывать

Copyright © 2006 ИРБИС64 Полнотекстовые базы данных. Все права защищены.

Рис. 9.23. Фрагмент дерева для рубрики «Ядерные реакторы»

Электронный каталог

Страница *Электронный каталог* открывает доступ к следующим ресурсам (базам данных).

- *Электронный каталог ГПНТБ России* содержит библиографические описания всех изданий, поступающих в фонд ГПНТБ России: авторефератов диссертаций (поступления с 1991 г.); неопубликованных переводов (поступления с 1992 г.); зарубежных

книг (поступления с 1990 г.); зарубежной периодики (поступления с 1993 г.); отечественных книг (поступления с 1990 г.); отечественных журналов (поступления с 1994 г.). Общее число записей Электронного каталога на конец 2007 г. – 480 тыс.; ежемесячное пополнение – около 3 тыс. записей.

- *Новые поступления в обработку (книги)* содержит сведения обо всех книгах, поступающих в фонд ГПНТБ, но еще не попавших в Электронный Каталог.

- *Авторефераты диссертаций.*

- *Российский сводный каталог по научно-технической литературе* содержит сведения о зарубежных и отечественных книгах и периодических изданиях по естественным наукам, технике, сельскому хозяйству, медицине, экологии, бизнесу и праву, поступивших в организации – участницы Автоматизированной системы Российского сводного каталога по научно-технической литературе (АС РСВК НТЛ). Объем базы данных сводного каталога на май 2008 г. составляет 763 000 записей (с ретроспективой), ежегодное пополнение – около 30 тыс. записей по всем видам изданий.

- *Каталог «Раритетные издания ГПНТБ России»* содержит 2837 библиографических описаний документов 17, 18 и 19 веков по математике, физике, естественным и точным наукам на русском, английском и немецком языках.

- *Каталог «Экология: ресурсы Интернета»* содержит около 500 снабженных библиографическими описаниями ссылок на официальные источники информации по экологии, представленные в Internet.

- *Каталог «Экология: наука и технология»* содержит библиографические описания книг, малотиражных изданий, статей более чем из 500 периодических изданий, имеющихся в фондах ГПНТБ России (около 12 тыс. библиографических записей).

- *Каталог электронных ресурсов, сериальных, книжных и справочных изданий по науке и технике* содержит сведения об электронных ресурсах, издательствах и отдельных изданиях (журналы, книги, справочные издания), доступных в электронном виде.

Рабочее пространство Электронного каталога разбито на три области: две из них (правая и левая) – управляющие, центральная область – для получения и просмотра результата (рис. 9.24).

Главная | Документация | WEB IRBIS 8.1 | Шлюз Z39.50

Государственная публичная научно-техническая библиотека России

Базы данных

- Электронный Каталог ГПНТБ
- Новые поступления в обработку (книги)
- Авторефераты диссертаций
- Российский сводный каталог по научно-технической литературе
- Раритетные издания
- Экология: ресурсы Интернета
- Экология: наука и технология
- Каталог электронных ресурсов, сериалов, фильмов

Электронный Каталог ГПНТБ - стандартный поиск

Ключевые слова

Искать

Поступления за 2008 год

2008		
Декабрь	Ноябрь	Октябрь
Сентябрь	Август	Июль
Июнь	Май	Апрель
Март	Февраль	Январь

Правила составления запроса при поиске

Выбрав необходимую область поиска (ключевые слова, автор, заглавие, год издания), введите запрос в поле ввода:

- При поиске по **ключевым словам** - фразу на естественном языке. Ключевые слова формируются практически из всех полей библиографического описания. Фраза разбирается на отдельные слова (цифры не учитываются!), которые морфологически нормализуются (только для русских слов!). При поиске по

Вид поиска

- Описание Базы данных
- Стандартный
- Расширенный
- Профессиональный
- По словарю
- ГРНТИ-навигатор
- УДК-навигатор
- ББК-навигатор
- Тематический навигатор
- Статистика обращений

Рис. 9.24. Рабочее пространство Электронного каталога

Главная | Документация | WEB IRBIS 8.1 | Шлюз Z39.50

Государственная публичная научно-техническая библиотека России

Базы данных

- Электронный Каталог ГПНТБ
- Новые поступления в обработку (книги)
- Авторефераты диссертаций
- Российский сводный каталог по научно-технической литературе
- Раритетные издания
- Экология: ресурсы Интернета
- Экология: наука и технология
- Каталог электронных ресурсов, сериалов, фильмов

Российский сводный каталог по научно-технической литературе - расширенный поиск

Расширенный поиск по ключевым словам:

Ключевые слова: торий ядерный реактор

в любом поле

логика: И

окончания слов: ☒ не учитывать / ☐ учитывать

Уточняющие поисковые элементы (объединяются логикой "И"):

Тематика поиска: Энергетика

Автор:

Вид издания:

ISSN/ISBN:

СИГЛА:

Год издания:

Формат показа: полный формат (новый ГОСТ)

Искать Обновить

Правила составления запроса при поиске

Вид поиска

- Описание Базы данных
- Стандартный
- Расширенный
- Профессиональный
- По словарю
- ГРНТИ-навигатор
- УДК-навигатор
- ББК-навигатор
- Тематический навигатор
- Статистика обращений

Рис. 9.25. Бланк запроса расширенного поиска

В левой управляющей области представлены пиктограммы, содержащие наименования доступных баз данных. Пиктограмма текущей БД выделена альтернативным цветом.

В правой управляющей области перечислены виды поисковых стратегий, обеспечивающих в совокупности различные способы доступа к содержимому электронных ресурсов:

- описание базы данных;
- поиск – стандартный, расширенный, профессиональный;
- вход в БД через словарь;
- вход в БД через классификационные системы – ГРНТИ, УДК, ББК, Тематический рубрикатор.

Виды поиска. Для проведения поиска в электронных ресурсах предлагается три формы поисковых запросов. *Стандартный* поиск позволяет в бланке запроса (рис. 9.24) выбрать область поиска из фиксированного списка (ключевые слова, автор, заглавие, год издания) и ввести через пробел термины запроса. При поиске по ключевым словам применяется алгоритм ранжирования найденных документов: результаты поиска сортируются в порядке убывания ранга документа. Ранг документа тем выше, чем больше в найденном документе терминов из запроса и чем эти термины ближе друг к другу.

Бланк запроса для *расширенного* поиска позволяет сконструировать запрос, задав условия поиска для нескольких полей. При этом одно поле выделяется для ввода ключевых слов (с отсечением морфологического окончания или без), которые могут быть объединены одним из логических операторов («И», «ИЛИ», «НЕ»), а условия поиска для уточняющих поисковых полей (*Тематика поиска, Вид издания, Автор* и т.д.) добавляются в поисковый запрос с помощью логического оператора «И» (рис. 9.25).

Режим *профессионального* поиска предназначен для ввода булевоподобного выражения запроса в соответствии с приоритетом логических операций и правилами булевой алгебры (рис. 9.26). Выражение запроса последовательно формируется в окне «Комплексный поисковый запрос» с помощью окон ввода «Область поиска» (выбор имени поля) и «Поисковый запрос» (задание терминов). При этом термины можно объединить в поисковое выражение

с помощью операторов поиска («\$» – правое усечение; «*» – логическое И, «^» – логическое НЕ, «+» – логическое ИЛИ).

По команде «Искать» в любом из поисковых режимов выполняется поиск документов в соответствии с введенным поисковым выражением. Отображение результатов поиска представлено на рис. 9.27. Результаты могут быть отсортированы по значению одного из полей: *Автор, Заглавие, Год издания, Тип документа*.

Использование словарей. Поиск документов можно проводить посредством словарей, сформированных для полей *Ключевые слова, Автор, Заглавие, Год издания* (рис. 9.28).

Термины словаря оформлены как поисковые гиперссылки, по которым можно перейти на документы, содержащие термин. К сожалению, число, стоящее слева от термина, показывает не количество документов, содержащих данный термин, а число поисковых ссылок.

The screenshot shows the Web IRBIS search interface. At the top, there is a navigation bar with links to 'Главная', 'Документация', 'WEB IRBIS 8.1', and 'Шаблоны Z39.50'. The main header features the 'Web IRBIS' logo and the text 'Государственная публичная научно-техническая библиотека России'. Below this, a subtitle reads 'Российский сводный каталог по научно-технической литературе - профессиональный поиск'.

The central search form is titled 'Поисковые поля, объединенные различной логикой'. It includes several input fields and buttons:

- 'Область поиска:' with a dropdown menu set to 'Ключевые слова'.
- 'Усечение:' with a checked checkbox.
- 'Формат представления:' with a dropdown menu set to 'полный формат (новый ГОСТ)'.
- 'Поисковый запрос:' with an empty text input field.
- 'Операторы присоединения:' with buttons for 'И', 'ИЛИ', and 'НЕТ'.
- 'Комплексный поисковый запрос:' with a text input field containing the example query: '<.>K=ториз<.>*(<.>K=ядериз<.>*<.>K=реактор\$<.>'. Below this input are 'Искать' and 'Обновить' buttons.

Below the search form, there is a section titled 'Правила составления запроса при поиске'. It contains two paragraphs of text explaining the search operators and the use of the '\$' symbol for term truncation. The first paragraph states that the 'Операторы присоединения' buttons allow for creating complex queries. The second paragraph explains that the '\$' symbol is used to truncate terms at the end.

On the left side of the interface, there is a sidebar with various database links:

- 'Базы данных' (Databases)
- 'Электронный каталог ГИИО' (Electronic catalog of GIIO)
- 'Новые поступления в обработку (онлайн)' (New arrivals for processing (online))
- 'Авторефераты диссертаций' (Abstracts of dissertations)
- 'Российский сводный каталог по научно-технической литературе' (Russian consolidated catalog of scientific and technical literature)
- 'Раритетные издания' (Rare editions)
- 'Экология: ресурсы Интернета' (Ecology: Internet resources)
- 'Экология: наука и технология' (Ecology: science and technology)
- 'Каталог электронных ресурсов, сериалов' (Catalog of electronic resources, serials)

On the right side, there is another sidebar with search options:

- 'Вид поиска' (Search type) with buttons for 'Описание базы данных', 'Стандартный', 'Расширенный', 'Профессиональный', 'По словарю', 'ГРНТИ-навигатор', 'УДК-навигатор', 'ББК-навигатор', 'Тематический навигатор', and 'Статистика обращений'.

Рис. 9.26. Бланк запроса профессионального поиска



Рис. 9.27. Отображение результатов поиска

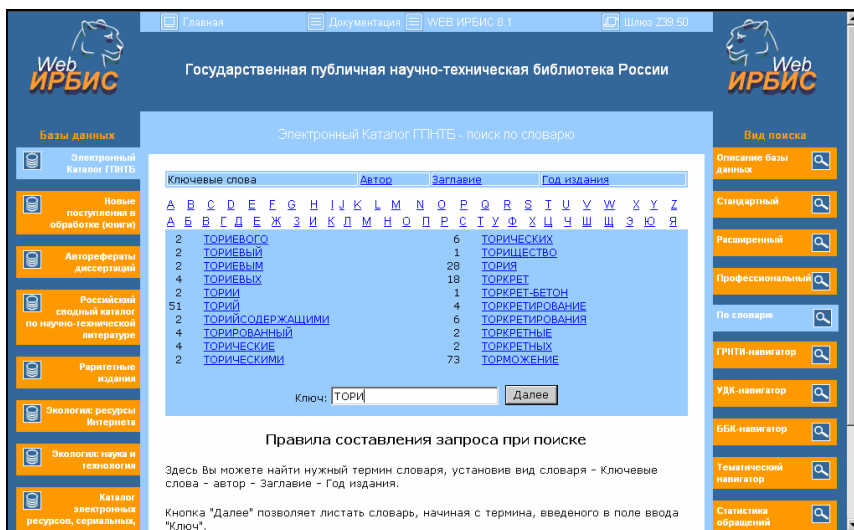


Рис. 9.28. Поиск по словарю

Использование классификаций. Для поиска документов предусмотрен режим входа в информационные ресурсы с помощью классификационных систем – УДК, ББК, ГРНТИ, Тематический рубрикатор (описание классификационных систем см. в главе 3). Рубрики классификаций снабжены гипертекстовыми отсылками, позволяющими перемещаться по графу классификации (на рис. 9.29 представлен верхний уровень УДК).

Для поиска нужной рубрики по ключевым словам можно использовать привычный бланк запроса, позволяющий ввести термины (при желании – с признаком усечения окончаний) и объединить их одной из предлагаемых логических операций.

Для перехода к просмотру документов необходимо отметить нужные рубрики и активизировать поиск с помощью кнопки «Искать» в нижней части страницы (рис. 9.30).

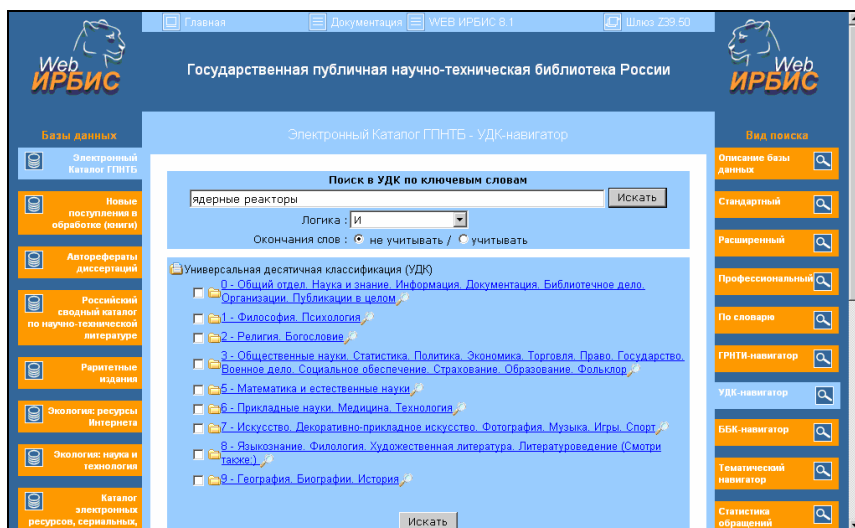


Рис. 9.29. Верхний уровень УДК

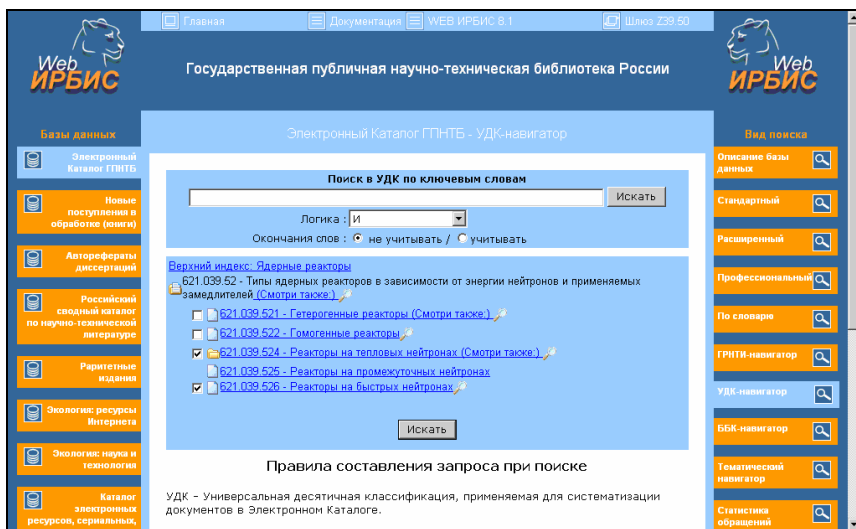


Рис. 9.30. Поиск документов с использованием УДК

Каталог электронных ресурсов сериальных, книжных и справочных изданий по науке и технике

Доступ к каталогу отличается от общих правил работы с базами данных и осуществляется с помощью поисковой страницы, представленной на рис. 9.31.

Вход в каталог с поисковой страницы возможен в двух режимах: просмотр заглавий в алфавитном порядке (в этом случае заглавия оформляются как гипертекстовые отсылки к библиографическим описаниям) либо поиск документов с использованием бланка запроса.

Бланк запроса, помимо ввода ключевых слов, поиск которых ведется только в заглавии либо во всем документе, позволяет выбрать тематику поиска и задать параметры вывода результатов (формат показа и размер порции). На рис. 9.32 представлены результаты поиска в полном формате.

Шлюз Z39.50

Страница Web ИРБИС в режиме Электронного каталога открывает на базе библиотечного протокола Z39.50 доступ к **Корпоративной сети Московских библиотек – КСМБ** (рис. 9.33).

Посредством режимов *быстрого* или *расширенного* поиска в распределенном каталоге можно получить сведения о наличии необходимых материалов в присоединенных ресурсах (рис. 9.34, 9.35).

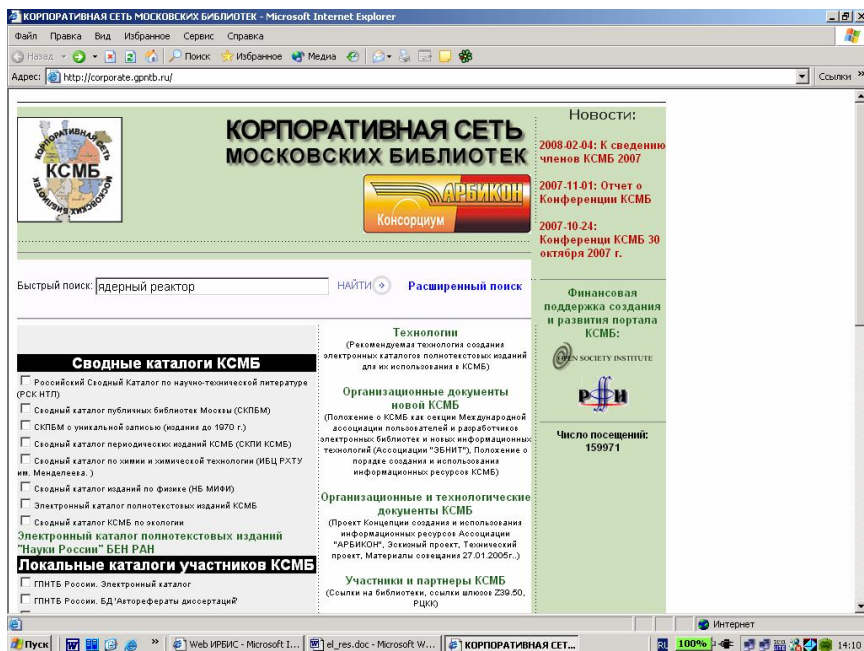


Рис. 9.33. Стартовая страница Корпоративной сети Московских библиотек

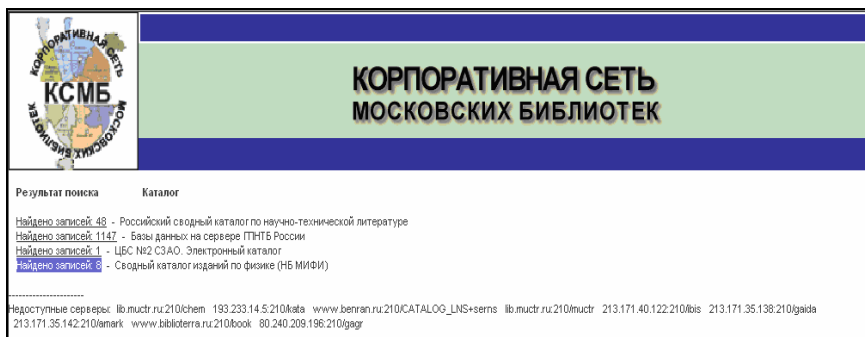


Рис. 9.34. Результаты поиска в КСМБ

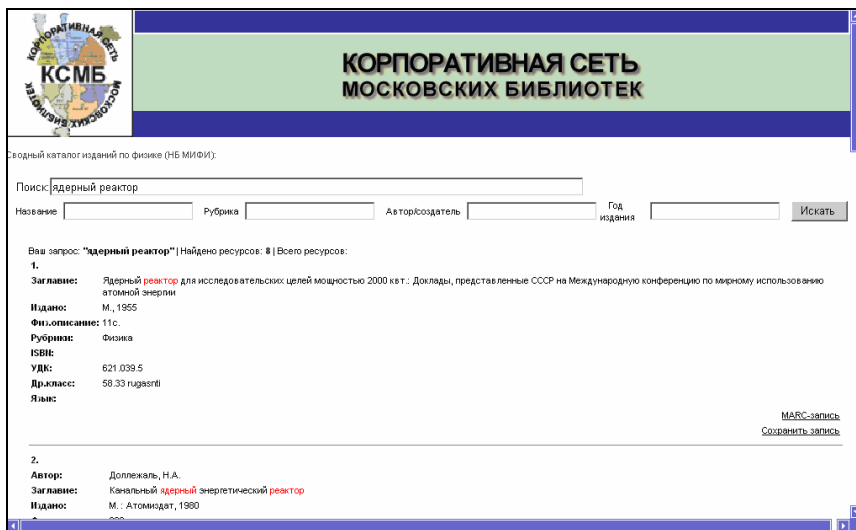


Рис. 9.35. Поиск в Сводном каталоге изданий по физике (НБ МИФИ)

9.3. Электронные ресурсы Библиотеки Конгресса США

Доступ к каталогам и базам данных Библиотеки Конгресса США (<http://www.loc.gov>) осуществляется со стартовой страницы сайта Библиотеки с помощью управляющей кнопки «Library Catalogs» (рис. 9.36).

На форме работы с каталогами (рис. 9.37) можно выбрать один из двух режимов поиска: *Простой поиск (Basic search)* и *Расширенный поиск (Guided search)*.

Простой поиск (Basic search). Бланк поискового запроса режима Простого поиска дает возможность сформировать поисковое выражение в виде последовательности терминов, отнесенных к одному из возможных полей документа.

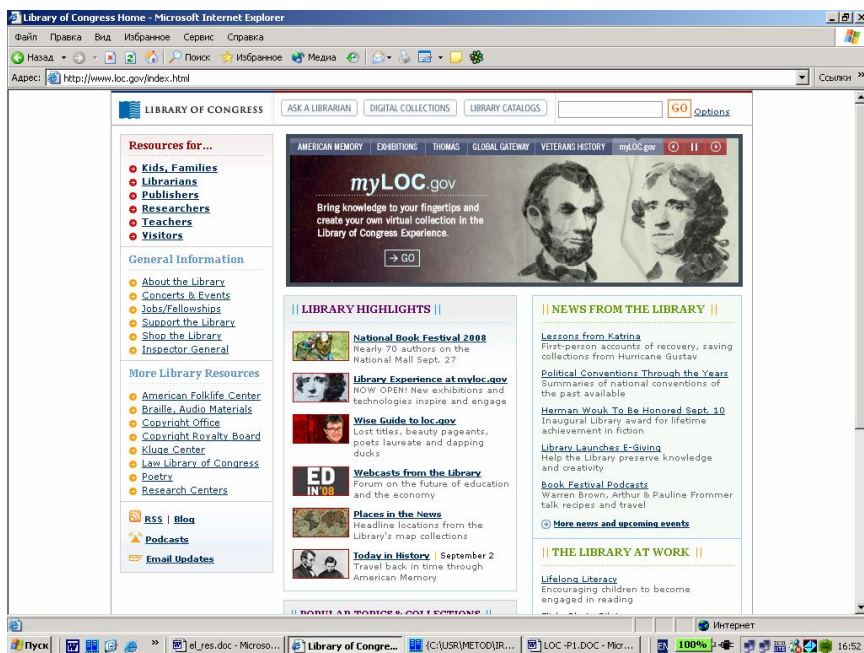


Рис. 9.36. Стартовая форма сайта Библиотеки Конгресса США

Основными поисковыми полями являются следующие.

- *Заглавие (Title)* – предполагает поиск документа по заглавию (либо его части, начиная с первого слова). Введенное в область «Search Text» значение для этого поля усекается справа автоматически.
- *Автор/создатель (Author/Creator Browse)* – поле включает не только непосредственно автора (создателя работы), но и редакторов, исполнителей, фотографов, иллюстраторов, официальные названия юридических лиц, органов власти, конференций и т.д.
- *Рубрика (Subject Browse)* – предполагает поиск документа по названию рубрики (либо его части, начиная с первого слова). Введенное в область «Search Text» значение для этого поля усекается справа автоматически.
- *Ключевые слова (Key Words)* – поле содержит термины, индексирующие тематическую направленность документа.

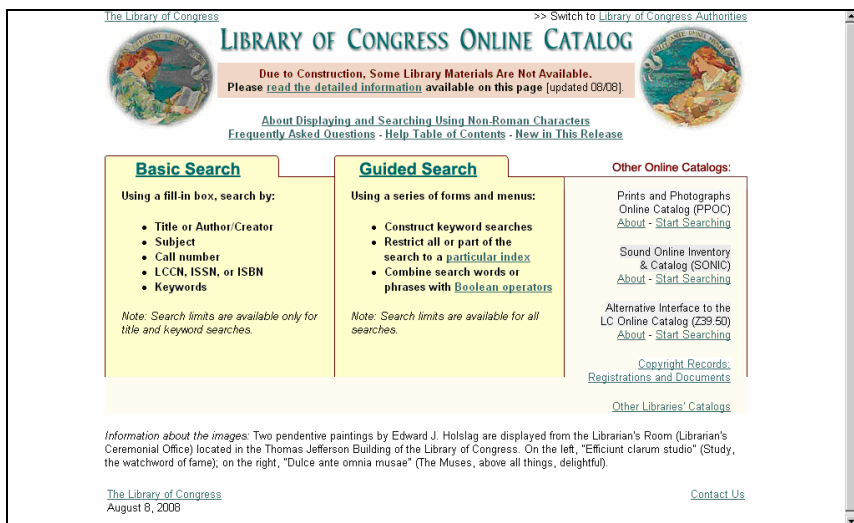


Рис. 9.37. Поиск в каталогах и базах данных

Для поиска необходимо ввести одно или несколько поисковых слов в строку поиска «Search Text», затем выбрать из списка тип поискового поля и количество документов результата, выводимых на страницу (25, 50, 75, 100). По нажатию кнопки «Begin Search» запрос передается поисковой машине (рис. 9.38).

В зависимости от типа поискового поля результат будет выведен либо в виде списка заглавий (например, для полей *Автор/создатель* и *Предметная область*), либо в виде списка библиографических карточек (например, для полей *Шифр*, *LCCN-ISBN-ISSN*, *Ключевые слова*). Выдача результата при поиске по ключевым словам содержит колонку уровня релевантности документа (для использования сортировки по релевантности) (рис. 9.39).

Поисковые результаты могут быть отсортированы также по значениям таких полей как *Полное заглавие*, *Основной автор*, *Дата*.

The screenshot shows the Library of Congress Online Catalog search interface. At the top, there's a navigation bar with links like 'Help', 'Search', 'Search History', 'Headings List', 'Titles List', 'Request an Item', 'Account Status', and 'Start Over'. Below this, the 'DATABASE: Library of Congress Online Catalog' is noted. A yellow banner states: 'Your search found no results. Refer to search examples, check spelling and punctuation, or try another type of search.' The 'Basic Search' tab is active. The 'Search Text' field contains 'nuclear reactor'. The 'Search Type' dropdown menu is open, showing a list of search options: 'keyword (match all words)*', 'Title Keyword*', 'Author/Creator Keyword', 'Subject Keyword', 'Name/Title Keyword', 'Series/Uniform Title Keyword', 'Expert Search (use index codes and operators)*', 'Title Begins With (omit initial article)*', 'Author/Creator Browse (enter last name first)', 'Subject Browse', 'Call Number Browse (LC Classification No.)', 'Call Number Browse (Other Shelving Nos.)', 'Number Search (LCCN-ISBN-ISSN)*', and 'Author/Creator Sorted by Title Browse'. At the bottom, there's a '100 records per page' selector, a 'Begin Search' button, a 'Clear Search' button, and a 'Set Search Limits' button.

Рис. 9.38. Бланк запроса режима *Простого поиска*

The Library of Congress

>> Go to Library of Congress Authorities

LIBRARY OF CONGRESS ONLINE CATALOG

[Help](#)

[New Search](#)

[Search History](#)

[Headings List](#)

[Titles List](#)

[Request an Item](#)

[Account Status](#)

[Start Over](#)

DATABASE: Library of Congress Online Catalog

YOU SEARCHED: Title Keyword = nuclear reactor

SEARCH RESULTS: Displaying 1 through 100 of 428.

< Previous 1 101 201 301 401 Next >

Resort results by: Date (Descending)

[Add Limits to Search Results](#)

#	Relevance	Name: Main Author, Creator, etc.	Full Title	Date
☐ [1]	●●●●●	Keilbach, K. B. (Kimberly B.), 1960-	Global warming is good for business : how savvy entrepreneurs, large corporations, and others are making money while saving the planet / by K.B. Keilbach.	2008
LIBRARY OF CONGRESS HOLDINGS INFORMATION NOT AVAILABLE				
☐ [2]	●●●●●	Lewis, E. E. (Elmer Eugene), 1938-	Fundamentals of nuclear reactor physics / E.E. Lewis.	2008
SELECT TITLE FOR HOLDINGS INFORMATION				
☐ [3]	●●●●●	Koch, Leonard J.	Experimental breeder reactor-II : (ebr-ii) an integrated experimental fast reactor nuclear power station / Leonard J. Koch.	2008
LIBRARY OF CONGRESS HOLDINGS INFORMATION NOT AVAILABLE				
		Kyōto Daigaku. Genshiro Jikkenjo.	Kyōto Daigaku Genshiro Jikkenjo dai 15-kai Genshiro Hōshasen Gijutsu Kenshūkai. Kyōto Daigaku Sōgō	

Рис. 9.39. Таблица поисковых результатов

The Library of Congress	>> Go to Library of Congress Authorities
LIBRARY OF CONGRESS ONLINE CATALOG	
Help New Search Search History Headings List Titles List Request an Item Account Status Start Over	
DATABASE: Library of Congress Online Catalog YOU SEARCHED: Title Keyword = nuclear reactor SEARCH RESULTS: Displaying 1 of 428.	
< Previous Next >	
Brief Record Subjects/Content Full Record MARC Tags	
<i>Global warming is good for business : how savvy entrepreneurs, large...</i>	
Relevance: ●●●●●	
LC Control No.: 2008033864	
LCCN Permalink: http://lccn.loc.gov/2008033864	
Type of Material: Book (Print, Microform, Electronic, etc.)	
Personal Name: Keilbach, K. B. (Kimberly B.), 1960-	
Main Title: Global warming is good for business : how savvy entrepreneurs, large corporations, and others are making money while saving the planet / by K.B. Keilbach.	
Published/Created: Fresno, CA : Quill Driver Books, 2008.	
Description: p. cm.	
Projected Publication Date: 0809	
ISBN: 9781884956881 1884956882	
Links: Publisher description	

9.40. Просмотр библиографической карточки (краткое описание)

SEARCH RESULTS: Displaying 1 of 428.

◀ Previous Next ▶

[Brief Record](#)
 [Subjects/Content](#)
 [Full Record](#)
 [MARC Tags](#)

Global warming is good for business : how savvy entrepreneurs, large...

Relevance: ●●●●●●

LC Control No.: 2008033864

LCN Permalink: <http://lcn.loc.gov/2008033864>

Type of Material: Book (Print, Microform, Electronic, etc.)

Personal Name: [Keilbach, K. B. \(Kimberly B.\), 1960-](#)

Main Title: Global warming is good for business : how savvy entrepreneurs, large corporations, and others are making money while saving the planet / by K.B. Keilbach.

Published/Created: Fresno, CA : Quill Driver Books, 2008.

Projected Publication Date: 0809

Description: p. cm.

ISBN: 9781884956881
1884956882

Contents: The LEAFHouse -- The power mix : wind & waves -- Paradigm for the future -- The hydrogen highway -- Eco-capitalists -- The nuclear reactor in our backyard -- The grandchildren's test -- The earth's heat -- Uncharter waters -- Here comes the sun -- Wind power : a pioneering technology -- Save the planet : use more trees -- Not your father's automobile -- From railroads to rapid transit -- Green buildings -- The moral equivalent of war -- The energy star -- Stewardship in the 21st century -- Sustainable systems -- From big oil to big wind -- Indigenous energy -- Think globally, act locally -- From stumptown to sustainable -- Smart growth.

Subjects: [Global warming --Economic aspects --United States.](#)
[Climatic changes --Economic aspects --United States.](#)

LC Classification: QC981.8.G56 K45 2008

Dewey Class No.: 363.738/74 22

Geographic Area Code: n-US---

9.41. Просмотр библиографической карточки (полное описание)

The Library of Congress >> [Go to Library of Congress Authorities](#)

LIBRARY OF CONGRESS ONLINE CATALOG

[Help](#)
 [Search](#)
 [Search History](#)
 [Headings List](#)
 [Titles List](#)
 [Request an Item](#)
 [Account Status](#)
 [Start Over](#)

DATABASE: Library of Congress Online Catalog

[Basic Search](#)
 [Guided Search](#)

Search: all of these ▾ As: ▾

☒ AND
 ☐ OR
 ☐ NOT

Search: any of these ▾ As: ▾

☒ AND
 ☐ OR
 ☐ NOT

Search: all of these ▾ As: ▾

100 records per page ▾ Begin Search Clear Search Set Search Limits

Guided Search Tips *Search Limits Are Available!*

Construct your search using the fill-in blanks and menus on the previous page:

1. **Optional** -- Set [Search Limits](#) before following the steps below.
2. **Search** -- Enter a word, several words, or a phrase in the box.
3. From the following drop-down menu, select **all of these**, **any of these** or as a phrase ("all of these" is the default, and is highly recommended).

9.42. Бланк запроса *Расширенного поиска*

При активизировании гиперссылки-заглавия открывается подробная библиографическая карточка, которую можно посмотреть в одном из четырех режимов:

- краткое описание (рис. 9.40);
- разделы/содержание;
- подробное описание (рис. 9.41);
- поля в соответствии со стандартом MARC.

Расширенный поиск (Guided Search). Бланк запроса *Расширенного поиска* дает возможность сформировать поисковые условия для трех различных областей поиска и объединить их логическими операциями И, ИЛИ, НЕ (рис. 9.42).

The Library of Congress
LIBRARY OF CONGRESS ONLINE CATALOG

Help New Search Search History Headings List Titles List Request an Item Account Status Start Over

DATABASE: Library of Congress Online Catalog
YOU SEARCHED: Guided Search = (thorium nuclear)[in Keyword Anywhere]AND(reactor cycle)[in Keyword Anywhere]
SEARCH RESULTS: Displaying 1 through 5 of 5.

◀ Previous Next ▶

Resort results by: [Date (Descending)] Add Limits to Search Results

#	Name: Main Author, Creator, etc.	Full Title	Date
☐ [1]	IAEA Technical Meeting on Fissile Material Management Strategies for Sustainable Nuclear Energy (2005 : Vienna, Austria)	Fissile material management strategies for sustainable nuclear energy : proceedings of an IAEA Technical Meeting on Fissile Material Management Strategies for Sustainable Nuclear Energy, held in Vienna, 12-15 September 2005.	2007
SELECT TITLE FOR HOLDINGS INFORMATION			
☐ [2]	Maroncelli, James M.	Traveler's guide to nuclear weapons [electronic resource] : a journey through America's cold war battlefields / James M. Maroncelli & Timothy L. Karpin.	2002
ACCESS: See Reference Staff. By Appt in Jefferson Main RR (MRC)		CALL NUMBER: U264 .3 [2005 01641]	
☐ [3]	Indo-Japan Seminar on Thorium Utilization (1990 : Bombay, India)	Thorium utilization : proceedings of the Indo-Japan Seminar on Thorium Utilization, December 10-13, 1990, Bombay, India / edited by Mahadeva Srinivasan ... [et al.] : sponsored by Indian Nuclear Society and Atomic Energy Society of Japan.	1991

Буфер обмена (24 из 24)

9.43. Просмотр результатов поиска

Отдельное поисковое условие представляет собой один или несколько терминов, связываемых либо логической операцией И («all of these»), либо логической операцией ИЛИ («any of these»), либо воспринимаемых как словосочетание («as a phrase»). Результаты поиска представлены на рис. 9.43. Если объем выдачи превышает 10 000 документов, то выдается сообщение об ошибке и пре-

дупреждение, что будут показаны только первые 10 000 релевантных документов.

The screenshot shows the 'LIBRARY OF CONGRESS ONLINE CATALOG' interface. At the top, there are navigation links: 'The Library of Congress', '>> Go to Library of Congress Authorities', and a menu bar with 'Help', 'New Search', 'Search History', 'Headings List', 'Titles List', 'Request an Item', 'Account Status', and 'Start Over'. Below the menu bar, it says 'DATABASE: Library of Congress Online Catalog'. A note states: 'When Adding Limits to Search Results, select all limits at once, then select Set Search Limits button to apply.' The main form is titled 'SET SEARCH LIMITS (select labels for context-sensitive help)'. It contains several fields with dropdown menus: 'Date' (set to 2000), 'Language' (with options Romanian, Hindi, Russian), 'Type' (with options Cartographic Material (Manuscript), Computer File, Manuscript Material), 'Location' (with options Business Reference Collection, Children's Literature Reference Collection, European), and 'Place of Publication' (with options Russia (Federation), Russian S.F.S.R. [used before 6/1992] -- see also: Russia (Federation), Rwanda). At the bottom of the form are buttons for 'Set Search Limits' and 'Clear Search Limits'. Below the form is a navigation bar with links: 'Help - Search - History - Headings - Titles - Request - Account - Exit'.

Рис. 9.44. Установка дополнительных ограничений для поисковой сессии

Установка дополнительных ограничений. С помощью кнопки «Set Search Limits» можно установить дополнительные ограничения на проведение поиска. Применение этих ограничений возможно как перед началом поиска, так и в выданных документах (рис. 9.44).

Длительность поисковой сессии не может превышать 5 мин (обратный счетчик расположен в нижнем левом углу страницы, а последние 60 секунд отсчитываются во всплывающем окне). По истечении этого времени выводится сообщение об окончании сессии (рис. 9.45).

Протокол поисковой сессии. На протяжении поисковой сессии ведет протокол («Search History») с возможностью просмотра и редактирования результатов предыдущего поиска (рис. 9.46).

The Library of Congress
>> Go to Library of Congress Authorities

Warning: Your session will end soon. Press "Enter" or click "Refresh Session" to continue.
Session timeout: 0 minutes, 49 seconds
Refresh Session

LIBRARY OF CONGRESS ONLINE CATALOG

Search Search History Headings List Titles List Request an Item Account Status Start Over

Basic Search Guided Search

Search: thorium nuclear all of these As: Keyword Anywhere (GKEY)
AND OR NOT
Search: reactor cycle all of these As: Keyword Anywhere (GKEY)
AND OR NOT
Search: all of these As: Keyword Anywhere (GKEY)
100 records per page Begin Search Clear Search Set Search Limits

Guided Search Tips Search Limits Are Available!

Construct your search using the fill-in blanks and menus on the previous page:

1. Optional -- Set Search Limits before following the steps below.
2. Search: -- Enter a word, several words, or a phrase in the box.
3. From the following drop-down menu, select all of these, any of these or as a phrase ("all of these" is the default, and is highly recommended).

Рис. 9.45. Предупреждение о завершении поисковой сессии

The Library of Congress
>> Go to Library of Congress Authorities

LIBRARY OF CONGRESS ONLINE CATALOG

Help New Search Search History Headings List Titles List Request an Item Account Status Start Over

Combine	EDIT SEARCH	# OF HITS	TYPE OF SEARCH select link to re-execute your search	YOUR SEARCH STRING select link to re-execute your search
☐	1	5	Guided Search	(thorium AND nuclear)[in Keyword Anywhere (GKEY)] AND (reactor OR cycle)[in Keyword Anywhere (GKEY)]
☐	2	1271	Keyword (match all words)	nuclear reactor

Help - Search - History - Headings - Titles - Request - Account - Exit

The Library of Congress
URL: <http://www.loc.gov/>
Mailing Address:
101 Independence Ave. S.E.
Washington, DC 20540

Catalog/authority record errors?
Use our Error Report Form
Questions about searching?
Ask a Librarian

Library of Congress Online Catalog
URL: <http://catalog.loc.gov/>
Library of Congress Authorities
URL: <http://authorities.loc.gov/>

Рис. 9.46. Протокол поисковой сессии

ГЛАВА 10. РЕФЕРАТИВНО-БИБЛИОГРАФИЧЕСКИЕ ДОКУМЕНТАЛЬНЫЕ БД

10.1. Электронные каталоги и базы данных ИНИОН РАН

Институт научной информации по общественным наукам Российской академии наук (ИНИОН РАН) обеспечивает доступ к электронным каталогам и базам данных по социальным и гуманитарным наукам с 1980 г.. Общий объем массивов – более 3 млн. записей, а ежегодное пополнение составляет около 100 тыс. записей. Доступ к электронным каталогам и базам данных находится в разделе «Продукты и услуги» главной страницы (<http://www.inion.ru>), которая представлена на рис. 10.1.

В базы данных включаются аннотированные описания поступивших в Фундаментальную библиотеку ИНИОН книг и статей из журналов и сборников на 140 языках мира. Описания статей и книг в базах данных могут быть снабжены динамическими ссылками на полные тексты источников из электронного хранилища ИНИОН и из Научной электронной библиотеки. Для создания и публикации баз данных ИНИОН используется АИС IRBIS, Web-доступ к электронным ресурсам ИНИОН осуществляется посредством модуля WebIRBIS¹ (Web-версии АИС IRBIS).

Для поиска предоставлены следующие реферативно-библиографические БД.

1. Сводная база данных поступлений литературы в библиотеку ИНИОН за 1993–1995 гг. (241280 документов);
2. Сводная база данных поступлений литературы в библиотеку ИНИОН за 1996–1998 гг. (271527 документов);
3. Сводная база данных поступлений литературы в библиотеку ИНИОН за 1999–2001 гг. (267825 документов);
4. Сводная база данных поступлений литературы в библиотеку ИНИОН за 2002–2004 гг. (240634 документа);
5. Сводная база данных поступлений литературы в библиотеку ИНИОН за 2005–2007 гг. (269351 документ);

¹ АИС IRBIS (©1992-2008), в отличие от библиотечной системы ИРБИС, является информационно-поисковой системой с развитым аппаратом поиска и обработки результатов.

6. База данных по экономике и демографии (294588 документов);
7. База данных по философии и социологии (357370 документов);
8. База данных по истории, археологии и этнологии (383107 документов);
9. База данных по языкознанию (232811 документов);
10. База данных по литературоведению (286595 документов);
11. База данных по науковедению (235113 документов);
12. База данных по государству и праву (225498 документов);
13. База данных по религиоведению (62019 документов);
14. Депонированные рукописи (14162 документа);
15. База данных по востоковедению (12000 документов);
16. Русская культура в мировом контексте (2702 документа);
17. Редкая книга (1691 документ);
18. Гендерные исследования (3818 документов);
19. Трофейная коллекция (1691 документ);
20. Список периодических изданий, обрабатываемых в ИНИОН (20000 документов);
21. Электронный каталог ИНИОН (57103 документа).

Помимо библиографических БД, в режимах поиска и просмотра пользователям доступны компоненты лингвистического обеспечения, оформленные как базы данных:

1. Рубрикатор ИНИОН (4901 документ)
2. Тезаурус «Экономика и демография» (2974 дескриптора)
3. Тезаурус «Философия» (2831 дескриптор)
4. Тезаурус «Социология» (2026 дескрипторов)
5. Тезаурус «Гендерные исследования» (1881 дескриптор)

Библиографическая запись БД ИНИОН содержит более 200 элементов (полей). При формировании поисковых запросов могут быть использованы словари следующих полей.

Ключевые слова [KW] – содержит ключевые слова и дескрипторы на русском и английском языках. Словосочетания в поисковый словарь для поля попадают полностью и в виде отдельных слов. Если необходимо провести поиск по словосочетанию, необ-

ходимо заключить его в поисковом выражении в одинарные кавычки.

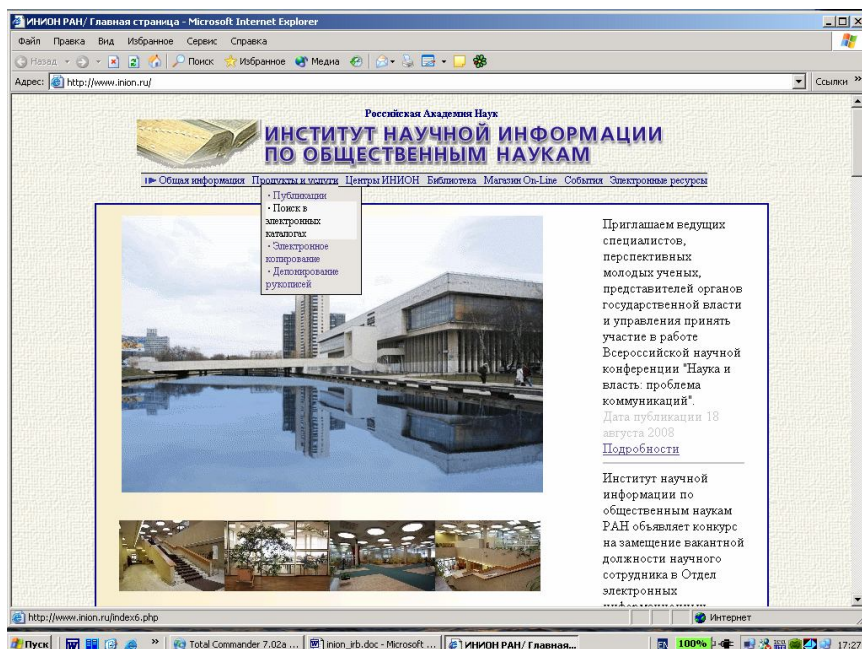


Рис. 10.1. Главная web-страница ИНИОН РАН

Общий словарь [ALL] – словарь содержит отдельные слова и части сложносочиненных слов, разделенных знаком дефис, собранные по всем значимым текстовым полям библиографической записи.

Тип документа [KD] – в данном варианте БД ИНИОН записи (документы) отнесены только к 4 типам: монография, много- томник, статья, рецензия (реферат).

Язык [LG] – язык текста документа (источника) по стандар- ту ISO. В словарь включены все варианты написания названия язы- ка (например, «голландский» и «нидерландский»).

Шифр [SV] – шифр хранения ИНИОН.

Авторы [AU] – словарь содержит фамилии и инициалы авторов, редакторов, составителей, переводчиков и т.п. (например, «Иванов, И.И.» или «Петров, П»).

Заглавие [TI] – словарь содержит отдельные слова из следующих полей библиографической записи: «Основное заглавие», «Параллельное заглавие», «Сведения, относящиеся к заглавию», «Сведения об ответственности».

Источник [SO] – словарь содержит названия периодических и продолжающихся изданий (газет, журналов, сборников статей).

Место издания [PB] – список городов – мест издания.

Год издания [DT] – перечень годов издания источников.

Серия [SER] – содержит отдельные слова из названий серий и подсерий.

Аннотация [AB] – в словаре содержатся слова из аннотаций на русском языке.

Рубрики ИНИОН [SI] – буквенно-цифровые рубрики в соответствии с рубрикаторм ИНИОН (первые 3 уровня соответствуют рубрикаторму ГРНТИ; всего 8 уровней). Рубрикатор содержит 5 фасетов (А – основной фасет; В – фасет стран и регионов; С – отраслевая детализация; D – хронология; Е – персоналия).

Аспект [ASP] – аспектный дескриптор.

Год поступления [EN] – год поступления источника в библиотеку ИНИОН (год поступления источника не всегда совпадает с годом издания).

Поиск в электронных каталогах

Поиск в базах данных осуществляется посредством поисковой системы **WebIRBIS**, предназначенной для многоцелевой обработки баз данных (в том числе и полнотекстовых), содержащих разнородные документы неограниченной длины с нерегулярной структурой. Система имеет развитые средства поиска, сортировки и вывода информации, обеспечивая гибкость и эффективность технологий информационного поиска.

Идентификация пользователя. Перед началом поисковой сессии пользователь должен идентифицировать свой вход в систему (рис. 10.2).

© Институт научной информации по общественным наукам РАН

Идентификация пользователя

Для ранее зарегистрировавшихся пользователей - введите Ваше учетное имя (вместо GUEST) и пароль

Для анонимных - нажмните "Ввод" (статус "гость")

Учетное имя (Login)

GUESY

Пароль

Ввод

Внимание! Статус "гость" (GUEST) не позволяет:

- восстанавливать и продолжать прерванную сессию;
- обращаться и заказывать электронные копии текстов;
- сохранять и повторно использовать запрос.

ИЛИ

Для первоначальной регистрации нового пользователя - Введите учетные данные

Фамилия И.О. Иванов Иван Петрович Заполняется обязательно

Учетное имя (Login) ivanip Не более 16 символов. Только латиница и цифры

Пароль ivanp1234 Не более 8 символов. Только латиница и цифры

Страна Россия Город Москва

Организация МИФИ

Электронная почта ivanip@mail.ru Заполняется обязательно для заказа копий документов

Ввод

Рис. 10.2. Страница идентификации и регистрации пользователя

Для зарегистрированных пользователей предоставлены дополнительные возможности, такие как:

- заказ электронных копий документов;
- восстановление и продолжение прерванной поисковой сессии;
- сохранение (с возможностью повторного использования) поисковых запросов.

Учетное имя GUEST («гость») с паролем по умолчанию позволяет незарегистрированному пользователю проводить поиск и просмотр документов в любой из представленных БД (например, для ознакомления с содержанием информационных массивов и с поисковыми возможностями среды).

Начало поисковой сессии. Страница «Начало сессии» предназначена для задания параметров поисковой сессии – типа поискового интерфейса, объема выдачи, режима просмотра (рис. 10.3).

Поисковый интерфейс может быть ориентирован на *простой* или *профессиональный* поиск.

Флаг *Многооконный режим* устанавливает дополнительную возможность вывода результатов работы некоторых функций в дополнительных окнах браузера.

Параметр *Объем выдачи* позволяет указать количество записей для просмотра на одном экране (5, 10, 20 или 50 записей).

Раскрывающийся список *Запрос* появляется только для зарегистрированных пользователей. В списке можно выбрать режим работы с новым запросом или с одним из поименованных и сохраненных пользователем ранее. В случае работы с сохраненным запросом его данные загружаются в *Протокол* системы и могут быть использованы для продолжения поисковой сессии.

Начало сессии

База данных: esop

Запрос: Новый

Объем выдачи: 20 Записей

Поисковый интерфейс

- ☒ Профессиональный поиск
- ☐ Простой поиск
- ☒ Многооконный режим

Начало поиска

Установка параметров и начало поиска

Рис. 10.3. Начало поисковой сессии

Кнопка *Установка параметров и начало поиска* инициирует функцию задания дополнительных параметров поиска и просмотра (стр. 10.4). При этом параметры поиска должны быть установлены до выполнения поисковых операций, а параметры просмотра могут быть изменены после поиска, перед просмотром порции документов.

Альтернативный флаг *Тип поискового запроса* позволяет установить форму ввода поискового запроса для интерфейса профессионального поиска (по простому вхождению, по логическому выражению или по образцу).

В результат поиска по очередному выражению (предложению запроса в терминах ИПЯ системы) могут быть включены либо все найденные по предложению запроса документы, либо только новые по отношению к предыдущему предложению (*исключая документы последнего поиска*) или ко всему запросу (*исключая все ранее найденные*).

База данных: **Экономика и демография 1991 - 2007 (Записей: 294588)**

Параметры поиска и просмотра

Тип поискового запроса

☐ по простому вхождению

☒ по логическому выражению

☐ по-образцу

Выдавать в результате поиска по предложению

☒ все найденные по предложению документы

☐ исключая док-ты последнего поиска

☐ исключая все ранее найденные

Искать также в БД: (можно отметить не более 4-х баз)

☐ Библиография ИНИОН 1999-2001 гг. [99]

☐ BSU 1917-1987 [bsu]

☒ Электронный каталог ИНИОН (с 2000 г.) [ecat]

☐ Поступления 1993-1995 [9395]

☐ Гендерные исследования [gender]

☐ Тезаурус по экономике и демографии [theco]

☐ Tesaurs Philosophy [tphil]

☐ Трофейная коллекция [gota]

☐ Euroscinty [euro]

☐ История. Археология. Этнология 1986-2007 [hist]

☐ Литературоведение 1986 - 2007 [liter]

☐ Государство и право 1986 - 2007 [law]

☐ Сводная база поступлений 2005-2007 [2005]

☐ Linguistics Thes [lingth]

☐ Редакц. книги [RETRO]

Формат

Biblio View (INION) [biblion]

Объем выдачи

10 Записей

Применить установки

Рис. 10.4. Установка дополнительных параметров поиска

Для просмотра документов может быть выбран один из следующих форматов:

- **Full Ref.(Eng)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника.
- **Full Ref. (Rus)** – названия элементов данных выводятся на русском языке. Запись выводится на языке первоисточника.
- **Full Ref.(Transl)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника. Кириллические символы в названии первоисточника, аннотации и терминах выводятся в латинской транслитерации (стандарт библиотеки Конгресса США). Этот формат можно использовать для вывода на печать записей на кириллице, если устройство вывода не имеет соответствующих шрифтовых возможностей.
- **Short Ref. (Eng)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника. Сокращенная форма вывода записи.
- **Short Ref. (Rus)** – названия элементов данных выводятся на русском языке. Запись выводится на языке первоисточника. Сокращенная форма вывода записи.

- **Short Ref. (Transl)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника. Кириллические символы в названии первоисточника, аннотации и терминах выводятся в латинской транслитерации. Сокращенная форма вывода записи.

- **FR & Subj. (Eng)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника. Вместе с кодами рубрикатора ИНИОН в записи выводятся названия рубрик на русском и английском языках.

- **FR & Subj. (Rus)** – названия элементов данных выводятся на русском языке. Запись выводится на языке первоисточника. Вместе с кодами рубрикатора ИНИОН в записи выводятся названия рубрик на русском и английском языках.

- **FR & Subj. (Transl)** – названия элементов данных выводятся на английском языке. Запись выводится на языке первоисточника. Кириллические символы в названии первоисточника, аннотации и терминах выводятся в латинской транслитерации. Вместе с кодами рубрикатора ИНИОН в записи выводятся названия рубрик на английском языке и транслитерация названий рубрик на кириллице.

- **Biblio View (INION)** – названия элементов данных выводятся на русском языке. Запись выводится на языке первоисточника. Формат вывода описания первоисточника на каталожную карточку ИНИОН (с добавлением аннотации, терминов и кодов рубрик).

- **Citation** – названия элементов данных выводятся на русском языке. Запись выводится на языке первоисточника. Сокращенный формат вывода описания первоисточника на каталожную карточку ИНИОН.

При установке параметров можно также дополнительно выбрать для поиска еще не более четырех баз данных. Результаты поиска по отдельным базам фиксируются в *Протоколе* поисковой сессии как отдельные предложения запроса.

Профессиональный поиск

База данных: **Экономика и демография 1991 - 2007** (Записей: 294588)

Искать также в БД: *theco scien*

2) Сформируйте строку запроса для поиска по словам в текстовых полях

Тезаурус по экономике и демографии

Рубрикатор ИНИОН

(Для усечения слов используйте символ звездочки - *)

Текст запроса:

Дополнительные условия отбора:

☐ Искать строго по указанным словам
 ☐ - по вариантам словоформ (автомаскирование)

☒ При выводе использовать краткий список по полю *ITJ*

Найти

Очистить

Показать словарь

для поля

Общий словарь [ALL]

начиная с

цикл

Начало сессии / Выбор базы данных

Завершить сеанс

Выход из поиска и завершение сессии

Рис. 10.5. Интерфейс простого поиска.

Поисковые интерфейсы

Интерфейс простого поиска. Страница простого поиска содержит строку формирования запроса и интерфейсные элементы для задания дополнительных условий отбора и выдачи документов (рис. 10.5, 10.6).

Запрос вводится как последовательность терминов, относимых при поиске ко всем полям документа и объединяемых логической операцией ИЛИ, а выдача ранжируется по количеству терминов запроса, найденных в документе. Термины включаются в выражение запроса либо в том виде, в котором они введены в строку поиска (символ «*» при этом указывает на усечение), либо в нормализованной форме с автоматическим усечением окончаний (*по вариантам словоформ*).

Флаг *При выводе использовать краткий список...* включает сокращенный формат выдачи документов, при котором содержимое документа урезается до одного поля.

При формировании запроса может быть использован словарь одного из предлагаемых в списке выбора полей. Термины словаря при выдаче оформлены как гиперссылки и снабжены числовым значением, характеризующим количество документов, содержащих

термин (например, при формировании поискового запроса, представленного на рис. 10.6, термин ЦИКЛ не был включен в выражение по причине высокой частотности).

Рис. 10.6. Формирование запроса при простом поиске

Результаты поиска (в режиме простого поиска) выводятся постранично. В верхней части первой страницы содержится справочная информация о количестве документов, найденных во всех отмеченных для поиска БД (рис. 10.7). В нижней части каждой страницы размещены элементы управления поисковым процессом, доступные в режиме простого поиска. Подробно возможности управления поисковым процессом будут представлены ниже.

Интерфейс профессионального поиска. Страница профессионального поиска содержит форму запроса в одном из трех вариантов.

Форма *по простому вхождению* аналогична форме запроса простого поиска.

Форма *по логическому выражению* содержит три области ввода поисковых условий, относимых к различным полям. Поиско-

вые условия традиционно объединяются в логическое (булевоподобное) выражение одной из логических операций (И, ИЛИ, НЕ). По умолчанию логика реализации поискового выражения соответствует приоритету логических операций. Если же отмечен флаг *Выполнение в порядке следования*, то поисковые условия будут выполняться в естественном порядке. Например, для поискового запроса, представленного на рис. 10.8, будут сформированы следующие выражения (в соответствии с ИПЯ системы):

- ALL: (ядер* И реактор*) ИЛИ (ALL: (ядерн* И цикл*) И LG: русский)
- если флаг не установлен и
- (ALL: (ядерн* И реактор*) ИЛИ (ALL: (ядерн* И цикл*))) И LG: русский
- если флаг установлен.

В первом случае требование русскоязычного документа будет относиться только к документам, содержащим модификации терминов **ядерн** и **цикл**.

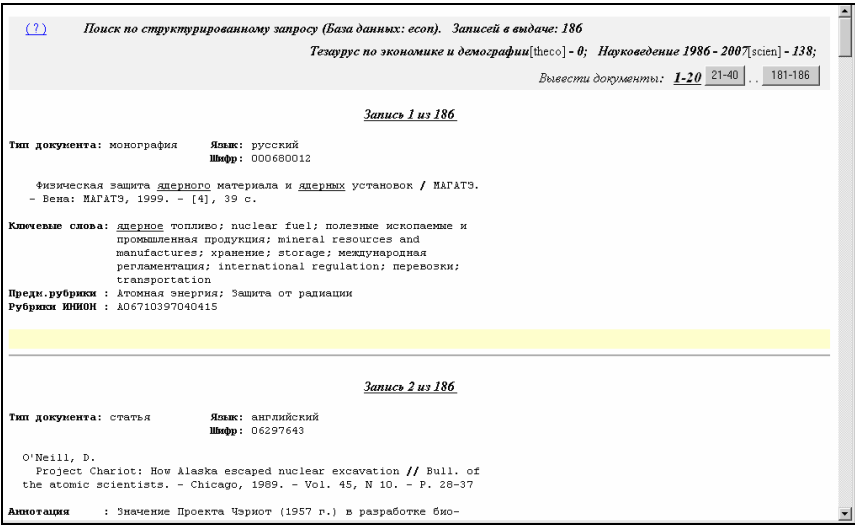


Рис. 10.7. Результаты поиска в режиме простого поиска

Форма *по образцу* содержит несколько строк для ввода отдельных поисковых условий, одни из которых относятся к фиксированным полям, а другие предлагают список из группы полей на выбор. Форма запроса задается *схемой* конкретной БД и может

быть различной для разных источников данных. На рис. 10.9 представлена форма запроса, в которой зафиксированы условия поиска для полей *Авторы*, *Язык*, *Тип документа*. Каждое из полей *Заглавие*, *Общий словарь* и *Год издания* задает свою группу полей, одно из которых можно выбрать в образец (например, группа *Год издания* содержит поле *Год поступления*). Для каждого из полей в образце можно сформировать условие из одного термина или нескольких, связав их логическими операторами. Условия для отдельных полей объединяются в выражение запроса логическим оператором И.

Простой поиск | Параметры поиска и просмотра

База данных: **Экономика и демография 1991 - 2007 (Записей: 294588)**

Искать также в БД: *theso* *scien*

Фрагмент словаря для поля **Язык [LG]**

РУССКИЙ [16447 док.]	ТАТАРСКИЙ [2 док.]
САНСКРИТ [4 док.]	ТУРЕЦКИЙ [40 док.]
СЕРБОХОРВАТСКИЙ [2076 док.]	УЗБЕКСКИЙ [27 док.]
СИАМСКИЙ [10 док.]	УКРАИНСКИЙ [2572 док.]
СЛОВАЦКИЙ [208 док.]	ФИНСКИЙ [306 док.]
СЛОВЕНСКИЙ [455 док.]	ФРАНЦУЗСКИЙ [18410 док.]
ТАДЖИКСКИЙ [1 док.]	ХИНДИ [2 док.]
ТАИ [10 док.]	Далее...

Показать словарь для поля [Общий словарь [ALL]] начиная с []

[\(?\)](#) **Сформируйте запрос по поисковым полям** [Тезаус по экономике и демографии](#)
[Рубрикатор ИНИОН](#)

(Для учета слов используйте символ звездочки - *)

[ИЛИ]	[Общий словарь [ALL]]	: ядерн* реактор*	[Словарь]
[И]	[Общий словарь [ALL]]	: ядерн* цикл*	[Словарь]
[И]	[Язык [LG]]	: русский	[Словарь]

☒ Выполнение в порядке следования (а не приоритета операторов: *НЕ-И-ИЛИ*)

Дополнительные условия отбора:

☒ Искать строго по указанным словам ☐ Нечеткий поиск (по части поисковых терминов)

☐ При выводе использовать краткий список по полю *ГТД* ☐ - по вариантам словоформ (автомаскирование)

[Найти] [Очистить]

Рис. 10.8. Форма запроса «по логическому выражению».

Вывод результатов поиска

Полные функциональные возможности работы с поисковыми результатами доступны пользователю в режиме профессионального поиска.

Верхняя часть первой страницы результатов позволяет инициировать следующие функции (рис.10.10):

Переотобразить результат – интерфейс выполнения функции включает раскрывающийся список *Сортировать по ключу* для выбора поля, по значению которого необходимо отсортировать выдачу, флаг альтернативного выбора последовательности сортировки (по убыванию/возрастанию), раскрывающийся список *Формат* для выбора формы отображения документа внутри страницы, раскрывающийся список *Объем выдачи* для выбора размера страницы и кнопку «Выполнить» для запуска функции на исполнение.

Новый поиск – интерфейс выполнения функции нового поиска содержит строку ввода/редактирования поискового запроса (*Поиск.условие/Имена полей*), в которой можно также и ввести новый запрос, используя синтаксис ИПЯ системы, флаг *Уточняя последнюю выдачу*, включающий режим нового поиска только среди документов предыдущего результата, и кнопку *Найти* для запуска функции.

Простой поиск

Параметры поиска и просмотра

База данных: **Экономика и демография 1991 - 2007 (Записей: 294533)**

Искать также в БД: *theco scien*

[Текстус по экономике и демографии](#)
[Рубрикатор ИНИОН](#)

(?) Введите поисковые реквизиты запроса

Авторы [AU]:

Язык [LG]:

Тип документа [KD]:

Заглавие [TI]:

Общий словарь [ALL]:

Год издания [OT]:

Дополнительные условия отбора:

☒ Искать строго по указанным словам

☐ - по вариантам словоформ (автомаскирование)

☒ При выводе использовать краткий список по полю **[TI]**

☐ (по части поисковых терминов)

для поля начиная с

(?) Протокол поиска
(Всего документов в выдачах - 1028)

4 док. - просмотрены
2 док. - отмечены как соот. теме
0 док. - заказаны копии текстов

Рис. 10.9. Форма запроса «по образцу».

При отображении на странице результатов поиска каждый документ снабжается альтернативным переключателем *Релевантный/Не релевантный*. Установка переключателя в одно из значений

позволяет дополнительно объявить документ «нужным», соответствующим поисковой потребности пользователя (релевантным) или «не нужным». Эти отметки в дальнейшем могут быть использованы при выполнении функций *Эвристический поиск* и *Новый поиск и Протокол* (управляющие кнопки для запуска функций расположены в нижней части каждой страницы результата – рис. 10.11).

В режиме просмотра при каждом документе размещается флаг *Заказ копии*, значение которого использует одноименная функция.

От каждого документа можно по гиперссылке *Аналоги* (в случае многооконного режима) перейти в новое окно просмотра документов-аналогов, т.е. документов, *похожих* на текущий (функция похожести выполняется как поиск документов, содержащих определенное количество общих терминов с текущим – рис.10.12).

(?) Поиск по структурированному запросу (База данных: esop). Записей в выдаче: 4

Научковедение 1986 - 2007[scien] - 52;

Сортировать по кнопке: [Не установлено] Формат: Full Ref. (Rus) [biblio2]

☒ по возраст. ☐ по убыванию

Поиск условие/Имена полей: [(ALL(ядер* реактор*) ИЛИ (ALL(ядер* цикл*))) И (

(?) ☐ Уточняя последнюю выдачу

Объем выдачи: [20] Записей

Запись 1 из 4

Язык: русский

Тип документа: многотомник

Идентификатор хранения: BR83M/4323K

Осн. заглавие: Ежегодный доклад

Ответственность: Междунар. агентство по атомной энергии

Место издания: [Вена]

Издательство: ИАГАТУ

Год издания: 1993

Номер тома: за 1992 год.

Страниц: 246 с.

Иллюстрация: ил.

Аннотация: Осн. разд.: атомная энергетика, ядерный топливный цикл, обращение с радиоактивными отходами, продовольствие и сельское хозяйство, здравоохранение, радиационная безопасность, гарантии.

Ключевые слова: международные экономические организации; international economic organizations; доклады; reports; атомная энергетика; atomic energetics; отрасли промышленности; specific industries; страны мира; countries of the world

Рубрики ИИНОИ: A06515735; A06710397040415

Аспект-код: M O

Рис. 10.10. Первая страница результата поиска

Заказ копии. Страница заказов копий документов, отмеченных флагом *Заказ копии*, доступна только зарегистрированным пользователям (рис. 10.13). Документы, для которых заказаны ко-

пии, еще раз выводятся на странице. Отметки заказов могут быть еще раз пересмотрены и сняты у отдельных документов. В случае, если количество отмеченных документов меняется, необходимо выполнить функцию *Пересчитать*, т.е. заново рассчитать стоимость заказа. Заказ передается на исполнение только после активации кнопки *Подтверждение заказа*.

Эвристический поиск. Кнопка эвристического поиска запускает функцию поиска документов, похожих на документы, отмеченные в выдаче как релевантные, в соответствии с эвристической процедурой. В основе процедуры лежит формирование кластеров документов, содержащих наборы терминов (из лексики релевантных документов), имеющих значение весового коэффициента выше порогового, и упорядочение кластеров для вывода. Результат эвристической процедуры ограничен (в рассматриваемой реализации – не больше 20 документов) и содержит документы, имеющие максимальный суммарный вес.

Язык: русский
Тип документа: статья
Страна изд.: 810
Идентификатор: 08917633
Авторы: Субботин, С.А.; Стукалов, В.А.; Чернылин, Ю.Ф.
Осн. заглавие: Атомная энергетика: экономические и физические особенности развития топливного цикла
Источник: Энергия = Energy
Выход. дан. изд.: М., 2002
Осн. порядк. ед.: N 3. - С. 10-17
Ключевые слова: атомная энергетика; atomic energetics; отрасли промышленности; specific industries
Рубрики ЮНКО: A06710397040415
Аспект-код: 0

☒ Релевантен ☐ Не релевантен

[Аналоги](#)

☒ Заказ копии

Результаты по БД:

scien (52)

Заказ копии

Оформление заказа копий текстов отмеченных записей

Эвристический поиск

Объединение и поиск по отм. документам

Новый поиск и Протокол

История и обработка результатов поиска

Завершить сеанс

Выход из поиска и завершение сессии

Начало сессии / Выбор базы данных

Рис. 10.11. Нижняя часть страницы результатов поиска.

Результатом процедуры эвристического поиска является также словарь терминов релевантных документов, снабженных частотой встречаемости во множестве релевантных документов. Отобразить лексику релевантных документов можно по соответствующей ссылке (Поиск по лексике релевантных документов) на странице результатов эвристического поиска (рис. 10.14).

Отметив термины из предложенного списка (рис.10.15) и инициировав функцию *Поиска по комбинациям отмеченных терминов*, пользователь запускает процедуру обратной связи по релевантности, которая строит кластеры документов и включает эти кластеры как предложения запроса в *Протокол* поисковой сессии (рис. 10.16).

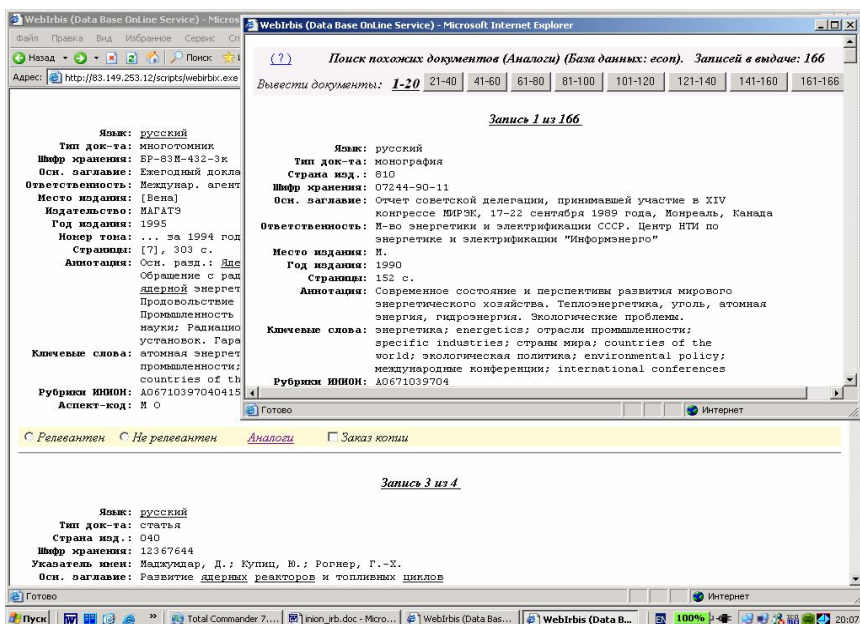


Рис. 10.12. Окно вывода результатов поиска аналогов

(?) *Регистрация заказа для исполнения (База данных: econ).*

Стоимость заказа(руб.): 80 Исполнение заказа(дней): 4

Запись 1 из 1

Language: Russian
 Type of material: component part
 Country of publ.: 040
 Call number: 12367644
 Name Index: Маджумдар, Д.; Кулиц, Ю.; Рогнер, Г.-Х.
 Title proper: Развитие ядерных реакторов и топливных циклов
 Host publication: Бюл. МАГАТЭ
 Host publ. area: Вена, 2000
 Numbering: Т. 42, N 2. – С. 51-59
 Key Words: атомные электростанции; nuclear power stations;
 безопасность; security; новая технология; new
 technology; атомная энергетика; atomic energetics;
 отрасли промышленности; specific industries;
 исследования и разработки; research and development
 Subject Index: A06710397040415
 Aspect-code: M O

☐ Релевантен ☐ Не релевантен [Аналоги](#) ☒ Заказ копии

Результаты по БД: scien (62)

 если отметка заказа снята

Электронная почта:
 FTP-адрес: ftp.inion.ru/PRIVATE/PETR
 Формат копии: Способ доставки:
 Регистрация заказа для исполнения

Рис. 10.13. Форма заказа электронных копий

(?) *Поиск по лексике релевантных док-тов (База данных: econ)*

Сортировать по полю: Формат: ☐ по возраст. ☐ по убыванию Объем выдачи: [Записей](#)

Запись 1 из 2

Язык: [русский](#)
 Тип док-та: статья
 Страна изд.: 810
 Шифр хранения: 08917633
 Авторы: Чернылин, Ю.Ф.; Жигульский, В.В.
 Осн. заглавие: Замедление роста атомной энергетики. Причины и последствия
 Источник: Энергия = Energy
 Выход. дан. изд.: М., 2004
 Осн. порядк. ед.: N 12. – С. 10-19
 Ключевые слова: атомная энергетика; atomic energetics; отрасли промышленности; specific industries; Россия; Russia
 Рубрики ЮНИС: A0691; B5990; C06710397040415
 Аспект-код: C O

☐ Релевантен ☐ Не релевантен [Аналоги](#) ☐ Заказ копии

Запись 2 из 2

Язык: [русский](#)
 Тип док-та: статья
 Страна изд.: 810
 Шифр хранения: 08917633
 Указатель имен: Смирнова, Л.С.; Стукалова, В.А.; Субботин, С.А.
 Осн. заглавие: Перспективы совместного развития газовой и атомной отраслей России
 Ответственность: Смирнова Л.С., Стукалова В.А., Субботин С.А. и др.

Рис. 10.14. Вывод результатов эвристического поиска

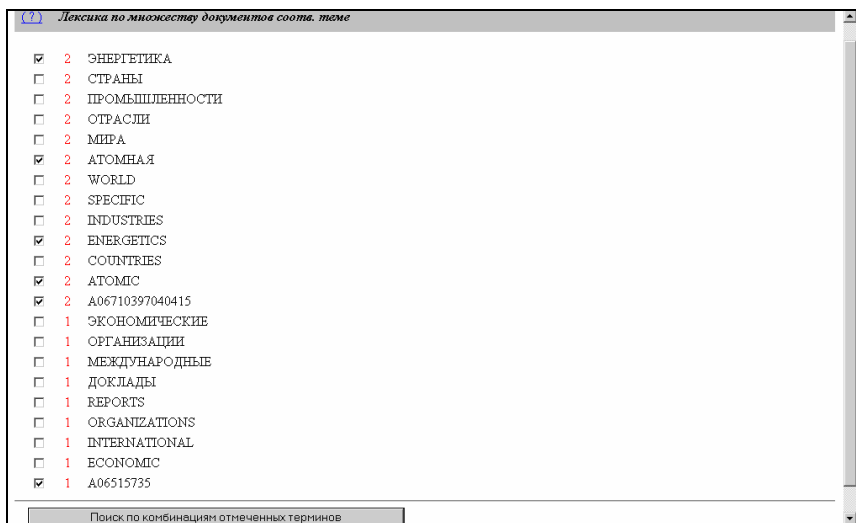


Рис. 10.15 Отображение лексики релевантных документов

Протокол поиска. На всем протяжении поисковой сессии система предоставляет возможность обращаться ко всем введенным и обработанным поисковым выражениям. Активизировав функцию *Новый поиск и Протокол* (см. рис. 10.11), пользователь попадает на страницу, которая (помимо формы запроса) содержит перечень предложений запроса, снабженных гипертекстовыми ссылками, позволяющими перейти к просмотру результата (см. рис. 10.16).

Каждое из предложений запроса имеет в рамках *Протокола* свой номер, который может быть использован при формировании нового поискового условия (в соответствии с синтаксисом ИПЯ системы).

Помимо порядкового номера, у каждого предложения есть флаг отметки. Значения флагов используются при выполнении функций по обработке поисковых результатов.

Интерфейс функций *Протокола* содержит управляющие кнопки, а также флаги и текстовую область для ввода дополнительных параметров обработки.

Альтернативный флаг *Выполнить* задает множество предложений запроса (последнее, все предложения запроса или только отмеченные), для которых будет выполняться одна из следующих функций:

- *Итог и эвристический поиск* – для заданного подмножества предложений запроса будет выполнена операция логического сложения результатов и затем проведен эвристический поиск с использованием отметок релевантности документов. Результат эвристического поиска будет записан в Протокол.

- *Итоговый результат* – для заданного подмножества предложений запроса будет сформировано множество объединения результатов, из которого будут выделены подмножества релевантных документов и документов, для которых были заказаны копии. В Протокол будет записан общий итоговый результат объединения и (если таковые найдутся) результаты отбора релевантных и (отдельно) заказанных документов.

- *Уничтожить* – заданное подмножество предложений запроса будет удалено из Протокола. После выполнения операции возврат удаленных предложений в Протокол невозможен.

(?) Протокол поиска (Всего документов в выдачах - 4428)		7 док. - просмотрены 5 док. - отмечены как соот. теме 3 док. - заказаны копии текстов
#1	☐ ((ALL(ядери* реактор*)) ИЛИ (ALL(ядери* цикл*))) И (LG(русский)) { 52 док. по БД ессон }	
#2	☐ ((ALL(ядери* реактор*)) ИЛИ (ALL(ядери* цикл*))) И (LG(русский)) { 4 док. по БД ессон }	
#3	☐ Результат По всем предложениям: Заказ копии { 2 док. по БД ессон }	
#4	☐ Результат По всем предложениям: Заказ копии { 1 док. по БД ессон }	
#5	☐ ((Аналоги 221809 док.)) { 184 док. по БД ессон }	
#6	☐ ((Аналоги 103155 док.)) { 166 док. по БД ессон }	
#7	☐ (Итог и эврист. поиск#2) { 2 док. по БД ессон }	
#8	☐ (АТОМІС and ENERGETICS and АТОМНАЯ and ЭНЕРГЕТИКА and A06710397040415 and A06515735) { 1 док. по БД ессон }	
#9	☒ (АТОМІС and ENERGETICS and АТОМНАЯ and ЭНЕРГЕТИКА and A06710397040415) { 77 док. по БД ессон }	
#10	☐ (АТОМІС and ENERGETICS and АТОМНАЯ and ЭНЕРГЕТИКА) { 290 док. по БД ессон }	
#11	☐ (ENERGETICS and ЭНЕРГЕТИКА) { 1903 док. по БД ессон }	
#12	☐ (АТОМНАЯ and A06710397040415) { 85 док. по БД ессон }	
#13	☒ (A06710397040415 and A06515735) { 2 док. по БД ессон }	
Выполнить ☐ По последнему ☐ По всем предложениям ☒ По отмеченным		
Итог и эврист.поиск Объединение и поиск по отм. документам		
Итоговый результат (Объединение и выделение отмеченных док-тов)		
Уничтожить (Удалить все/отмеченные предложения из запроса)		
Найти Комбинация результатов по условию		

Рис. 10.16. Протокол поисковой сессии

Предложения запроса (а точнее – их результаты) можно использовать для формирования нового поискового запроса. Текстовая область *по условию* позволяет сформировать (в синтаксисе ИПЯ) выражение запроса, где вместо терминов могут быть указаны номера предложений Протокола. Например, выражение запроса **#10 not #9** позволит получить документы, отнесенные к рубрикам, отличным от рубрики, по которой проведен поиск в предложении #9.

Поиск в базах данных ИНИОН с использованием тематического рубрикатора и тезаурусов

Для формулировки поискового запроса можно использовать Рубрикатор ИНИОН и Тезаурус по тематике БД. Доступ к объектам лингвистического обеспечения возможен через соответствующие гипертекстовые ссылки в правом верхнем углу формы поискового запроса (см., например, рис. 10.9).

Дополнительные объекты лингвистического обеспечения представлены в той же структуре доступа, что и основные БД. В базах данных Рубрикатора и Тезауруса можно проводить поиск в режиме логического выражения (рис. 10.17, 10.19).

Тезаурус. В БД Тезаурус для формирования поискового запроса могут быть использованы следующие поля:

- дескриптор;
- вышестоящий дескриптор;
- нижестоящий дескриптор;
- ассоциативный дескриптор;
- рубрика ИНИОН (цифровой индекс).

Поисковые словари полей «Дескриптор», «Вышестоящий дескриптор», «Нижестоящий дескриптор», «Ассоциативный дескриптор» содержат дескрипторы тезауруса – слова и словосочетания без разбиения на отдельные слова.

Для поиска по отдельным словам из дескрипторов-словосочетаний служит словарь «Все поля», который включает термины-слова из содержимого всех перечисленных выше полей, кроме поля «Рубрики ИНИОН».

База данных: **Тезаурус по экономике и демографии (Записей: 2974)**

Фрагмент словаря для поля **Дескриптор [DR]**

АТОМНЫЕ ЭЛЕКТРОСТАНЦИИ [1 док.]	БАЛАНС ДОХОДОВ И РАСХОДОВ ПРЕДП [1 док.]
АТТЕСТАЦИЯ РАБОТНИКОВ [1 док.]	БАЛАНС НАРОДНОГО ХОЗЯЙСТВА [1 док.]
АУДИТ [1 док.]	БАЛАНСОВЫЕ МОДЕЛИ [1 док.]
АУДИТОРСКИЕ СЛУЖБЫ [1 док.]	БАЛАНСОВЫЙ МЕТОД [1 док.]
АУКЦИОН [1 док.]	БАНКИ [1 док.]
АЭРОПОРТЫ [1 док.]	БАНКИ ДАННЫХ [1 док.]
БАЛАНС [1 док.]	БАНКИ РАЗВИТИЯ [1 док.]
БАЛАНС ДОХОДОВ И РАСХОДОВ НАСЕЛЕНИЯ [1 док.]	Далее...

Показать словарь для поля [Дескриптор [DR]] начиная с []

[\(?\)](#) **Сформируйте запрос по поисковым полям** (Для усечения слов используйте символ звездочки - *)

И	Дескриптор [DR]:	атомн*	Словарь
И	Рубрика [SI]:		Словарь
И	Все поля [ALL]:		Словарь

☐ Выполнение в порядке следования (а не приоритета операторов: **НЕ-И-ИЛИ**)

(Всего документов в выдачах - 1)

0 док. - просмотрены
0 док. - отмечены как соотв. теме
0 док. - заказаны копии текстов

DR:'ЭНЕРГЕТИКА' { 1 док. по БД thes }

Рис. 10.17. Форма поискового запроса Тезауруса

Поисковая система ищет словарные статьи тезауруса, содержащие термины, указанные в запросе, и выводит их на экран (рис. 10.18).

Дескриптор:

☒ атомные электростанции

Выше стоящие:

☐ промышленные предприятия;

☐ электростанции

Рубрика: A06710395

и поиск по отмеченным терминам

История и обработка результатов поиска

Выход из поиска и завершение сессии

Рис. 10.18. Вывод результата поиска в Тезаурусе

Каждый дескриптор при выводе снабжается флагом отметки. При выполнении функции *Возврат в основную БД* из отмеченных дескрипторов тезауруса формируется поисковый запрос (дескрипторы в запросе связываются логической операцией И), по которому

проводится поиск в библиографической БД. Если при поиске система не найдет документов, соответствующих данному запросу, включается механизм эвристического поиска: операция И между терминами запроса меняется на логическое ИЛИ. Результаты поиска в этом случае ранжируются по степени соответствия содержания документов поисковому запросу.

Рубрикатор ИНИОН. Для поиска рубрик в тематическом Рубрикаторе ИНИОН (рис. 10.19) можно использовать поля, содержащие:

- цифровой код рубрики;
- заглавие на русском языке;
- заглавие на английском языке.

Словари полей построены по принципу разбиения содержимого поля на отдельные слова. Цифровые коды рубрик при выводе поискового результата снабжены флагами отметки (рис. 10.20). Коды отмеченных рубрик (аналогично поиску в Тезаурусе) используются для формирования запроса и проведения поиска в библиографической БД (функция *Возврат в основную БД*).

База данных: Рубрикатор ИНИОН (Записей: 4911)

Фрагмент словаря для поля **Subj.Heading (RU) [CR]**

АТОМНАЯ [2 док.]	АФРИКААНС [1 док.]
АТОМНОЕ [1 док.]	АФРИКАНСКИЕ [6 док.]
АТОМНОЙ [3 док.]	АФРИКАНСКИЙ [1 док.]
АТТЕСТАЦИЯ [1 док.]	АФРИКИ [15 док.]
АФАР [1 док.]	АЭРОКОСМИЧЕСКАЯ [2 док.]
АФГАНИСТАН [1 док.]	БАГАМСКИЕ [1 док.]
АФГАНИСТАНСКАЯ [1 док.]	БАЗА [7 док.]
АФГАНИСТАНСКИЙ [1 док.]	Далее...

для поля начиная с

[\(?\)](#) **Сформируйте запрос по поисковым полям** (Для усеечения слов используйте символ звездочки - *)

Code [SI]	Subj.Heading (RU) [CR]	Subj.Heading (EN) [CE]	Словарь
И	И	И	Словарь
И	И	И	Словарь

☐ Выполнение в порядке следования (а не приоритета операторов: **НЕ-И-ИЛИ**)

Дополнительные условия отбора:

☐ Искать строго по указанным словам ☐ Неточный поиск (по части поисковых терминов)

☐ - по вариантам словоформ (автомаскирование)

Рис. 10.19. Форма поискового запроса Рубрикатора

☐	A065151073531 Международное агентство по <u>атомной</u> энергии International atomic energy agency Стрелка: Рубрика не используется с 1993 г.
☐	A065717073525 Европейское сообщество по <u>атомной</u> энергии (Евратом) European atomic energy community (EAEC) Стрелка: Рубрика не используется с 1993 г.
☒	A06710397040415 <u>Атомная</u> промышленность Atomic industry
☒	A10875137 Международное сотрудничество по использованию <u>атомной</u> энергии. Международное <u>атомное</u> право International cooperation in peaceful uses of atomic energy. International atomic law
☒	С06710397040415 <u>Атомная</u> промышленность Atomic industry
Возврат в основную БД и поиск по отмеченным терминам	
Новый поиск и Протокол История и обработка результатов поиска	
Завершить сеанс Выход из поиска и завершение сессии	
Начало сессии / Выбор базы данных	

Рис. 10.20. Вывод результата поиска в Рубрикаторе

10.2. Базы реферативно-библиографической информации ВИНТИ РАН

Всероссийский институт научной и технической информации ВИНТИ РАН – крупнейший информационный центр, обеспечивающий с 1952 г. на основе анализа потока отечественных и зарубежных публикаций российское и мировое сообщество научно-технической информацией по проблемам точных, естественных и технических наук.

База данных ВИНТИ РАН включает материалы Реферативного журнала (РЖ) по естественным, точным и техническим наукам с 1981г. Общий объем БД – более 26 млн документов. База формируется по материалам периодических изданий, книг, фирменных изданий, материалов конференций, тезисов, патентов, нормативных документов, депонированных научных работ, 30% которых составляют российские источники. БД ВИНТИ пополняется ежемесячно, а тематический фрагмент Химия – 2 раза в месяц. Документы БД содержат библиографию, ключевые слова, рубрики и

реферат первоисточника на русском языке. Информационный ресурс ВИНТИ включает 29 тематических фрагментов и более 230 выпусков БД, а также генерируемую с 2001 г. единую политематическую БД, объединяющую все тематические фрагменты, и отдельный выпуск «Экономия энергии».

Вход в информационные ресурсы ВИНТИ осуществляется через, приведенную на рис. 10.21 главную страницу (www.viniti.ru).



Рис 10.21. Главная страница ВИНТИ РАН

Начало поиска в БД инициируется через ссылку «База данных», размещенную на главной странице. В поисковой форме, представленной на рис. 10.22, пользователь выбирает тематическую базу данных и поисковое поле, для которого в окне запроса задается выражение для отбора, в том числе на естественном языке. Для ограничения выдачи можно дополнительно указать нужный выпуск, год и язык публикации или другие библиографические поля.

Результат поиска (выражение запроса и количество найденных документов) будет зафиксирован в протоколе – «Истории поиска», который отображается в нижней части формы. Результаты поиска, полученные по отдельным запросам, могут быть использованы в новых запросах: множества документов, соответствующих отмеченным запросам, будут объединены или пересечены, в зависимости от выбранной опции «Объединение маркированных запросов».

БД ВИННИТИ | Словарь | Информация о БД | Помощь

Поиск. Демо-версия | **Выполнить запрос** | **Удалить запрос**

Тематика БД: Поштематическая БД ВИННИТИ ?

Запрос ? | искать в: | Поле ?

реакторы ядерные

Выпуск БД ?

Год: | ?

Язык, страна ?

☒ включать морфологический анализ

История поиска | **Вставить** | **Снять маркировку** | **Условия вывода**

При вставке на страницу "Поиск" объединять отмаркированные запросы через: ☒ И ☐ ИЛИ

Запросы	Дата поиска	БД	Найдено
☐ 20 (реакторы ядерные)	2008-09-02	DEFO	79 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 16 ((торий реактор) in T1,K90)	2008-09-02	DEFO	0 показать
☐ 10 (хлопковые пези)	2008-09-02	DEMT	2 показать
☐ 10 (хлопковые пези)	2008-09-02	DEMT	2 показать

Рис. 10.22. Форма запроса

Для отображения в краткой форме найденных по запросу документов необходимо в окне «История поиска» для соответствующего запроса инициировать ссылку «Показать», после чего в открывшемся окне для каждого документа будут выведены список авторов, заглавие, ключевые слова и шифры рубрик. Термины, по которым был отобран документ, будут подсвечены (рис. 10.23).

Для вывода полного описания документов на экран или загрузки в файл, необходимо использовать функциональные кнопки в верхней части формы.

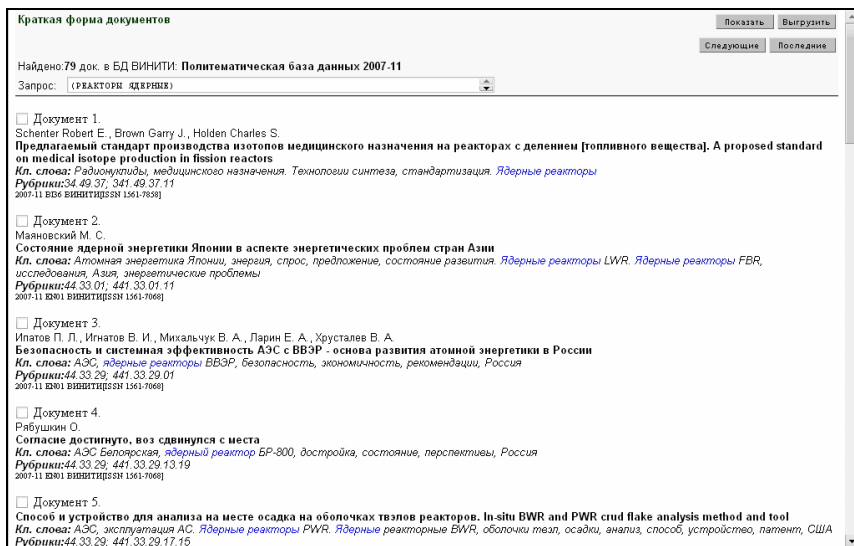


Рис. 10.23. Результаты поиска по запросу

Рубрикатор отраслей знаний ВИНТИ. Рубрикатор отраслей знаний ВИНТИ является классификационной схемой универсального охвата по естественным и техническим научным дисциплинам. Он предназначен для систематизации информационных продуктов и поиска в базе данных ВИНТИ, позволяя полно и точно сформулировать информационный запрос пользователя.

В состав рубрикатора входят отдельные классификации по следующим 30 областям знания: математика и вычислительные науки; механика; физика; астрономия, геодезия и космические исследования; география; геология; горное дело; геофизика; охрана окружающей среды; биология; физико-химическая биология и биотехнология; психология; медицина и здравоохранение; сельское и лесное хозяйство; химия и химическая технология; машиностроение; электротехника и энергетика; металлургия; автоматика и тепломеханика; транспорт; метрология и измерительная техника; радиотехника; электроника; вычислительная техника; экономика и управление; обеспечение безопасности в чрезвычайных ситуациях; связь; техническая кибернетика; издательское дело и полиграфия; информатика.

Рубрикатор ВИНИТИ построен на основе углубления Государственного рубрикатора НТИ России (ГРНТИ) до 9-го уровня – по мере потребности отдельных отраслей. Рубрики имеют названия на русском и английском языках; предметные описания (списки ключевых слов с указанием частотности их использования) сформированы по пятилетнему массиву баз данных ВИНИТИ и работам экспертов. Отрасли знания представлены в виде разделов. В системе классификационных схем есть отображения рубрик ВИНИТИ на другие классификации.

Для многих разделов классификация ВИНИТИ фактически является общегосударственным стандартом. Рубрики ВИНИТИ индексированы кодами классов УДК и тем самым включены в общемировую систему тематического упорядочения знаний.

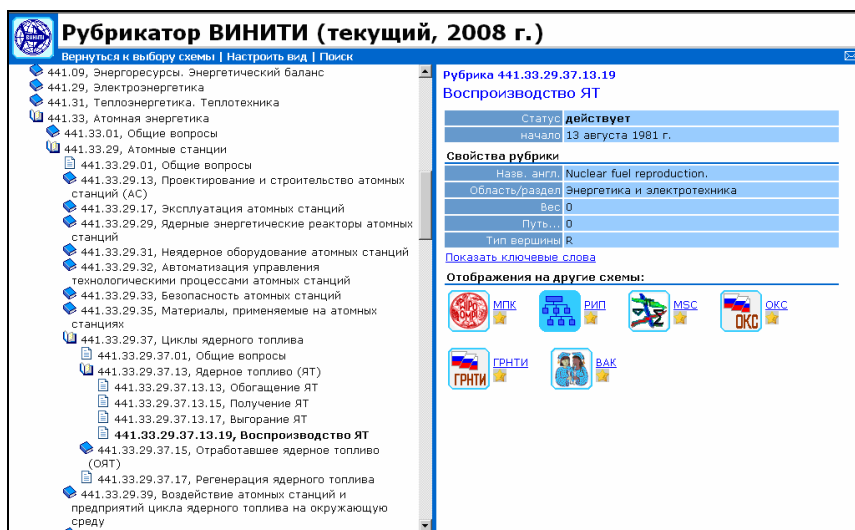


Рис. 10.24. Форма просмотра Рубрикатора ВИНИТИ

Вход в рубрикатор осуществляется через ссылку на главной странице. На рис. 10.24 представлена форма рубрикатора, где он отображается в виде последовательно раскрывающегося дерева, каждая позиция которого содержит шифр и наименование рубрики. Для выделенной рубрики в правой части формы выводятся харак-

теристики, такие как, статус, раздел наук, наименование на английском языке, дата введения в действие, значимость. Здесь же размещается ссылка на список дескрипторов, наиболее связанных с этой предметной областью, а также ссылке, позволяющие вывести разделы других классификационных систем, соответствующих по наполнению данной рубрике.

10.3. Базы информационных карт по диссертациям и НИР ВНТИЦentra

Всероссийский научно технический информационный центр (ВНТИЦ) осуществляет формирование и поддержку национального библиотечно-информационного фонда РФ по непубликуемым источникам научно-технической информации – отчётам о научно-исследовательских и опытно-конструкторских работах (НИОКР), кандидатским и докторским диссертациям, описаниям алгоритмов и программ. Комплектование, регистрация и хранение обязательного экземпляра регламентируются Федеральным законом РФ «Об обязательном экземпляре документов», который предусматривает, что все организации страны, ведущие НИР и ОКР, независимо от их организационно-правовых форм, форм собственности и видов финансирования научных разработок, обязаны направлять во ВНТИЦ обязательный бесплатный экземпляр отчётных документов.

В настоящее время фонд ВНТИЦ содержит более 7 млн документов, охватывающих все области знаний. Ежегодные поступления составляют более 50 тысяч регистрационных и информационных (вторичных) документов и около 10 млн страниц полнотекстовых отчётов о НИОКР и диссертаций (первичных документов). Политематичность фонда позволяет получать полную картину состояния научных исследований и разработок в стране, выполнять кроссдисциплинарные исследования и запросы, формировать справочники персоналий и организаций и словари русскоязычной научно-технической лексики по всем научным дисциплинам.

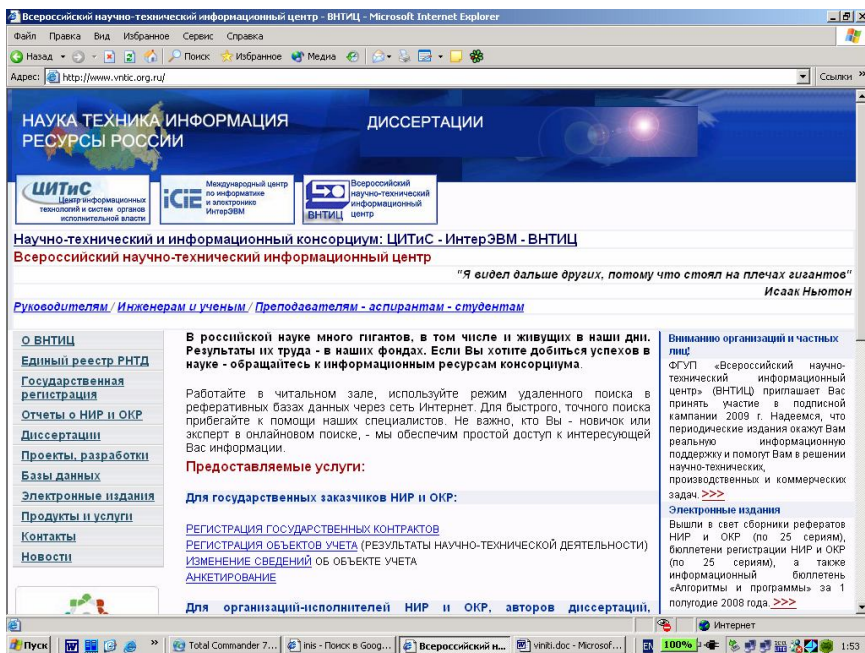


Рис. 10.25. Главная страница сайта ВНТИЦентра

Информация о поступивших в федеральный фонд документах распространяется через следующие электронные издания:

- бюллетени регистрации научно-исследовательских и опытно-конструкторских работ (библиографическая информация) – 25 серий;
- сборники рефератов научно-исследовательских и опытно-конструкторских работ и диссертаций – 25 серий;
- бюллетени регистрации алгоритмов и программ.

На основе поступающих во ВНТИЦ вторичных документов формируются справочно-информационные базы данных:

- «Организации научно-технической сферы». Представлена реестром более 6000 организаций России, работы которых в соответствии с Федеральным законом РФ «Об обязательном экземпляре документов» зарегистрированы во ВНТИЦ. База данных обновляется ежегодно;

- «Научные кадры высшей квалификации 21 века». Содержит информацию о ведущих учёных и специалистах России. Сведения получены на основе ведущегося во ВНТИЦ федерального фонда докторских диссертаций, защищённых по всем отраслям знаний. База данных обновляется ежегодно.

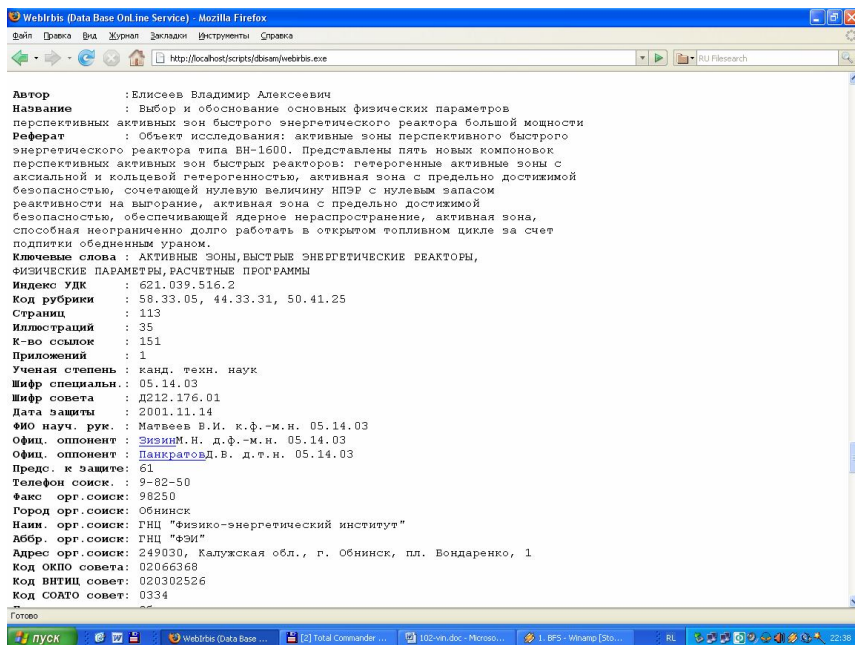


Рис. 10.26. Информационная карта диссертации.

Индивидуально, по запросам заказчиков создаются и поддерживаются проблемно-ориентированные базы данных о НИОКР, отраслевые базы данных по всем областям науки и техники. Объем определяется заказчиком.

Главная страница сайта ВНТИЦцентра (www.vntic.org.ru) представлена на рис. 10.25.

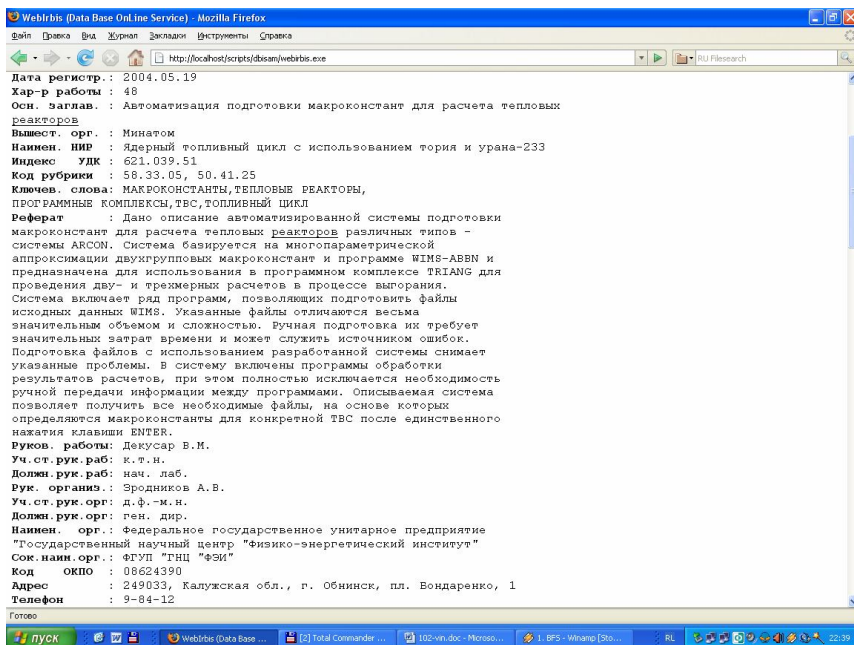


Рис. 10.27. Информационная карта НИОКР.

Через сеть Internet доступны следующие базы данных:

- реферативной информации о научно-исследовательских и опытно-конструкторских работах (ИК), проводимых в России (до 1991 г. – в СССР) и финансируемых из госбюджета. Ретроспектива базы данных 25 лет (с 1982 г.), объем – более 1,2 млн документов;
- реферативной информации о кандидатских и докторских диссертациях (ИКД) по всем отраслям знаний, защищенных в России (до 1991 г. – в СССР). Ретроспектива базы данных 25 лет (с 1982 г.), объем – более 500 тыс. документов;
- оперативной информации о вновь начинаемых научно-исследовательских работах (РК) во всех областях науки, техники и культуры. Ретроспектива базы данных 18 лет (с 1990 г.), объем – более 300 тыс. документов.

На рис. 10.26 и 10.27 приведены примеры информационных карт диссертаций и НИОКР.

Формирование, распространение на CDROM и доступ через Internet осуществляются с помощью АИС IRBIS, функциональные возможности и основные интерфейсные решения были описаны применительно к информационным ресурсам ИНИОН РАН.

10.4. Базы патентной информации ФИПС

Федеральное государственное учреждение «Федеральный институт промышленной собственности» (ФИПС) Федеральной службы по интеллектуальной собственности, патентам и товарным знакам обеспечивает формирование и ведение баз данных по отечественным патентам и товарным знакам. Главная страница ФИПС (www.fips.ru) представлена на рис. 10.28.

Пользователям доступны несколько баз, в том числе следующие.

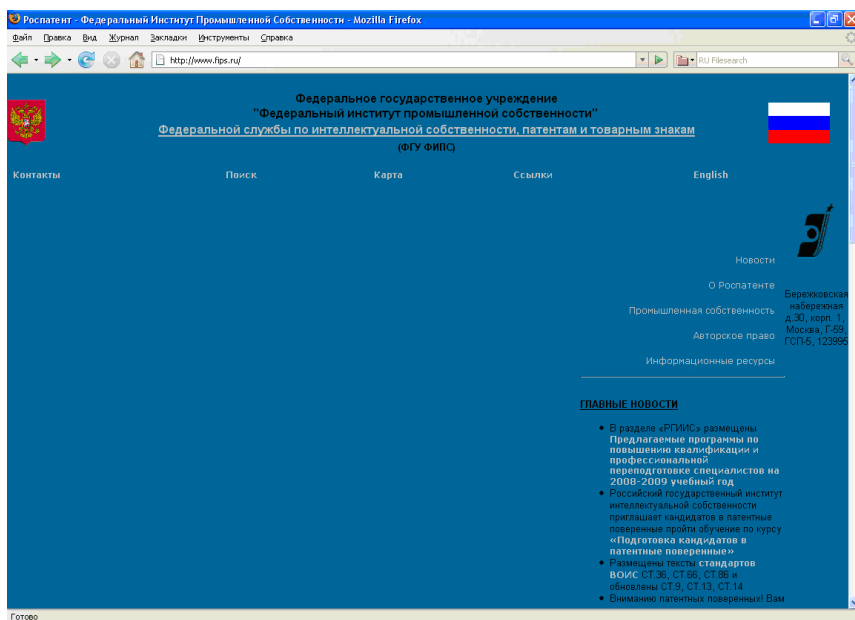


Рис. 10.28. Главная страница ФГУ ФИПС

- Реферативная база данных по изобретениям на русском языке, которая содержит информацию о патентах Российской Федерации и заявках на изобретения с 1994 по 2007 гг. с текущим пополнением. В состав документа входят библиографические данные, реферат, основной чертеж.

- Полнотекстовая база данных по изобретениям, которая содержит информацию о патентных документах СССР и России на изобретения с 1924 г. в факсимильном виде. В состав полнотекстового документа входят библиографические данные, реферат, формула изобретения, описание изобретения, чертежи и иные материалы, необходимые для понимания сущности изобретения.

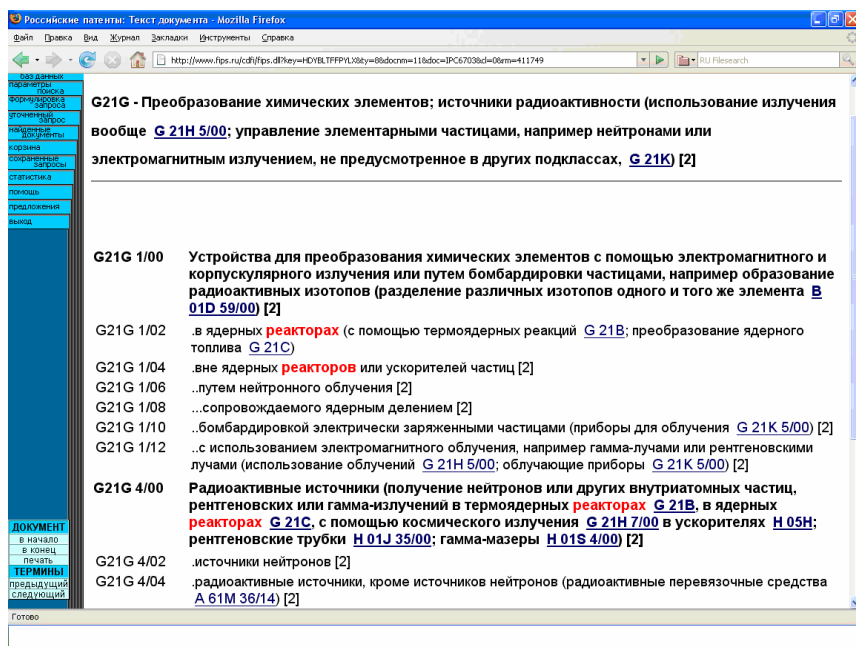


Рис 10.29. Форма отображения МПК

- База данных по товарным знакам, которая содержит информацию о товарных знаках, зарегистрированных в Российской Федерации с 1991 по 2007 гг., с текущим пополнением за 2008 г. В состав документа входят библиографические данные (включая

описание классов и словесное воспроизведение знака), графическое воспроизведение знака, извещения.

- База данных наименований мест происхождения товаров, которая содержит информацию о наименованиях мест происхождения товаров, зарегистрированных в Российской Федерации с 1991 по 2007 гг., с текущим пополнением за 2008 г. В состав документа входят сведения о регистрации и предоставлении права пользования наименованием места происхождения товара (библиографические данные и др.), извещения, включая сведения об изменениях в правовом статусе свидетельств, и др.

ФИПС также обеспечивает доступ к текстам международных патентных классификаций и международных классификаций по товарам и услугам: Пример раздела МПК приведен на рис. 10.29.

Данные в БД стандартизованы. Состав и наполнение полей представлены в табл. 10.1.

Табл. 10.1.

Состав и наполнение описания патента

Код поля	Код ИНИД	Содержание поля	Пример
Основной текст	54 57	Название, реферат, описание, формула	Устройство предназначено для
F110_num	11	Номер документа	2087100
F130	13	Вид документа	A A1 A2 A3 A4 C C0 C1
F140_date	14	Дата публикации	1997.12.20
F190	19	Страна публикации	RU SU
F210RU_num	21	Регистрационный номер заявки	95108597
F220_date	22	Дата подачи заявки	1997.12.20
F240RU	24	Дата начала действия патента	1997.12.20
F300RU	30	Приоритетные данные	US/21.05.87/53215
F310	31	Номер конвенционной заявки	M 191; A 0001432 (возможны варианты)

Код поля	Код ИНИД	Содержание поля	Пример
F320	32	Дата подачи конвенционной заявки	1992.03.24
F330	33	Страна приоритета	DE US и т.д.
F430	43	Дата публикации заявки	1998.02.10
F460	46	Дата публикации формулы изобретения	1998.02.10
F511	51	Основной индекс МПК	C07J41/00
F512	51	Дополнительные индексы МПК	A61K31/565
F516	51	Номер редакции МПК	7
Title	54	Название	Машина для посадки картофеля
F560	56	Аналоги изобретения	Авторское свидетельство СССР N 1106109 кл. В 63С 9/08, 1983
F711	71	Имя заявителя	Хохлов Виктор Васильевич
F721RU	72	Имя изобретателя	Федоров В.В.; Горольд Э.Вудроу[US]
F731	73	Имя патентообладателя	Институт проблем механики РАН
F741	74	Патентный поверенный	Матвеева Н.А.
F850	85	Дата соответствия ст.22/39 РСТ	1995.01.20
F860	86	Номер и дата международной или региональной заявки	РСТ/EP 89/101049 110989; FI 91/00392 (17.12.91)
F870	87	Номер и дата публикации	WO 94/00435 (06.01.94)
F960	96	Данные публикации ЕАПО	-----
F980	98	Адрес для переписки	Почтовый адрес

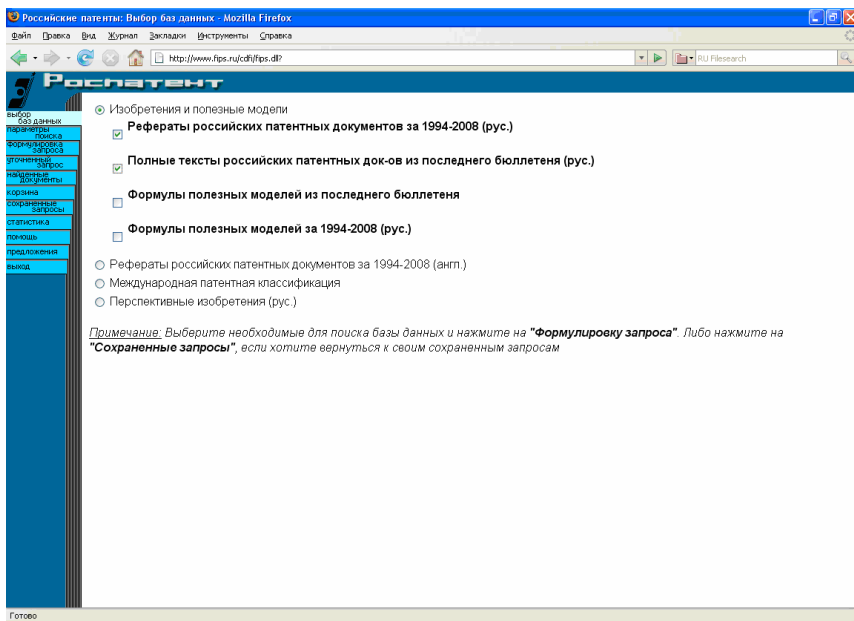


Рис 10.30. Форма выбора баз данных

Перед началом поиска необходимо выбрать базы данных, в которых будет производиться отбор документов по запросу. Форма установки параметров поиска приведена на рис 10.30.

ИПС позволяет осуществлять три вида поиска: нечеткий, словарный, логический.

Нечеткий поиск основан на сравнении не слов, а «битовых образов» запроса и документа, т.е. происходит сравнение набора нулей и единиц, представляющих собой битовый образ запроса, с наборами нулей и единиц, представляющих собой битовые образы документов, по определенному оригинальному алгоритму, позволяющему находить наиболее похожие сочетания.

В результате пользователь ИПС может найти документы, содержащие слова с ошибками в написании (например, после распознавания средствами OCR или ошибочного ручного ввода), с неправильными транслитерациями и т.д.

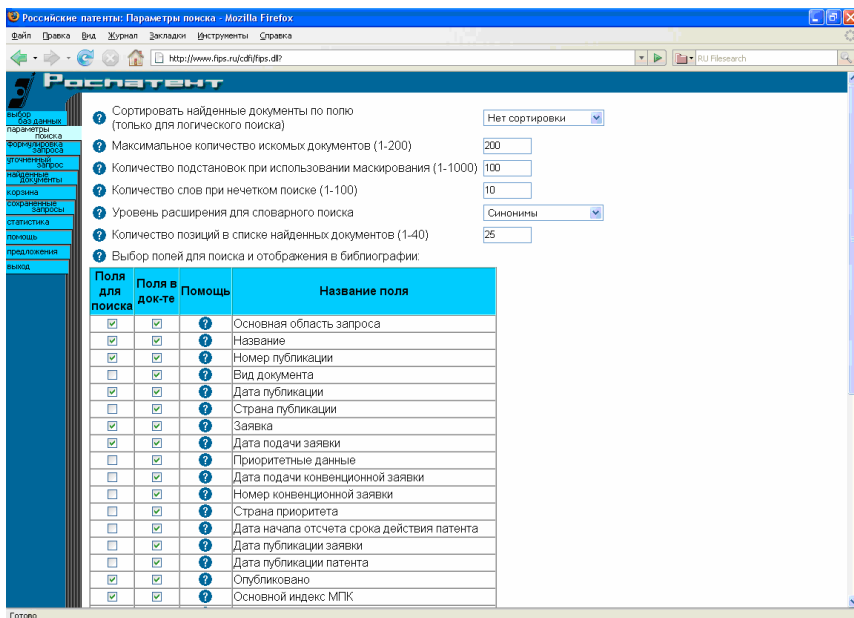


Рис 10.31. Форма установки параметров поиска

Словарный вид поиска использует семантическую сеть русского или английского языков. Кроме слов и/или словосочетаний, введенных пользователем в запрос, попадают слова и/или словосочетания из тезауруса, связанные с искомыми каким-либо видом семантической связи (синонимы, выше, ниже, ассоциация и т.д.).

Для данного вида поиска можно использовать маскирование символов слов запроса. Для слов с маскированием поиск семантически связанных терминов не производится. Параметр *Уровень расширения для словарного поиска* позволяет определить глубину использования тезауруса при поиске.

Логический поиск позволяет производить отбор по словам текста документа с использованием логических и/или контекстных операторов и скобочных конструкций.

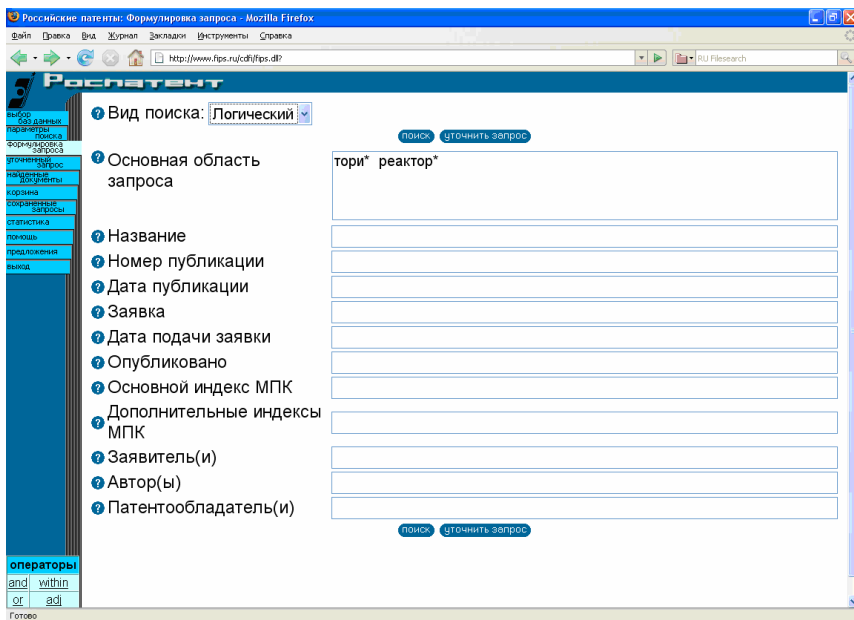


Рис 10.32. Форма подготовки запроса

Допускается маскирование символов слов запроса следующими операторами:

@ – один символ алфавита;

– одна цифра;

* – много или не одного символа;

? – точно один символ ;

[^] – какой-либо символ за исключением;

[] – может включать один из перечня символов в скобках.

Параметры поиска (рис. 10.31) определяют точность поиска и количество найденных документов; позволяют задавать набор библиографических (форматных) полей, используемых для поиска и/или выводимых в документе при его просмотре; задавать порядок сортировки списка найденных документов.

Форма подготовки запроса приведена на рис 10.32.

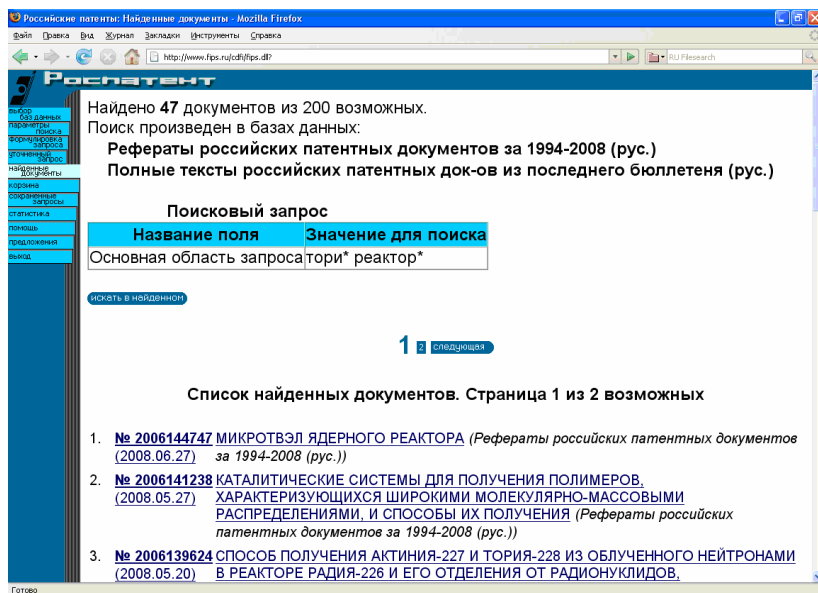


Рис 10.33. Форма просмотра результатов поиска

В результате проведения поиска ИПС выдает список найденных документов (рис. 10.33). Каждый документ списка содержит номер документа, дату его публикации, название документа (по товарным знакам - словесное воспроизведение товарного знака) и название БД. Галочка в окне слева от номера указывает на то, что документ уже помещен (отобран) в корзину. Общее количество документов, найденных в результате поиска, можно ограничивать, используя параметр *Максимальное количество искомых документов*.

В верхней части экрана помещена информация о количестве найденных в результате поиска документов, список БД, в которых проводился поиск, и приведена формулировка запроса. Каждый экран списка содержит (по умолчанию) информацию о 25 документах. Количество документов, выводимых на экран, можно изменить, используя параметр *Количество позиций в списке найденных документов*.

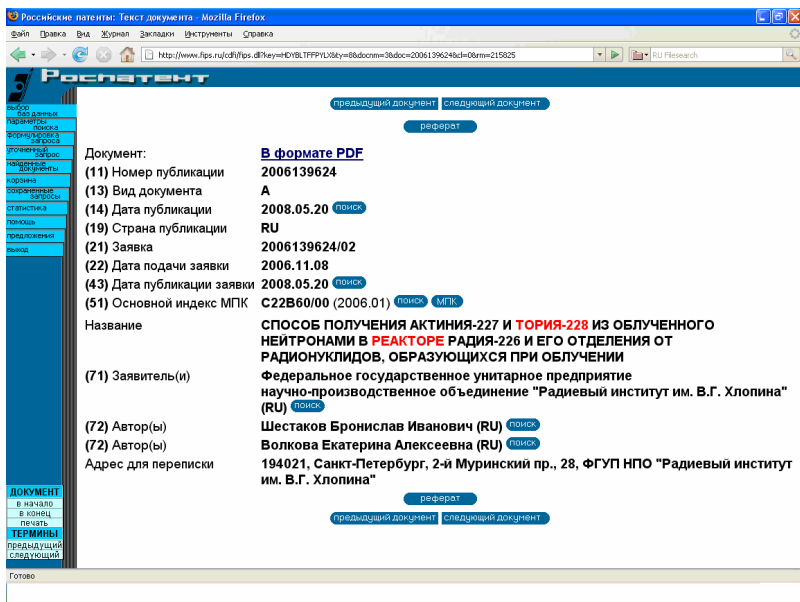


Рис. 10.34. Текст документа – описания патента

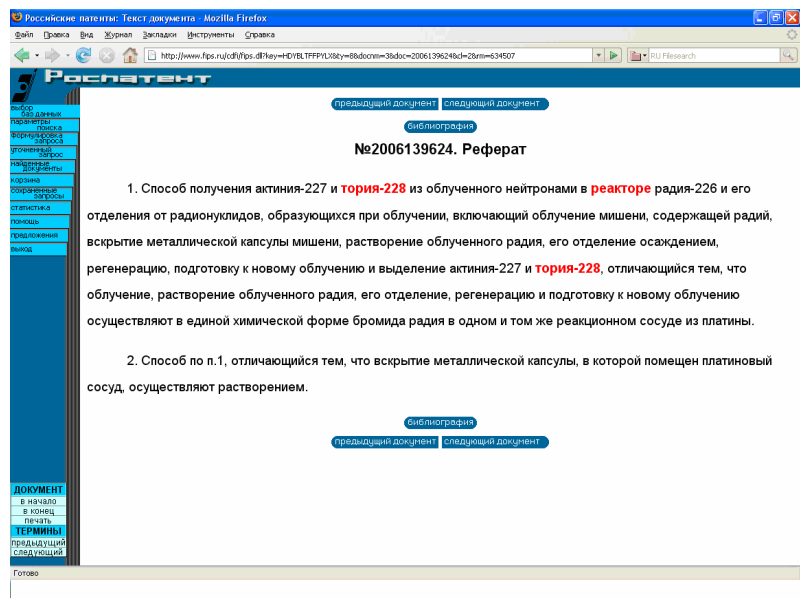


Рис. 10.35. Форма просмотра реферата

По гиперссылке от названия или номера можно просмотреть текст документа – описания патента (рис. 10.34).

Для просмотра реферата необходимо инициировать кнопку «Реферат», расположенную в верхней части формы (рис. 10.35).

Для просмотра электронной копии документа необходимо инициировать ссылку в поле «Документ», расположенную в первой строке описания (рис. 10.36).

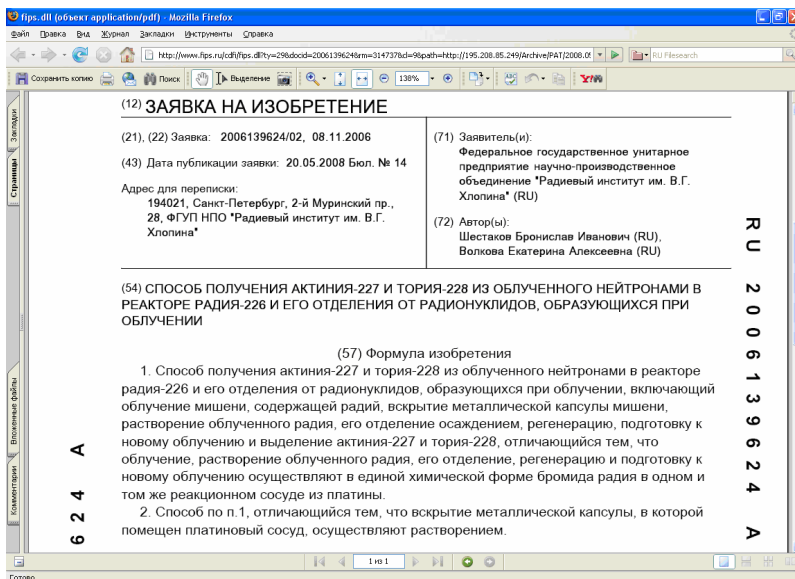


Рис. 10.36. Полный текст документа

10.5. Базы данных Международного агентства по атомной энергии

Информационная служба по ядерной информации INIS Международного агентства по атомной энергии, главная страница которого представлена на рис. 10.37, ведет формирование и распространение баз данных научной информации в области ядерной науки и техники.



Рис. 10.37. Главная страница INIS

База данных INIS на 2008 г. содержит 2.9 млн библиографических ссылок и рефератов журнальных статей, научных и технических отчетов, материалов конференции, книг, патентов, стандартов и Web-документов, изданных на 63 языках и имеющих отношение к ядерной физике. Записи включают ключевые слова и реферат на английском языке

Тематическое распределение публикаций, представленных в БД INIS, иллюстрируется диаграммой на рис. 10.38.

База данных также включает более 850 000 полнотекстовых документов, в том числе «серую» литературу.

Средний годовой прирост более 100 000 записей.

ФГУП ЦНИИАтоминформ, web-страница которого представлена на рис. 10.39, является официальным представителем INIS на территории РФ.

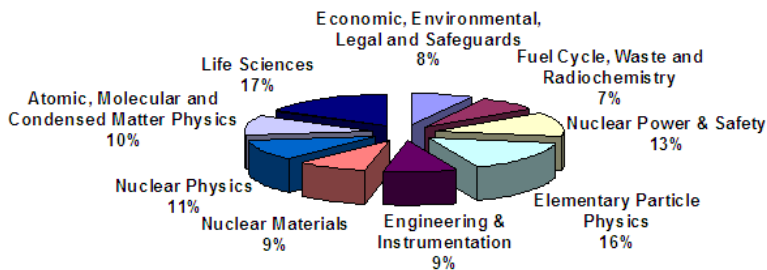


Рис. 10.38. Тематическое распределение публикаций

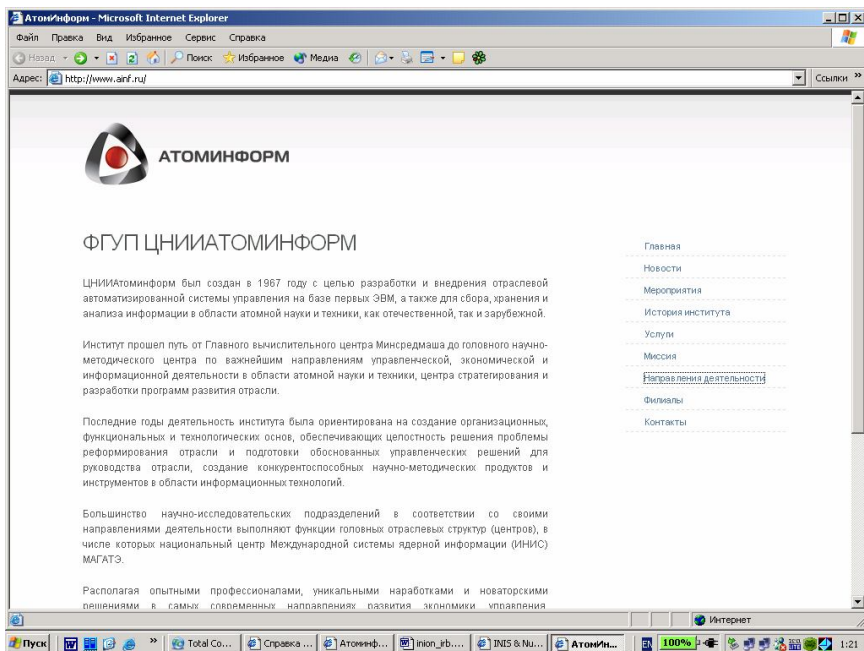


Рис. 10.39. Главная страница ЦНИИАтоминформ

Функциями ЦНИИАтоминформ являются: предоставление доступа к полной библиографической базе данных ИНИС на CD-ROM, к текущим пополнениям базы банных INIS Atomindex, к

полнотекстовой базе данных труднодоступной литературы ИНИС, к базе данных ИНИС на сайте МАГАТЭ в Internet, к полнотекстовым базам электронных публикаций ряда зарубежных научных организаций и издательств научно-технической литературы, сотрудничающих с Секцией ИНИС МАГАТЭ, к банку данных компьютерных программ Агентства по ядерной энергии Организации экономического сотрудничества и развития, к учебным, методическим, нормативным и рекламным разработкам ИНИС.

Институт осуществляет ввод отечественной информации в ИНИС, получает и распространяет получаемую продукцию среди отечественных потребителей.

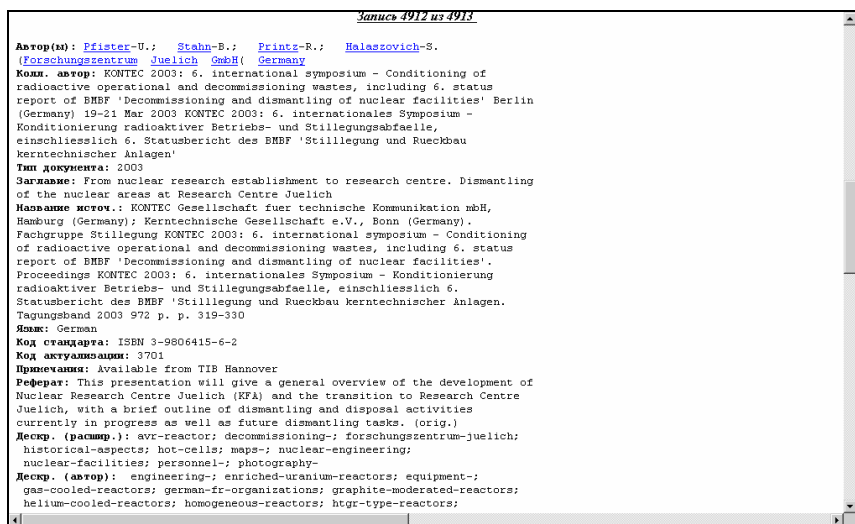


Рис 10.40. Пример документа БД INIS

Институт ведет работу по созданию Отраслевой системы стандартных справочных данных о свойствах веществ и материалов в различных направлениях атомной науки, техники и технологии (ОСССД).

Доступ к базам данных INIS и САПИ через Internet в ЦНИИ-Атоминформ осуществляется с помощью АИС IRBIS, функциональные возможности и основные интерфейсные решения которой

были описаны применительно к информационным ресурсам ИНИ-ОН РАН.

На рис 10.40 и 10.41 представлены примеры документов баз данных INIS и CARI соответственно.

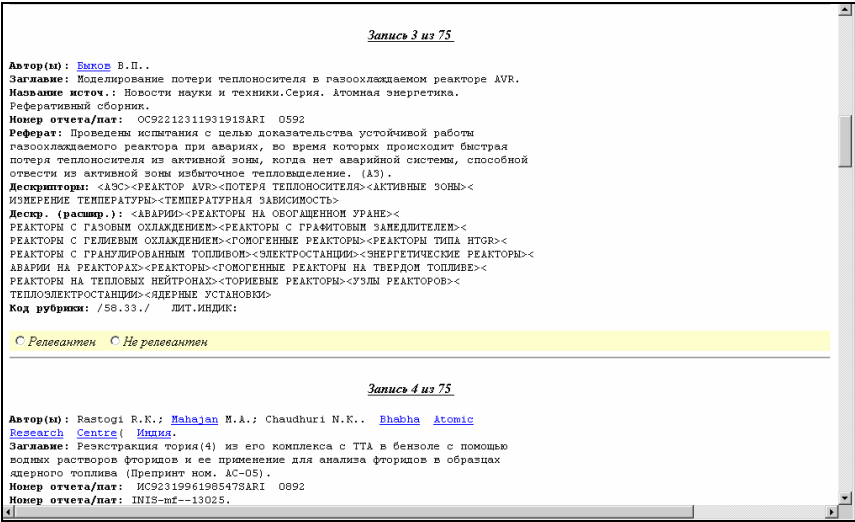


Рис 10.41. Пример документа БД CARI

ГЛАВА 11. ЭЛЕКТРОННЫЕ РЕСУРСЫ ИЗДАТЕЛЬСТВ

Издательства, традиционно выпускающие литературу на бумажном носителе, обычно обеспечивают их доступность в электронной форме через Internet. Для этого электронные копии размещаются на собственном сайте или на сайте доверенной организации.

Обычно доступность материалов ограничена: для каких-то категорий пользователей, например, доступна только справочная информация и ограничен доступ к полным текстам, для других доступны материалы только по ограниченной тематике и т.д.

Далее проиллюстрируем основные возможности, предоставляемые сервисами Science Direct и eLibrary.

11.1. Сервис Science Direct

Science Direct обеспечивает доступ к одной из крупнейших в мире электронной коллекции библиографической и полнотекстовой информации по различным разделам науки и техники, медицине, биологии и т.д. Система обеспечивает не только поиск и получение публикации, но позволяет зарегистрированному пользователю сохранять результаты поиска, создавать собственные профили и сохранять ссылки на другие ресурсы.

Главная страница Science Direct (www.sciencedirect.com), приведенная на рис. 11.1, включает форму запроса «Быстрого поиска», где можно задать терминологическое выражение для поиска публикации по заглавиям, аннотациям, ключевым словам, дополнительно указать автора, название, год, номер выпуска.

На рис 11.2. приведена форма для профессионального поиска, позволяющая использовать в выражении запроса логические операторы.

Альтернативой поиску отдельной публикации по запросу является режим просмотра выпусков. Для этого под формой «Быстрого поиска» в левой стороне размещается алфавитный и предметный указатели, позволяющие вывести список наименований выпусков в алфавитном порядке (рис. 11.3), или в соответствии с принадлежностью выбранной рубрике.

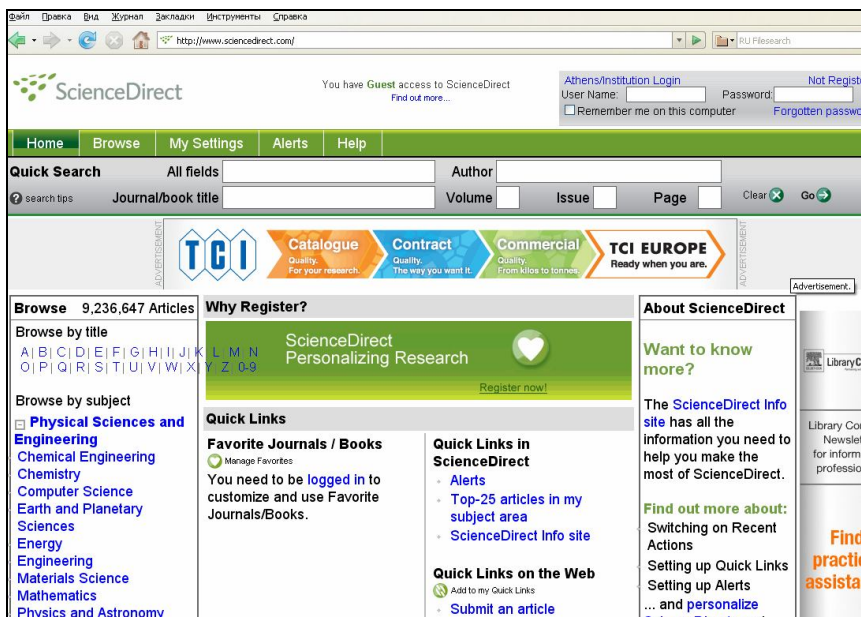


Рис 11.1. Главная страница

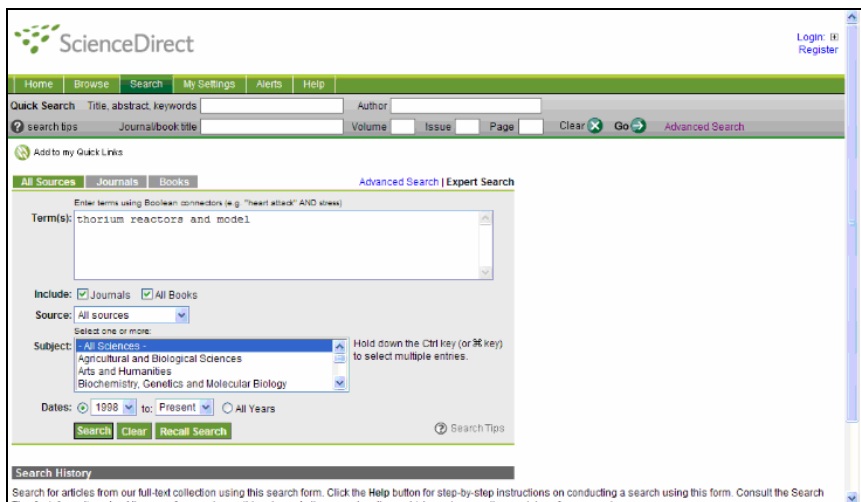


Рис 11.2. Форма запроса для профессионального поиска

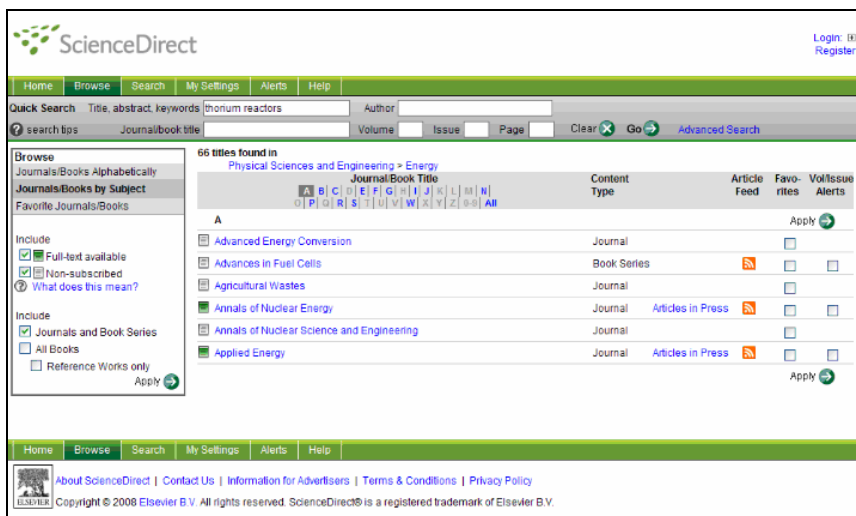


Рис 11.3. Форма вывода изданий в алфавитном порядке

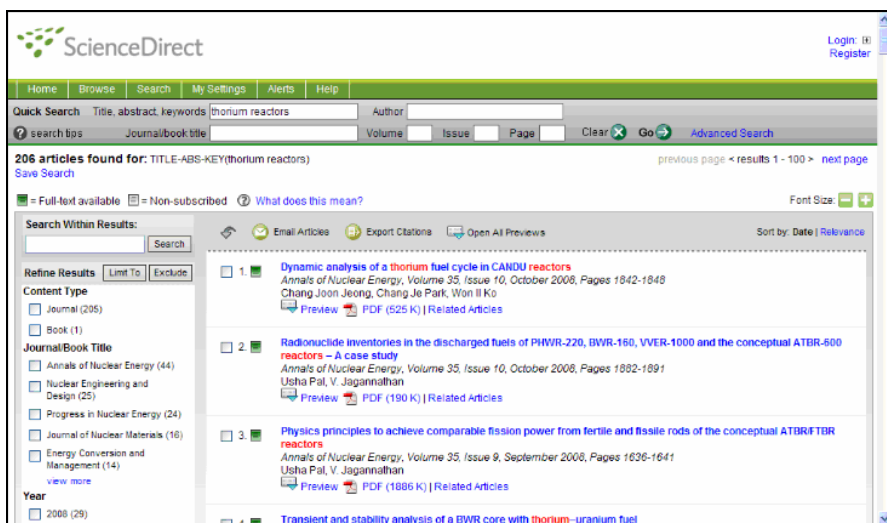


Рис 11.4. Результаты поиска по запросу

На рис. 11.4 представлена форма просмотра результатов поиска.

Форма позволяет вывести аннотацию публикации (рис. 11.5). Форма также обычно имеет ссылку на электронную копию, размещенную под соответствующим описанием. Инициирование ссылки позволяет загрузить копию, пример чего приведен на рис.11.6.

Кроме того, форма просмотра результатов поиска позволяет инициировать поиск похожих документов, инициировав ссылку «Related Articles», размещенную под соответствующим описанием. Результат такого поиска приведен на рис.11.7.

Форма просмотра аннотации позволяет вывести таблицы и рисунки, а также пристатейные ссылки (библиографию) документа (рис.11.8), от которых можно перейти в текст к месту их цитирования, (рис.11.9) или экспортировать в файл (рис.11.10).

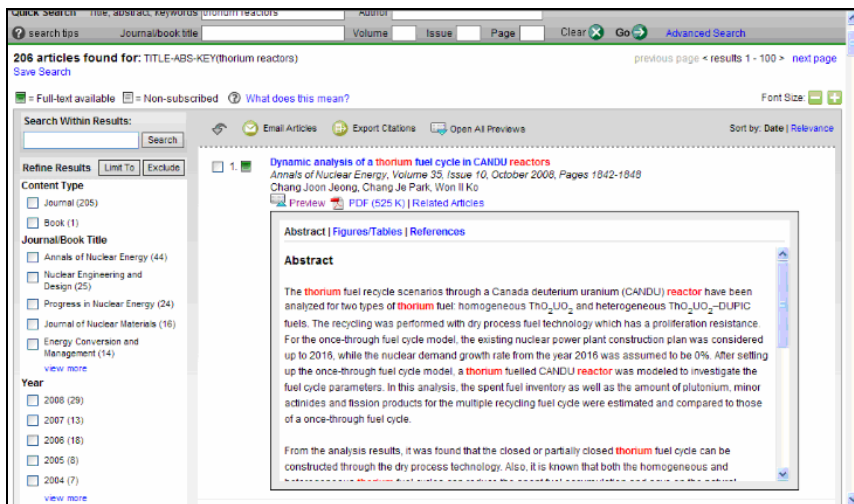


Рис 11.5. Аннотация документа

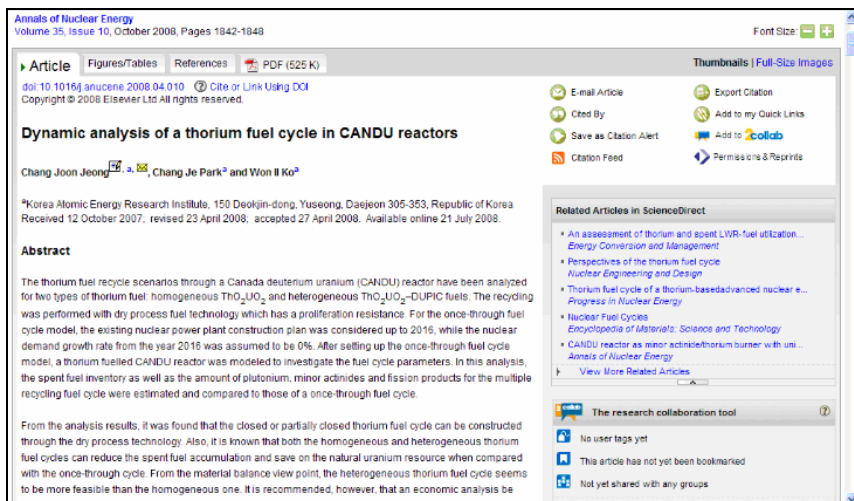


Рис 11.6. Электронная копия документа

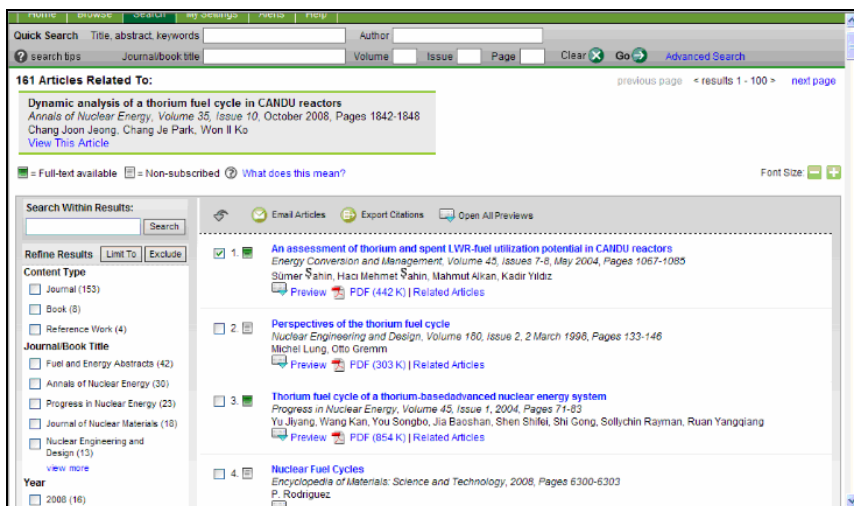


Рис 11.7. Результаты поиска похожих документов

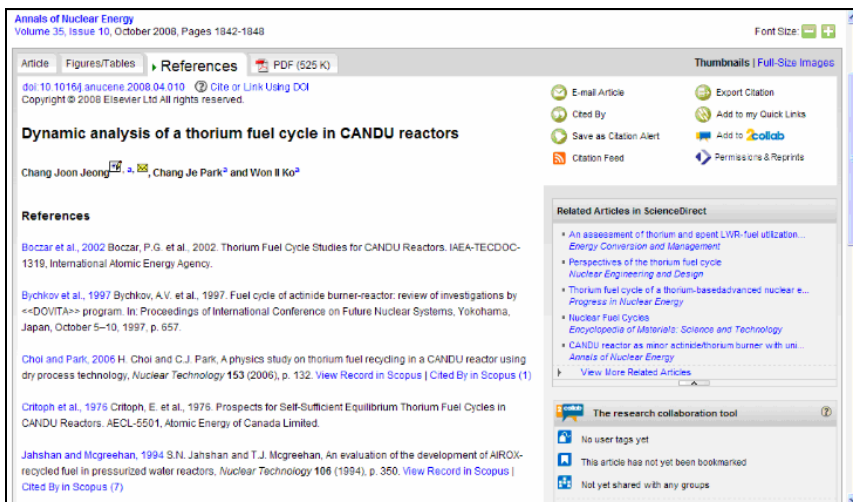


Рис 11.8. Пристатейные ссылки (библиография) документа

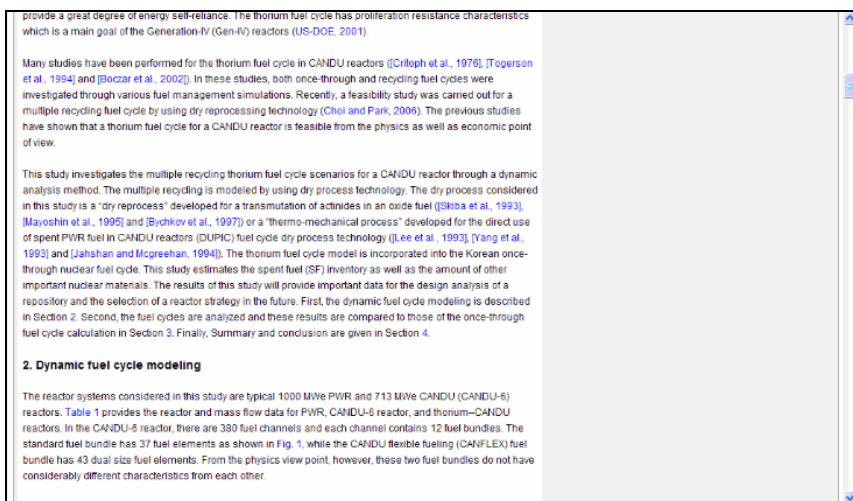


Рис 11.9. Переход по пристатейной ссылке

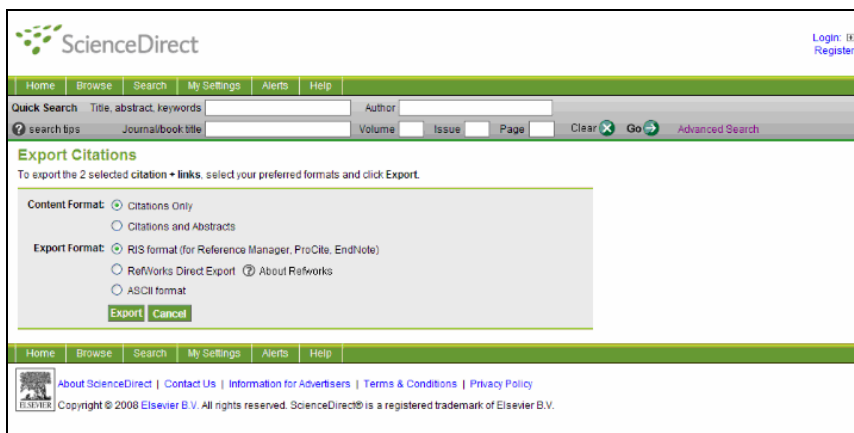


Рис 11.10. Экспорт приставочной ссылки

11.2. Научная электронная библиотека eLibrary

Научная электронная библиотека eLibrary.ru, созданная в 1998 г. при финансовой поддержке РФФИ, в настоящий момент является крупнейшей научной электронной библиотекой на территории Российской Федерации. Ее фонды содержат около 200 тыс. выпусков научных журналов ведущих зарубежных издательств. Среди них такие издательства, как Elsevier Science, Kluwer Academic Publishers, Springer, Blackwell, Royal Society Of Chemistry, Institute Of Physics (London), Academic Press и др.

Библиотека также содержит базы данных научного цитирования Института научной информации США Science Citation Index Expanded, Social Science Citation Index, базы данных по математике и медицине, журналы некоторых российских издательств.

Главная страница сайта библиотеки (<http://www.elibrary.ru>) представлена на рис. 11.11.

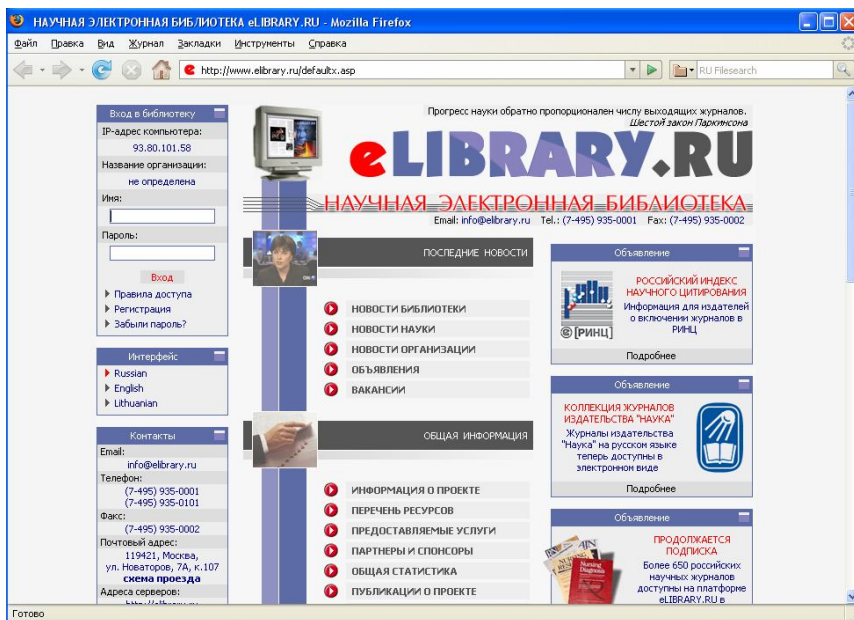


Рис. 11.11. Главная страница сайта электронной библиотеки eLibrary.ru

Регистрация и начало работы. Читательским билетом, разрешающим пользоваться библиотекой, в eLibrary.ru является логин и пароль пользователя. Для их получения необходимо заполнить регистрационную форму, содержащую сведения о читателе и об организации (рис. 11.12). Попасть на данную форму можно, воспользовавшись гипертекстовой ссылкой «Регистрация» в меню «Вход в библиотеку». После успешной регистрации вход в библиотеку возможен с любого компьютера, подключенного к сети Интернет.

e-LIBRARY		Новости библиотеки		Help	
 ПЕРСОНАЛЬНАЯ КАРТОЧКА					
Фамилия		Имя		Отчество	
Пол		Дата рождения			
Для Российской Федерации необходимо, кроме страны, указать также регион					
Страна		Российская Федерация		Регион	
Название населенного пункта указывается без добавления "г.", "пос." и т.д. Если у Вас нет электронной почты, поставьте в поле "E-mail" прочерк.					
Город		E-mail			
Выберите из списка зарегистрированных организаций организацию, в которой Вы работаете или учитесь. Если Вы не обнаружите организацию в списке, то введите ее полное (без сокращений) название вручную.					
Организация					
Должность должна быть указана полностью, без сокращений					
Должность		Статус			
Нажмите на кнопку "Добавить" справа и выберите разделы рубрикатора, отражающие Ваши научные интересы. Для удаления кода выделите его в списке и нажмите на кнопку "Удалить"					
Разделы рубрикатора		Добавить Удалить			
Укажите имя пользователя и пароль (не менее 6 символов)					
Имя		Пароль			
Сохранить введенную информацию			Выход без сохранения		
Информация, указанная Вами при заполнении данной анкеты, никогда не будет передана третьим лицам без Вашего разрешения или использована для рассылки нежелательных рекламных материалов.					

Рис. 11.12. Форма регистрации пользователя

Заходя в библиотеку по своему личному логину и паролю, пользователь, фактически, попадает в *свою собственную* библиотеку и получает возможность менять ее внешний облик, создавать персональные группы журналов, проводить поиск во всех без исключения ресурсах библиотеки, сохранять условия и результаты поиска и т.п.

Каталог журналов. Поиск нужной статьи в Каталоге журналов начинается с выбора пункта «Каталог журналов» в разделе «Навигатор» Главной страницы.

Пользователю предоставляется список журналов, в котором названия журналов оформлены как гипертекстовые ссылки, инициируя которые можно далее получить информацию о выпусках журнала (рис. 11.13). Имеется возможность сортировать журналы по алфавиту в прямом и обратном порядке, по ISSN, по названию издательства, по количеству выпусков. Слева от списка журналов находятся окна управления. С их помощью можно перейти на журналы, начинающиеся на конкретную букву, или провести поиск журнала по ключевым словам в его названии.

The screenshot displays the 'e-LIBRARY' interface for journal catalogs. On the left, there is a 'Навигатор' (Navigator) sidebar with options like 'Поиск по названию' (Search by name), 'На английском языке' (In English), and 'На русском языке' (In Russian). The main area features a table of journals under the heading 'НАУЧНЫЕ ЖУРНАЛЫ' (Scientific Journals). The table has columns for 'ISSN', 'Название журнала' (Journal Name), 'Издательство' (Publisher), and 'Выпусков' (Issues). The first few rows of the table are as follows:

ISSN	Название журнала	Издательство	Выпусков
1122-1151	Axiomathes	KAP	4
0929-5593	Autonomous Robots	KAP/B	34
1387-2532	Autonomous Agents and Multi-Agent Systems	KAP/B	19
1474-8665	Autonomic & Autacoid Pharmacology	Blackwell Science	1
0926-5805	Automation in Construction	Elsevier	48
0005-1098	Automatica	Pergamon	104
0928-8910	Automated Software Engineering	KAP/B	25
1385-1306	Austrian Review of International and European Law	KAP	5
0045-0766	Australian Occupational Therapy Journal	Blackwell Science	8
1038-5282	Australian Journal of Rural Health	Blackwell Science	14
0312-8962	Australian Journal of Management	University of New South Wales	2
1326-6756	Australian Journal of Entomology	Blackwell Science	10
0812-0099	Australian Journal of Earth Sciences	Blackwell Science	13
0004-0505	Australian and New Zealand Journal of	Blackwell Science	10

Рис. 11.13. Работа с Каталогом журналов

По гипертекстовой ссылке от названия журнала происходит переход к отсортированному по годам списку выпусков журнала и далее аналогичным образом – на содержание нужного номера (рис.11.14, 11.15).

проса либо ключевые слова из названия статьи, либо фамилию автора.

Информация о статье в оглавлении представляет собой набор гипертекстовых ссылок, по которым можно перейти к дополнительной информации, например, по названию издания можно получить краткий обзор этого издания по годам; по году публикации можно посмотреть обзор данного года.

По гипертекстовой ссылке от названия статьи отображается полная информация о статье (рис. 11.16): название, фамилии авторов, название журнала, год выпуска, номер тома, язык, страницы, тип, код статьи и аннотация. В зоне «Полный текст» будут представлены ссылки, по которым можно получить полный текст (если он доступен для организации). Если статья не доступна, то будут даны ссылки на условия доступа и на ближайшую организацию, в которой доступна искомая статья.

Режим поиска по запросу. Чтобы получить подборку статей, отвечающих определенным поисковым условиям, нужно перейти в раздел *Персональный профиль* и выбрать пункт *Поисковые запросы*. Форма поискового запроса представлена на рис. 11.17.

Поисковое выражение формируется по правилам запроса по образцу: условие, введенное для каждого следующего поля, уточняет запрос. Для поиска используются поля: Ключевые слова (слова из текста), Авторы, Статьи, Журналы, Выпуски.

Ключевые слова могут объединяться логическими операциями **AND**, **OR**, **NOT** и контекстной операцией **NEAR** (термины, объединенные данным оператором, должны присутствовать в найденном документе в одном и том же предложении).

При составлении поискового запроса может использоваться символ маскирования - «*», означающий любое количество символов.

Результаты поиска могут быть отсортированы. Может быть также определен размер страницы для вывода результатов.

Текст запроса может быть сохранен и использован при дальнейшем поиске.

e-LIBRARY		Результаты запроса		Help	
ИНФОРМАЦИЯ О СТАТЬЕ					
Название статьи	ДВУХЧАСТОТНЫЙ ТВЕРДОТЕЛЬНЫЙ ЛАЗЕР С ПОЛУПРОВОДНИКОВОЙ НАКАЧКОЙ: СОЗДАНИЕ, РАЗРАБОТКА ТЕОРИИ И ЕЕ ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА				
Авторы	И.В. Евсеев, Р.А. Карле-Шананин, Н.В. Наумов, В.Н. Петровский, Е.А. Шатрова				
Журнал	Информационный бюллетень РФИ				
ISSN		Издательство	РФИ		
Год выпуска	1996	Дата	январь		
Том	4	Номер	2		
Полный текст	Доступ закрыт. Вы можете ознакомиться с условиями получения доступа к журналам данного издательства. Вы можете также обратиться в ближайшую организацию , имеющую доступ к полному тексту данной статьи.				
Язык	русский	Тип	Научный отчет		
Страницы	98	Код статьи	РФИ 96-02-16288-а		
Адрес	Московский государственный инженерно-физический институт (МИФИ) 115409, г. Москва, Каширское ш., 31 Тел.: (095)3243384 Факс: 3242111				
Аннотация статьи	Создана теоретическая модель двухчастотного твердотельного лазера - система дифференциальных уравнений, описывающих поведение инверсной населенности, интенсивности и частоты генерации в режиме перестройки одной или обеих мод резонатора в той части контура линии усиления, где усиление превышает потери. Анализ системы уравнений показал, что она описывает основные физические процессы, влияющие на генерацию двухчастотных твердотельных лазеров: усиление и поглощение света, спонтанный распад инверсии и ее насыщение, а также конкуренцию мод, связанную с выжиганием пространственных дырок инверсии. Эта система уравнений переходит в традиционные уравнения Статса-де Марса в пределе одночастотного режима генерации. Разработан и создан двухчастотный YAG:Nd мини-лазер с торцевой накачкой полупроводниковым лазером. Генерация лазера реализована на длине волны 1.06 мкм. Получен режим устойчивой генерации двух линейных и взаимно ортогонально поляризованных аксиальных мод в диапазоне изменения межмодовых расстояний от 5 МГц до 9 ГГц. Реализовано максимальное значение области двухмодового режима генерации 9 ГГц. Для создания устойчивого двухчастотного режима генерации использован фазоанізотропный резонатор, традиционно применяемый научной группой при создании двухчастотных лазеров. Разработанные в процессе работы над проектом уравнения использовались для теоретического анализа конкуренции аксиальных мод различных поляризаций твердотельного лазера. Сравнение рассчитанных зависимостей с соответствующими экспериментальными позволило сделать вывод о слабой связи между аксиальными взаимно ортогонально поляризованными модами; это обстоятельство позволило реализовать максимально возможные значения области двухмодового режима генерации. Полученные экспериментально значения межмодовых расстояний позволяют реализовать сигнал межмодовых биений для широкого набора частот, вплоть до диапазона СВЧ. Полученные в работе над проектом результаты сравнимы с зарубежными аналогами, а по отношению сигнала к шуму превосходят их.				
Добавить статью в подборку		Новая подборка			

Рис. 11.16. Информация о статье

The screenshot displays the 'e-LIBRARY' search interface. At the top, there are navigation tabs: 'Подборки журналов', 'Подборки статей', and 'Help'. Below these is a section titled 'ВАШИ ПОИСКОВЫЕ ЗАПРОСЫ' (Your search queries) with a 'Поиск' (Search) button and a dropdown menu set to 'Новый поиск' (New search). To the right are buttons for 'Удалить' (Delete) and 'Переименовать' (Rename).

The main section is titled 'УСЛОВИЯ ЗАПРОСА' (Search conditions). It includes a dropdown for 'Поиск в названии, ключевых словах или аннотации' (Search in title, keywords or annotations). Below this is a text input field containing 'fractal aggregation'. To the right of the input field are buttons for 'AND', 'NOT', 'OR', and 'NEAR'. Below the input field are buttons for 'Добавить' (Add) and 'Удалить' (Delete).

There are several other input fields and buttons: 'Авторы' (Authors) with 'Добавить' and 'Удалить' buttons; 'Статьи' (Articles) with a dropdown; 'Журналы' (Journals) with 'Добавить' and 'Удалить' buttons; 'Выпуски' (Issues) with dropdowns for 'за указанный период' (for the specified period), 'начиная с' (starting from), 'и заканчивая' (and ending with), and 'январь' (January); 'Сортировка' (Sorting) with dropdowns for 'по степени соответствия' (by degree of correspondence), 'Порядок' (Order), 'по убыванию' (by descending), and 'Вывод' (Output) set to 'по 20 статей' (by 20 articles).

At the bottom, there are buttons for 'Поиск' (Search), 'Очистить' (Clear), 'Сохранить как' (Save as), and a dropdown for 'Новый запрос' (New query).

Рис. 11.17. Форма поискового запроса

Вывод результата. На рис. 11.18 представлен результат запроса. В верхней части выводятся данные о количестве найденных статей и об общем количестве статей. Немного ниже находится полоса управления списком статей. Если количество статей велико и не помещается на одной странице, при помощи этой полосы осуществляется переход к определенной, следующей, предыдущей, первой или последней странице.

Поиск по ключевым словам. В режиме «Навигатор» существует возможность просмотреть статьи по заданному ключевому слову или его фрагменту. Воспользоваться поиском по ключевому слову можно, выбрав пункт «Ключевые слова» на панели «Навигатор».

e-LIBRARY

Условия запроса

Новый запрос

Искать в найденном

Help

РЕЗУЛЬТАТЫ ЗАПРОСА

(всего найдено статей: 207 из 4698123)

В начало

<<

1

2

3

4

5

6

7

8

9

10

>>

В конец

1

Fractal aggregation in aqueous solutions of porphyrins

☐

N. Micali, F. Mallamace, L. Monsù Scolaro, A. Romeo

Physica A: Statistical Mechanics and its Applications, 249 (1998), 1-4 (январь 02), 501-510

2

Nucleation-limited aggregation of crystallites in fractal growth

☐

X.-Y. X., Y. Liu, D.-W. Li, C.S. Strom, N.-b. Ming, M. Wang

Journal of Crystal Growth, 208 (2000), 1-4 (январь 01), 687-695

3

Growth laws and spinodal decomposition type of scaling in fractal aggregation of colloids

☐

G. Ramirez-Santiago, A.E. Gonzalez

Physica A: Statistical Mechanics and its Applications, 236 (1997), 1-2 (февраль 15), 75-84

4

Aggregation of digital terrain data using a modified fractal interpolation scheme

☐

R. Bindlish, A.P. Barros

Computers & Geosciences, 22 (1996), 8 (октябрь), 907-917

5

Formation of fractal structures by aggregation of anisometric iron (III) hydroxide particles

☐

R. Brunner, S. Gall, W. Wilke, M. Zrinyi

Physica A: Statistical Mechanics and its Applications, 214 (1995), 2 (март 01), 153-161

6

Adapted diffusion limited aggregation model for the effects of magnetic fields on fractal electrodeposits

☐

T.R. N Mhochain, J.M.D. Coey

Journal of Magnetism and Magnetic Materials, 226 (2001), (май), 1281-1283

7

Large scale fractal aggregates using the tunable dimension cluster-cluster aggregation

Рис. 11.18. Вывод результатов поиска

e-LIBRARY

Рубрикатор

Авторский указатель

Поиск

Справка

КЛЮЧЕВЫЕ СЛОВА

	Ключевое слово	Статей
1	05.45.Df Fractals	1
2	1/f fractal	1
3	3D fractal analysis	1
4	47.53.+n Fractals	4
5	61.43.Hv Fractals; macroscopic aggregates	1
6	61.43.Hv Fractals; macroscopic aggregates (including diffusion-limited aggregates)	5
7	64.60.Ak Renormalization-group, fractal, and percolation studies of phase transitions	9
8	64.60.Ak Renormalization-group, fractal, and	1

Рис.11.19. Список Ключевых слов

Открывшаяся страница слева, над панелью навигатора, будет содержать строку для ввода ключевых слов – «Ключевое слово». Введенное в строке слово (или фраза) обрабатывается после нажа-

тия на кнопку «Поиск». Результатом служит список предопределенных ключевых слов и выражений, в которых содержится введенное слово (рис. 11.19).

Полученный список можно сортировать по ключевым словам или по количеству статей на каждое ключевое слово.

Ключевое слово в списке служит ссылкой для просмотра статей, содержащих конкретное ключевое слово (рис. 11.20).

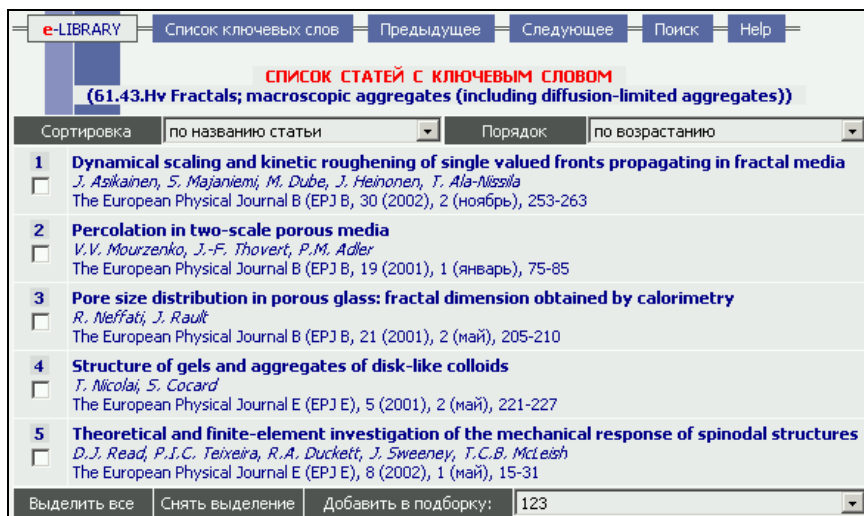


Рис.11.20. Результат навигации по Ключевым словам

Авторский указатель. Пункт «Авторский указатель» меню «Навигатор» открывает доступ к списку статей по автору.

После ввода фамилии автора (или ее начальной части) в строку поискового условия и вызова функции «Поиск» пользователь получает список авторов, в фамилии которых входит введенное слово (рис. 11.21).

Фамилии авторов в списке снабжены информацией о количестве написанных ими статей. Полученный список можно отсортировать по автору или по количеству статей, располагая их по возрастанию или убыванию (аналогично списку ключевых слов или рубрик).

e-LIBRARY

Рубрикатор

Ключевые слова

Поиск

Help

АВТОРСКИЙ УКАЗАТЕЛЬ

		Автор	Статей
1	☐	Gontijo, A.F.	1
2	☐	Gontijo, B.F.	2
3	☐	Gontijo, C.M.F.	1
4	☐	Gontijo, CM	1
5	☐	Gontijo, G.S.	1
6	☐	Gontijo, I.	1
7	☐	Gontijo, J.A.R.	1
8	☐	Gontijo, N.F.	1
9	☐	Gontijo, Possidonia F.D.	1
10	☐	Gontijo de Amorim, N.F.	1
11	☐	Gontijo de Amorim, Natale Ferreira	1

Выделить все

Снять выделение

Рис. 11.21. Страница Авторского указателя

Для просмотра статей искомого автора необходимо активизировать строку с его фамилией (как гиперссылку) в полученном списке авторов (рис. 11.22).

Если необходимо посмотреть статьи нескольких авторов, их фамилии нужно отметить флажком, расположенным слева, и нажать кнопку «Поиск» в верхнем меню.

Работа со списком полученных таким образом статей полностью аналогична работе со статьями при поиске по запросу или по Ключевым словам.

Тематический рубрикатор. Вход в пространство статей библиотеки возможен также и через Тематический рубрикатор. Для этой цели необходимо воспользоваться пунктом «Рубрикатор» в панели «Навигатор». Рубрикатор постоянно пополняется новыми

журналами, однако не все журналы в каждый момент времени могут быть внесены в него.



Рис. 11.22. Вывод статей через Авторский указатель

В тематическом рубрикаторе собраны все основные рубрики, представленные в библиотеке, с указанием количества журналов, отнесенных к рубрике (рис. 11.23).

Список рубрик можно отсортировать по коду, названию или по количеству журналов (по убыванию или по возрастанию).

e-LIBRARY			
Каталог журналов		Авторский указатель	Поиск
		Справка	
ТЕМАТИЧЕСКИЙ РУБРИКАТОР			
Код	Название рубрики	Журналов	
02.00.00	Философия	37	
03.00.00	История. Исторические науки	6	
04.00.00	Социология	53	
05.00.00	Демография	12	
06.00.00	Экономика. Экономические науки	71	
10.00.00	Государство и право. Юридические науки	64	
11.00.00	Политика. Политические науки	18	
12.00.00	Науковедение	8	
13.00.00	Культура	11	
14.00.00	Народное образование. Педагогика	33	
15.00.00	Психология	58	
16.00.00	Педагогика	10	

Рис. 11.23. Тематический рубрикатор

Название рубрики (как гиперссылка) позволяет перейти в полный список входящих в нее журналов, работать с которым можно так же, как и с каталогом журналов.

Рубрикатор имеет трехуровневую иерархическую структуру. Слева от каждой рубрики находится ее шестизначный код. Синими цифрами показывается текущий уровень отображенных рубрик. Красные цифры справа от синих служат отсылками к списку под-рубрик, а красные цифры слева возвращают в вышестоящую рубрику.

Нужную рубрику можно найти по названию. В строку запроса можно ввести одно или несколько слов или фрагмент из названия требуемой рубрики. Альтернативный переключатель определяет, будут ли показаны пустые рубрики, т.е. такие, в которых нет ни одного журнала. Результат поиска рубрик по названию представлен на рис. 11.24.

e-LIBRARY

Каталог журналов

Авторский указатель

Поиск

Справка

ТЕМАТИЧЕСКИЙ РУБРИКАТОР

Код	Название рубрики	Журналов
02.31.31	Философские и методологические проблемы математических, естественных и технических наук	3
06.35.51	Экономико-математические методы и модели	8
15.21.69	Методы психологического исследования. Математическая психология	1
27.00.00	Математика	133
27.01.00	Общие вопросы математики	15
27.01.09	История математики. Персоналия	1
27.01.45	Преподавание математики	2
27.03.00	Основания математики и математическая логика	10
27.03.19	Математическая логика	4
27.23.00	Математический анализ	4
27.35.00	Дифференциальные и интегральные уравнения математических моделей естественных наук	16
27.35.21	Математические модели гидродинамики	1

Рис. 11.24. Результат поиска рубрики по названию

Просмотр новых поступлений. Специальный пункт Навигатора («Новые поступления») позволяет вывести список журналов, в которых были обновления за определенный период.



Рис. 11.25. Просмотр новых поступлений

Слева от списка журналов находится панель управления, где можно задать период, за который нужно искать обновления (рис. 11.25). Период можно выбрать одним из альтернативных переключателей, а можно ввести вручную. Здесь же можно указать, по какой дате нужно проводить поиск: по дате выпуска журнала или по дате его установки (публикации) в электронной библиотеке. Над списком журналов находится раскрывающееся меню «Журналы», в котором можно выбрать какую-либо подборку (если она существует) и тогда поиск обновлений будет проходить только в тех журналах, которые входят в эту подборку.

Работать с полученными журналами можно так же, как и с каталогом журналов, включая управление подборками.

Функциональная кнопка «Поиск» над меню «Журналы» на странице новых поступлений позволяет провести поиск только среди новых статей.

Создание подборок статей и журналов. По мере работы с результатами поиска пользователь может отложить просмотр отдельных статей, поместив их в персональную подборку. Такая под-

борка будет сохранена в персональном профиле пользователя и ее содержимое можно будет посмотреть в любой момент без какого-либо поиска. Создать подборку можно прямо со страницы с результатом запроса. Для этого служит меню внизу страницы.

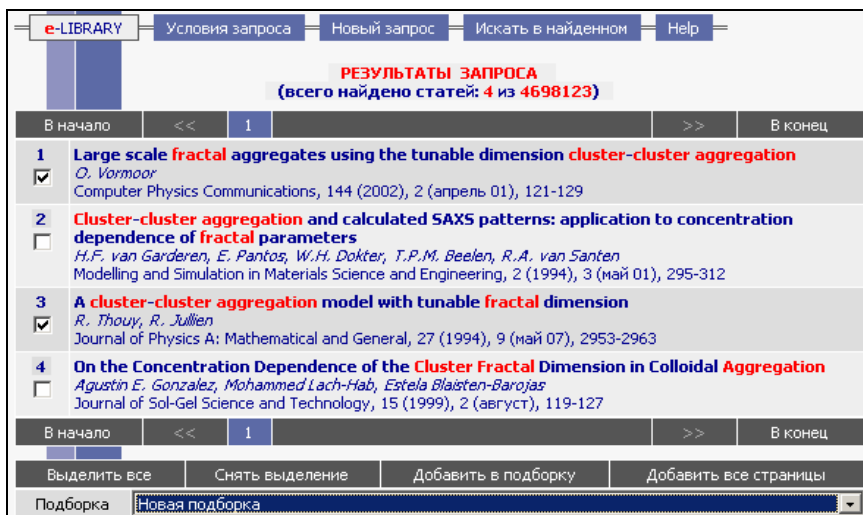


Рис. 11.26. Отметка статей для формирования подборки

Статьи, предназначенные для добавления в подборку, отмечаются слева от названия галочкой (рис. 11.26).

Кнопки «Выделить все» и «Снять выделение» отмечают или сбрасывают отметки статей. Кнопка «Добавить в подборку» добавляет в подборку отмеченные статьи, а кнопка «Добавить все страницы» добавляет в подборку все найденные статьи. В раскрывающемся меню можно выбрать подборку, в которую необходимо добавить статьи или же создать новую подборку. Для новой подборки нужно будет ввести ее название.

Персональные подборки можно создавать не только для статей, но и для журналов. Составить подборку журналов можно в каталоге журналов. Выбор журналов происходит аналогично выбору статей.

ГЛАВА 12. СПЕЦИАЛИЗИРОВАННЫЕ БД И ИР

12.1. Портал NUCLEUS

Примером специализированного портала в области использования ядерной энергии в мирных целях является портал NUCLEUS (МАГАТЭ). Портал разработан с целью облегчения доступа к базам данных международного агентства по использованию атомной энергии. Электронный адрес портала NUCLEUS – <http://nucleus.iaea.org> (рис. 12.1).

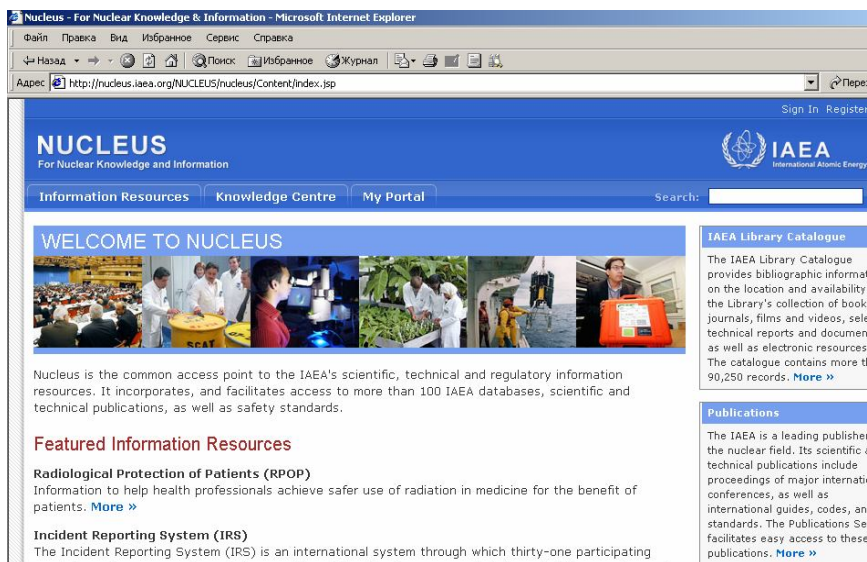


Рис. 12.1. Главная страница портала NUCLEUS

Через NUCLEUS можно выйти на базы данных, которые содержат постоянно обновляемую информацию о состоянии дел в ядерной отрасли. Базы данных разбиты на группы:

- Наука и технологии:
 - Ядерная наука и приложения;
 - Техническое сотрудничество;
 - Ядерная энергетика;

- Безопасность и физическая защита;
- Гарантии нераспространения ядерных технологий.

Кроме обеспечения выхода на различные сайты и базы данных на портале NUCLEUS в рамках центра знаний можно осуществить поиск нужной информации по следующим направлениям (рис. 12.2):

- эксперты в области использования ядерных технологий;
- различные материалы по ядерной тематике;
- конференции и рабочие встречи по ядерным технологиям;
- базы знаний по ядерным реакторам.

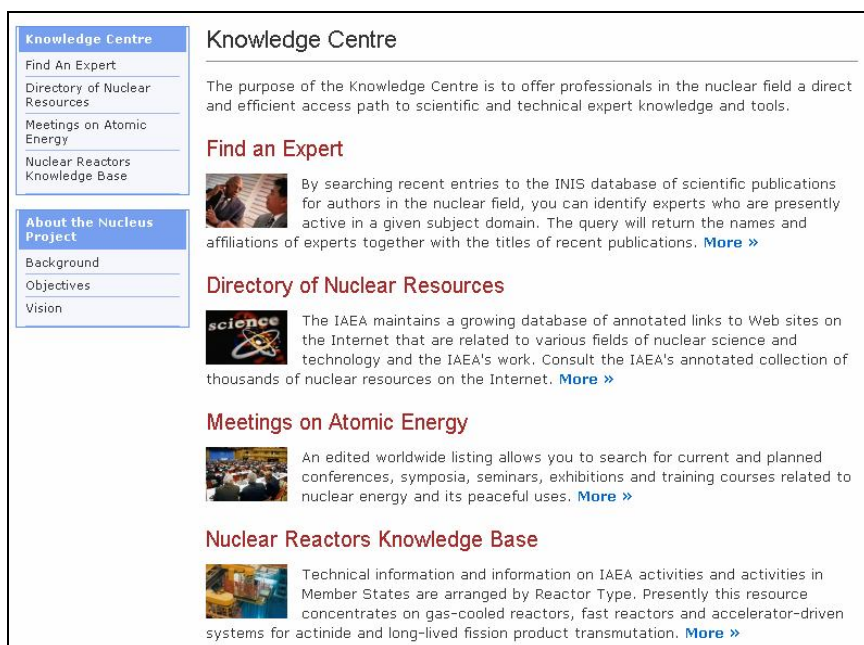


Рис. 12.2. Фрагмент страницы «Центра знаний» портала NUCLEUS

К достоинствам портала NUCLEUS можно отнести понятный интерфейс, регулярное обновление материалов, информационная полнота. Практически все материалы, издаваемые под контролем МАГАТЭ, можно найти и посмотреть или заказать, если зарегист-

рироваться на данном портале. Для получения некоторых материалов требуется произвести оплату.

В учебном процессе данный портал можно использовать несколькими способами. Во-первых, студентам можно ставить задачи по поиску обзорных материалов по конкретной тематике при написании курсовых и дипломных работ. Во-вторых, можно ставить задачи по скачиванию и изучению конкретной информации из баз данных, к которым можно получить доступ через данный портал. Например, при изучении специальных курсов по направлению подготовки 140300 – ядерная физика и технологии можно использовать следующие базы данных:

- по энергетическим ядерным реакторам (PRIS);
- по технологиям топливного цикла и переработки отходов;
- по инновационным реакторам и топливным циклам

(ИНПРО).

12.2. Сайты российских предприятий и организаций ядерной отрасли

Сегодня практически все предприятия и организации ядерной отрасли имеют свои сайты в Интернете. На этих сайтах часто представлена информация не только о непосредственной деятельности предприятия или организации, но и материалы, которые могут быть использованы в образовательных целях. На многих сайтах представляются и комментируются новости в области использования ядерной энергии, имеется подборка электронных материалов обзорного характера, можно найти ссылки на специальные материалы. Знание данных сайтов и регулярное посещение некоторых из них, на наш взгляд, является необходимым элементом обучения студентов и аспирантов.

В учебном процессе информацию о ядерных сайтах преподаватели передают студентам на своих лекциях и индивидуальных занятиях (УИР, консультации КП и т.п.). Особый интерес к информации, представленной на этих сайтах, возникает у студентов при выборе места прохождения преддипломной практики, дипломного проектирования и будущего места работы.

Таблица 12.1.

Сайты российских предприятий и организаций ядерной отрасли.

	Ссылка	Предприятие/Организация	Анонсы сайта	Интересные материалы
1.	http://www.minatom.ru	Государственная корпорация по атомной энергии (Росатом)	Материалы по деятельности предприятий Росатома	Новости Росатома Нормативно-правовая база Направления деятельности АЭС России и других стран Электронные адреса министерств РФ и многих предприятий Росатома
2.	http://www.ruselectroatom.ru	Пресс-центр атомной энергетики и промышленности	Новости и комментарии в области атомной энергетики	Новости атомной отрасли Пресс-релизы Журнал «Росэнергоатом»
3.	http://www.ruselectroatom.ru КОНЦЕРН "РОСЭНЕРГОАТОМ"		Основные направления деятельности концерна: эксплуатация АЭС, экономическое, финансовое и коммерческое обеспечение выполнения функций эксплуатирующей организации, централизованный обмен по единому тарифу производимой электроэнергией, инвестиционная деятельность, междугородное сотрудничество в области повышения безопасности АЭС, подготовка персонала. Обеспечение безопасности АЭС на всех этапах жизненного цикла - приоритетная задача концерна.	
4.	http://www.nuclear.ru	Интегральный проект, посвященный ядерной отрасли РФ	Nuclear Ru - Интернет-проект, осуществленный ЗАО "Национальное инвестиционное агентство".	Новости Каталог предприятий атомной отрасли Каталог продукции предприятий атомной отрасли

5.	http://nuclearno.ru	Российский сайт ядерного нераспространения	Материалы по проблемам радиационной безопасности	Проблема обращения с РАО в РФ, статьи, новости
6.	http://www.graph.rnnet.ru	Государственный региональный образовательный центр МИА Атома РФ в Санкт-Петербурге	Материалы по деятельности предприятий МИА Атома в северном регионе РФ	История атомной энергетики, описание АЭС, типов реакторов, Проведение конференций
7.	www.wanoms.ru	ВАО АЭС - МЦ	Всемирная ассоциация организаций, эксплуатирующих атомные электростанции (ВАО АЭС), объединяет всех операторов АЭС мира. Организация создана для обмена опытом эксплуатации и совместной работы, направленной на достижение наивысшего уровня безопасности и надежности при эксплуатации АЭС.	Публикации, ссылки Карта АЭС мира
8.	www.kiae.ru	РНЦ «Курчатовский институт»	Направления работ и новости РНЦ КИ Краткая информация о подразделениях, членстве из них имеют свои сайты	Новости: Ядерная энергия, человек и окружающая среда, Структура РНЦ Информация о направлениях работ по подразделением Институт Ядерного Синтеза (ВАНТ: сер. Термоядерный синтез)
9.	http://www.dgae.ac.ru/	Институт проблем безопасного развития атомной энергетики РАН	Основной деятельности Института является комплексный анализ безопасности объектов атомной энергетики и топливно-энергетического комплекса	Интересные материалы по безопасности
10.	http://www.nibac.ru	Научно-исследовательский институт атомных реакторов, г. Димитровград	Описание деятельности, структуры, истории НИИАР	Каталог радиоактивной продукции, публикации
11.	http://www.ifp.nsk.su/	Институт Ядерной физики им. Г.И.Будкера	История, структура и основные направления деятельности института	Препринты в pdf формате

12	http://www.vniitf.ru/	Российский федеральный ядерный центр - Всероссийский научно-исследовательский институт теоретической физики (РФЯЦ - ВНИИТФ) г.Снежинск (бывш. Челябинск-70), расположенном на севере Челябинской области.	Информация об истории и деятельности РФЯЦ - ВНИИТФ	Информация об истории создания и современном состоянии ядерного щита РФ.
13	http://www.nikiizt.ru/	ФГУП Научно-исследовательский и конструкторский институт энерготехники им Н.А.Доллежале (НИКИЭТ)	Основные направления деятельности института	Материалы по БРЕСТ ам и ВК-300 Доклады представителей НИКИЭТ на различных форумах и заседаниях Возможен заказ книг (Белая книга АЭ, англо-русский словарь по АЭ)
14	http://www.jinx.ru/	Объединенный Институт Ядерных Исследований (ОИЯИ) г.Дубна	Почти все публикации доступны в электронной форме. Хороший прообраз электронной библиотеки, к которой должна стремиться каждая научная организация. Английский язык.	Нейтронная физика (теоретические вопросы)
15	http://www.gidropress.podol.su/	Федеральное государственное унитарное предприятие опытное конструкторское бюро "Гидропресс" (ОКБ "Гидропресс")	На севере отражена информация о развитии: история его развития, данные о выполненных ранее разработках и ведущихся в настоящее время, структура организации, экспериментально – исследовательская и производственная базы. Присутствует перечень публикаций и отчетов НИОКР, подготовленных ОКБ «Гидропресс».	Информация о ВВЭР, новых разработках

16.	http://www.shk.tomsk.ru/	Сибирский химический комбинат (СХК) г.Северск, Томской обл.	Описание основных заводов: разделения изотопов, реакторный, радиохимический, сублиматный, химико-металлургический, НИКИ. Перечень направлений конверсии (стабильные изотопы, магнитные сплавы и магниты, ультрадисперсионные порошки).	Статьи, фотографии, новости, История СХК
17.	www.bockvar.ru	Всероссийский научно-исследовательский институт неорганических материалов им. А.А.Бочвара (ВНИИНМ)	ФГУП ВНИИНМ - головная организация Министерства РФ по атомной энергии по проблемам материаловедения и технологий, выполняет исследования по ряду важнейших научно-технических федеральных и отраслевых программ.	История, публикации, инновации
18.	www.niiefa.spb.su	Научно-исследовательский институт электрофизической аппаратуры (НИИЭФА) им. Д.В. Ефремова	Институт внес решающий вклад в создание ускорительных комплексов от циклотронов и высоковольтных ускорителей до крупнейших установок мирового уровня в Харькове, Дубне, Серпухове, Гатчине.	Фотографии и описания установок, Описания деятельности научно-технических центров на базе института.
19.	http://www.ippe.ru/index.htm	Федеральное государственное унитарное предприятие "Государственный научный центр Российской Федерации – Физико-энергетический институт имени А.И.Лейпунского"	История, направления деятельности, экспериментальная база, разработки, научная информация ФЭИ	Фотоальбом, описание Обнинской АЭС, описания станций и установок, препринты Полезные ссылки

В таблице 12.1 представлены сайты некоторых предприятий и организаций ядерной отрасли, информация о которых полезна студентам, которые обучаются по направлению подготовки 140300 – ядерная физика и технологии.

12.3. Специализированные базы данных

Как уже было отмечено в разделе о портале NUCLEUS, МАГАТЭ поддерживает ведение нескольких баз данных по ядерной тематике.

База данных по энергетическим ядерным реакторам. Рассмотрим в качестве примера базу данных по энергетическим ядерным реакторам (PRIS). На рис. 12.3 представлено главное окно базы данных по энергетическим ядерным реакторам.

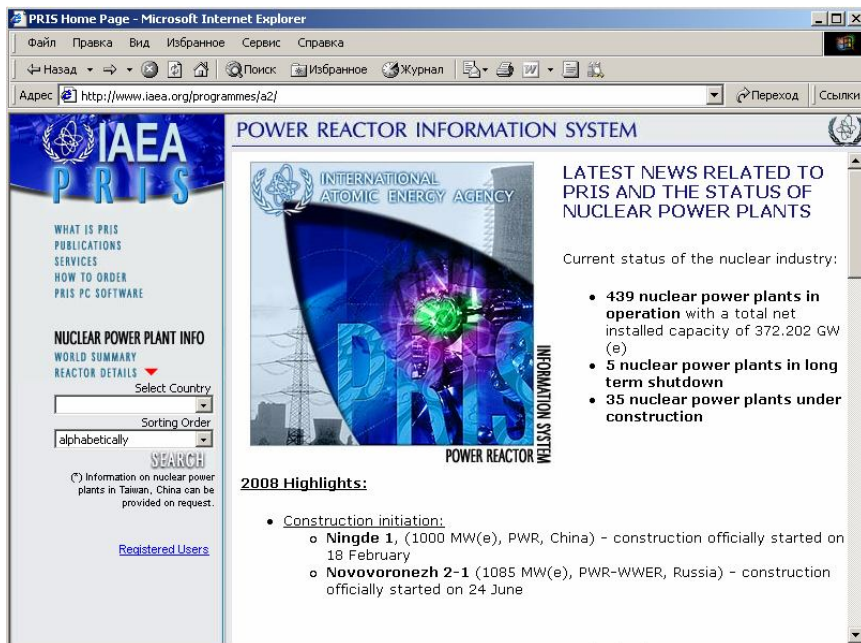


Рис. 12.3. Главное окно базы данных по энергетическим ядерным реакторам (PRIS)

С помощью базы данных PRIS можно узнать о текущем состоянии ядерной энергетики мира. Сколько ядерных реакторов работает, строится и находится на реконструкции в различных странах? Реакторы каких типов работают и строятся в настоящее время? На данные и другие подобные вопросы можно получить ответы, обратившись в данной базе. Запросы на поиск нужной информации осуществляются путем выбора из соответствующих списков.

Информация представляется в виде таблиц, некоторые поля которых являются активными кнопками. Используя активные кнопки можно получить дополнительную информацию.

На рис. 12.4 представлен фрагмент таблицы, которая появилась на запрос о ядерной энергетике России по типам реакторов.

POWER REACTOR INFORMATION SYSTEM

Russian Federation: Nuclear Power Reactors - By Type

Under Construction	7	Operational	31
Shutdown	5		

Annual Electrical Power Production for 2007	
Total Power Production (including Nuclear)	Nuclear Power Production
925200 GWh(e)	147770 GWh(e)

Click on the name of a reactor to view its full details.

Name	Type	Status	Location	Capacity (MWe)		Date Connected
				Net	Gross	
BELOYARSKY-3 (BN-800)	FBR	Operational	SVERDLOVSK	560	600	1980/04/08
BELOYARSKY-4 (BN-800)	FBR	Under Construction	SVERDLOVSK	750	800	
APS-1 OBNINSK	LWGR	Permanent Shutdown	KALUGA	5	6	1954/06/26
BELOYARSKY-1	LWGR	Permanent Shutdown	SVERDLOVSK	102	108	1964/04/26
BELOYARSKY-2	LWGR	Permanent Shutdown	SVERDLOVSK	146	160	1967/12/29

Рис.12. 4. Информация о ядерной энергетике России.

Как видно из таблицы, в 2007 г. в России работал 31 энергетический реактор, 7 реакторов строятся и 5 реакторов не эксплуатируются длительное время (остановлены). Также из приведенного

фрагмента таблицы видно, что блок №3 Белоярской АЭС (БН-600): является реактором на быстрых нейтронах (FBR); имеет электрическую мощность 600 МВт, из которых 560 МВт отдает в сеть; был подключен к сети 8 апреля 1980 г. Название реактора является в данной таблице активной кнопкой. Используя активную кнопку можно более подробно узнать и работе реактора по всем годам эксплуатации. Например, для реактора БН-600 можно узнать: строительство началась 1 января 1969 г.; за все время эксплуатации, включая 2007 год, передано в сеть 96920 ГВт·ч электрической энергии; средний КИУМ за все время работы составил 73,35%.

Базу данных PRIS, на наш взгляд, можно эффективно использовать в учебном процессе при изучении истории и текущего состояния мировой ядерной энергетики. Например, в рамках в обучения по направлению подготовки 140300 – Ядерная физика и технологии, студенты могут получать задания найти в базе PRIS информацию о ядерной энергетике какой-нибудь страны и/ли сравнить планы развития ядерной энергетике нескольких стран между собой. Выполняя подобные задания, студенты не только ознакомятся с современной ядерной энергетикой, но и познакомятся с ядерным порталом NUCLEUS, информация с которого может понадобиться им в будущей деятельности.

Файлы оцененных ядерных данных. Другим примером баз данных, которые активно используются и поддерживаются в ядерной отрасли, являются файлы оцененных ядерных данных (ФОЯД). В ФОЯД содержится информация о ядерных свойствах нуклидов, которая является симбиозом экспериментальных данных, теоретических моделей и аппроксимационных зависимостей. Данная информация обновляется по мере появления новой экспериментальной информации и новых теоретических моделей. Ведущие страны имеют собственные ФОЯД. Наиболее известные из которых: ENDF/B (США), JENDL (Япония), JEFF (ЕС), BROND (Россия). Информация в ФОЯД содержится в текстовом виде в определенном формате. Поэтому получение данной информации для анализа представляет определенную трудность. Существуют специальные программы, которые облегчают доступ к этой информации. Например, при поддержке европейского ядерного энергетического агентства (NEA/OECD) в университете Бирмингема при сотрудни-

честве с французским центром ядерной спектроскопии (CSNSM-Orsay) была разработана система JANIS. (<http://www.nea.fr/janis>).

JANIS (Java-based nuclear information software) – программа визуализации ядерных данных. Старое название программы JEP-PC. С помощью JANIS обеспечивается доступ к информации, содержащейся в трех банках данных: оцененных ядерных данных (ENDF формат), экспериментальных ядерных данных (EXFOR формат) и библиографических данных (CINDA).

Интерфейс программы JANIS прост, и для работы с ней требуются только знания из ядерной физики о тех данных, которые нужно получить. На рис. 12.5 изображено главное окно программы JANIS.

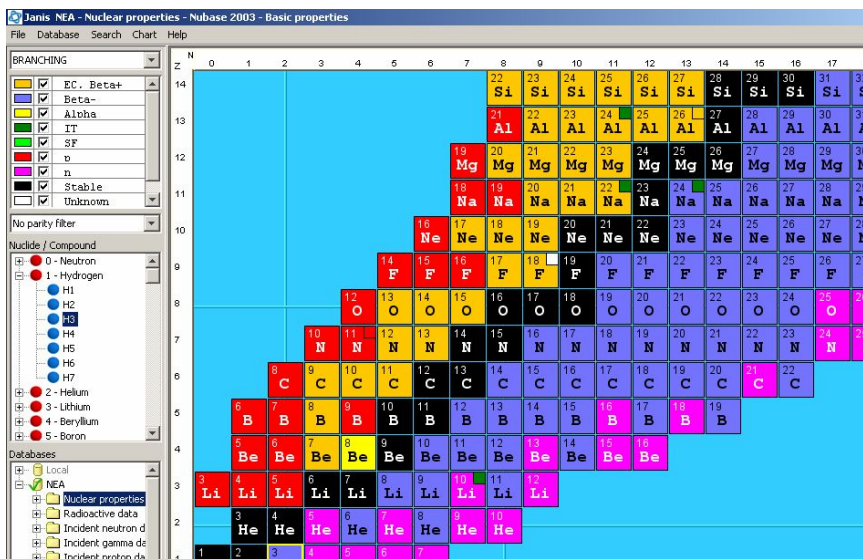


Рис. 12.5. Главное окно программы JANIS

Пользователь должен выбрать базу данных. Например, он планирует найти данные о сечениях взаимодействиях нейтронов с ядрами изотопа ^{235}U в ENDF/B-VI.8. Для этого он выбирает соответствующие ячейки меню и получает картинку, изображенную на рис. 12.6. Выбирая изотоп ^{235}U в карте нуклидов, он попадает в ме-

ню выбора процесса взаимодействия нейтрона с данным ядром. Выбрав нужное сечение (номер процесса MT в формате ENDF/B), он получает график и таблицу значений сечения в зависимости от энергии нейтрона. На рис. 12.7 изображено окно программы JANIS с графиком зависимости полного сечения (MT=1) изотопа ^{235}U .

С помощью программы JANIS можно получить быстрый доступ к различной информации, хранящейся в ФОЯД. Например, можно:

- просматривать данные о свойствах атомных ядер;
- сравнивать данные из различных ФОЯД между собой;
- просматривать сечения взаимодействия различных частиц с ядрами среды;
- сравнивать сечения различных изотопов между собой;

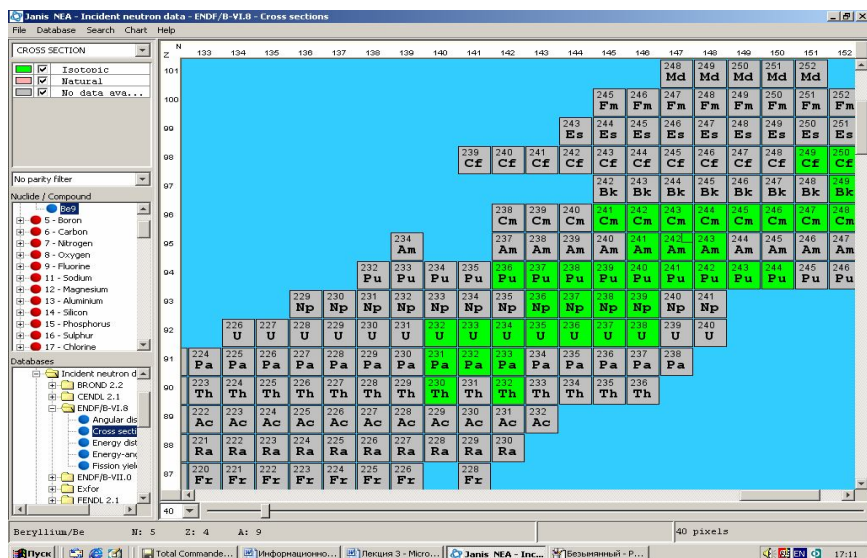


Рис. 12.6. Окно программы JANIS для получения данных о нейтронных сечениях

Программу JANIS можно эффективно использовать в учебном процессе по различным направлениям подготовки, в которых необходимо изучение ядерной физики. Некоторые преподаватели,

читающие специальные курсы («Нейтронные эффективные сечения», «Нейтронно-физические процессы» и т.п.), используют программу JANIS не только для иллюстраций к лекционному материалу, но и для проведения лабораторных работ и выполнения специальных заданий. Кроме учебных целей, программа JANIS может использоваться как справочное пособие по ядерным данным специалистами в области использования ядерной энергии.

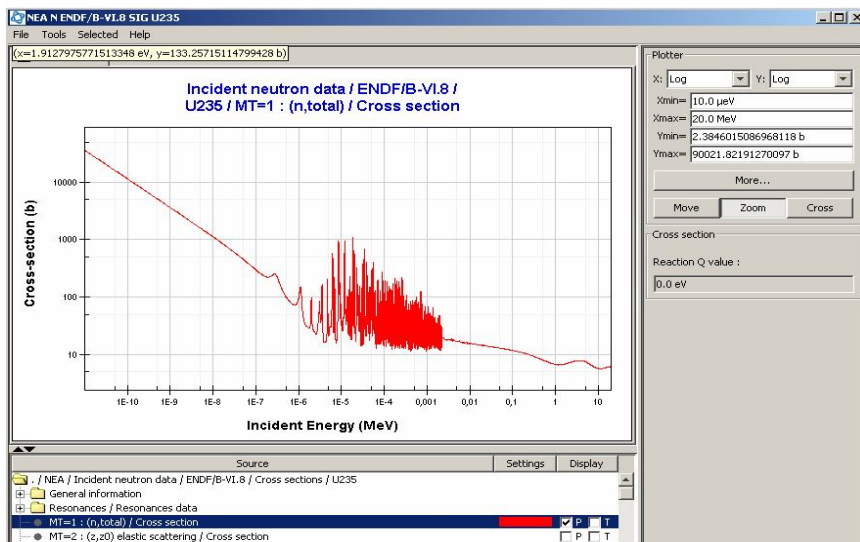


Рис. 12.7. Окно программы JANIS с графиком полного сечения изотопа ^{235}U

Задание по разделу

1. Для нескольких выбранных информационных ресурсов проведите сравнительный анализ по следующим параметрам:

- доступность ресурса (характер ограничений);
- виды поисковых документов;
- доступность (и способ получения) электронной копии первоисточника;

- наличие нескольких поисковых входов (поисковое выражение, рубрикационные структуры, алфавитные указатели и т.д.);
- тип поискового языка;
- наличие и способ использования справочных терминологических систем (рубрикаторов, тезаурусов);
- использование системой лингвистических процессоров;
- наличие средств развития запроса;
- экспорт результатов поиска;
- взаимодействие с другими ИР.

2. Дайте качественную оценку ресурса с точки зрения:

- полноты и достоверности информации в предметной области;
- удобства работы (эргономичности);
- показателей поиска (полнота, точность выдачи);
- гибкости и управляемости поиска.

Заключение: еще раз о человеко-машинном поиске информации в распределенных ИР

Как уже понятно читателю, данное пособие не является фундаментальным исследованием в области теории информационных систем, как и не является полным справочником по информационным ресурсам. Его задача состоит в том, чтобы показать, что разнообразие предлагаемых пользователю информационных систем, технологий и ресурсов скорее кажущееся. Их структура, свойства и функциональное назначение (и возможности), как показано в первой главе, достаточно ограничены и, что самое главное, предопределяются закономерностями процессов воспроизводства знаний и распределения информации.

В частности, приведенные в четвертом разделе описания функциональных возможностей и наполнения нескольких (различных по видам документов, источникам информации, интерфейсам, условиям предоставления и т.д.) информационных ресурсов является, как надеются авторы, хорошей иллюстрацией всеобщности свойства рассеяния, которое характерно как для всех элементов основной и информационной деятельности (см. рис. 1.2) в цикле генерации/распространения информации, так и для всех этапов жизненного цикла идеи. Действительно, объекты исследований (или относящиеся к нему) «рассеяны» как по всем этапам жизненного цикла, так и среди различных субъектов – индивидуальных или коллективных авторов настоящих или будущих публикаций; описания объектов, процессов и результатов исследований «рассеяны» в лексическом пространстве – для описания даже одного и того же объекта разные авторы могут (в том числе одинаково эффективно!) использовать разную лексику; сведения о публикациях «рассеяны» среди разных справочных изданий, баз данных и сайтов, а сами экземпляры публикаций – среди разных хранилищ (например, традиционных и электронных библиотек, полнотекстовых БД, отдельных электронных коллекций и т.д.); форма и способ представления информации «рассеяны» среди различных международных, национальных и фирменных стандартов на наполнение и формат документов; функции поиска и обработки информации «рассеяны» среди многих реализаций ИПС, а их интерфейсные

представления – среди различных метафор и визуальных компонентов, используемых разработчиками систем.

Именно поэтому полноценный поиск информации может обеспечить только интерактивный и итеративный процесс (как, например, представленный на рис. 1.3). Следуя технологии поиска и развития запроса (см. рис. 2.4), можно достигнуть необходимой полноты выдачи в отдельном ресурсе, после чего лексику запроса, развитую в этом ресурсе, используют для поиска (руководствуясь, в общем случае, той же технологией) в следующем ресурсе, далее переходят в очередной ресурс, или в те, которые исследовали ранее, в том случае, если выявлен какой-либо важный аспект или терминология, по каким-то причинам не использованные. В итоге реальная поисковая «навигация» оказывается достаточно ресурсоемкой, так как включает достаточно много процедур, требует использования вспомогательных справочных средств и знания конкретных информационных ресурсов. Именно поэтому в больших критически важных проектах до 10% ресурсов может выделяться на информационное обеспечение, а поиском и систематизацией информации занимаются аналитики, а не исключительно разработчики.

Важно понимать, что, по большому счету, информационной системой может называться только *совокупная* информационная система, объединяющая генераторы (в роли которых выступает, обычно, человек) и поставщиков-посредников (в роли которых выступает автоматизированная система) информации. Без средств систематизации (классификаций, рубрикаторов), позволяющих не только упорядочить массив, но и обобщенно отразить его содержание в системе наук, языка запросов, позволяющего «позиционировать» индивидуальность точки зрения пользователя, и, наконец, человека, который, комбинируя и систематизируя получаемые данные и наличные знания, синтезирует новую «информацию», любой информационный фонд АИС – это только хранилище данных.

В пособии основное внимание уделено «инженерным» вопросам: организации данных, технологиям поиска, языкам представления документальной информации, интерфейсам и т.д. Но, поскольку пользователь является неотъемлемой частью системы, обратим внимание на некоторые существенные с точки зрения эффективности совокупной основной/информационной деятельности

(и, соответственно технологии поиска) особенности человека, хотя они являются достаточно очевидными¹.

1. Человек обладает способностью к мышлению, которая при этом делает из него очень ненадежное средство регистрации информации, не гарантирующее длительное сохранение в неизменном состоянии. При этом он «интерпретирует» получаемые сведения, «виртуально» применяя их к собственным обстоятельствам, то есть сообщаемая им информация уже будет не точным отражением наблюдаемого явления, а его «личной» оценкой.

2. Человек обладает памятью, но при однообразии сообщений он может случайно (эффект *deja vu*) вместо нужного сообщения повторить/выбрать какое-либо похожее предыдущее.

3. Человек *понимает* наблюдаемое явление, знает или предполагает, как оно должно развиваться, и поэтому он склонен пойти по пути «угадывания» события, теряя связь с реальным его состоянием.

4. Человек может забывать и уставать, что сказывается не только на его производительности, но и на остальных способностях, в частности на его восприятии, понимании и интерпретации.

Эти особенности являются ключом к пониманию того, что технология поиска, сколь интеллектуальным бы не был механизм отбора документов, не может быть простым одноактным процессом типа «запрос-ответ». Система должна предоставлять пользователю разнообразные механизмы отбора, средства «входа» в предметную область и управления процессом поиска, а интерфейс системы должен иметь и другие элементы, кроме строки для выражения запроса и окна для просмотра документов. Достаточно полное представление о разнообразии таких средств дает технологическая схема поиска, реализованная в АИС IRBIS (см. рис. 2.4). Но при этом рисунок наглядно демонстрирует и, можно сказать, парадоксальную особенность поискового процесса: схема подобна лабиринту, имеющему много, явно заметных, входов, но выход в явной форме не указан – его должен определять пользователь индивидуально в зависимости от собственных целей, наличных ресурсов и способностей.

¹ ...и вполне соответствующими известной мудрой мысли, что наши недостатки – это продолжение наших достоинств!

Возвращаясь к механизмам и технологиям поиска еще раз надо резюмировать, что каждый из механизмов поиска имеет свои достоинства и недостатки. В АИПС же особенно важно обеспечить возможность использования различных механизмов поиска и их комбинаций для реализации всех типов поисковых задач. При этом «неточность» механизмов будет компенсирована за счет избыточности выдачи, и, в целом, будет обеспечено последовательное расширение терминологического и документального пространства. Это, как было показано в первом разделе, может быть достигнуто следующими путями:

- 1) реформулировкой и обогащением выражения запроса, в том числе с использованием других терминологических систем;
- 2) использованием нескольких поисковых механизмов с разными критериями отбора;
- 3) использованием итеративных процедур поиска, основанных на технологии обратной связи по релевантности.

Кроме этих путей, ориентированных на развитие и адаптацию запроса, не менее эффективным является поиск «источников» (генераторов) информации, когда в качестве поискового признака используется имя индивидуального или коллективного автора, проекта или исследовательской группы, ведущих работы в данной тематике, а также пристатейные ссылки на публикации, материал которых был использован или имеет отношение к теме работы.

Всё это позволяет утверждать, что форма и способ представления запроса принципиально имеет двойственную природу: при стремлении к «завершённой» вербальной форме выражения запроса в силу неопределенности, присущей реальной ИП (и потому невозможности построения адекватного выражения в вербальной форме), часть или весь запрос может быть представлен отдельными документами или их кластерами, что соответствует дискретной, «мозаичной» *кластерной* форме запроса. Таким образом, свойство комбинативности используется для последовательного снятия неопределенности за счет введения процедурной избыточности – механизмов порождения и упорядочения кластеров документов и терминов. При этом определение корреляционных связей между кластерами внутри отдельной выдачи и кластерами различных выдач позволяет определить относительные полноту и точность каждой итерации и оценить сходимость процесса поиска. А использо-

вание нескольких информационных ресурсов, материал которых может относиться к разным стадиям жизненного цикла идеи (обычно, это поставляемые информационными службами специализированные БД, такие как БД научной информации, БД отчетов о НИР, патентные БД и т.д.), позволяет обеспечить полноту представления о состоянии проблемы в конкретной предметной области.

В целом и сам процесс поиска информации имеет двойственную природу: на начальном этапе необходимо выявить список терминов, характеризующих предметную область. Анализ получаемых по запросам, использующим эти термины, документов, позволяет не только оценить возможность целевого практического использования их содержания, но и обогатить запрос, а также расширить терминологию предметной области¹. В свою очередь, систематизация терминологии порождает уже понятийную модель предметной области, являющуюся не только необходимым элементом познавательного процесса, но и одной из форм представления, сохранения и распространения знаний.

Особенностью сложившейся ситуации в части распределения функций управления между человеком и АИПС является то, что основные определяющие функции управления – выбор цели, определение критерия полезности, оценка и принятие решения, а, главное, генерация новой информации, осуществляются человеком. Система выступает как средство для быстрого отбора, в том числе, в очень больших массивах, упорядочения найденного, возможно, генерации вариантов представления. Но алгоритмы и критерии отбора, упорядочения и генерации предопределены заранее. Только человек знает, что ищет. Только он может образовывать или выделять *полезные* ассоциативные связи между различными информационными объектами и выводить на их основе новые свойства, которые не были зафиксированы в информационной системе.

Человек знает больше, чем может сказать. Только итерационное взаимодействие человека и информационной системы позволит «вытянуть» информацию об объекте исследования, которая представлена не только в информационной системе, но и в созна-

¹ Пример такого развития запроса приведен в Приложении 5.

нии человека (не вербализованная составляющая). Система, фиксируя траекторию его поисков и вербализованные образы (возможно, ею сгенерированные, но выбранные человеком), позволяет ему в любой момент вернуться к любому информационному объекту и пойти по другой траектории. В частности, именно поэтому развитие автоматизированных информационных систем идет по направлению не только создания «интеллектуальных» средств отбора документов и адаптации информационных ресурсов, но и создания информационной среды, интегрирующей процессы поиска и анализа информации и обеспечивающей системность информационных представлений. Такой подход позволит обеспечить общность информационного пространства, а также преемственность и развитие информационной поддержки на всех этапах жизненного цикла знания.

Действительно, «...чтоб умно поступать - одного ума мало».

Список литературы

1. Белоозеров В.Н. Опыт разработки словаря с разветвленной системой тезаурусных связей / Белоозеров В.Н., Косарская Ю.П. // НТИ. Сер. 2, 2001. - № 8.
2. Бениаминов Е.М.. Система представления знаний Ontolingua - принципы и перспективы / Бениаминов Е.М., Болдина Д. М. // НТИ, сер. 2, 1999, №10 - М.:ВИНИТИ.
3. Браславский П. И. Тезаурус как средство описания систем знаний. / Браславский П. И., Гольдштейн С. Л., Ткаченко Т. Я. //НТИ, Сер. 2, №11, 1997.
4. Гладун А.Я. «Онтологии в корпоративных системах» / Гладун А.Я., Рогушина Ю.В. // Корпоративные системы, 2006, №1.
5. Голицына О.Л.. Информационные системы: учеб. пособие / Голицына О.Л., Максимов Н.В., Попов И.И. – М.: ФОРУМ: ИНФРА-М, 2007.
6. Горькова В.И. Критерии оценки структурных связей понятий классификационных систем / Горькова В.И., Зотова Л.А. // НТИ, сер. 2, 1979, № 9.
7. Громов Г.Р. От гиперкнижки к гипермозгу: информационные технологии эпохи Интернета. / Громов Г.Р. – М.: Радио и связь, 2004.
8. Добров Б.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска./ Добров Б.В., Лукашевич Н.В.
9. Дубров А.М. Многомерные статистические методы./ Дубров А.М., Мхитарян В.С., Трошин Л.И. – М: «Финансы и Статистика», 1998.
10. Иванкин В.И. Опыт анализа семантических связей между текстами запросов и релевантных документов. / Иванкин В.И. // НТИ, Сер. 2, 1975, №1
11. Информационно-поисковый тезаурус по информатике. / Сост. Пашенко Н.А., Ксенофонтова Е.Б., Скоробогатая В.Ф., научный редактор Черный А.И. - М.:ВИНИТИ, 1987.
12. Колин К.К. Информационные проблемы социально-экономического развития общества / Колин К.К. // Проблемы социальной информатики. Вып.1. - М. 1995

13. *Криницкий Н.А.* Автоматизированные информационные системы / Криницкий Н.А., Миронов Г.Д., Фролов Г.Д. / Под ред. Дородницына А.А. - М.: Наука, 1982.

14. *Кулик А.Н.* Информационные сети и языковая совместимость дескрипторных информационно-поисковых систем. / Кулик А.Н. - М.: «Сов. Радио», 1977.

15. *Мартин Дж.* Организация баз данных в вычислительных системах. / Мартин Дж. – М.: Мир, 1980.

16. *Маршакова Н.Б.* Построение информационно-поискового тезауруса методом дистрибутивно-статистического анализа. / Маршакова Н.Б. // НТИ, сер. 2, 1977, № 5.

17. Методика индексирования документов по «Тезаурусу по атомной науке и технике» для системы автоматизированного распределения информации. – М.: ЦНИИАТОМИНФОРМ, 1977.

18. *Михайлов А.М.* Основы информатики. / Михайлов А.М., Черный А.И., Гиляревский Р.С. – М.: Наука, 1968.

19. *Москович В.А.* Информационные языки. / Москович В.А. - М.: Наука, 1971.

20. *Муранивский Т.В.* Теоретические основы научно-технической информации. / Муранивский Т.В. – М.: МГИАИ, 1982.

21. *Плотников Б.А.* Об использовании лексико-графических данных при построении тезауруса. / Плотников Б.А. // НТИ, сер. 2. 1975, № 9.

22. *Покрас Ю.Л.* Об одном способе установления парадигматических отношений при построении информационно-поискового тезауруса. / Покрас Ю.Л. // НТИ, сер. 2, 1971, №3.

23. *Попов И. И.* Моделирование информационных систем. / Попов И. И., Романенко А. Г. // Итоги науки и техники. Серия Информатика, Том 5 - М.: ВИНТИ, 1981.

24. *Ратцева И.И.* Проблема выбора значения слова и смысловые расстояния. / Ратцева И.И. // НТИ, 1966, №5.

25. *Шайкевич А.Я.* Дистрибутивно-статистический анализ в семантике. / Шайкевич А.Я. // Принципы и методы семантических исследований - М.: Наука, 1976.

26. *Скороходько Э.Ф.* Лингвистические проблемы обработки текстов в автоматизированных информационно-поисковых системах. / Скороходько Э.Ф. // Вопросы информационной теории и практики. Сб. №25, - М.: ВИНТИ. 1974.

27. Солтон Дж. Динамические библиотечно-информационные системы. / Солтон Дж. / Пер. с англ. -М.: Мир, 1979.
28. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. / Сэлтон Г. - М.: Советское радио, 1973.
29. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. Государственный стандарт Союза ССР. ГОСТ 7.25-80. (СТ СЭВ 174-85). // Государственный комитет СССР по стандартам. Москва. 1988.
30. Шалфеева Е.А. Классификация свойств онтологий. Свойства онтологий и их классификации / Шалфеева Е.А. // Научно-техническая информация, серия 2, 2005, № 11.
31. Buckley C. Automatic routing and retrieval using SMART: TREC-2. / Buckley C., Allan J., Salton G. // Inf. Proc.& Manag., V. 31, №3, 1986.
32. Chen Hsinchun. A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System. / Chen Hsinchun [и др.]. // Journal of the American Society for Information Science. January 1997, Vol. 48, No.1
33. Gruber T. R. A translation approach to portable ontologies. / Gruber T. R. // Knowledge Acquisition, 5(2):199-220, 1993.
34. MacGregor R. Inside the LOOM classifier / MacGregor R // SIGART bulletin, Vol.3, No.2, pages 70-76, 1991.
35. Krooks David A., Lancaster F.W. The Evolution of Guidelines for Thesaurus Construction. / Krooks David A., Lancaster F.W. // Libri, 1993, Vol. 43, No. 4.
36. Rijsbergen C.J. Information Retrieval. / Rijsbergen C.J. - London-Boston, Butterworths, 1975.
37. Salton G., Zhang Y. Enhancement of text representations using related document titles. / Salton G., Zhang Y. // Inf. Proc.& Manag. V. 22, №5, 1986.
38. Svenonius E. Bibliographical Classification / Svenonius E. // Int. Conf. Libr. Classif. and its Funct., Edmonton, June 20-21, 1989. - Edmonton, 1989.

Глоссарий

HTML (Hypertext Markup Language) – язык высокого уровня для определения структуры документов. Разработан в CERN и является одним из применений SGML.

XML – eXtensible Markup Language, «расширяемый язык разметки» – инструмент для создания и обработки документов.

Абстрагирование – способ *упрощения* совокупности фактов, относящихся к реальному объекту. При этом некоторые свойства объекта игнорируются, поскольку считается, что для решения данной прикладной задачи (или совокупности задач) они не являются определяющими и не влияют на конечный результат.

Автоматизированная информационно-поисковая система (АИПС) – совокупное название как для программных оболочек, ориентированных на ввод, хранение, поиск и выходное представление документов (структур данных сложного или неопределенного формата), так и для конкретных систем определенного наполнения и предметной ориентации, реализованных на основе таких оболочек (или иными программными методами).

Агрегат данных – именованная совокупность элементов данных, представленных простой (векторной) или иерархической (группы или повторяющиеся группы) структурой.

Алфавит – любая конечная совокупность знаков (букв, цифр и т.п.), используемых в языке.

Аналитико-синтетическая переработка – преобразование документов в процессе их анализа и извлечение необходимой информации, а также оценка, сопоставление, обобщение и представление информации в виде, соответствующем запросу.

Архитектура документа – структурное описание документа, включающее в себя все входящие в него виды информации (текст, векторная и растровая графика, таблицы).

Атрибут – поле данных содержащее информацию об объекте.

База данных (БД) – именованная совокупность взаимосвязанных данных, отображающая состояние объектов и их отношений в некоторой предметной области, используемых несколькими пользователями и хранящимися с минимальной избыточностью.

Библиографические данные (записи) – выходные данные (включают сведения об авторах, заголовков, классификационный

индекс, место публикации и пр.), иногда реферат документа (публикации).

Браузер – прикладная программа клиента, которая позволяет извлекать и просматривать содержание документов, находящихся на Internet-серверах.

Вид документа – элемент классификации множества документов, циркулирующих в организации.

Гипертекст – организация текстовых документов, любой фрагмент которых может содержать (и можно использовать как) ссылку на фрагмент этого или другого документа, связанного с исходным по тематическому или какому-либо другому признаку.

Грамматика – система способов и средств построения слов и предложений в рассматриваемом языке. Грамматика состоит из *морфологии* и *синтаксиса*.

Данные – 1) сигналы, факты, сообщения, выраженные средствами формальной системы, обеспечивающей их хранение, обработку и передачу; 2) информация, обработанная и представленная в формализованном виде для дальнейшей обработки.

Денотат – класс предметов, обладающих некоторыми общими свойствами.

Дескриптор – предназначенное для координатного индексирования документов и информационных запросов нормативное ключевое слово, по определенным правилам отобранное из основного словарного состава того или иного естественного языка.

Дескрипторный язык – информационно-поисковый язык, словарный состав которого состоит из дескрипторов, а использование основано на принципе координатного индексирования.

Документ – агрегат данных в документальных системах, имеющий иерархическую структуру и, кроме форматных полей (элементы или агрегаты данных фиксированной длины), обычно содержащий текстовые поля, или символные последовательности неопределенной длины, логически подразделяющиеся на параграфы, предложения, слова.

Документ вторичный – документ, являющийся результатом аналитико-синтетической переработки одного или нескольких первичных документов.

Естественный язык – язык, словарь и грамматические правила которого обусловлены практикой применения и не всегда формально зафиксированы.

Запись логическая – идентифицируемая (именованная) совокупность элементов или агрегатов данных воспринимаемая прикладной программой как единое целое при обмене информацией с внешней памятью. Запись – это упорядоченная в соответствии с характером взаимосвязей совокупность полей (элементов) данных, размещаемых в памяти в соответствии с их типом.

Запись физическая – совокупность данных, которая может быть считана или записана как единое целое одной командой ввода-вывода.

Знак – материальный предмет (явление, действие, событие), который выступает в процессе коммуникации как представитель другого предмета, свойства или отношения и используется для передачи, переработки и хранения информации.

Значение – информация, присвоенная знаку в процессе коммуникации.

Идентификация – установление тождества объектов на основе совпадения их признаков.

Инвертированный файл (список) – файл, предназначенный для быстрого произвольного поиска записей по значениям ключей, организованный в виде независимых упорядоченных списков (индексов) ключей – значений определенных полей записей основного файла.

Индекс – таблица ссылок на объекты, используемая для определения адреса записи.

Индексирование (документа) – формирование описания документа как совокупности дескрипторов, выбираемых из заранее созданных словарей понятий, либо из текстов документов.

Интерпретации правила – определяют способы перевода терминов и предложений (фраз) формального языка на соответствующий естественный язык.

Интерфейс пользователя – совокупность правил, методов и программно-аппаратных средств, обеспечивающих взаимодействие пользователя с компьютером.

Информатизация – комплекс мер, направленных на обеспечение оперативного доступа к информационным ресурсам.

Информационная система – система, предназначенная для хранения, обработки, поиска, распространения, передачи и предоставления информации.

Информационная технология – совокупность методов, производственных процессов и программно-технических средств, объединенных в технологический комплекс, обеспечивающий сбор, создание, хранение, накопление, обработку, поиск, вывод, копирование, передачу и распространение информации.

Информационно-поисковая система (ИПС) – программная система для хранения и поиска данных по неформатированным запросам. Для общения пользователя с ИПС разработчики системы стремятся применять упрощенный естественный язык.

Информационно-поисковый язык (ИПЯ) – искусственный язык, обеспечивающий компактную, строго алгоритмизированную запись содержания документов и запросов в ИПС. ИПЯ можно определить как специализированную семантическую систему, состоящую из алфавита, правил образования (грамматики) и правил интерпретации (семантики).

Информационный анализ – выявление в документах и фиксация в виде данных информации, относящейся к определенной предметной области.

Информационные ресурсы – совокупность накопленной информации, зафиксированной на материальных носителях в любой форме, обеспечивающей ее передачу во времени и пространстве. В контексте автоматизированных информационных систем под информационными ресурсами обычно подразумевают информационные массивы и базы данных, рассматриваемые совместно с информационными технологиями, обеспечивающими их доступность.

Информационный запрос – записанный на естественном языке текст, выражающий некоторую информационную потребность.

Информационный поиск – процесс отыскания в поисковом массиве таких записей, которые соответствуют признакам, указанным в информационном запросе.

Информация – сведения, воспринимаемые человеком и (или) специальными устройствами как отражение фактов материального или духовного мира в процессе коммуникации.

Информация научная – логически организованная информация, получаемая в процессе научного познания и отображающая явления и законы природы, общества и мышления.

Информация научно-техническая (НТИ) – информация, получаемая и (или) используемая в области науки и (или) техники.

Информация документальная – информация, содержащаяся в документах.

Информационный язык – искусственный язык, созданный в целях лучшей реализации основной функции языка – коммуникативной, то есть функции передачи информации.

Классификация – процесс соотнесения содержания документов с понятиями, зафиксированными в заранее составленных систематических схемах.

Ключ – значение (элемент данных) используемый для идентификации или определения адреса записи.

Ключевое поле – поле в структуре записи. Поле определяют как ключевое (или индексированное) для убыстрения или упрощения операций поиска и/или для модификации операций обработки данных.

Ключевое слово – предметное слово, выбираемое из некоторого текста (документа) и используемое для координатного индексирования этого текста (документа).

Когнитивный процесс – (лат. *Cognitio* – знание, познание) соответствующий познанию.

Кодирования система – система представления информации в виде данных, состоящая из набора условных знаков и правил присвоения им значений.

Коммуникативные (обменные) форматы данных – соглашения о представлении агрегатов информации при передаче.

Концепт – совокупности признаков, однозначно определяющих денотат.

Координатное индексирование – индексирование, при котором основное содержание документа представляется в виде сочетания дескрипторов.

Лексика (или словарный состав) – вся совокупность слов, входящих в состав языка.

Лемматизация – нахождение начальной формы слова по любой его словоформе. Например, дано слово «погоняемый» - найдено слово «погонять».

Лингвистическая переменная – переменная, значение которой зависит от контекста ближайшего окружения – значения соседних переменных.

Логическая структура БД – определение БД на физически-независимом уровне.

Логическая структура документа – предназначена для представления произвольных иерархических видов организации информации. Например, рубрикация документа, включение в текст таблиц и рисунков.

Метаданные – информация, которая описывает другие данные, с помощью таких атрибутов как их структура, ассоциации, типы и диапазоны.

Методы поиска – совокупность моделей и алгоритмов реализации отдельных технологических этапов, таких, как построение поискового образа запроса, отбор документов (сопоставление поисковых образов запросов и документов), расширение и реформулирование запроса, локализация и оценка выдачи.

Механизмы поиска – реализованные в АИПС модели и алгоритмы процесса формирования выдачи документов в ответ на поисковый запрос.

Модель – искусственно созданный объект, который отображает, воспроизводит в ином, обычно более простом, уменьшенном виде, структуру, свойства, взаимосвязи и отношения исследуемого объекта, непосредственное изучение которого связано с большими затратами средств или недоступно.

Модель данных – базовый инструментарий, обеспечивающий на формальном абстрактном уровне конкретные способы представления объектов и связей.

Модель документа – совокупность представлений, которая охватывает аспекты создания, преобразования, хранения, поиска, передачи и отображения документов.

Морфология – совокупность действующих в языке способов и средств построения слов. Наименьшая неделимая без потери данного качества единица системы выражения, непосредственно соотносимая с соответствующим ей элементом содержания, называется

морфемой. Морфемы делятся на *грамматические* (аффиксы) и *лексические* (лексемы). Аффиксы предназначены для видоизменения значения основной части слова (базы). Лексема – это корневая или словообразовательная морфема, выражающая основное значение слова.

Морфологический поиск – поиск с учётом морфологии (всех возможных форм слова). Например, в строке поиска введено «хрюкать» – в результатах поиска присутствуют «хрюкающий», «хрюкало», и другие производные от заданного слова.

Мультимедиа – среда, материал, состоящий из комбинации текста, графики, видео, мультипликации и звука, представляющий таким образом информацию в более понятном и удобном для обработки виде.

Навигация поисковая – целенаправленная, определяемая стратегией, последовательность использования методов, средств и технологий конкретной АИПС для получения и оценки результата поиска.

Наименование – слово или словосочетание, используемое для идентификации какого-либо существа, предмета или класса.

Носитель информации (данных) – средства регистрации, хранения, передачи информации (данных).

Образ – результат отражения объекта 1) в гносеологии – в сознании человека; 2) в искусственных системах – функционально-ориентированная модель (процедуры и данные) в памяти ЭВМ.

Объект – осязаемая реальность, проявляющая четко выделяемое поведение.

Оригинал – объект, для которого строится образ.

Парадигма – совокупность всех словоформ данного слова. Если дано некое существительное (например, «котенок»), его парадигму составят слова, образуемые при изменении по падежам и числам («котенок», «котятя», «котенка», «котят» и т. д.).

Парадигматические отношения (или – аналитические, базисные отношения) – логические отношения, существующие между лексическими единицами языка независимо от контекста их конкретного употребления.

Поисковый образ документа (ПОД) – описание документа, выраженное средствами ИПЯ и характеризующее основное смы-

словое содержание или какие-либо другие признаки этого документа, необходимые для его поиска по запросу.

Поисковый образ запроса (ПОЗ) – записанный на ИПЯ текст, выражающий смысловое содержание информационного запроса и содержащий указания, необходимые для наиболее эффективного осуществления информационного поиска.

Полнотекстовые документы (записи) – полный (или почти полный) исходный текст журнальной статьи или другого документа.

Понятие – форма мышления, отражающая существенные свойства, связи и отношения предметов и явлений.

Предметная область (Про) – набор объектов, представляющих интерес для актуальных или предполагаемых пользователей, когда реальный мир отображается совокупностью конкретных и абстрактных понятий, между которыми фиксируются определенные связи.

Признак – свойство объекта, обуславливающее его различие или общность с другими объектами

Профиль документа – в ODA набор свойств документа, которые относятся к документу в целом.

Разметка – дополнительная информация, включаемая в документ и выполняющая функции выделения логических элементов данного документа и задания процедур обработки выделенных элементов.

Распознавание документа – построение редактируемой электронной копии бумажного документа. Как правило, проводится в два этапа; сначала при помощи сканера получают электронную «фотографию» страницы, затем обрабатывают её специальной программой распознавания символов.

Релевантность – свойство некоторой информации (документ, факт, и пр.) удовлетворять информационную потребность пользователя АИС (relevant – относящийся к делу).

Сведения – 1) познания в какой-либо области; 2) известие, сообщение; 3) знание, представление о чем-либо.

Семантика – раздел языкознания, исследующий смысловую сторону слов и выражений, отношения между знаками. Логико-лингвистическая дисциплина, исследующая отношение между

формально построенным исчислением и той областью действительности, которая в нем отражается.

Семиотика – наука о знаках и знаковых системах.

Символ – элемент согласованного набора материальных предметов, используемых для представления знаков.

Синтагматические отношения (или – текстуальные, синтетические, синтаксические отношения) – отношения между лексическими единицами языка, выражающие действительные конкретные логические связи между соответствующими понятиями в тексте сообщения, которые, в свою очередь, отражают конкретные ситуативные отношения между обозначаемыми ими объектами.

Синтаксис – совокупность действующих в языке способов и средств построения предложений.

Словарь данных – исчерпывающий набор таблиц или файлов, представляющий собой каталог всех описаний данных (имен, типов). Может содержать также информацию о пользователях, привилегиях и т. д., доступную только администратору базы данных.

Словообразование – определяется как совокупность способов и средств образования слов на базе уже существующих.

Смысл – содержание, сущность, значение чего-нибудь. Для человека – значение, которое объект, событие или слово приобретают для него в результате его жизненного опыта.

Содержание документа – представляет собственно информацию документа: текст, рисунки и т.п.

Сообщение – 1) информация, переданная и (или) полученная в процессе коммуникации; 2) физическое состояние, определенным образом отличающееся от других состояний в цепи управления.

Стратегия поиска – общий план (концепция, предпочтение, предрасположенность, установка) поведения пользователя для выражения и удовлетворения информационной потребности, обусловленный характером цели и типом поиска, архитектурой БД, а также методами и средствами поиска конкретной АИПС.

Структура данных – атрибутивная форма представления свойств и связей предметной области, ориентированная на выражение описания данных средствами формальных языков, и таким образом учитывающая возможности и ограничения конкретных средств, с целью сведения описаний к стандартным типам и регулярным связям.

Структура данных линейная – структура, порядок следования элементов данных которой имеет линейный характер и соответствует порядку расположения элементов в памяти.

Таблица – основная единица информации в системе управления реляционной базой данных. Состоит из одной или более единиц информации (строк), каждая из которых содержит значения некоторого вида (столбцы).

Тезаурус – семантическая сеть, в которой понятия связаны регулярными и устойчивыми семантическими отношениями – иерархическими (например, род-вид, целое-часть), ассоциативными, а также отношениями эквивалентности.

Термин – слово или словосочетание, являющееся точным обозначением определенного понятия какой-либо области знания.

Терминологическая система – организованная совокупность терминов в специальном языке определенной области знания.

Технологии поисковые – унифицированные (оптимизированные в рамках конкретной АИПС) последовательности эффективно-го использования в процессе взаимодействия пользователя с системой отдельных средств поиска для устойчивого получения конечного и, возможно, промежуточных результатов.

Типы данных – совокупность соглашений о программно-аппаратурной форме представления и обработки а также ввода, контроля и вывода элементарных данных к типам данных.

Уровни представления данных в БД – концептуальный, внутренний и внешний. Внутренний уровень - глобальное представление БД, определяет необходимые условия в первую очередь для организации хранения данных на внешних запоминающих устройствах. Представление на концептуальном уровне – это обобщенный взгляд на данные с позиций предметной области. Внешний уровень – представляет потребности пользователей и прикладных программ.

Формат – способ расположения и представления данных на носителе информации.

Численные данные – записи, которые содержат таблицы статистических, финансовых и других сведений в числовой форме.

Электронный документ – документ, носителем которого является электронная среда – МД, компакт-диск и т. д.

Элемент данных (элементарное данное) – неделимое именованное данное, характеризующееся типом (напр., символьный, числовой, логический, и пр.), длиной (в байтах). Это минимальная адресуемая (идентифицируемая) часть памяти – единица данных, на которую можно сослаться при обращении к ним.

Элемент текста – часть текста, ограниченная начальной и конечной метками (При использовании SGML).

Язык – 1) система знаков, обеспечивающая коммуникацию и включающая набор знаков (словарь терминов) и правила их употребления и интерпретации (грамматика); 2) знаковая система любой физической природы, выполняющая познавательную и коммуникативную функцию в процессе человеческой деятельности.

Язык естественный – язык, словарь и грамматические правила которого обусловлены практикой применения и не всегда формально зафиксированы.

Язык искусственный – язык, специально созданный и регулируемый на основе согласованных принципов.

Язык манипулирования данными (ЯМД) – обычно включает в себя средства запросов к базе данных и поддержания базы данных (добавление, удаление, обновление данных, создание и уничтожение БД, изменение определений БД, обеспечение запросов к справочнику БД).

Язык описания данных (ЯОД) – средство внутрисистемного определения данных, представляющего обобщение внешних взглядов. Описание представляет собой модель данных и их отношений, т. е. структур, из которых образуется БД.

Список сокращений

АИПС – автоматизированная информационно-поисковая система

АИС – автоматизированная информационная система

АИТ – автоматизированные информационные технологии

АСУ – автоматизированная система управления

ББК – Библиотечно-библиографическая классификация

БД – база данных

ЕЯ – естественный язык

ИД – информационная деятельность

ИП – информационная потребность

ИПС – информационно-поисковая система

ИПТ – информационно-поисковый тезаурус

ИПЯ – информационно-поисковый язык

ИР – информационный ресурс

ИС – информационная система

ИСр – информационная среда

ИТ – информационная технология

ЛЕ – лексическая единица

ЛО – лингвистическое обеспечение

МПК – Международная патентная классификация

НИР – научно-исследовательская работа

НТД – научно-техническая документация

НТИ – научно-техническая информация

ОД – основная деятельность

ПО – программное обеспечение

ПОЗ – поисковый образ запроса

ПОД – поисковый образ документа

ПОТ – поисковый образ темы

ПрО – предметная область

СУБД – система управления базами данных

УДК – Универсальная десятичная классификация

ЭБ – электронная библиотека

ЭВМ – электронно-вычислительная машина

ЭДД – электронная доставка документов

**Фрагмент Методики индексирования документов по
«Тезаурусу по атомной науке и технике» для
системы автоматизированного распределения
информации**

**5. ОБЩАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ СТАДИЙ
ОБРАБОТКИ ИНФОРМАЦИОННЫХ ДОКУМЕНТОВ НА
ЭТАПЕ ИНДЕКСИРОВАНИЯ**

Настоящая методика регламентирует следующий порядок работ индексировщика по общему анализу документа:

- определение вида и тематической направленности;
- составление библиограммы обрабатываемого документа;
- разделение (в случае необходимости) многотемного документа по аспектам содержания на отдельные информационные блоки, каждый из которых выступает как самостоятельный документ;
- перевод заголовка, если он дан на иностранном языке, и обязательное составление краткой аннотации для отечественных материалов;
- отбор ключевых слов из заголовка, аннотации и текста обрабатываемого документа;
- тождественная замена отобранных ключевых слов на дескрипторы информационно-поискового языка, поиск в тезаурусе дескрипторов, дословно совпадающих с ключевыми словами;
- этап нетождественных замен ключевых слов на дескрипторы, поиск в тезаурусе дескрипторов, не совпадающих с ключевыми словами, но характеризующих (с расширением понятия) содержание документа ;
- формирование дескрипторного ПОД;
- этап контрольной самопроверки: восстановление содержания документа по составленному ПОД.

**6. АНАЛИЗ СОДЕРЖАНИЯ ИНФОРМАЦИОННОГО
ДОКУМЕНТА НА ЭТАПЕ ИНДЕКСИРОВАНИЯ**

Непосредственному анализу текста документа предшествует этап Определения его общей тематической направленности и составление по принятой форме библиограммы.

6.1. Этап определения общей тематической направленности документа

Индексировщик должен:

- определить, принадлежит ли данный документ к областям тематического охвата САРИ, определяемого отраслевым рубрикатом;
- решить, к какой предметной рубрике (рубрикам) должен быть отнесен документ;
- решить, состоит ли документ из отдельных частей, которые можно рассматривать как достаточно различные по содержанию; в этом случае разделение может быть выполнено на библиографическом уровне и каждая часть может быть представлена как отдельный документ.

6.2. Порядок составления библиограммы документа

В соответствующие графы информационной карты вносятся:

- поисковый номер документа;
- обозначение вида первичного информационного материала
- (книга, отчет, статья, патент, стандарт, фирменный проспект, чертеж);
- обозначение формы первичного информационного материала (оригинал, микрокарта, диамикромкарта, микрофильм, фотокопия);
- фамилия и инициалы автора (авторов);
- заголовок;
- название страны, в которой издан первичный информационный материал, с указанием языка издания;
- название источника информации, например, организация, выпустившая отчет; издатель (издательство), выпустивший книгу; страна, издавшая патент и т.д.;
- характеристики издания (номер тома; номер журнала и страницы; номер отчета; номер патента);
- дата издания;
- место хранения первичного информационного документа и инвентарный номер документа;
- гриф документа.

Если материал опубликован на иностранном языке, индексировщик должен указать название документа на языке оригинала и перевести заголовок.

Непосредственно после перевода заголовка желательно дать краткую аннотацию, передающую основное содержание документа. Для отечественных труднодоступных материалов это выполняется обязательно.

6.3. Анализ содержания текста документа

Сначала анализируется заголовок, затем реферат, оглавление и полный текст документа.

Из заголовка, реферата, оглавления и текста документа индексирующий выбирает ключевые слова, наиболее полно отражающие содержание документа. Перечень этих ключевых слов должен являться свернутой реферативной моделью содержания информационного документа.

Анализ содержания индексируемого документа должен проводиться с учетом следующих правил:

Правило I. При анализе содержания документа следует выбирать только полезную информацию.

Интерпретация документов заключается в том, что индексирующий устанавливает наиболее важное информационное содержание документа. Критерием, которым следует руководствоваться при этом выборе, является то, что полезная информация наиболее информативна и интересна с точки зрения пользователя системы. Как показывает практика информационного поиска, документ может быть адекватно определен одним, двумя и, возможно, тремя наиболее важными аспектами, которые обычно являются основанием для опубликования этого документа. В процессе подготовки поискового образа индексирующий должен представить каждый из этих аспектов одним или несколькими понятиями и затем согласовать эти группы для отражения содержания рассматриваемого документа. При выборе ключевых слов индексирующий должен помнить, что следует использовать только те ключевые слова, которые действительно отражают понятия, рассматриваемые в документе, и могут отразить полезную с точки зрения пользователя системы информацию, содержащуюся в документе, т.е. каждое выбранное ключевое слово должно нести достаточную информативную нагрузку. Например, при индексировании сообщения «Для определения уровня жидкости использовался сцинтилляционный счетчик с кристаллом NaI» информационная единица «Йодиды натрия» (NaI) должна быть опущена; в то же время она должна быть использова-

на для сообщения «Эффективность сцинтилляционных счетчиков с кристаллами NaI большого диапазона измерений». В первом случае не представляет интереса, какой тип счетчика использовался, а во втором случае устройство счетчика само по себе является понятием, которое обсуждается.

Правило 2. При анализе содержания документа следует игнорировать избыточную информацию, выявлять скрытую информацию и уточнять неполную информацию.

Так, избыточная информация содержится во фразах, указывающих, что такие-то явления не рассматривались (отрицательная информация) или рассматривались в других работах; указывающих на факт аналогии между данным явлением и каким-то другим; при упоминании параллельно работающих (помимо авторов) над темой исследования групп ученых.

Иногда имеется тенденция к индексированию слишком большого числа информационных единиц, приведенных в полном тексте документа. Надо помнить, что избыточное индексирование и использование расплывчатых понятий препятствуют поиску релевантных документов, ухудшают параметры всей информационно—поисковой системы. Примеры расплывчатых информационных сообщениях: «Этот эффект был открыт группой из Окриджской национальной лаборатории», «Эти эксперименты по облучению проводились на реакторах для испытания материалов», «Резонансный захват обнаружен не был». Эти сообщения не содержат фактов, представляющих научный интерес. Не имеет научного значения, кем открыт эффект и на каком реакторе проводились эксперименты. Такая информация не должна индексироваться. Для сообщения «Толстые пластины из нержавеющей стали сваривались» дескриптор ТОЛЩИНА будет расплывчатым понятием, и использование его явится избыточным индексированием. В то же время для сообщения «Тепловые и радиационные испытания были выполнены на топливных частицах с покрытием для определения оптимальной толщины свободного от топлива покрытия из графита» необходимо при индексировании использовать дескриптор ТОЛЩИНА.

Если в документе рассматриваются различные научные аспекты, не все из которых входят в тематический охват информационно-поисковой системы, определяемый для САРИ отраслевым

рубрикатором, то индексируются только те проблемы, которые попадают в тематический охват САРИ. Аналогично, если в документе обсуждаются некоторые химические процессы для определенных химических элементов, то индексируются только те процессы и элементы, которые представляют интерес для атомной науки и техники. Например, в сообщении «Химический анализ циркония, гафния и меди» должны быть заиндексированы только цирконий и гафний.

Неполная информация имеется во фразах, описывающих «различные излучения», «некоторые ядра», «три сплава» и т.п. Если тема документа этого требует, по возможности следует выяснить, какие именно явления и предметы имеются в виду, и внести в поисковый образ соответствующие дескрипторы. Например, если реферат содержит информацию «Были измерены эффективные сечения для быстрых нейтронов некоторых тяжелых ядер», то индексирующий может найти перечень конкретных ядер в полном тексте документа.

Некоторые документы содержат скрытую информацию. Сообщение «Реакция протекает интенсивно в присутствии платины» содержит понятие «катализ»; сообщение «Пассивация металла уменьшается после термической обработки» содержит понятие «коррозия»; сообщение «Долгоживущий изотоп углерода» относится к углероду-14. Если в документе рассматривается «фотометрическое определение циркония с пиридилазонафтолом», то в этом сообщении подразумевается образование «комплексов циркония».

Правило 3. В документе могут быть выделены простые и сложные понятия, описываемые одним или несколькими терминами.

Если одно понятие описывается посредством терминов A и B , а другое - совокупностью терминов C и D , то в описании документа перечисленные термины A , B , C , D могут быть зафиксированы на информационной карте один за другим. Но отношения между ключевыми словами A и B иные по своей природе, чем отношения между ключевыми словами A и C : первое отношение отражает действительную связь явлений, второе – случайную, ложную связь. Документы, описанные терминами A , B , C , D , будут с одинаковой вероятностью выданы и на запрос, содержащий термины $A + B$, и на запрос с терминами $A + C$. Вторая выдача будет

ложной, «шумовой». Это явление может наблюдаться и при употреблении одиночных, не связанных друг с другом понятий, когда термины А и В, являясь самостоятельными, при соположении в ПОД могут образовать ложное составное понятие А+В, что приведет в свою очередь к «шуму» и потерям при поиске.

Во избежание образования ложных ассоциаций содержание источника необходимо разбивать на отдельные части - информационные блоки, каждый из которых является независимой единицей при анализе. Таким образом, информация, содержащаяся в блоке, может стать самостоятельным объектом поиска. Понятия, общие для всего документа, повторяются в каждом блоке.

7. ИНДЕКСИРОВАНИЕ ИНФОРМАЦИОННЫХ ДОКУМЕНТОВ С ПРИМЕНЕНИЕМ «ТЕЗАУРУСА ПО АТОМНОЙ НАУКЕ И ТЕХНИКЕ»

Перевод содержания документа, выраженного кратким перечнем терминов естественного языка - ключевыми словами, выбранными из текста документа, на дескрипторный язык осуществляется с помощью тезауруса, использование которого обеспечивает выражение смыслового содержания документа и информационной потребности абонента системы на одном языке едиными терминами, что гарантирует высокую полноту и релевантность поиска.

Термины естественного языка иногда имеют несколько значений, а когда они используются в различных областях науки и техники, то и различный смысл. Для целей индексирования может быть принято только одно значение. Большинство дескрипторов в тезаурусе не имеют неопределенности смысла.

Дескрипторы, для которых некоторая неопределенность может существовать (это относится в основном к терминам более или менее общего характера), подвергаются лингвистическому контролю. Это семантическое определение в тезаурусе осуществляется следующими путями:

- дескриптор располагается в точном семантическом контексте со всеми терминами, которые связаны с ним в словарной статье, т.е. значение дескриптора обеспечивается его связями с запрещенными, родовыми, видовыми и ассоциативными терминами;

- для некоторых дескрипторов, когда все-таки может существовать вероятность неоднозначности их смысла, замечания в скобках уточняют значение термина.

Следует заметить, что вышеуказанными способами точные определения также даны для всех аббревиатур, используемых как дескрипторы. Большинство принятых аббревиатур относятся или к сложным органическим соединениям, или к конкретным названиям реакторов. Полные названия реакторов приводятся в тезаурусе как запрещенные термины и имеют отсылку к аббревиатуре, принятой как дескриптор. Следовательно, в словарной статье для аббревиатуры появляется с индексом «исп. вместо» полная расшифровка аббревиатуры.

Пример лингвистического контроля смысла дескриптора с помощью словарной статьи.

Для сообщения «Средняя продолжительность времени жизни углерода-14» дескриптор ПРОДОЛЖИТЕЛЬНОСТЬ ЖИЗНИ использовать неправильно, так как в его словарной статье имеются отсылка к запрещенному термину СОКРАЩЕНИЕ СРОКА ЖИЗНИ и ассоциативные термины ВОЗРАСТ, СМЕРТЬ и т.д., что ограничивает значение этого термина только биологическим аспектом. Правильным дескриптором в данном случае является дескриптор ВРЕМЯ ЖЗНИ.

Обобщая опыт работы поисковых систем, можно сказать, что чем более специфические термины выбраны для поиска, тем более высока релевантность документов при поиске. Индексировщик должен помнить о процедуре обогащения ПОД («up-posting»), в результате которой к каждому дескриптору, использованному непосредственно при индексировании документа, автоматически приписываются все связанные с ним более широкие (родовые) термины. При поиске это позволяет достичь высокой полноты выдачи по поисковому предписанию (запросу), включающему достаточно общие поисковые термины. Однако, когда требуется высокая релевантность по очень специальным вопросам, необходимо формулировать запрос с использованием очень специальных (видовых) дескрипторов. Поэтому и индексирование должно вестись более специфическими дескрипторами. Отсюда можно сформулировать правило: необходимо всегда использовать более специфические (видовые) соответствующие дескрипторы. Например:

Не следует использовать более общий дескриптор, РАССЕЯНИЕ, УСКОРИТЕЛИ, ЭЛЕКТРОННЫЕ СХЕМЫ, СПЛАВЫ МАРГАНЦА, АЗОТСОДЕРЖАЩИЕ ОРГАНИЧЕСКИЕ СОЕДИНЕНИЯ.

Для нахождения соответствующего наиболее специфического дескриптора индексирующий должен просмотреть словарную статью выбранного дескриптора и обдумать, какой из связанных с ним видовых терминов наиболее приемлем. Когда в документе обсуждается очень специфическое понятие, а в тезаурусе нет дескриптора на этом уровне специфичности, понятие должно быть заиндексировано путем использования дескриптора следующего, более высокого уровня иерархической шкалы. С другой стороны, если такое специфическое понятие кажется важным в области атомной науки и техники и считается, что оно обладает достаточно высокой степенью вероятности быть полезным при поиске, то оно должно быть предложено как новый дескриптор, однако применение ближайшего родового термина в этом случае также необходимо. Таким образом, индексирующий должен включать в набор дескрипторов, которыми он описывает документ, дескрипторы, наиболее конкретные и специфичные для данного документа, если они имеются в тезаурусе.

Благодаря наличию автоматической процедуры обогащения ПОД («up-posting») от индексирующего не требуется приписывать при индексировании документов родовые понятия, связанные с выбранным дескриптором, за исключением следующих случаев:

I. Родовые термины используются, когда в документе детально рассматриваются вопросы, относящиеся отдельно к родовому понятию и к понятию, являющемуся видовым по отношению к первому, т.е. когда оба эти понятия являются равноправными предметами изучения. Например, в статье могут детально обсуждаться информация, которая релевантна высокотемпературным реакторам с газовым охлаждением, а также специфические детали реактора DRAGON, принадлежащего к этому типу реакторов. В этом случае индексирующий должен использовать оба дескриптора: ВЫСОКОТЕМПЕРАТУРНЫЕ ГАЗОГРАФИТОВЫЕ РЕАКТОРЫ и РЕАКТОР DRAGON, хотя первый дескриптор является родовым термином для второго и автоматически припишется по программам обогащения. Однако наличие обоих дескрипторов в информацион-

ной карте будет более глубоко отражать содержание документа и при необходимости даст возможность провести поиск только по дескрипторам, выбранным индексировщиком.

2. Родовые термины используются, когда в документе обсуждаются несколько узких понятий, принадлежащих одному и тому же классу, но лишь некоторые из них присутствуют в тезаурусе. Индексировщик должен включить в набор дескрипторов те видовые дескрипторы, которые имеются в тезаурусе, но также и более широкий (родовой) термин, идентифицирующий этот класс. Таким образом, будут отражены все специфические понятия этого класса, хотя некоторых специфических (видовых) дескрипторов в тезаурусе нет. Например, если в статье обсуждаются соединения гептан, октан, нонан и декан, то должны быть использованы дескрипторы ГЕПТАН и ОКТАН, так же, как и их родовой термин АЛКАНЫ, использование которого покрывает понятия нонан и декан, не настолько существенные, чтобы их представлять видовыми дескрипторами в тезаурусе.

8. ПОСЛЕДОВАТЕЛЬНОСТЬ СТАДИЙ СОСТАВЛЕНИЯ ПОИСКОВОГО ДЕСКРИПТОРНОГО ОБРАЗА ДОКУМЕНТА

Настоящая методика регламентирует следующую последовательность записи дескрипторов в ПОД:

- дескриптор, обозначающий тематическую рубрику, к которой относится анализируемый документ;
- дескрипторы, обозначающие объект или цель исследования, разработки;
- дескрипторы, описывающие исследуемый процесс;
- дескрипторы, описывающие методику и характер исследования;
- дескрипторы, отражающие условия, в которых исследование проводится;
- дескрипторы, описывающие применяемое для исследования оборудование;
- дескрипторы, указывающие область использования результатов исследования, разработки;
- дескриптор, обозначающий страну, в которой производится исследование, разработка;

- дескрипторы, отражающие аспекты, не вошедшие в данную свернутую формализованную модель содержания документа, но с точки зрения индексирующего представляющие значительный интерес.

Порядок расположения дескрипторов на информационной карте не влияет на машинный поиск информации, однако он помогает индексирующему правильно описать индексируемый документ, не упустить важных моментов, провести контроль индексирования и является своего рода формализованным рефератом, что особенно важно для документов, на информационной карте которых нет аннотации или реферата. Потребитель информации, зная порядок расположения дескрипторов на информационной карте, легко может оценить полезность полученного оповещения.

Фрагмент таблицы УДК

**004 Информационные технологии. Вычислительная техника.
Теория, технология и применения вычислительных
машин и систем.**

*Все подклассы 004 подразумевают цифровую обработку
данных, кроме тех, которые включают индексы 004.386 или
004.387.*

=> [621.3.049.77](#) Микроэлектроника. Интегральные схемы

=> [621.39](#) Электросвязь

Основные деления

004.2 Архитектура вычислительных машин

004.3 Аппаратные средства. Техническое обеспечение

004.4 Программные средства

004.5 Человеко-машинное взаимодействие. Пользовательский
интерфейс

004.6 Данные

004.7 Связь компьютеров. Сети ЭВМ. Вычислительные сети

004.8 Искусственный интеллект

004.9 Прикладные информационные (компьютерные) технологии

Специальные определители

004.01 Документация

004.02 Методы решения задач

004.021 Алгоритмы

=> [004.421](#) Алгоритмы составления программ

004.023 Эвристические методы

004.03 Типы и характеристики систем

004.031 Типы систем

004.031.2 Автономные системы. Системы с пакетной обработкой

004.031.4 Неавтономные системы. Онлайн-овые системы

004.031.42 Интерактивные системы

004.031.43 Системы реального времени. Системы обработки
транзакций

004.031.6 Встроенные системы

004.032 Характеристики систем

- 004.032.2 Режим обработки данных
 - 004.032.22 последовательный
 - 004.032.24 параллельный, одновременный
 - => [004.272](#) Архитектуры параллельной обработки
 - 004.032.26 Нейронные сети
- 004.032.3 Согласование по времени. Задание времени цикла
 - => [004.074.34](#) Время цикла памяти
- 004.032.32 Синхронные процессы
 - => [004.451.23](#) Синхронизация
- 004.032.322 Синхронизирующая частота. Тактовая частота
- 004.032.324 Период синхронизирующих импульсов. Такт.
 - Машинный цикл. Цикл внутреннего тактирования
- 004.032.34 Асинхронные процессы
- 004.032.6 Мультимедиа
 - => [004.357](#) Акустическая и мультимедийная периферия.
 - Устройства ввода данных с голоса
 - => [004.427](#) Средства разработки мультимедиа
- 004.032.8 Поколения компьютеров
 - Подразделять путем добавления к индексу номера поколения, например*
- 004.032.84 Компьютеры четвертого поколения
- 004.04 Ориентация процесса обработки данных
 - 004.041 процедурная
 - 004.042 на поток данных
 - 004.043 на структуру данных
 - 004.045 объектная
 - 004.046 функциональная
 - 004.047 логическая
 - 004.048 на реализацию искусственного интеллекта
 - => [004.8](#) Искусственный интеллект
 - .
 - .
 - .
- 004.07 Характеристики памяти
 - 004.072 Функционирование памяти
 - 004.072.2 Чтение
 - 004.072.3 Запись
 - 004.072.4 Доступ

004.072.5 Адресация
 004.072.6 Поблочная передача
 004.074 Эффективность памяти
 004.074.2 Плотность записи
 004.074.3 Время доступа
 004.074.32 Время позиционирования. Время установки головок
 004.074.34 Время цикла памяти
 => [004.032.3](#) Согласование по времени. Задание времени
 цикла
 004.076 Энергозависимость
 004.076.2 Энергозависимая память
 004.076.4 Энергонезависимая память
 004.08 Носители вводимых и выводимых данных. Запоминающие
 среды
 => [621.377.6](#) Цифровые накопители, резисторы и
 запоминающие устройства

Основной ряд индексов

004.2 Архитектура вычислительных машин
 004.22 Представление данных
 => [621.3.037.3](#) Виды представления информации
 004.222 Численные данные
 004.222.2 Представление чисел с фиксированной запятой (точкой)
 004.222.3 Представление чисел с плавающей запятой (точкой)
 004.222.5 Переполнение. Потеря значимости
 004.223 Символьные и другие подобные данные
 004.223.2 Алфавитно-цифровые данные
 Отдельные алфавиты обозначать при помощи :003.33...,
 например
 004.223.2:003.332.5 Представление арабской письменности
 004.223.3 Графические знаки. Включая: Двухбайтовое и
 трехбайтовое представление
 Отдельные письменности обозначать при помощи
 :003.32..., например
 004.223.3:003.324.1 Представление китайской письменности
 004.223.5 Специальные символы
 004.223.6 Управляющие символы
 004.223.7 Переключающие символы. Ключи перехода

- 004.23 Структура системы команд
 - 004.231 Виды систем команд
 - 004.231.2 Вычислительные машины с полной системой команд (CISC)
 - 004.231.3 Вычислительные машины с сокращенным набором команд (RISC)
 - 004.232 Формат команд
 - 004.233 Виды команд
 - 004.233.2 Команды ветвления
 - 004.233.3 Команды обработки данных
 - 004.233.5 Команды ввода-вывода
 - 004.234 Регистры
 - 004.235 Схемы адресации
 - 004.236 Подпрограммы в системе команд
 - 004.237 Прерывания
 - 004.238 Состояние процесса
 - 004.239 Защита памяти
- 004.25 Системы памяти
 - 004.252 Иерархия памяти
 - 004.254 Кэш-память
 - 004.255 Виртуальная память
 - 004.258 Система управления памятью
- 004.27 Перспективные архитектуры. Нефоннеймановские архитектуры
 - 004.272 Архитектуры параллельной обработки
 - => [004.032.24](#) Параллельный, одновременный режим обработки данных
 - 004.272.2 Методы параллельной обработки
 - 004.272.22 Конвейерное управление
 - 004.272.23 Использование присоединенных вспомогательных процессоров
 - 004.272.25 Векторная обработка
 - 004.272.26 Многопроцессорная обработка
 - 004.272.3 Архитектурные решения для параллельной обработки
 - 004.272.32 ОКМД-архитектура (одионочный поток команд и множественный поток данных, SIMD)
 - 004.272.33 МКОД-архитектура (множественный поток команд и одионочный поток данных, MISD)

- 004.272.34 МКМД-архитектура (множественный поток команд и множественный поток данных, MIMD)
- 004.272.4 Виды процессорных систем для параллельной обработки
- 004.272.42 Системы матричных процессоров
- 004.272.43 Многопроцессорные системы
- 004.272.44 Поточковые процессоры. Системы, управляемые потоком данных
- 004.272.45 Архитектура сети взаимодействующих процессоров
- 004.273 Программно-ориентированная архитектура
- 004.274 Динамическая архитектура
- 004.3 Аппаратные средства. Техническое обеспечение

Специальные определители

Здесь применяются специальные определители -1/-8 из класса 62 Инженерное дело. Техника в целом и определитель -9 из класса 66 Химическая технология, а также нижеследующие:

- 004.3'1 Производство вычислительных устройств
- 004.3'12 Принципы конструирования. Проектные соображения
- 004.3'122 Проектирование снижения шума. Проектирование снижения помех
- 004.3'124 Проектирование теплового режима. Технология охлаждения
- 004.3'14 Технология сборки ЭВМ. Компоновка компьютеров
- 004.3'142 Корпусы и конструктивное оформление
- 004.3'142.2 Компоновка на уровне устройств
- 004.3'142.22 Объединение компонентов в блоки. Крепление кристаллов на подложке
- Включая: Проводное соединение. Автоматическая сборка на ленточном носителе
- => [621.792](#) Соединение материалов с помощью адгезии
- 004.3'142.23 Перевернутые кристаллы (связывание за одну операцию с металлизированными межсоединениями подложки)
- 004.3'142.24 Корпусы с однорядным расположением выводов. Однорядные корпуса
- 004.3'142.25 Корпусы с двухрядным расположением выводов. Двухрядные корпуса

- 004.3'142.26 Корпусы с матричным расположением выводов.
Матричные корпуса
- 004.3'142.27 Плоские корпуса
- 004.3'142.4 Компоновка на уровне плат. Установка блоков на
панелях
Включая: Штырьковый монтаж. Поверхностный
(планарный) монтаж. Гибридная компоновка
- 004.3'142.6 Компоновка на уровне стойки
- 004.3'144 Компоненты компьютеров
Подразделять при помощи : (знак отношения), например
- 004.3'144:621.3.049.75 Печатные схемы компьютеров
- 004.3'144:621.3.049.771.15 Сверхбольшие интегральные схемы
компьютеров
- 004.3'144:621.314 Блоки питания компьютеров
- 004.3'144:621.316.54 Переключатели
- 004.3'144:621.318.5 Реле
- 004.3'2 Компьютерные установки. Установка компьютеров
Например
- 004.3'2:692.5 Конструкция пола для установки компьютеров
- 004.3'2:697.9 Вентиляция и кондиционирование воздуха для
компьютерных установок

Фрагмент классификации наук ГРНТИ

20 ИНФОРМАТИКА

УДК [002](#)

ВАК 05.25.00; 05.13.17

Примечание. Информационная деятельность в отдельных областях науки, техники и отраслях экономики отражается в соответствующих разделах с окончанием кода XX.01.29

20.01 Общие вопросы информатики

УДК [002](#)

ВАК 05.25.00; 05.13.17

20.01.01 Руководящие материалы

УДК [002](#)(094)

ВАК 05.25.00; 05.13.17

20.01.04 Информатизация общества. Информационная политика

УДК [002](#); [002:338.2](#)

ВАК 05.25.00; 05.13.17

См. также [12.41](#) Организация науки. Политика в области науки

Отс. от [26.11](#) Глобальные проблемы

20.01.07 Теория и методология информатики

УДК 002.001; [002:001.8](#)

ВАК 05.13.17

20.01.09 История информатики и информационной деятельности. Персоналия

УДК [002](#)(091); [002](#)(092)

ВАК 05.25.00; 05.13.17; 07.00.10

20.01.13 Научные и технические общества, конгрессы, конференции, симпозиумы, семинары, выставки

УДК [002:061.2/.4](#)

ВАК 05.25.00; 05.13.17; 05.26.05

20.01.17 Международное сотрудничество, деятельность международных организаций по информатике

УДК [002+02\].009\(100\)](#); [002.61](#)

ВАК 05.25.00; 05.13.17

20.01.33 Терминология информатики. Справочная литература. Учебная литература

УДК [002:001.4](#); [002\(03\)](#); [002\(075\)](#)

ВАК 05.25.00; 05.13.17; +13.00.02

20.01.37 Стандартизация в научно-информационной деятельности

УДК [002+02\].006](#)

ВАК 05.25.05, 08.00.05

20.01.45 Преподавание информатики

УДК [002:372.8](#)

ВАК 05.25.00; 05.13.17; 13.00.02

20.01.79 Кадры

УДК 002.6.08

ВАК 05.25.00; 05.13.17; 13.00.02

20.01.80 Правовые вопросы

УДК [002:34](#)

ВАК 05.25.00; 05.13.17; 12.00.03

20.15 Организация информационной деятельности

УДК [002.6](#); [021](#)

ВАК 05.25.00, 08.00.05

(Развитие рубрики введено с 1995 г.)

Библиотечное дело

см. [13.31](#) Библиотечное дело. Библиотечное дело

Архивное дело

см. [13.71](#) Архивное дело. Архивоведение

См. также [12.41](#) Организация науки. Политика в области науки

20.15.05 Информационные службы, сети, системы в целом

УДК [002.6](#)

20.15.06 Наднациональные и международные органы информации

УДК [002.61](#)

20.15.07 Национальные органы информации

УДК [002.63](#)

20.15.09 Отраслевые и ведомственные органы информации

УДК [002.63](#)

20.15.11 Региональные, локальные органы информации

УДК [002.63](#)

20.15.13 Информационные службы на предприятиях и в учреждениях

УДК [002.66](#)

20.15.31 Научные и технические библиотеки и библиотечные сети

УДК [026](#)

См. также [13.31](#) Библиотечное дело. Библиотековедение

20.15.71 Архивы, службы перевода и др. информационные органы

УДК [930.25](#); 651.927.07

Издательские организации

см. [19.51.61](#) Издательское дело

Распространение книжной продукции

см. [19.51.65](#) Книжная торговля. Пропаганда и распространение печати

См. также [13.71](#) Архивное дело. Архивоведение

20.17 Документальные источники информации

ВАК +05.25.02; +05.25.03; +05.25.04

20.17.01 Общие вопросы

УДК [002.2](#)

20.17.15 Виды источников информации

УДК [002.2](#)

20.17.17 Комплектование, учет и хранение фондов источников информации

УДК 002.52/.59

20.19 Аналитико-синтетическая переработка

документальных источников информации

УДК 002.53/.55

ВАК 05.25.05

Вопросы автоматического перевода текста

см. [16.31.21](#) Автоматическая обработка текста.

Автоматический перевод. Автоматическое распознавание речи

См. также [13.41](#) Библиография. Библиографоведение

20.19.01 Общие вопросы

УДК [002.53/.55](#)

ВАК [05.25.05](#)

20.19.15 Библиографическое описание источников информации

УДК [025.32](#)

ВАК [05.25.05](#), [05.25.03](#)

20.19.17 Предметизация и индексирование

УДК [025.4.025](#)

ВАК [05.25.05](#), [05.25.03](#)

20.19.19 Аннотирование и реферирование

УДК [002.53/.55:001.814](#)

ВАК [05.25.05](#)

20.19.21 Составление обзоров

УДК [002.53/.55:001.891.32](#)

ВАК [05.25.05](#)

20.19.23 Перевод научных текстов

УДК [651.926](#)

ВАК [05.25.05](#), [10.02.19](#)

20.19.27 Автоматизация знаковой обработки текста

УДК [002.53:\(084\)](#); [002.53:\(086\)](#); [004.91](#)

ВАК [05.25.05](#), [+05.11.18](#)

20.19.29 Обработка изобразительных и аудиовизуальных документов

(Введено с 1998 г.)

20.23 Информационный поиск

УДК [025.4.03](#)

ВАК [05.25.05](#)

20.23.01 Общие вопросы

УДК [025.4.03](#)

20.23.15 Информационно-поисковые языки

УДК [025.4](#)

См. также [16.21.47](#) Лексикология. Терминоведение

[16.31.31](#) Информационные и формализованные языки

20.23.17 Информационно-поисковые массивы. Базы данных. Манипулирование данными и файлами

УДК [002.53](#); [002.53:004.65](#); [002.53:004.62/.63](#)

(Содержание уточнено с 1998 г.)

20.23.19 Процессы информационного поиска

УДК [025.4.03](#)

20.23.21 Информационно-поисковые системы. Банки данных

УДК [025.4.03](#); 002.53:[004.65](#)

20.23.25 Информационные системы с базами знаний

УДК 002.53:[004.89](#)

См. также [28.23.35](#) Экспертные системы

20.51 Информационное обслуживание

УДК 002.55; [024](#)

ВАК 05.25.05

См. также [12.41.55](#) Информационное обеспечение научной деятельности

20.51.01 Общие вопросы

УДК 002.55; [024](#)

20.51.15 Потребители информации

УДК [002-052](#)

20.51.17 Информационные потребности и запросы

УДК 002.009.7:[330.163](#); [025.4.03](#)

20.51.19 Виды информационного обслуживания

УДК 002.55

20.51.21 Научно-техническая пропаганда

УДК [001.92](#)

20.51.23 Эффективность информационного обслуживания

УДК 002.55.003.13

20.53 Технические средства обеспечения информационных процессов

УДК [002+02\].002.5](#); [002:004.8](#); [002:004](#)

ВАК 05.25.05; 05.13.14

См. также [13.20.31](#) Техническое оснащение библиотек

20.53.01 Общие вопросы

УДК [002+02\].002.5](#); [002:004](#)

20.53.15 Средства ввода информации

УДК [004.35](#)

20.53.17 Средства хранения информации

УДК 004.33.07/.08; [004.33](#); [004.08](#)

20.53.19 Средства обработки и поиска информации
УДК 002.5:[004](#)

20.53.21 Средства выдачи информации
УДК [004.35](#); [004.08](#)

20.53.23 Средства передачи информации
УДК [004.71](#); 621.39.002.5

20.53.25 Средства копирования информационных материалов
УДК 002.513.3:[681.621.12](#)

20.53.27 Средства тиражного размножения информационных материалов
УДК [681.6](#)

20.53.29 Средства микрофильмирования информационных материалов
УДК [778.14](#)

20.53.31 Средства оргтехники в научно-информационной деятельности
УДК [002:651](#)

20.53.33 Здания информационных центров и их оборудование
УДК 002.6.006.002.5; [022](#)

Системы метаданных

«Дублинское ядро метаданных (Dublin-Core – DC) – средство формирования метаданных для описания широкого класса цифровых объектов. Ряд стран (Австралия, Швеция) объявили о принятии Дублинского ядра в качестве национального стандарта. Набор таких элементов, созданный группой, известен под именем Дублинского ядра по названию города Dublin в штате Огайо, где расположена штаб-квартира OCLC и где состоялась первая рабочая встреча. В результате было выделено пятнадцать элементов, составляющих набор метаданных Дублинского ядра. Все элементы являются необязательными и могут повторяться.

Преимущество Дублинского ядра заключается в том, что использовать его весьма просто, в то время, как традиционные методы каталогизации требуют профессиональной подготовки. Однако простота конфликтует с точностью. Для того, чтобы сохранить совместимость с простейшим описанием из 15 элементов и, увеличить детализацию и сложность описаний разрабатывают расширения и дополнительные *квалификаторы*.

Недостатки – невозможность создания универсальной системы метаданных, одинаково пригодной для различных функциональных задач. Особенно критикуется Дублинское ядро за слабое отражение аспектов, связанных с управлением правами на цифровые объекты.

MARC – машиночитаемый каталог. Один из старейших и самых известных и распространенных в России и в мире стандартов метаданных. Отличается детальным составом элементов данных, универсальностью, развитой структурой. Ориентирован на библиотечную практику, имеет конкурирующие версии.

GILS – Глобальная (правительственная) служба поиска информации. Являясь частью Национальной информационной инфраструктуры, GILS обеспечивает доступ частным лицам и организациям к федеральным информационным ресурсам, через общедоступный каталог этих ресурсов. GILS базируется на международных стандартах информационного поиска с использованием протокола

доступа Z39.50 и использует систему метаданных в рамках этого протокола. Развитие системы GILS предполагает создание системы взаимосвязанных каталогов для поиска метаданных, возможно, различных типов.

VCard – стандарт для представления информации о персоналиях, разработанный для унификации обмена данными между программными приложениями и информационными системами. Основными понятиями формата являются: *поток данных* и *объект vCard*. Поток данных – это набор vCard-объектов, которым обмениваются системы. vCard-объект – это набор предопределенных атрибутов, описывающих некоторый ресурс. Атрибут определяется своим именем, возможно, набором параметров, и значением. Параметры конкретизируют значение атрибута – например, для атрибута “телефон” параметры уточняют его значение: домашний, рабочий. Набор элементов и их параметров четко определен в стандарте и не расширяем. Формат позволяет создавать “электронные визитные карточки” для объектов типа персона.

CSDGM (Content Standards for Digital Geospatial Metadata) – стандарт для цифровых геопространственных данных. Стандарт призван представить достаточный набор терминов и определений для документации данного профиля. Стандарт CSDGM устанавливает имена элементов данных и групп элементов данных, используемых при обмене данными, определения этих элементов и групп, а также информацию о значениях, которые должны присваиваться элементам данных. Проведено различие между обязательными, ограниченно обязательными и необязательными терминами. Документ был одобрен в 1994 г., что обязывает американские федеральные организации использовать этот стандарт, начиная с 1995 г. Стандарт метаданных GSDGM содержит 334 элемента данных, более 100 из которых служат для описания связей между элементами.

ЕАД – кодировка архивных описаний. Набор изначально текстовых метаданных на базе языка разметки SGML, разработанных для нужд архивов и используемых для стандартизации и классификации уникальных архивных материалов, прежде всего рукописей. Версия 1.0 1998 г. совместима с форматом XML. Поддержи-

вается Американским архивным обществом и Библиотекой Конгресса США.

TEI – предложения по кодированию текстов <http://www-tei.uic.edu/orgs/tei/>): разработана в Центре электронных текстов Вирджинии в 1989 г. как инструмент при процессе оцифровке, который идентифицирует электронный ресурс и его печатный источник посредством метаданных, размещаемых внутри самого электронного ресурса.

IAFA/WHOIS – шаблонно ориентированные метаданные для описания сетевых ресурсов, первоначально использовавшиеся для описания списков электронной почтовой рассылки, других ftp-архивов, а позднее распространенные на другие ресурсы. Наиболее широкое применение наблюдалось в рамках ранних проектов британской программы по электронным библиотекам eLib (ROADS и т. д.), но даже сейчас эта схема метаданных считается одной из самых употребительных.

ONIX – основное назначение – выполнять функцию метаданных для представления информации об определенной группе товаров, а именно, полиграфической продукции, электронных изданий, аудио и видеопродукции с точки зрения книготорговли.

ONIX представляет собой реализацию схемы DTD (Definition Type Documents) данной группы товаров. Товар описан группой элементов, начинающейся с тега XML <Product> и заканчивающейся </Product>. Вся группа элементов, заключенная между двумя тегами, содержит описание товара.

Семантическое наполнение формата ONIX включает 25 групп элементов данных (полей). В некоторых случаях группы включают повторяющиеся подгруппы, которые называются «Области».

ПРИЛОЖЕНИЕ 5¹

Пример развития поиска и терминологической системы запроса²

Тема запроса: «Исследования в области использования тория в атомной энергетике»

Построим граф понятий предметной области, выделяя для направления **«Физика ядерных реакторов»** три аспекта:

- воспроизводство урана-233;
- замкнутый ториевый цикл;
- тяжеловодные реакторы.

Проведем поиск по полю «Ключевые слова» в реферативно-библиографической базе САРИ ЦНИИАтомИнформ (см. гл. 10) последовательно для всех аспектов, используя на первом шаге в поисковом запросе слова, составляющие обозначение основного понятия.

1. Аспект «Воспроизводство урана-233»

1.1. Проведем поиск по словам основного понятия

<i>Запрос</i>	<i>Всего найдено</i>	<i>Из них релевантных</i>
KW:(УРАН* и ВОСПРОИЗВОДСТ*)	18	15

Полученные 15 документов, описывают различные аспекты воспроизводства топлива в ядерных реакторах.

1.2. Выделим при просмотре из текста документов характерные термины (ключевые слова). Получили следующий список:

- ЗОНЫ ВОСПРОИЗВОДСТВА
- КОЭФФИЦИЕНТ ВОСПРОИЗВОДСТВА
- ОТРАБОТАННОЕ ТОПЛИВО
- РАСШИРЕННОЕ ВОСПРОИЗВОДСТВО
- ИЗОТОПНОЕ ОТНОШЕНИЕ

¹ Пример подготовлен Болотиным Е.И.

² Цель примера, в следствие ограничений объема пособия, состоит только в том, чтобы проиллюстрировать эффективность использования лексики релевантных документов.

- ОБРАБОТКА РАДИОАКТИВНЫХ ОТХОДОВ
- **ПРЕВРАЩЕНИЕ ИЗОТОПОВ**
- ВЫГОРАНИЕ
- ВРЕМЯ ВЫДЕРЖКИ ТОПЛИВА
- ПРОДУКТЫ ДЕЛЕНИЯ
- ЗАВОДЫ ПО ПЕРЕРАБ. ОБЛУЧЕННОГО ТОПЛИВА
- ОТРАБОТАННЫЕ ТВЭЛЫ
- **ПОВТОРНОЕ ИСПОЛЬЗОВАНИЕ УРАНА**

1.3. Значимые для нашего аспекта слова (выделено **жирным**) включим в выражение запроса и проведем поиск.

<i>Запрос</i>	<i>Всего найдено</i>	<i>Из них релевантных</i>
KW:(УРАН* и (ВОСПРОИЗВОДСТ* или (ПРЕВРАЩЕН* и ИЗОТОП*) или (ПОВТОРН* и ИСПОЛЬЗОВАН*)))	53	19

Среди 19-ти релевантных документов, 15 описывают *методы повторного использования уранового топлива*.

1.4. Выделим при просмотре из текста документов характерные термины (ключевые слова). Получили следующий список:

- **ЗАМКНУТЫЙ ТОПЛИВНЫЙ ЦИКЛ**

1.5. Значимые для нашего аспекта слова (выделено **жирным**) включим в выражение запроса и проведем поиск.

<i>Запрос</i>	<i>Всего найдено</i>	<i>Из них релевантных</i>
KW:(УРАН* и (ВОСПРОИЗВОДСТ* или (ПРЕВРАЩЕН* и ИЗОТОП*) или (ПОВТОРН* и ИСПОЛЬЗОВАН*) или (ЗАМКНУТ* и ТОПЛИВН* и ЦИКЛ*)))	71	30

Среди 30-ти релевантных документов, 11 описывают *разработки замкнутых топливных циклов для различных реакторов*.

1.6. Выделим при просмотре из текста документов характерные термины (ключевые слова). Получили следующий список:

- **ПЕРЕРАБОТКА ОБЛУЧЕННОГО ТОПЛИВА**
- ОБРАБОТКА РАДИОАКТИВНЫХ ОТХОДОВ

1.7. Значимые для нашего аспекта слова (выделено **жирным**) включим в выражение запроса и проведем поиск.

<i>Запрос</i>	<i>Всего найдено</i>	<i>Из них релевантных</i>
KW:(УРАН* и (ВОСПРОИЗВОДСТ* или (ПРЕВРАЩЕН* и ИЗОТОП*) или (ПОВТОРН* и ИСПОЛЬЗОВАН*) или (ЗАМКНУТ* и ТОПЛИВН* и ЦИКЛ*) или (ПЕРЕРАБОТК* и ОБЛУЧЕНН* и ТОПЛИВ*)))	112	47

Среди 47-ти релевантных документов, 17 описывают *методы переработки облученного топлива, а также информацию о перерабатывающих заводах.*

1.8. При просмотре документов новых понятий не обнаружено

1.9. Проведем эвристический поиск (реформулировка запроса по обратной связи), по множеству истинно релевантных документов

<i>Запрос</i>	<i>Всего найдено</i>	<i>Из них релевантных</i>
ЭВРИСТИЧЕСКИЙ ПОИСК	25	24

Среди выданных **НОВЫХ ДОКУМЕНТОВ НЕ ОБНАРУЖЕНО.**

Это подтверждает полноту поиска и *построенного в процессе поиска терминологического представления* (рис. 1). Таким образом, поиск по этому полю для ланного аспекта можно завершить¹.

2. Повторим аналогичные действия для аспекта

Коэффициент воспроизводства

(Описание процесса опускаем)

3. Повторим аналогичные действия для аспекта

Замкнутый ториевый цикл

(Описание процесса опускаем)

4. Построим дерево понятий предметной области по значимой лексике, обнаруженной в процессе поиска.

¹ При практическом поиске процесс было бы целесообразно повторить для других информативных полей, таких, как «Заглавие» и «Реферат».

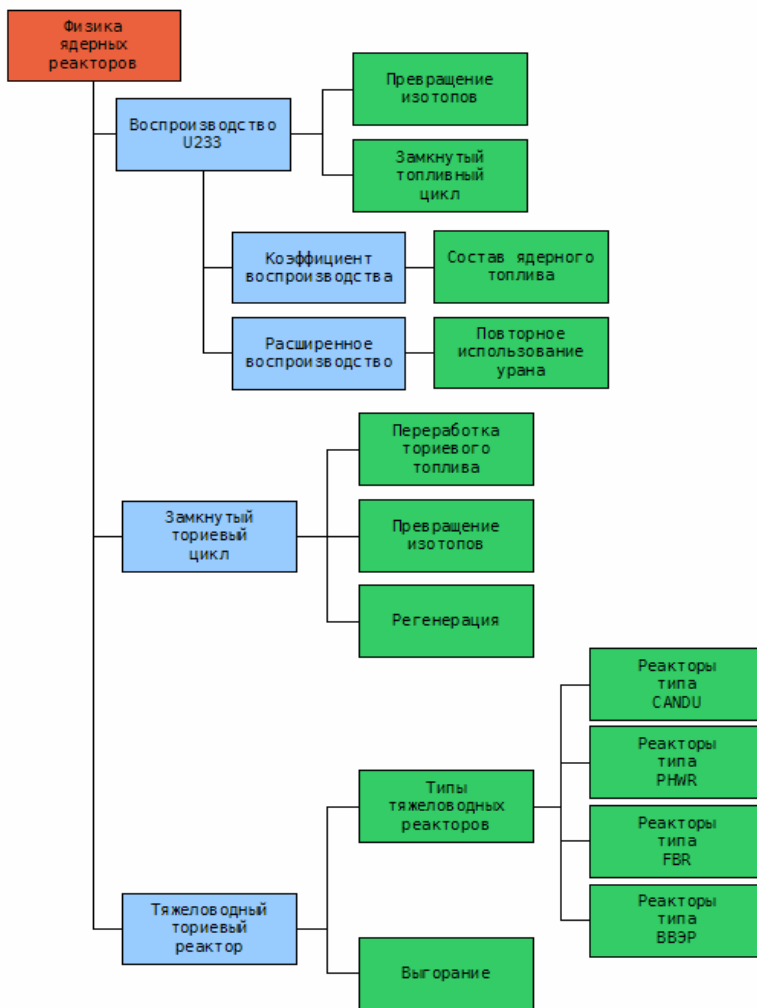


Рис. П.5.1. Итоговое дерево понятий по направлению «Исследования в области использованию тория в атомной энергетике»

Оглавление

Введение	3
РАЗДЕЛ 1. ИНФОРМАЦИОННЫЕ РЕСУРСЫ И АВТОМАТИЗИРОВАННЫЙ ДОКУМЕНТАЛЬНЫЙ ПОИСК.....	13
Глава 1. Концептуальные основы информационного поиска	13
1.1. О понятиях «информация» и «информационные ресурсы»	13
1.2. Введение в информационный поиск	19
1.3. Информационные коммуникации и свойства информации	26
1.4. Обобщенная схема информационного обслуживания.....	35
1.5. Типология поисковых задач и форма выражения запроса	39
1.6. Обобщенная схема поиска.....	42
Глава 2. Технологические и лингвистические основы информационного поиска	46
2.1. Организация данных и критерии поиска	46
2.2. Функциональная обработка запросов и документов в АИПС.....	57
2.3. Лингвистическое обеспечение и обработка информации в АИПС.....	60
2.4. Поисковый интерфейс	75
РАЗДЕЛ 2. ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ИС И АВТОМАТИЧЕСКАЯ ОБРАБОТКА ДОКУМЕНТАЛЬНОЙ ИНФОРМАЦИИ	82
Глава 3. Классификации в задачах информационного поиска.....	82
3.1. Библиотечно-библиографические классификации	85
3.2. Международная патентная классификация	106
3.3. Отраслевые классификационные системы.....	115
Глава 4. Терминологические структуры	127
4.1. Линейные терминологические структуры	128
4.2. Иерархические терминологические структуры	131
4.3. Терминологические структуры с сетевой организацией ..	135
Глава 5. Методы автоматической обработки и лексического анализа документальной информации	160
5.1. Методы выделения информативных терминов	160

5.2. Статистические меры близости терминов	164
5.3. Методы классификации и кластеризации документов.....	169
5.4. Автоматическое реферирование и аннотирование	175

РАЗДЕЛ 3. РАСПРЕДЕЛЕННАЯ ОБРАБОТКА ИНФОРМАЦИИ

Глава 6. Распределенные файловые системы	181
6.1. Файловый сервис	184
6.2. Сервис каталогов	186
6.3. Семантика разделения файлов	187
Глава 7. Распределенные вычисления	191
7.1. Архитектура распределенных вычислений.....	192
7.2. Виды параллелизма в распределенных вычислениях.....	193
7.3. Примеры проектов распределенных вычислений	197
7.4. Андеграунд распределенных вычислений	205
Глава 8. Архитектура распределенных ИР	207
8.1. Типология и структура распределенных ИР	207
8.2. Распределенная обработка в поисковых машинах Internet	212
8.3. Организация доступа к распределенным документальным ИР	213
8.4. Обобщенная схема доступа к ресурсам ЭБ.....	217
8.5. Доступ к полным текстам документов.....	219
8.6. Интероперабельность в распределенных ИР	222

РАЗДЕЛ 4. ДОКУМЕНТАЛЬНЫЕ РЕСУРСЫ НАУЧНОЙ ИНФОРМАЦИИ

Глава 9. Электронные каталоги и библиотеки.....	228
9.1. Электронные ресурсы Библиотеки по естественным наукам РАН	228
9.2. Электронные ресурсы Государственной публичной научно-технической библиотеки России (ГПНТБ).....	243
9.3. Электронные ресурсы Библиотеки Конгресса США.....	258
Глава 10. Реферативно-библиографические документальные БД	266
10.1. Электронные каталоги и базы данных ИНИОН РАН.....	266
10.2. Базы реферативно-библиографической информации ВИНИТИ РАН.....	288

10.3. Базы информационных карт по диссертациям и НИР ВНТИЦентра	293
10.4. Базы патентной информации ФИПС.....	297
10.5. Базы данных Международного агентства по атомной энергии	306
Глава 11. Электронные ресурсы издательств	311
11.1. Сервис Science Direct.....	311
11.2. Научная электронная библиотека eLibrary.....	317
Глава 12. Специализированные БД и ИР.....	332
12.1. Портал NUCLEUS.....	332
12.2. Сайты российских предприятий и организаций ядерной отрасли	334
12.3. Специализированные базы данных	339
Закключение: еще раз о человеко-машинном поиске информации в распределенных ИР	346
Список литературы.....	352
Глоссарий.....	355
Список сокращений	366
ПРИЛОЖЕНИЕ 1. Фрагмент Методики индексирования документов	367
ПРИЛОЖЕНИЕ 2. Фрагмент УДК.....	377
ПРИЛОЖЕНИЕ 3. Фрагмент Рубрикатора ГРНТИ.....	383
ПРИЛОЖЕНИЕ 4. Основные системы метаданных.....	389
ПРИЛОЖЕНИЕ 5. Пример развития поиска и терминологической системы	392

**Максимов Николай Вениаминович
Голицына Ольга Леонидовна
Тихомиров Георгий Валентинович
Храмцов Павел Брониславович**

Информационные ресурсы и поисковые системы

Учебное пособие

Редактор Т.В. Волвенкова

Подписано в печать 20.11.08. Формат 60×84¹/₁₆

Объем 25,0 п.л. 25,0 уч.-изд. л. Изд. № 1/42.

Заказ

Тираж 150 экз.

Московский инженерно-физический институт
(государственный университет)
115409, Каширское шоссе, 31.

Типография издательства «Тривант»
г. Троицк Московской обл.

