Московская финансово-промышленная академия

Голицына О.Л.
Максимов Н.В.

Информационные системы

Москва, 2004

Голицына О.Л, Максимов Н.В. Информационные системы / Московская финансово-промышленная академия. - М.: 2004. - 329 с.

Рекомендовано Учебно-методическим объединением по образованию в области прикладной информатики (по областям) в качестве учебного пособия для студентов высших учебных заведений, обучающихся по специальности 351400 «Прикладная информатика (по областям)» и другим междисциплинарным специальностям.

Содержание

ВВЕДЕНИЕ

С того времени, когда научные исследования стали индустрией, проблемы информационного обеспечения преобразовались в самостоятельное направление, в значительной степени ориентированное на вычислительную технику и электронные средства коммуникаций.

Достижения очевидны и ощутимы. Практически обеспечена возможность организации всемирного каталога публикаций через объединение[1] каталогов и баз данных крупнейших национальных библиотек и информационных центров практически всех развитых стран, включая РФ. Глобальная сеть Internet сделала реальностью открытость и доступность в реальном масштабе времени информационные ресурсы (ИР) самого разного объема и содержания, от частной коллекции до национальных архивов.

Однако даже широкое внедрение информационных систем и баз данных в сочетании с сетевыми решениями, тем не менее, остается пока лишь очередным этапом в попытках человека справиться с проблемами получения и переработки информации. Интенсивное развитие вычислительной техники и кибернетических методов управления информационными потоками по существу дало только новые (электронные) носители и сверхбыстрый доступ к хранилищам. Это позволило на порядки увеличить скорости получения данных и объемы оперативно доступного информационного пространства, однако ситуация с использованием собственно информации практически не изменилась.

Причины здесь кроются не столько в финансовой, организационной или технической сфере, сколько в том, что человек, как система переработки и генерации самой информации, принципиально не изменился. Рассматривая перспективы развития науки, физик Дж. Томсон по существу характеризовал и особенности использования информации: «…совершенно неизбежно, что по мере расширения наших знаний та их доля, которой в состоянии овладеть один человек, будет убывать. Поскольку для работы человеку надо знать очень многое, он испытывает величайший соблазн учить как можно меньше из того, что ему в работе непосредственно не пригодится. У него создается однобокое представление о мире, в котором он живет. Второе из нежелательных последствий проявляется на более поздней стадии специализации. Многие достижения науки и техники являются следствием внедрения тех или иных идей в областях, для которых они первоначально не предназначались. Однако реализация подобной возможности требует все труднее и труднее достижимой широты знаний. Вполне может случиться, что эта особенность воздвигает предел прогрессу науки, но здесь многое можно сделать для того, чтобы отодвинуть приход этой катастрофы» [Томсон1958].

---

[1] Например, с использование унифицированных средств доступа на основе протокола Z39.50

Количество литературы быстро растет практически во всех областях. И для науки индустриального периода развития общества стало характерно разделение труда: исследователь все больше нуждается в помощи специальной службы, призванной ориентировать его в потоке информации. В начале индустриального периода такой службой являлись органы научной информации, в задачи которых входили поиск и концентрация информации зачастую непосредственно для исследователя, не только на уровне проблемам, но на уровне отдельной задачи. Это вполне отражало экстенсивную оценку ситуации, данную Дж. Томсоном: «В науке неминуемо должно произойти то, что произошло в армии. В целях обеспечения боеспособности солдата на передовой линии приходится в тылу ставить за ним все больше людей». Развитие информационных технологий и глобализация информационных коммуникаций, свой ственные постиндустриальному периоду развития, как кажется, обеспечили возможность взаимодействия исследователей минуя информационных посредников. Парадокс современности состоит в том, что общество получив технические возможности непосредственного взаимодействия исследователей, обрело при этом такие проблемы как, например, ограничения авторского и имущественного права на передачу информации[2]. Организации, занимающиеся информационным обеспечением науки и производства, практически исчезли. Функции поиска перешли к потребителю информации (исследователям), а задачи систематизации и концентрации информации по большей части берут на себя издатели[3].

Библиотеки и автоматизированные банки данных это внешняя, глобальная, но пассивная память, хранящая разнородную и по-разному представленную информацию, а Internet и разнообразные средства вычислительной техники являют собой высокоэффективные, но, тем не менее, узко специализированные средства доступа к *данным*, хранимым в машиночитаемом виде. Причем, объемы данных и темпы их прироста, а также разнообразие форм и динамика представления информации настолько велики, что в некоторых случаях «…по сути дела легче открыть новый факт или создать теорию, чем удостовериться, что они еще не были созданы или выведены» [Воробьев1966].

Однако упомянутый тезис должен восприниматься, конечно же, не в качестве отрицания информационных технологий, а скорее как фактор, стимулирующий исследования и разработки направлений, связанных с изучением глубинных процессов переработки информации в человеко-машинных системах и, прежде всего, для задач генерации нового знания. Само развитие информационных технологий и их повсеместное внедрение практически во все сферы деятельности человека (по крайней

---

[2] То есть, «люди тыла» переместились из сферы, приближенной к исследователю, в сферу управления.

[3] При этом пока сохранились и продуценты (издатели) вторичной информации – информационные службы, производящие не только аналитическую обработку проблемно-ориентированных потоков информации, но и его систематизацию, что фактически и обеспечивает для исследователя возможность эффективного поиска информации без информационного посредника.

мере, той, которая связана или завершается документом или коллекцией данных) уже привело к тому, что электронная форма представления информации принята официально (и законодательно, например, в СССР еще в 1980г.). Радикальность и необратимость этого процесса признана, в том числе, в такой консервативной области, как архивоведение, что констатировалось на Международном совещании архивистов (Италия, Масерата, 1991г., Москва, 27-28 ноября 1997г).

Развитие информационных технологий в области документалистики и информационного обеспечения научных исследований получило развитие в многочисленных проектах создания и внедрения электронных библиотек (ЭБ) как глобального, так и локального масштаба. По результатам исследований Института развития информационного общества [45], в той или иной форме идея электронной библиотеки уже работает во многих университетах и крупных библиотеках ведущих стран мира. Например, электронная "библиотека XXI века" создается в Японии путем соединения усилий Агентства по внедрению новых технологий, Национальной парламентской библиотеки, ряда министерств, библиотек и культурных центров. Несколько лет назад Библиотека Конгресса США начала реализацию национальной программы создания электронной библиотеки. Начиная с 1994 г. по инициативе NSF, DARPA и NASA в США была развернута исследовательская программа Digital Libraries Initiative по электронным библиотекам. На второй стадии развития в начале 1998 г. эти программы были объединены в единую межведомственную программу (DLI - Phase 2), в которой, кроме того, участвуют Национальная медицинская библиотека, Агентство по статистике США, Национальный гуманитарный фонд, Национальный архив США и другие федеральные агентства. Начиная с 1995 г., осуществляется проект Bibliotheca Universalis создания электронных библиотек для стран «семерки». С 1995 г. осуществляется национальная программа eLib в Великобритании. В других странах (Канада, Германия и т.д.) многочисленные разрозненные проекты в последние годы также стали превращаться в национальные и международные программы создания электронных библиотек.

В России реализация проектов по созданию электронных библиотек начата сравнительно недавно. С 1998 г. по инициативе Российского фонда фундаментальных исследований и Российского фонда технологического развития осуществляется программа "Российские электронные библиотеки", в рамках которой ведутся работы по общесистемным вопросам создания и функционирования электронных библиотек, развитию инфраструктуры, разработке инструментальных средств, а также создание конкретных электронных библиотек по областям науки, культуры и образования. В настоящее время успешно функционируют электронные библиотеки ИНИОН РАН, ВИНИТИ РАН, ВНТИЦентра, РГБ, электронные каталоги БЕН, ГПНТБ и др.

Наряду с упомянутыми примерами, представляющими, в основном, масштабные проекты и, главное – имеющими преимущественно информационную историю и вобравшими лучшие традиции информационной деятельности, ориентированной на обслуживание пользователя, заметной составляющей (по крайней мере, Internet-ресурсов) стали коллекции информационных объектов различного вида и назначения. Особенностью таких ресурсов[4] помимо распределенности является гетерогенность - практически неограниченное разнообразие форм их представления (форматов и сред хранения), а также разнокалиберность условий и методов доступа. Сюда относятся ресурсы самого разного масштаба: от отдельных электронных документов, размещенных на авторской Internet-странице, до электронных коллекций и библиотек крупнейших издательств. Но, следует отметить, что в любом случае информационный ресурс, в отличие от набора данных, идентифицируется не только адресом хранения, но и содержанием, имеющим информационную природу (практически каждый ресурс создается изначально ориентированным на адресное хотя, возможно, и неоднозначное восприятие, что и является предпосылкой создания новой информации), а его доступность обеспечивается встроенным или внешним, более или менее развитым поисковым инструментом, *избыточном* по отношению к самому ресурсу.

В общем случае можно заметить, что в качестве компонентов здесь выступают электронные каталоги (библиографические и реферативные базы данных), полнотекстовые массивы (электронные журналы, фактографические базы данных, коллекции электронных документов или копий первоисточников и т.д.), справочно-нормативные файлы (рубрикаторы, тезаурусы, авторские, предметные, географические и другие указатели), возможно связанные между собой ссылками, указателями хранения или условиями поиска. Например, записи электронных каталогов содержат указания местоположения книг, а справочно-нормативные файлы традиционно используются в качестве "точек входа" в библиографические и реферативные базы данных. С появлением технических возможностей создания полнотекстовых баз данных справочно-поисковый аппарат и собственно массив информации технологически становятся единым целым, и на первый план выходит задача организации такой взаимосвязи, чтобы переход по ссылке от компонентов одного ресурса к компонентам другого, а также от компонентов одного уровня к компонентам другого, воспринимался пользователем как простейший одноактный процесс, подобный перевороту страницы книги.

Поскольку конечной целью построения любой информационной системы является обеспечение пользователю условий получения нужной

[4] Информационные ресурсы в [Попов1996] определяются как совокупность накопленной информации, зафиксированной на материальных носителях в любой форме, обеспечивающей ее передачу во времени и пространстве. Таким образом, в контексте автоматизированных информационных систем под информационными ресурсами можно подразумевать информационные массивы и базы данных, рассматриваемые *совместно* с информационными технологиями, обеспечивающими их доступность.

информации, немаловажную роль играет форма и процедура подачи этой информации. Для того чтобы пользователь мог легче воспринимать большие объемы информации, разработано множество форм и методов ее представления, что выражается, например, в создании «фирменных» стандартов хранения и методов поиска, а также интерфейсов, адаптируемых каждым конкретным пользователем для себя. В то же время наблюдается, что часть пользователей Internet полагают достаточным использование в качестве средств поиска стандартных программ общего назначения, как например Internet Explorer или Netscape Navigator.

Другой особенностью современности является наблюдаемый режим "информационного самообслуживания". Пользователь, привыкший к интуитивному освоению программных сред (в основном стандартных средств операционной системы, большинство из которых имеет существенно более простой и дружественный интерфейс), часто неадекватно оценивает состояние и результаты поиска. Показательными примерами являются такие ситуации, как:

- принятие безаппеляционного решения о "плохой" базе данных или поисковой системе после получения неудовлетворительного или нулевого результата по первому же запросу, иногда даже не являющемуся правильным с точки зрения поискового языка;
- прекращение пользователем развития запроса, если он получает известные или собственные публикации, т.е. когда происходит подмена критерия остановки процесса поиска по условию нахождения нужной новой информации или остановки по условию отсутствия новой информации в каждой следующей выдаче фактом подтверждаемости "результативности" выражения запроса.

Кроме того, пользователь рискует обрести некоторую убежденность в том, что поисковые системы (особенно когда речь идет о поисковых средствах Internet) всемогущи и вездесущи, а их способности извлекать информацию (знания) из текстов и массивов документов бесконечно выше человеческих. Такой подход в самом безобидном случае приводит к некритичному отношению к результату поиска, т.е. пользователь удовлетворяется уже фактом получения выдачи (а современные поисковые системы часто устроены так, чтобы практически всегда пользователю выдавались какие-нибудь документы, пусть даже и в минимальной степени формально соответствующие запросу).

Поскольку система является всего лишь *инструментом, используемым человеком при поиске*, а не *интеллектуальным автоматом для поиска* информации, эффективность ее использования зависит от того, насколько хорошо человек знает природу объектов и свойства инструмента, посредством которого он с этими объектами работает. Таким образом, можно сказать, что процесс информационного обеспечения (поиска и предоставления информации по проблеме) предполагает опреде-

ление (построение) *информационного пространства (среды)*, включающей:

- информационные ресурсы, содержащие данные, сведения и знания, зафиксированные на соответствующих носителях информации;
- организационные структуры, обеспечивающие функционирование и развитие информационного пространства, в частности, сбор, обработку, хранение, распространение, поиск и передачу информации;
- средства поддержки информационного взаимодействия, в том числе программно-технические средства и организационно-нормативные документы, обеспечивающие доступ к информационным ресурсам на основе соответствующих информационных технологий, рассматриваемых в контексте условия наследования особенностей существующего положения и требования преемственности будущих решений.

Изложенные выше аспекты позволяют выделить три уровня задач организации, создания и использования информационных ресурсов:

1) Уровень взаимодействия пользователей с ресурсами – задачи организации работы пользователей с информационными ресурсами (свойства и характер используемых ресурсов, интерфейсы и технологии поиска, справочно-обучающая поддержка и т.д.).

2) Уровень системной организации информационного пространства – задачи, связанные со структурой информационного пространства.

3) Организационно-технологический уровень - задачи функционирования и сопровождения информационного ресурса на протяжении всего его жизненного цикла (выбор источников, средства создания и ведения баз данных, выбор стандартов хранения информации, протоколов взаимодействия и доступа и т.д.).

Необходимо отметить, что все компоненты и задачи тесно связаны между собой и должны рассматриваться параллельно и в контексте конечной цели – такой организации взаимодействия компонент совокупной человеко-машинной системы (взаимодействия пользователя с распределенным информационным ресурсом), которая обеспечит эффективность процесса генерации нового знания. Причем, автоматизация информационной деятельности и управления информационными ресурсами на всех уровнях обуславливает необходимость разработки общих принципов и теоретических основ моделирования информационных ресурсов, которые, во-первых, охватывали бы максимальное количество типов и уровней информационных процессов и технологий управления ИР, а во-вторых – были бы работоспособны с позиций реализации конкретных систем.

Информационные системы, как и базы данных, составляющие их основу – это уже достаточно хорошо проработанная научная дисциплина. Существует множество, в том числе и фундаментальных, работ и

учебников (на материал которых авторы опирались при подготовке этого учебника, и убедительно рекомендуют их тем, кто серьезно интересуется этой проблематикой), среди которых в первую очередь необходимо выделить такие монографии, как «Основы информатики» Михайлова А.И, Черного А.И., Гиляревского Р.С., «Динамические библиотечно-информационные системы» Дж. Солтона, «Теоретические основы научно-технической информации» Т.В. Муранивского.

В своей работе авторы руководствовались и тем, что материал должен не только представлять существо конкретной темы, но и подвести читателя к пониманию обоснованности (или условности) того или иного решения. Авторы сознательно избегали описаний языков и технологий, применяемых в конкретных системах, предполагая, что полноценное освоение материала курса связано с практикой и, соответственно, с неизбежным изучением конкретных подходов, языков и технологий, свойственных выбранной системе, и изложенных в специальных пособиях, учебниках и руководствах.

Материал курса, представленный в восьмии главах и приложении, условно можно отнести к следующим разделам:

- введение в информатику и информационные системы;
- теоретические основы информационно-поисковых систем;
- лингвистическое обеспечение;
- когнитивные модели и особенности человеко-машинного взаимодействия в процессах информационного поиска;
- технологии и средства информационного поиска.

В первой главе введены основные понятия, относящиеся к информатике: информация, данные, знания; определены основные свойства информации.

Во второй главе определены состав и структура информационной системы, рассматриваемой как средство автоматизированной обработки данных. С точки зрения общей теории систем рассмотрена информационная модель управления в системах материальных преобразований и систем генерации знаний.

В третьей главе определены основные понятия, относящиеся к базам данных. Рассмотрен важнейший вопрос семантики баз данных в контексте информационных систем и определено соотношение понятий «информация» и «данные». Представлены базовые технологии машинной обработки данных и рассмотрены ключевые моменты, определяющие эффективность процессов управления данными. Приведены примерные схемы управления данными в документальных информационных системах.

В четвертой главе приведены математические модели поиска информации. Вводятся формальные меры оценки технической эффективности поиска документальной информации. Представлена система ма-

тематических моделей механизмов поиска в документальных базах данных.

В пятой главе рассматривается лингвистическое обеспечение информационных систем. Даны основные понятия и типология методов описания содержания документов. Описаны классификационные и дескрипторные языки. Рассмотрены особенности использования терминологических структур в технологиях документального поиска.

В шестой главе в рамках обобщенной модели воспроизводства информации определены требования к представлению информации на разных уровнях. На основе типологии информационной неопределенности и видов поисковых задач определены особенности реализации поисковых процессов. Вводятся определения основных компонент информационного поиска.

В седьмой главе рассматриваются модели и технологии поиска в документальных базах данных научной информации. Процесс поиска, представляемый как управляемое средствами интерфейса взаимодействие двух систем (человек – АИС), рассматривается на двух уровнях: внутрисистемного представления информации и особенностей восприятия информации пользователем с учетом когнитивных и поведенческих аспекты деятельности человека. Вводится схема типизации поведения пользователей при поиске. Подробно рассмотрены основные компоненты и технология информационного поиска.

В восьмой главе приведены примерные реализации интерфейсных решений процессов подготовки и развития поискового запроса.

1. Информация. Основные свойства и определения

Информатизация общества является стратегическим фактором развития цивилизации, который, благодаря особым свойствам информации, дает человечеству определенные шансы решить глобальные проблемы и перейти к новой парадигме устойчивого развития. При этом информационные системы и технологии становятся теми средствами, которые человек может использовать как для расширения, так и для развития своих собственных способностей: памяти, логики, пространственного воображения.

1.1. Информация как основной предмет информатики

В качестве источников информатики как теоретической платформы информационных систем обычно называют две науки — *документалистику* и *кибернетику,* возникновение которых было тесно связано с бурным развитием производственных систем и технологий. Основным предметом документалистики стало изучение рациональных средств и методов повышения эффективности документооборота как информационной основы накопления и поиска информации. Понятие информации составило также и основу кибернетики, как науки о методах анализа и синтеза систем эффективного управления.

Развитие средств вычислительной техники и широкое ее использование в различных областях человеческой деятельности привело к тому, что с понятием «информатика» стала тесно связана и другая область — *наука о средствах вычислительной техники (Computer Science)*. И сегодня предмет информатики, рассматриваемой уже как совокупность информационных ресурсов и технологий, в общем случае составляют такие понятия, как:

– средства вычислительной техники;

– программное обеспечение средств вычислительной техники;

– методы взаимодействия человека с вычислительной техникой и программными средствами (программным обеспечением);

– информационные ресурсы[5] (ИР), в том числе средства создания, хранения, поиска информации;

– средства итехнологии доступа к распределенным информационным ресурсам;

– методы и средства взаимодействия человека с информационными ресурсами на базе вычислительной техники с использованием программного обеспечения;

– инструментальные технологии, обеспечивающие жизненный цикл ИР.

---

[5] В [Громов, Попов96] к информационным ресурсам относят не только информационные продукты, но и средства и технологии их создания и использования

Информатика занимается не собственно вычислениями, а обработкой информации (представленной, преимущественно, в числовой и символьной форме), а само содержание термина «вычисления» в информатике расширяется, охватывая наряду с собственно вычислительными процессами также и те, которые связаны с кодированием и обработкой текстов, их поиском и передачей по различным каналам связи. Это обеспечивает возможность для уменьшения разнообразия и сложности знаковых средств, используемых для представления информации при сохранении ее количества. Современные формы представления информации (искусственные языки, модели, коды, символы, формулы и т.п.) позволяют минимумом знаковых средств выразить максимум содержания информации. Одна и та же по содержанию и объему информация может быть представлена более емко или менее емко в зависимости от того, какими знаками она выражена [Аветисян99].

Однако концентрация информации - это не только выбор системы знаков для ее более компактного выражения. Непрерывный процесс концентрации информации - это процесс постоянного ее обобщения, сведение разрозненных фактов и данных в более емкие и вместе с тем более богатые, глубокие, зачастую фундаментальные и методологические знания, из которых могут выводиться конкретные следствия или предположения. Это позволяет отказаться от запоминания и фиксирования исходных или выводимых составных элементов концентрированной информации, т.к. они всегда могут быть получены из более общего знания, более общей информации.

Приведем определение основного предмета информатики (определяющего основные объекты информационных коммуникаций), данное в предисловии к первому изданию одного из первых фундаментальных трудов в области автоматизированных информационных систем – «Основы научной информации» [Михайлов68]: «…научная информация захватывает три совершенно различные области человеческой деятельности. Первая – это мало исследованная область творческого мышления человека и организация умственного труда. Проблема поиска необходимых специалисту сведений может быть успешно решена лишь при условии, что будут изучены логические операции, производимые ученым при поиске нужной ему информации. … Вторая часть предмета – это довольно широкий комплекс вопросов, связанных с различными научными документами. Существующие виды научных документов сложились эмпирически, …. однако основания думать, что ученые располагают иными возможностями для распространения и сохранения во времени результатов своего труда, являются в значительной степени умозрительными. … Третий комплекс вопросов относится к созданию технических средств, необходимых для успешной научной деятельности».

Определение предмета информатики, приведенное выше, не потеряло актуальности. В рамках обобщенной информационной системы «поставщик – потребитель информации», где взаимосвязь понятий «по-

ставщик», «потребитель» и «информационная система» укрупненно отражаются схемой на рис. 1.1, произошло только некоторое изменение содержания понятий и смещение акцентов.

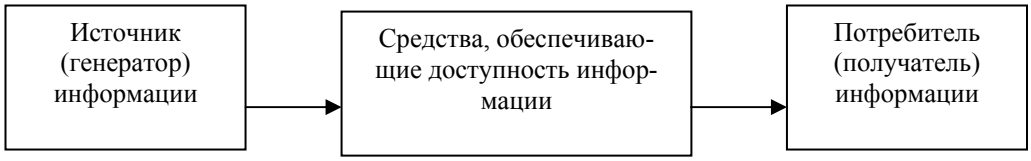


Рис. 1.1. Обобщенная схема взаимосвязи «поставщик – потребитель информации»

Источник, понимаемый на начальной стадии развития информатики (документалистике), как документ - физический носитель информации, обрел обобщенную форму – «генератор информации», что совокупно отражает не только форму представления информации, но и процесс или контекст ее получения. Практически не изменилось содержание блока средств, обеспечивающих доступность информации: это те же информационно-поисковые системы и лингвистическое обеспечение. В то же время «потребитель информации» определяется уже не только тематикой предметной области, но и особенностями восприятия и особенностями когнитивного процесса преобразования полученной информации. Т.е. с точки зрения процесса человеко-машинного взаимодействия не менее важными факторами, чем эффективность организации данных, становится эффективность организации интерфейса, не только адекватно представляющего потенциально полезные возможности системы, но и учитывающего психологические особенности восприятия информации человеком.

1.2. Соотношение понятий «информация», «данные», «знания»

Понятие "информация " достаточно широко используется в обычной жизни современного человека. Значение информации в жизни общества стремительно растет, меняются методы работы с информацией, расширяются сферы применения информационных технологий. Динамизм информатики как науки отражается и в постоянном появлении новых определений и толкований основного понятия информатики - информации.

Исходя из того, что определение научного понятия – это рабочая модель исследуемого объекта, отражающая основные составляющие и связи этого объекта (которые собственно и являются исследуемыми предметами, а в последствии – и в практической деятельности), рассмотрим некоторые определения понятий «информация», «данные», «знания» и их взаимосвязь.

1.2.1. Информация

Наиболее часто термин «информация» употребляется в его исходном значении (от латинского слова informatio) - это сведения, сообщения о каком-либо событии, деятельности и т.д. При этом в других областях знаний вводятся и другие определения этого понятия.

Следует, однако, отметить, что разнообразие определений тем не менее соответствует двум следующим концепциям: 1) *Атрибутивная концепция* рассматривает информацию как фундаментальную естественнонаучную категорию, находящуюся рядом с такими категориями как "вещество" и "энергия", 2) *Функционально–кибернетическя* - как неотъемлемый элемент управляемых или самоуправляемых систем (технических, биологических, социальных), как функцию этих систем. То есть, в зависимости от области знаний, где вводится понятие информации, его определение будет отражать специфику как области, так и задачи исследования. Рассмотрим далее наиболее извесные определения.

В естественных науках информация выступает в качестве меры сложности структур (Моль) и меры разнообразия (Эшби): чем выше упорядоченность (организованность) системы (объекта), тем больше в ней содержится "связанной" информации.

В физике информация определяется, как отрицание энтропии - меры неопределенности, учитывающей вероятность появления тех или иных сообщений (Бриллюэн).

В генетике понятие информации определяется как программа (генетический код) биосинтеза белков, материально представленная полимерными цепочками ДНК.

В теории информации, как науки об оптимальном кодировании сообщений и передачи сигналов по техническим каналам связи, понятие информации определяется как коммуникация, связь, в процессе которой устраняется неопределенность (Шеннон). Каждому сигналу приписывалась априорная вероятность его появления. Чем меньше вероятность появления того или иного сигнала, тем больше информации он несет для потребителя (т.е. чем неожиданнее сигнал, тем больше его информативность). Шеннон же предложил и единицу измерения информации – «бит», а количество информации определяется по формуле $I = - \sum p_i Log\ p_i$ где p_i - вероятность появления i-го сигнала из n возможных. Количество информации равно нулю, когда возможно только одно событие. С ростом числа событий количество информации увеличивается и достигает максимального значения, когда события равновероятны.

Отметим, что такой подход позволил ввести и другое определение: информация - это результат (вероятность) выбора из набора возможных альтернатив (Яглом).

Дальнейшее развитие математического подхода к определению понятия "информация " нашло в работах логиков (Карнап, Бар-Хиллел) и математиков (А.Н. Колмогоров). Здесь понятие информации не связы-

вается ни с формой, ни с содержанием сообщений, передаваемых по каналам связи, и определяется как абстрактная величина, не существующая в физической реальности, также как не существует мнимое число или точка, не имеющая линейных размеров.

Информация в кибернетических системах – это основа функционирования самоуправляемых систем (технических, биологических, социальных), и она рассматривается как обозначение содержания сигнала, полученного системой из окружающего мира в процессе взаимодействия системы с ним (Н. Винер).

Объединяющим (по крайней мере, с философской точки зрения) определением является следующее: «Информация – это отраженное разнообразие» (А.Д. Урсул). Разнообразие и отражение в развивающемся материальном мире неразрывно связаны и взаимно определяют друг друга: чем выше внутреннее разнообразие системы, тем более адекватно отражение ею внешнего мира. А чем больше возможность отражения (восприятия и понимания взаимодействия с окружающей средой), тем больше у системы возможностей адаптироваться – развиваться и увеличивать свое разнообразие.

Информация является одним из основных универсальных свойств материи. То есть, в самом общем смысле информация есть обозначение некоторой формы связей или зависимостей объектов, явлений, мыслительных процессов. Информация - это понятие, абстракция, относящееся к определенному классу закономерностей материального мира и его отражения в человеческом сознании.

Таким образом, можно сделать вывод, что в естественных науках понятие информации отождествляется с сигналами, которые наблюдаются в технических и биологических системах, и могут быть измерены, т.е. представлены как «рабочее тело», которое можно обрабатывать и хранить. Это тело состоит из упорядоченных дискретных или непрерывных сигналов, с которыми и имеет дело информационная технология.

Приведем далее несколько следующих «социально-коммуникационных» и гуманитарных определений понятия "информация":

– информация - сведения, передаваемые одними людьми другим людям устным, письменным или каким-нибудь другим способом [БСЭ1980]

– информация - содержание сообщения или сигнала; сведения, рассматриваемые в процессе их передачи или восприятия, позволяющие расширить знания об интересующем объекте [Терминологический словарь1991].

С правовой точки зрения информация определяется как "некоторая совокупность различных сообщений о событиях, происходящих в правовой системе общества, ее подсистемах и элементах и во внешней по отношению к данным правовым информационным образованиям среде,

об изменениях характеристик информационных образований и внешней среды, или как мера организации социально-экономических, политических, правовых, пространственных и временных факторов объекта. Она устраняет в правовых информационных образованиях, явлениях и процессах неопределенность и обычно связана с новыми, ранее неизвестными нам явлениями и фактами" [Рассолов1998].

В экономике информацию рассматривают как сведения, которые необходимо фиксировать, передавать, хранить и обрабатывать для использования в управлении как хозяйством страны в целом, так и отдельными его объектами. Информация позволяет получить решение, как эффективнее и экономически выгоднее организовать производство товаров и услуг.

Таким образом, можно сказать, что определения понятия информации, представленные в гуманитарных областях, не противоречат приведенным ранее естественнонаучным точкам зрения.

В заключение приведем *научно-методологическое* определение информации, данное Д.И. Блюменау и А.В.Соколовым, важное для понимания предмета информатики: "информация - это продукт научного познания, средство изучения реальной действительности в рамках, допустимых методологией одного из информационных подходов к исследованию объектов различной природы (биологических, технических, социальных). Подход предполагает описание и рассмотрение этих объектов в виде системы, включающей в себя источник, канал и приемник управляющих воздействий, допускающих их содержательную интерпретацию".

1.2.2. Данные

Рассмотрим понятие "данные", которое, например, в [Информатика1999] вводится следующим образом: "Мы живем в материальном мире. Все, что нас окружает, и с чем мы сталкиваемся, относится либо к физическим телам, либо к физическим полям. Все объекты находятся в состоянии непрерывного движения и изменения, которое сопровождается обменом энергией и ее переходом из одной формы в другую. Все виды энергообмена сопровождаются появлением сигналов. При взаимодействии сигналов с физическими телами в последних возникают определенные изменения свойств – это явление называется регистрацией сигналов. Такие изменения можно наблюдать, измерять или фиксировать теми или иными способами - при этом возникают и регистрируются новые сигналы, т.е. образуются данные".

Известны также и другие трактовки, как, например, д**анные -** это:

– факты, цифры, и другие сведения о реальных и абстрактных лицах, предметах, объектах, явлениях и событиях, соответствующих определенной предметной области, представленные в цифровом, символьном, графическом, звуковом и любом другом формате;

– информация, представленная в виде, пригодном для ее передачи и обработки автоматическими или автоматизированными средствами (при возможном участии человека).

1.2.3. Знания

Переходя к рассмотрению роли понятия «информация» в человеко-машинных и социальных системах, необходимо определить понятие "знания".

В [Урсул1976] понятие «знания» определяется следующим образом: «Научное знание – вся совокупность сведений, являющаяся результатом отражения материальной и нематериальной действительности в человеческом сознании».

С другой стороны, как вводится в [Муранивский1982], *научно-техническая информация*- это задокументированное научное знание, введенное в оборот, участвующее в функционировании и развитии общества». То есть, знание, являющееся достоянием чьего-либо сознания и не получившее «толчка» для циркулирования в обществе, не может рассматриваться как информация. Откуда следует, что информация не существует без материального носителя, обеспечивающего ее передачу. Существование информации не зависит от вида носителя и формы представления, однако от этого зависит возможность и эффективность ее использования. Например, информация, представленная в знаковой системе, не знакомой получателю, или на носителе который не может быть доступен, не будет использована.

Основываясь на приведенных трактовках, можно констатировать условность превращения знания в это информацию и информации в знание. Информация выступает как форма знания, отчужденная от его носителя (сознания субъекта), и обобществляющая его для всеобщего использования: информация - это динамическая форма существования знаний, обеспечивающая его распространение и социальное функционирование. Получая информацию, пользователь превращает ее путем интеллектуального усвоения (информационно-когнитивного процесса) в свои новые личностные знания, т.е. происходит воссоздание знаний на основе информации.

Таким образом, фиксируемые/воспринимаемые сигналы (факты) окружающего мира представляют собой объективно существующие данные.

Информация появляется при использовании данных в процессе решения конкретных задач – формирования нового знания субъекта. Результаты решения задач, обобщения в виде законов, теорий, совокупностей взглядов и представлений, выступающие как истинная, проверенная информация, образуют **обобществленные знания**, отчужденные от субъекта их сформировавших и представленные обычно в форме доку-

ментов и сообщений, которые, в свою очередь, могут рассматриваться как объективно существующие данные.

Функциональное соотношений этих понятий иллюстрируется схемой, приведенной на рис. 2.

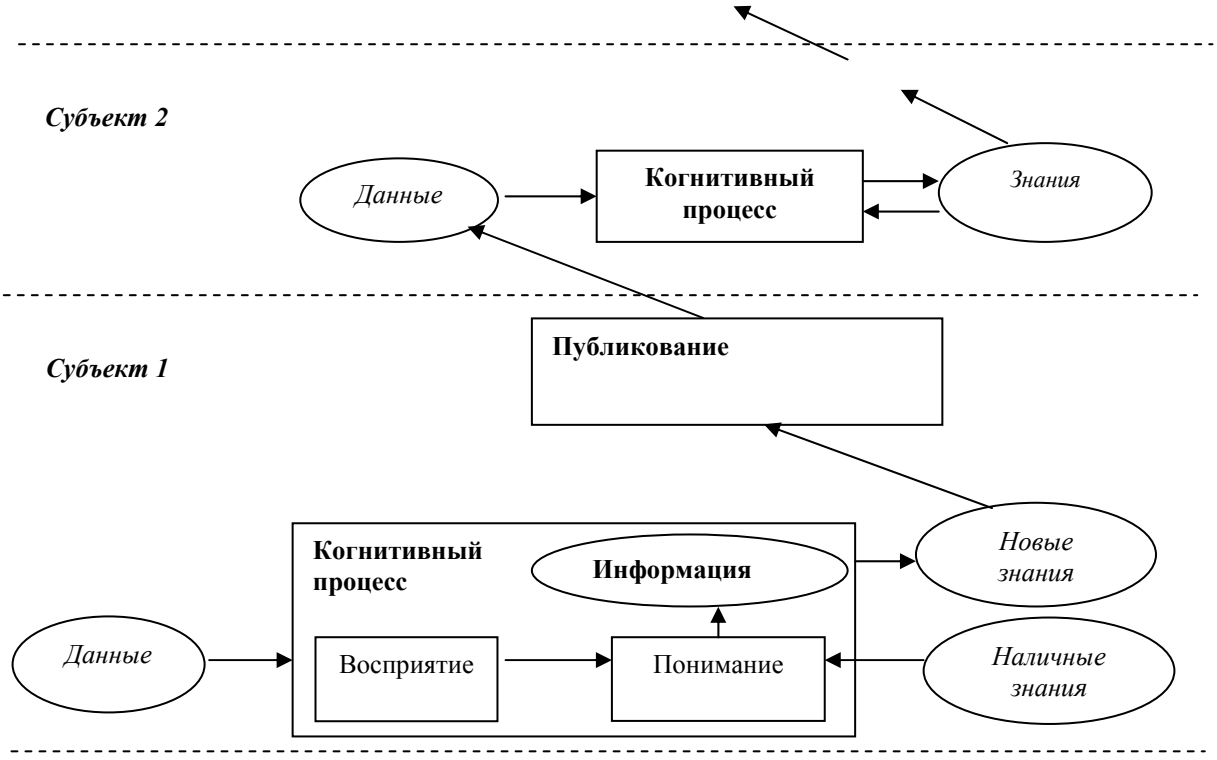


Рис. 2. Соотношение понятий «информация», «данные», «знания»

Станут ли данные информацией, зависит от того, известен ли метод преобразования (отражения) данных в новые или уже известные понятия. То есть, чтобы извлечь информацию из данных необходимо иметь метод получения информации, адекватный форме представления данных[6]. Причем необходимо учитывать тот факт, что информация не является статичным объектом – она динамична и существует только в момент взаимодействия данных и методов. Можно сказать, что все прочее время она пребывает в состоянии «потенциальном» - и представлена как данные.

Кроме того, одни и те же данные могут представлять разную информацию в зависимости от степени адекватности взаимодействующих с ними методов, к которым надо отнести и условия ее извлечения (например, наличного знания субъекта).

Таким образом, в отличие от данных, которые по своей природе являются объективными (так как это результат регистрации объективно существующих сигналов, вызванных изменениями в материальных телах или полях), методы являются субъективными. В основе искусствен-

[6] Данные, составляющие информацию, обычно имеют свойства, однозначно предопределяющие метод получения информации.

ных методов лежат алгоритмы (упорядоченные последовательности команд), составленные и подготовленные субъектами (людьми), а в основе естественных методов лежат биологические свойства субъектов. То есть, информация возникает и существует в момент взаимодействия объективных данных и субъективных методов.

Соответственно, будем называть субъекты (люди, предметы или устройства), от которых может быть получена информация, источниками информации, а субъекты (люди, предметы или устройства), которые получают информацию - приемниками информации.

1.2.4. Научно-техническая информация

В приведенном выше определении научно-технической информации, как органической составной части социальной информации, отражены следующие особенности и требования к ней:

- в качестве научно-технической информации следует рассматривать знания, которые создаются и используются учеными, инженерами, и другими специалистами в их практической полезной деятельности;
- знания превращаются в информацию только при условии их включения в социально-коммуникативные процессы, когда «первообладатель» знания делает его доступным другим людям;
- информация не может существовать и функционировать без средства, обеспечивающего ее включение и циркулирование в социально-коммуникативных процессах, т.е. без какого-либо материального носителя в виде знака, вещественного материала или технического оборудования;
- непременным требованием к форме представления (или выражения) информации (например, к языку и другой знаковой системе выражения информации или ее материальному носителю) является общественный, массовый характер, формы представления информации, ее доступность другим людям;
- только превращение знания в информацию обеспечивает его передачу в пространстве и времени и возможность практического использования, что, в итоге, является одним из важных условий превращения науки в непосредственную производительную силу.

Как объект содержательной обработки понятие «информация» в основном отражается терминами «научная информация» и «научно-техническая информация». В [Муранивский1982] отмечено, что впервые эти понятия были введены как различительные, как попытка дать такое определение информации, создаваемой и используемой в технических и естественных науках, чтобы ее можно было отличить от обыденной, не охватываемой сферой информационной деятельности. Однако, происходящее в последние годы и десятилетия сближение естественных и гуманитарных наук, охват информационными технологиями практически

всех отраслей производства, управления, культуры и даже быта привело к тому, что эти термины утратили изначально точное значение, а термин стал охватывать при этом все более широкий круг различного вида *социальной* информации. Поэтому представляется конструктивным в дальнейшем содержание понятия «информация» связывать (во многом отождествляя) с понятием «научно-техническая информация», как его определил Т.В. Муранивский: «Научно-техническая информация – это научно-технические знания, включенные в социально-коммуникативные процессы в форме, обеспечивающей передачу этих знаний в пространстве и времени, а также их восприятие и использование людьми в процессе научно-технической деятельности или в других сферах общественно-исторической практики». Здесь термин «научно-техническая» употребляется в универсальном значении, включая не только «научный» и «технический», как характеристики области применения (в определении явно указаны и все другие сферы деятельности человека и, кроме того, границы между «научностью» и «ненаучностью», как известно, достаточно условны). Однако с точки зрения методологии применения этого понятия как рабочей модели в рамках задач информационных систем, необходимо отметить принципиальность использования *научного* подхода к анализу объектов и процессов преобразования информации и *технические* аспекты использования средств автоматизации при ее дальнейшей автоматической или автоматизированной обработке. Кроме того, необходимо акцентировать внимание на *осознанности* (организованности) процесса получения *пользы* при обработке полученной информации.

1.3. Свойства информации

Как и всякий объект, информация обладает свойствами. На свойства информации влияют как свойства данных, так и свойства методов, взаимодействующих с данными в ходе информационного процесса. По окончании процесса обработки свойства информации переносятся на свойства новых данных, то есть свойства методов могут переходить в свойства данных.

Спектр свойств информации существенно шире того, которым обладают другие, например, физические объекты. Известно высказывание Б. Шоу: "Если у тебя и меня имеется по одному яблоку, и мы ими обменялись, то у каждого из нас осталось по одному яблоку; если у тебя и меня имеется по одной идее и мы ими обменялись, то у каждого из нас будет по две идеи". Информация специфична и с точки зрения старения (информация устаревает не со временем, а с появлением новой, отрицающей или уточняющей информации).

Информация, как отмечалось ранее, является объективным социальным явлением. Она представляет собой различные знания, сообщения, сведения, данные и т.п., включенные в коммуникативные процессы

и используемые в различных сферах практической деятельности людей. Когда знание при помощи каких-либо материальных средств (языка, различных средств массовой информации и коммуникации и т.п.) превращается в информацию, оно обретает самостоятельное, независимое от создателя и относительно независимое от материального носителя существование. Относительный характер независимости информации от материального носителя объясняется тем, что, внутренняя ее сущность не зависит от носителя, но от выбора носителя нередко зависит эффективность функционировании информации, ее хранение, поиск, восприятие, использование и т.п.

Из объективности информации вытекает объективный (т.е. независимый от создателей и потребителей информации) характер ее свойств и закономерностей движения и развития в науке, технике, производстве и обществе в целом.

Наиболее очевидными свойствами информации являются *прагматические* - те, которые характеризуют степень полезности информации для пользователя в его практике, т.е. проявляются в процессе ее использования. К данной группе свойств относятся не только смысл и новизна, но и полезность (ценность), полнота, достоверность, адекватность, доступность (мера возможности получить ту или иную информацию), актуальность (степень соответствия информации текущему моменту времени), объективность и субъективность.

С другой стороны, свойства информации необходимо рассматривать[7] в их органическом единстве, не только в контексте ее использования в сфере информационной деятельности, но и на других этапах работы и в других областях деятельности. С точки зрения исследования и

---

[7] При выявлении свойств, которые должны исследоваться и могут быть рекомендованы к использованию в информационной деятельности, необходимо руководствоваться следующими правилами:

- выделять такие свойства, которые носят универсальный характер, т.е. присущи любой информации, независимо от ее тематической, принадлежности;
- отбирать для изучения такие свойства, которые при соответствующем их обозначении облегчили бы регулирование движения и развития информации, а также разработать символику и методы обозначения определенных свойств информации с целью последующего управления этими свойствами и их практического использования;
- выявлять структуру связей между этими свойствами, т.е. построить единую систему универсальных объективных свойств информации;
- выбираемая для изучения и использования система универсальных свойств информации должна не исключать, а предполагать использование в практике информационной деятельности конкретных, специфических свойств, присущих различным отдельным видам и единицам информации, а также применяемых в настоящее время методов обработки и систем классификации документов;
- предлагаемая символика и методы обозначения и использования общих свойств информации должны предполагать потенциальную возможность алгоритмизации и автоматизации процессов обработки и поиска информации.

создания эффективных методов и средств обработки информации эти *атрибутивные* свойства делятся [Муранивский1982] на две группы:

- свойства, определяющие объективные закономерности, связанные с информацией и преимущественно в пределах отдельной предметной области науки, техники, производства (условно эти свойства можно назвать "внутренними");
- свойства, определяющие закономерности движения информации в межотраслевом масштабе ("внешние" свойства).

К первой группе относятся: кумулятивность, эмерджентность, неассоциативность и старение информации. Ко второй – стохастичность, межотраслевой характер, информирование по ассоциации.

1.3.1. Кумулятивность информации

Применительно к научной информации это свойство обычно связывают с ее непрерывным накоплением, которое резко усилилось за последние 250-300 лет, приняв экспоненциальную форму.

Необходимо, однако учитывать, что росту публикаций присуща большая избыточность информации: кроме новых сведений, они зачастую содержат старые, дублирующиеся и т.п. Поэтому следует различать накопление информации и накопление источников информации.

Если простое накопление количества информационных документов и объемов информации рассматривать как первую ступень кумулятивности информации, то вторая, более высокая ступень ее кумулятивности непосредственно связана с преемственностью в развитии науки, техники, производства и других сфер полезной деятельности человека, которая проявляется, по крайней мере, в следующих направлениях: историческом, отраслевом, межотраслевом и международном.

Историческая преемственность хорошо иллюстрируется словами Исаака Ньютона. На вопрос, как ему удалось сделать великие открытия, он ответил: "Я видел дальше других, потому что стоял на плечах гигантов". Но преемственность не означает простое восприятие всего прошлого, что накоплено в процессе исторического развития науки, а подразумевает его критическое освоение и переработку.

Преемственность информации проявляется как на отраслевом, так и на межотраслевом уровне, что тесно связано с интеграцией в науке, производстве, а также с усилением связей между наукой и производством, и ведет к синтезу нового знания и нового информации.

Отраслевая и межотраслевая преемственность информации отражает преемственность в развитии системы "наука – техника - производство". С информационной точки зрения эта преемственность заключается в том, что объемы информации, циркулирующей в сфере науки, превышают те объемы, которые используются в технике, а информация в сфере техники больше ее объемов, используемых в производстве. Пре-

емственность в развитии системы "наука-техника-производство" отражает также и закономерность опережающего характера развития науки перед техникой и техники перед производством.

Преемственность информации на международном уровне непосредственно связана с общечеловеческим, интернациональным характером науки, производства и общественного труда в целом. Это ее свойство проявляется, в частности, в том, что достижения науки, техники и производства в короткий срок становятся достоянием других стран и народов.

1.3.2. Концентрация информации

Процесс концентрации информации (иногда этот процесс называют свертыванием, обобщением, агрегированием и т.п.) отражает свойство, которое также проявляется как закономерность развития научного и других видов человеческого знания (формами концентрации информации являются, например, законы и категории науки, пословицы, поговорки и т.п.). Эта закономерность проявляется в том, что по мере накопления (в отрасли, группе отраслей науки или техники и т.д.) определенных единиц информации они имеют тенденцию к объединению в более информационно-емкие формы, высшей из который является эмерджентная[8] система.

Концентрация информации проявляется, по крайней мере, в трех формах: документационной, фактографической и теоретико-концептуальной [Муранивский1982].

К первой (документационной) форме относятся те виды концентрации информации, большинство из которых применяется в библиотечно-библиографической и документационной практике, 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 информационные признаки документа (автор, заголовок, выходные данные, краткая характеристика содержания в виде аннотации и т.п.). Более емким видом концентрации информации является реферирование, которое, как известно, предполагает извлечение из документа и представление в виде реферата определенных единиц информации.

Близким к этому виду концентрации информации является индексирование, в результате которого на специальном информационном языке (как правило, формализованном) отражаются с большей или меньшей степенью глубины и полноты тематические признаки содержания документов. Примыкают к перечисленным видам концентрации информации также некоторые виды систематизации документов (например, по тематической принадлежности или по индексам информацион-

[8] Эмерджентность рассматривается ниже как одно из свойств информации.

ного языка, по авторам и т.п.), которые являются завершающим этапом библиографической обработки документов. Все эти виды концентрации информации предполагают неизбежные потери информации в следствие отображения только основных аспектов содержания документов и, соответственно абстрагирования от деталей.

Вторая форма - это информационно-фактографическая концентрация информации, где в качестве основной единицы обработки и концентрированного представления выступает не документ, а 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) накопленных на предыдущем этапе единиц информации, которое обеспечивает создание более емких единиц информации и переход научного знания к более высоким уровням абстракции. Такие формы обычно реализуются в ходе исследовательской творческой деятельности человека.

Важность концентрации информации в процессе развития человеческого знания обобщается словами Н.И. Моисеева [Моисеев1977]: «Любое человеческое знание начинается с накопления фактов, с помощью наблюдения или направленного эксперимента. Но не превращенная в систему, река новых знаний не утолит жажду. Пока хаос новых фактов не структуризован, пока человек не может окинуть взглядом явления в целом, он не может эти знания использовать для практики. Поэтому второй этап - это переработка информации, представление ее в такой форме, которая уже может быть "переварена" человеком. Ну а третий этап - это возвращение к практике, использование знаний для тех целей, ради которых они были созданы».

Обращаясь к примерам технологий обработки данных, отметим, что документы базы данных, рассматриваемые не как отдельные единицы знаний, изолированные друг от друга, а как семантические компоненты, уже позволяют на основе их совместного использования генерировать принципиально новую информацию, что получило практическое воплощение в автоматизированных технологиях «извлечения знаний» (Data mining). Таким образом, после накопления больших объемов первичных данных («сырой» информации) производится анализ информации по различным «срезам», позволяя информационным центрам работать по замкнутому циклу, синтезируя новую информацию.

1.3.3. Эмерджентность и неассоциативность информации

Термин «эмерджентность» (от англ. emerge - появляться, выясняться, возникать) заимствован из теории систем. Эмерджентными называют такие свойства сложных систем, которые порождаются взаимодействием элементов и не наблюдаются ни в одном из элементов, если они рассматриваются отдельно. В этом смысле система больше суммы своих частей.

Эмерджентность информации проявляется в том, что производные единицы-континуумы обладают свойствами, не присущими ни одной из формирующих их условно-дискретных исходных единиц информации.

Эмерджентность информации проявляется во всех формах и видах кумуляции: чем выше кумуляция информации, тем выше ее эмерджентность. Например, совокупность научных фактов при их систематизации позволяет обнаружить значительно больше важных свойств информации, чем, скажем, совокупность документов. Наиболее высокой степени эмерджентности информация достигает на третьем - теоретико-концептуальном этапе ее концентрации.

Важным аспектом эмерджентности любой системы является то, что по мере увеличения количества ее элементов, используемых для контроля и регулирования внутрисистемных процессов, для управления системой на определенном этапе ее развития требуются качественно новые методы и средства. Например, при ограниченном количестве документов для поиска нужной информации можно было применить метод простого перебора и просмотра этих документов. Однако, по мере роста количества документов простым перебором документов ничего не достигнешь. Нужны вспомогательные средства: каталоги, картотеки, указатели и т.п.

Эмерджентность - это конкретное проявление перехода количества в качество. Этот принцип постоянно проявляется в практике управления.. Например, по мере расширения функций и задач, решаемых в организации или на предприятии, руководитель вынужден отказываться от контроля всех типов технологического процесса и следить лишь за конечными результатами, содержание которых также постоянно меняется.

С эмерджентностью информации тесно связано свойство *неассоциативности*. В самом общем виде это означает, что любая единица информации - это не просто арифметическая сумма (или произведение) составляющих ее элементов, причем эти элементы нельзя без искажения смысла располагать в любой последовательности или группировать в разные сочетания,.и с точки зрения различных вариантов относительно материального носителя или знаковой формы, и с точки зрения воздействия (например, результат воздействия на потребителя сообщений А, В, С, Д ... неравнозначен результату воздействия тех же сообщений на того же потребителя, если они поступают в иной последовательности или другом сочетании.

1.3.4. Старение информации

Часто понятие "старение информации" связывают с понятием ценности. Например, устаревшей часто считается информация, которая уже не представляет ценности для потребителя. При этом старение рассматривается с точки зрения новизны, актуальности и связывается с констатацией утраты практической полезности для достижения поставленной цели.

Такое представление о старении информации имеет определенный смысл: свойствами, противоположными ее старению являются ее новизна, актуальность и практическая полезность для решения научных, технических и иных задач. Понять, устарела информация или нет, можно только в сравнении с другой - новой, актуальной, полезной.

Н. Винер писал, что основной причиной старения информации является не само время, а появление новой информации [Винер1958].

Вопрос о старении информации быдет не полным, если не отметить, что понятие "объект информации" как и сама информация, носит многоуровневый характер: объектом информации может быть как непосредственный объект или процесс действительности, так и определенный уровень модели, которая выступает в качестве исходного элемента для образования моделей более высоких уровней.

Например, совокупность научных фактов и связей между ними может выступать в качестве внешнего объекта для образования информации в виде научной концепции, теории и т.п. В результате агрегирования в более емкие информационные единицы исходные единицы информации или, по крайней мере, их часть оказываются ненужными, они утрачивают новизну и ценность, т.е. относительно стареют. Наконец, при рассмотрении проблемы старения информации не следует игнорировать связь между собственно информацией и ее потребителем, которая зачастую в практике выступает как единственный наблюдаемый вид связи.

Старение информации может носить как абсолютный, так и относительный характер. Абсолютно устаревшей считается информация, которая с появлением новой информации оказалась недостоверной. Относительный характер старения информации можно рассматривать с точки зрения ее новизны не только по временным параметрам, но и по отношению к совокупному или индивидуальному знанию. Если, например, в физике открыта новая частица, то информация об этом будет новой и для физики, и для любого, кто об этом прочитает или услышит. В то же время сообщение об известных ранее частицах в физике (по отношению к совокупному знанию представителей этой науки) будет считаться относительно устаревшей, а по отношению к отдельному индивидуальному знанию конкретного человека может быть новой.

1.3.5. Межотраслевые свойства и рассеяние информации

Проблему межотраслевого характера информации различного уровня следует рассматривать, по крайней мере, в двух аспектах:

- во-первых, из одних и тех же единиц информации могут «конструироваться» различные по характеру и назначению гипотезы, решения и т.п., как из одних и тех же химических элементов - различные вещества или из одних и тех же слов - различные фразы;

- во-вторых, информация, созданная для решения какой-либо одной научной или практической задачи, может быть использована для решения других, в том числе тех, которые создатель информации предвидел.

Практически любая единица информации обладает потенциальной возможностью превращения в межотраслевую, т.е. может оказаться полезной за пределами той сферы, в которой она создавалась. По некоторым данным до 50% всей научной информации, получаемой современным исследователем, извлекается из смежных областей знания, причем именно в ней он нуждается наиболее остро.

Межотраслевой характер информации проявляется в различных формах, среди которых особый интерес представляет информация, получаемая по ассоциации. Из жизненного опыта известно, что связи между различными объектами и процессами реальной действительности могут иметь такой характер, когда восприятие или воспоминание об одном объекте или явлении влечет за собой представление о других объектах или явлениях. Физиологической основой этого процесса служат временные, но достаточно прочные связи, которые образуются в коре головного мозга человека под влиянием объективных связей, существующих между различными объектами и процессами. "Временная нервная связь, - отмечал И.П. Павлов, - есть универсальное физиологическое явление в животном мире и в нас самих. И вместе с тем оно же и психическое - то, что психологи называют ассоциацией, будет ли это образование соединений из всевозможных действий, впечатлений или из букв, слов или мыслей".

Психологи различают два рода ассоциаций: простые и сложные. Простые – это ассоциации по смежности, по сходству, по контрасту.

Ассоциации по смежности являются отражением связей предметов и явлений в пространстве и времени (день – ночь, работа - отдых).

Ассоциации по сходству возникают в результате обобщения (генерализации условной связи). Они выступают тогда, когда представления одних предметов или явлений вызывают в сознании человека представления других предметов и явлений, сходных с первыми (внешне или по какому-либо существенному признаку, например, квадрат - прямоугольник).

Ассоциации по контрасту характеризуются тем, что представление одного предмета или явления может вызвать представление другого предмета или явления, обладающего по сравнению с первым противоположными свойствами. Примеры ассоциаций по контрасту: тяжелый - легкий, прозрачный - непрозрачный и т.д.

Кроме простых ассоциаций (по смежности, сходству и контрасту) существуют также сложные - по смыслу. Главными из смысловых, составляющих основу наших знаний, являются следующие ассоциации вид - род, часть - целое и причина - следствие.

Таким образом, единицы информации обладают свойством вызывать по ассоциации в сознании воспринимающих их индивидуумов новые единицы информации, хотя результат восприятия (в зависимости от интересов, образования, рода занятий, психического склада ума и многих других факторов) может быть различными у разных людей. Это относится не только к словам, терминам, понятиям, но также к единицам информации более высоких уровней.

Рассеяние информации - это свойство информации, вытекающее из ее межотраслевого характера и противоположно концентрации информации. Как и концентрация, рассеяние информации - сложный, диалектически противоречивый процесс: по мере концентрации информации в более емкие информационные единицы возрастает ее межотраслевое значение и, следовательно, усиливается тенденция к ее рассеянию.

В строгом смысле рассеяние информации означает, что информация, которая была бы полезной для решения данной проблемы, может оказаться в документах, относящихся к совершенно другой предметной области.

Другим проявлением свойства рассеяния может быть рассредоточение информации по документам, относящимся к разным предметным областям, т.е. дублирование информации. 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 $1 : n : n^2 \ldots$».

Закон Брэдфорда не содержит объяснения причин рассеяния информации. Он отражает лишь то, что уменьшение количества научных публикаций по мере удаления от профильного ядра есть следствие ослабления связей, которые существуют между профильной для данной проблемы областью и другими областями. Рассеяние собственно ин-

формации основывается не на ослаблении, а на усилении межотраслевых связей в условиях научно-технического прогресса. Поэтому при изучении характера рассеяния информации следует рассматривать не столько рассредоточение профильных для какой-либо отрасли науки или техники публикаций, сколько возможность появления потенциально полезной информации для решения той или иной задачи в других, даже весьма отдаленных отраслях. Причем, такая информация может существовать как в явной, так и в латентной форме.

1.4. Информационные единицы

Связывание понятий информация и знание в информатике означает, что при поиске единицы информации, следует идти по пути поиска единицы знания, представленной в виде информационного сообщения. То есть, единица научно-технической информации может быть определена следующим образом [Муранивский1982]: «… в качестве единицы информации может рассматриваться любое (сколь угодно широкое или узкое) информационное сообщение или знание, которое имеет самостоятельное значение и выражено в доступной для восприятия и практического использования форме». Например, единицами информации являются сообщения о научных фактах, описания экспериментов, их методов или результатов, определения, закономерности, суждения, умозаключения, гипотезы, теории, законы, а также чертежи или описания конструкций, узлов, приборов, машин, технологических процессов и т.п. Другими словами, единица информации — это единица знания, выраженная в форме письменных или каких-либо других документов или их фрагментов.

В качестве единицы научной информации может выступать и краткое или пространное сообщение, и отдельный научный факт, и научная теория, и совокупность фактов, гипотез, концепций, теорий и законов, образующих основу какой—либо отрасли науки. Самостоятельными единицами технической информации в такой отрасли, как машиностроение, могут быть описания конструкции детали, узла, машины или устройства, системы машин и т.п. Какую единицу выбирать, зависит от конкретных условий: от отрасли знания, интересов потребителя информации и других факторов, В одних случаях достаточно сообщить об отдельных результатах экспериментального исследования, а в других потребуется обоснование тенденций развития целой отрасли или группы отраслей.

Предлагаемое толкование единицы информации основывается на методологическом принципе относительности элементарного.

Из истории развития науки можно привести немало примеров, когда на место дискретного элемента, выступающего в качестве объекта научного исследования, со временем приходит непрерывная единица. Например, вплоть до XVII века математика ограничивалась изучением

постоянных величин и фиксированных зависимостей между ними. Когда запросы астрономии и механики выдвинули проблему математического отображения процессов и движения, она стала исследовать переменные величины. А когда их изменяемость была распространена до бесконечно малого и бесконечно большого, возникло дифференциальное и интегральное исчисление.

В современной науке сложная структура элементарных образований различных материальных уровней, своеобразие законов их функционирования и развития приводят к уточнению общего понятия элемента. Простоту *элемента* следует понимать не абсолютно, а относительно изучаемой целостной системы. Множественности форм и уровней организации в природе соответствует множественность элементарных образований, каждое из которых является элементарным лишь по отношению к системам более высокого уровня (например, в живой природе – это клетка и клеточный организм, организм и популяция, вид популяции и биоценоз и т.д.).

В семиотике, как языковой системе, осуществляющей функции обмена информацией, рассматривается несколько структурно-взаимосвязанных уровней информации: уровень букв и элементарных знаков; уровень слов, чисел, классификационных шифров; уровень высказываний (показателей и других единиц языка, которые имеют форму наименований некоторых величин и их значений); уровень документов. Каждый из этих уровней, соответственно, может иметь свою единицу информации.

Любой процесс (событие, действие) существует не сам по себе, а непременно во взаимосвязи с другими процессами, причем связи эти многолики: причина и следствие, прошлое - настоящее - будущее, укрупнение или дробление и т.п. Закономерности взаимосвязей и непрерывного изменения объектов и процессов полностью распространяются и на научно-техническую информацию: одни и те же единицы информации могут быть включены в различные процессы и способны выполнять различные функции, непрерывно адаптируясь к новым условиям существования.

Следовательно, в качестве единицы информации в информационной теории целесообразно рассматривать *единицу знания*, которая имеет относительный характер и может менять свои функции в зависимости от конкретных условий ее практического использования.

Из относительного характера единицы информации следует ее дискретно-непрерывный характер. Это свойство проявляется, например, в текстах: отдельные слова одновременно образуют предложения, из которых составляются отрывки текста и т.п., причем в каждом случае относительно элементарные единицы одновременно образуют более сложные.

Точно такие же трудности наблюдаются при изучении единиц информации, являющихся составной частью более широких систем (например, фактические данные в теоретической концепции или гипотезе). Кроме того, одни и те же факты не привязаны жестко к одному и тому же материальному носителю могут входить во многие теоретические обобщения и играть в каждом из них различную роль, проявляя новые свойства (как, например, слова в различных контекстах).

Таким образом, дискретно-непрерывный характер информации заключается в том, что элементарные единицы информации одновременно могут быть составными частями более широких систем, которые в свою очередь могут также образовывать новые единицы, по отношению к которым они сами имеют условно-дискретную, элементарную форму, причем эти единицы информации в различных системах могут играть разную роль. Наконец, в ряде случаев бывает трудно (а иногда нецелесообразно или даже невозможно) четко вычленить элементарные единицы информации из сложных образований более высоких уровней.

Принятие в качестве единицы информации единицы знания является принципиальным, однако трудно измеримым. Во многих случаях единица информации совпадает с содержанием документа (например, описание объекта в авторском свидетельстве или патенте, учебник по какому-либо предмету и т.п.).

Например, в библиотековедении и документоведении в качестве единицы информации была предложена натуральная единица (НИТ) - число сообщений, документов, слов, букв, символов. В математической теории информации в качестве единицы измерения используется бит - информация, содержащаяся в одном двоичном разряде, определяемая как количество информации, получаемой при выборе одного из двух равновероятных состояний.

С другой стороны, содержание информации, как уже отмечалось, не зависит ни от языка, ни от вида носителя: оно может быть выражено на различных языках и представлено в виде самых разнообразных документов. Другими словами информация инварианта по отношению к ее носителю. Поэтому отношение документа, как физической единицы, и единицы информации, содержащейся в нем, — это отношение формы и содержания, явления и сущности. Соответственно, существуют и разрабатываются методы и средства извлечения информации из данных, например, средства аналитической обработки баз данных, методы автоматического реферирования и индексирования и т.д. Отметим, что работа этих автоматизированных методов аналогична тому, что делает человек: путем всестороннего изучения явления (текста документа) проникает в сущность, выявляя и извлекая из текста нужную информацию.

Контрольные вопросы

1. Охарактеризуйте соотношение понятий «информация», «данные», «знания».
2. Дайте определение понятия «информация».
3. Охарактеризуйте прагматические свойства информации.
4. Перечислите атрибутивные свойства информации.
5. Назовите и охарактеризуйте формы концентрации информации.
6. Приведите примеры проявления свойства старения информации.
7. Охарактеризуйте свойство рассеяния информации.
8. Дайте определение единицы информации.

2. Информационные системы и технологии обработки информации

Широко используемое понятие «информационные системы», тем не менее, практически не имеет единого концептуального определения. Наиболее часто это понятие трактуется согласно определению, данному в [Криницкий1982], как «…комплекс, состоящий из информационного фонда и процедур: управляющей, обновления, информационного поиска и завершающей обработки, - позволяющего накапливать, хранить, корректировать и выдавать информацию».

Процедуры обработки информации могут быть автоматизированными или нет. В настоящее время наиболее эффективным является вариант, основанный на использовании вычислительных машин. С другой стороны, практически любая целенаправленная деятельность человека в той или иной мере связана с обработкой информации. Таким образом, автоматизированная информационная система (АИС), как специализированная подсистема эффективной обработки данных, является обязательной составной частью практически любой системы управления, будь то управление производством, научные исследования или проектные работы, экономические или социальные системы. И, как следует из существа самого понятия «информация», именно сфера управления является главной областью применения АИС.

Приведенное выше определение информационной системы «вытекает» и связано с устоявшейся и уже привычной, но, тем не менее, особой формой целенаправленной деятельности человека - обработкой информации, как сведений о чем-то, материально представленных на традиционных или машиночитаемых носителях, и обеспечивающих эффективность решения задач его основной деятельности. Причем понятие «системности» здесь отражает существо функциональности: состав и структура ИС определяется, исходя из требований к уровню *эффективности обслуживания информационных потребностей* конечных пользователей, прежде всего в части нахождения в накопленных массивах тех записей (документов), которые содержат сведения, нужные для организации эффективного управления процессами в сфере основной деятельности.

2.1. Состав и структура информационной системы

Для определения состава и взаимосвязей компонентов системы приведем предварительно определения следующих основных понятий.

Система (от греч. systema – целое, составленное из частей соединение) – совокупность элементов, взаимодействующих друг с другом и образующих определенную целостность.

Элемент системы – часть системы, имеющая определенное функциональное назначение. Сложные элементы систем, в свою очередь состоящие из более простых взаимосвязанных элементов, часто называют *подсистемами*[9].

Организация системы – внутренняя упорядоченность, согласованность взаимодействия элементов системы (проявляющаяся, в частности, в ограничении разнообразия состояний элементов системы)

Структура системы – состав, порядок и принципы взаимодействия элементов системы, определяющие основные свойства системы.

Архитектура системы – совокупность свойств системы, существенных для организации взаимодействия ее составляющих.

Целостность системы – принципиальная несводимость свойств системы к сумме свойств отдельных ее элементов (эмерджентность) и, в то же время, зависимость свойств каждого элемента от его места и функции внутри системы.

С точки зрения формы существования системы выделяют абстрактные и материальные системы.

Абстрактные системы – это системы, которые имеют в качестве операционных объектов преимущественно идеализированные, например, знания, теории, гипотезы.

Материальные системы подразделяются на технические, эргатические и эргатехнические (смешанного типа). Именно эргатехнические системы - материальные системы «человек-машина», состоящие из эргатического элемента (человека-оператора) и технического элемента (машины), будут составлять основной предмет изучения.

Информационная система (ИС) – материальная система, организующая, хранящая и преобразующая информацию. Это система, основным предметом и продуктом труда в которой является информация.

Автоматизированная **система обработки данных (СОД) -** комплекс взаимосвязанных методов и средств преобразования данных, необходимых пользователю. Здесь данные понимаются как информация, представленная в формализованном виде, пригодном для автоматической обработки при возможном участии человека.

---

[9] Существенным фактором с точки зрения абстрактности (идеальности природы) определения понятия «система» является то, что отдельный элемент какой-либо системы (как и сама система) может также быть элементом другой системы.

Системы обработки знаний (СОЗ) - автоматизированные СОД, имеющие специальное программное обеспечение для анализа семантики информации и ее гибкой логической структуризации.

В основу построения эффективных автоматизированных СОД (как систем наиболее эффективного типа) положены следующие принципы:

– **принцип интеграции** – обрабатываемые данные, однажды введенные в систему, многократно используются для решения возможно большего числа задач, чем максимально устраняется дублирование данных и операций их преобразования;

– **принцип системности** – обработка данных в различных «разрезах» с целью получения информации, необходимой для принятия решений на всех уровнях и во всех функциональных подсистемах управления;

– **принцип комплексности** – механизация и автоматизация процедур преобразования данных на всех стадиях технологического процесса.

2.1.1. Классификация информационных систем

Понятие системы тесно связано с эффективностью обработки информации, что, в свою очередь, базируется на специализированности компонентов и процессов. Такая специализированность (ориентация на определенный класс задач) обусловлена обычно свойствами обрабатываемых объектов. Информация, как основной объект обработки, может быть классифицирована следующим образом:

– по характеру, определяемому свойствами объектов предметной области – четкие или лингвистические[10] переменные;

– по принадлежности к подсистеме управления - информация по целевой подсистеме, по сопровождению системы, по связи с внешней средой, по обеспечивающей, управляемой и управляющей подсистемам;

– по форме представления – числовая или вербальная (словесная);

– по изменчивости во времени - условно-постоянная и условно-переменная;

– по способу передачи (используемому носителю) - письменная или электронная, спутниковая или телефонная и т.д.;

– по режиму передачи - в нерегламентные сроки, по запросу или принудительно;

– по назначению - экономическая, техническая, социальная, организационная и т.д.;

---

[10] Лингвистическими переменными называют величины, которые могут иметь несколько значений. Впервые это понятие было введено в лингвистике для обозначения свойств слов иметь несколько смыслов в зависимости от контекста, например, *ключ* для дверного замка и *ключ* – водный источник.

– по стадиям жизненного цикла объекта – этап научно-исследовательской или опытно-конструкторской работы, организационно-технологическая подготовка производства, эксплуатация и т.д.;

– по отношению объекта управления к субъекту - между системой и внешней средой, по вертикали и горизонтали, между управляющей и управляемой подсистемой.

С точки зрения применения информационные системы могут классифицироваться:

1) по характеру использования результатной информации:

– информационно-поисковые (сбор, хранение, выдача информации по запросу пользователя);

– информационно-советующие (системы поддержки принятия решений);

– информационно-управляющие (результатная информация непосредственно участвует в формировании управляющих воздействий);

2) по функциональному назначению:

– производственные ИС;

– коммерческие ИС;

– маркетинговые ИС;

– финансовые ИС и т.д.

3) по объектам управления

– ИС автоматизированного проектирования;

– ИС управления технологическими процессами;

– ИС управления предприятием и т. д.

4) по степени автоматизации процессов обработки

– ИС с ручной обработкой информации;

– ИС механизированной обработки информации;

– ИС автоматизированной обработки информации;

– ИС автоматической обработки информации.

Другим аспектом типологии ИС является «информационный», что позволяет выделить следующие классы, определяющиеся, по существу, технологией создания и использования информации, основанной на концепции **баз данных**, свойства которых и переносятся на информационные системы.

По *форме представляемой информации* можно выделить фактографические, документальные, мультимедийные, что в той или иной степени соответствует цифровой, символьной и другим формам представления информации в вычислительной среде.

По *типу хранимой (исключая мультимедийную) информации* можно выделить фактографические, документальные, лексикографические БД. Лексикографические базы – это классификаторы, кодификаторы, словари основ слов, тезаурусы, рубрикаторы и т.д., которые обычно ис-

пользуются в качестве справочных совместно с документальными или фактографическими БД. Документальные базы подразделяются по уровню представления информации – полнотекстовые (обрабатывающие так называемые «первичные» документы) и библиографическо-реферативные (обрабатывающие «вторичные» документы, отражающие на адресном и содержательном уровне первичный документ).

По *типу используемой модели данных* выделяют традиционно три класса БД: иерархические, сетевые, реляционные. Развитие технологий обработки данных привело к появлению постреляционных, объектно-ориентированных, многомерных БД, которые в той или иной степени соответствуют трем упомянутым классическим моделям.

По *топологии хранения* данных различают локальные и распределенные ИС.

По *типологии доступа* и *характеру использования* хранимой информации ИС могут быть разделены на специализированные и интегрированные[11].

По *функциональному назначению* (характеру решаемых с помощью ИС задач и, соответственно, характеру использования данных) можно выделить операционные и справочно-информационные. К последним можно отнести ретроспективные ИС (электронные каталоги библиотек, БД статистической информации и т.д.), которые используются для информационной поддержки основной деятельности и не предполагают внесение изменений в уже существующие записи, например, по результатам этой деятельности. Операционные ИС предназначены для управления различными технологическими процессами. В этом случае данные не только извлекаются из БД, но и изменяются (в том числе добавляются), в том числе в результате этого использования.

По *сфере возможного применения* можно различать универсальные и специализированные (или проблемно-ориентированные) системы.

По *степени доступности* можно выделить общедоступные и ИС с ограниченным доступом пользователей. В последнем случае говорят об управляемом доступе, индивидуально определяющем не только набор доступных данных, но и характер операций, которые доступны пользователю.

Следует отметить, что представленная классификация не является полной и исчерпывающей. Она в большей степени отражает исторически сложившееся состояние дел в сфере деятельности, связанной с разработкой и применением баз данных.

---

[11] В последнем случае правильнее говорить об интегрированных информационных системах, объединяющих в общей среде разнородные данные, хранимые возможно в разнотипных базах, но используемые для решения одной прикладной задачи.

2.1.2. Основные компоненты ИС[12]

Как было сказано, ИС – это составная часть некоторой большей системы, обеспечивающая достижение какой-либо реальной цели в деятельности человека. Это уже предполагает, что всякая информационная система имеет некоторую материальную основу – носитель информации. Такими носителями, составляющими *физический компонент* системы (технические средства), являются как среда внешней памяти, так и технические и вычислительные средства, обеспечивающие непосредственно обработку и взаимодействие пользователя с АИС.

Второй компонент – это процедуры, обеспечивающие функционирование системы (*программные средства*), и в первую очередь – подсистемы управления данными, а также процедуры специализированной обработки, отражающие требования предметной области.

Однако существо информационной системы в наибольшей степени выражается третьим компонентом - *информационным фондом*, который характеризуется не только содержащейся информацией, но и способом ее организации (*модель данных*) и формой представления. Последнее в свою очередь определяется языком представления и управления информацией (*лингвистическое обеспечение*).

Примерная организационно-функциональная классификация АИС приведена на рис. 2.1.

---

[12] Практически все современные ИС включают в свой состав вычислительные машины и поэтому являются **информационно-вычислительными системами (ИВС).** Обычно функциями ИВС, управляющей крупным предприятием, являются следующие: вычислительная, коммуникационная, запоминающая, следящая, регулирующая, оптимизационная, прогнозирующая, анализирующая, контролирующая, документирующая.

Рис. 2.1. Организационно-функциональный состав ИС

Функциональные подсистемы ИС реализуют и поддерживают модели, методы и алгоритмы получения управляющей информации. Состав функциональных подсистем зависит от предметной области использования ИС. Их назначение достаточно очевидно. Отметим только, что подсистема *научно-технической подготовки* отвечает за выполнение научно-исследовательских (в том числе маркетинговых) работ, конструкторскую и технологическую подготовку производства.

Состав **обеспечивающих систем** достаточно стабилен и по большей части мало зависит от предметной области использования ИС. Отметим следующие компоненты:

– *Программное обеспечение* – совокупность программ регулярного применения, необходимых для решения функциональных задач, и программ, позволяющих наиболее эффективно использовать вычислительную технику, обеспечивая пользователям наибольшие удобства в работе.

– *Математическое обеспечение* – совокупность методов, моделей и алгоритмов обработки информации, используемых в системе.

– *Лингвистическое обеспечение* – совокупность языковых средств. Обычно включает комплекс языков, отражающий многоуровневость представления информации в АИС, и в том числе язык запросов и отче-

тов, обеспечивающий удобство работы конечному пользователю, специальные языки определения и управления данными, обеспечивающие адекватность внутреннего представления и согласование внутреннего и внешнего представлений. В наибольшей степени определяется особенностями предметной области.

Организационные подсистемы также относятся к обеспечивающим подсистемам, но направлены в первую очередь на обеспечение эффективной работы персонала и поэтому могут быть выделены отдельно.

Отметим, что разработка ИС должна начинаться именно с создания организационного обеспечения: экономического обоснования целесообразности системы, состава экономических показателей, определяющих ее деятельность, состава функциональных подсистем, организационной структуры управления, технологических схем преобразования информации, порядка проведения работ и т.д.

2.2. Информационные компоненты в системах управления

Управление – это процесс обработки информации, направленный на достижение определенной цели. С другой стороны, управление – это функция системы, обеспечивающая либо сохранение ее основных свойств, либо ее развитие в заданном направлении. Управление осуществляется для достижения определенной цели, вполне конкретной для каждого отдельного объекта управления и связанной с состояниями объекта и среды, в которой он находится. Обобщенная схема информационных потоков и место ИС в процессе управления представлены на рис. 2.2.

Для исследования характера взаимодействия управляемого процесса и информационной системы через определение структуры и характера информационных потоков между ними будем рассматривать эти элементы с позиции системного подхода, принимая, что «…система представляет собой отражение материального образования с точки зрения единства его поведения и строения, обусловленность поведения этого целого определяется спецификой внутреннего строения, спецификой его элементов и особенностями взаимодействия между ними. Т.е., система это такое строение, которое осуществляет преобразование причинных воздействий из окружающей среды и изнутри системы в соответствующие изменения объекта как целого» [Смирнов1978]. Здесь понятие «поведение» отражает связь изменений в окружающей среде и/или самой системе с внешними или внутренними причинными воздействиями, вызвавшими эти изменения, а «строение» системы, как противоположное свойство, определяется единством множества элементов и структуры, осуществляющей их интеграцию в целостное образование.

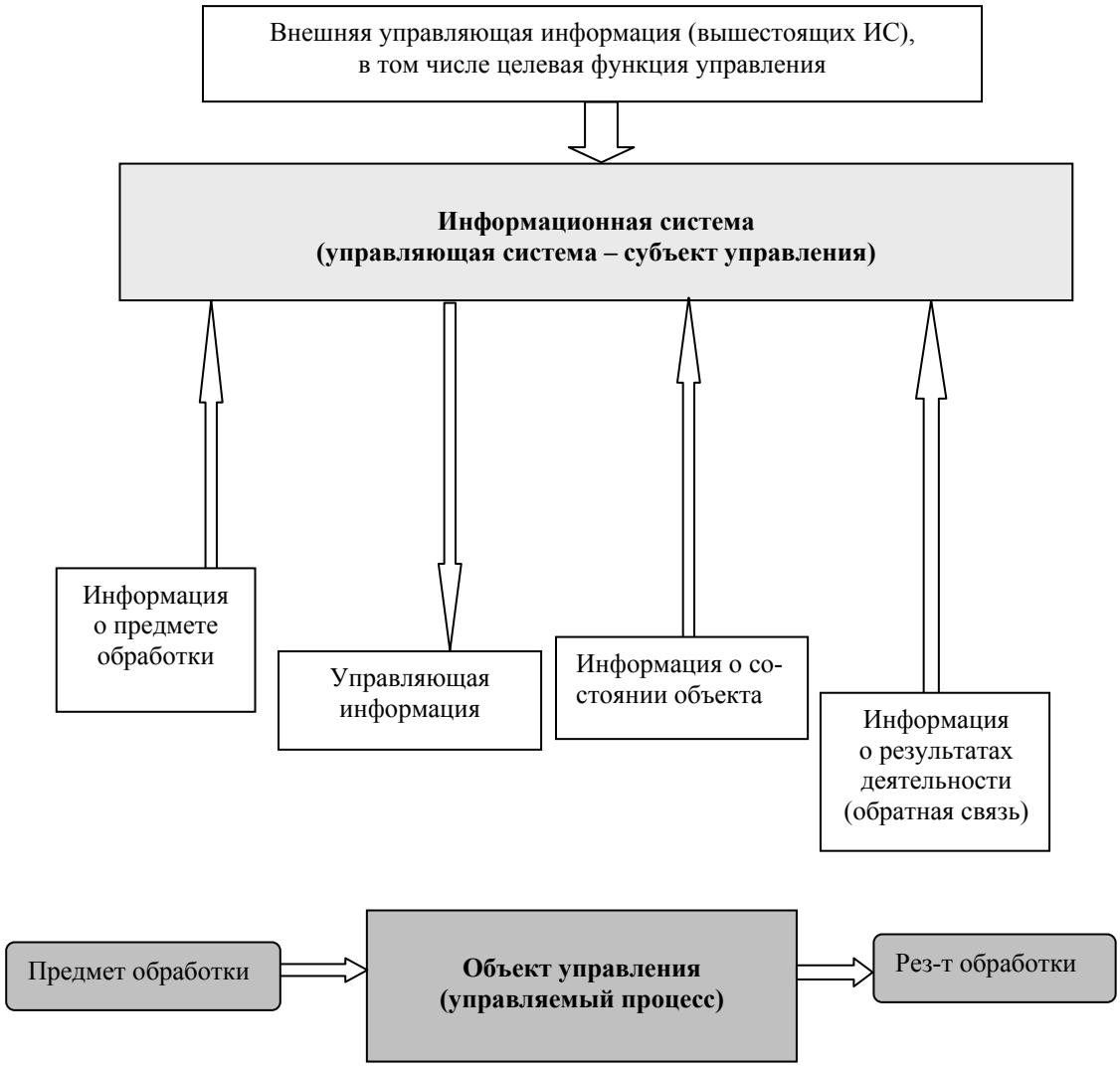


Рис. 2.2. Место ИС в процессе управления

2.2.1. Информационная модель управления в системах материальных преобразований

На основании результатов анализа генезиса механизма управления, представленных в [Дружинин1976, Криницкий1982, Поспелов1975], обобщенная информационная модель механизма управления, как средства, реализующего «поведение» системы в сфере материальных преобразований, может быть представлена в виде, приведенном на рис. 2.3.

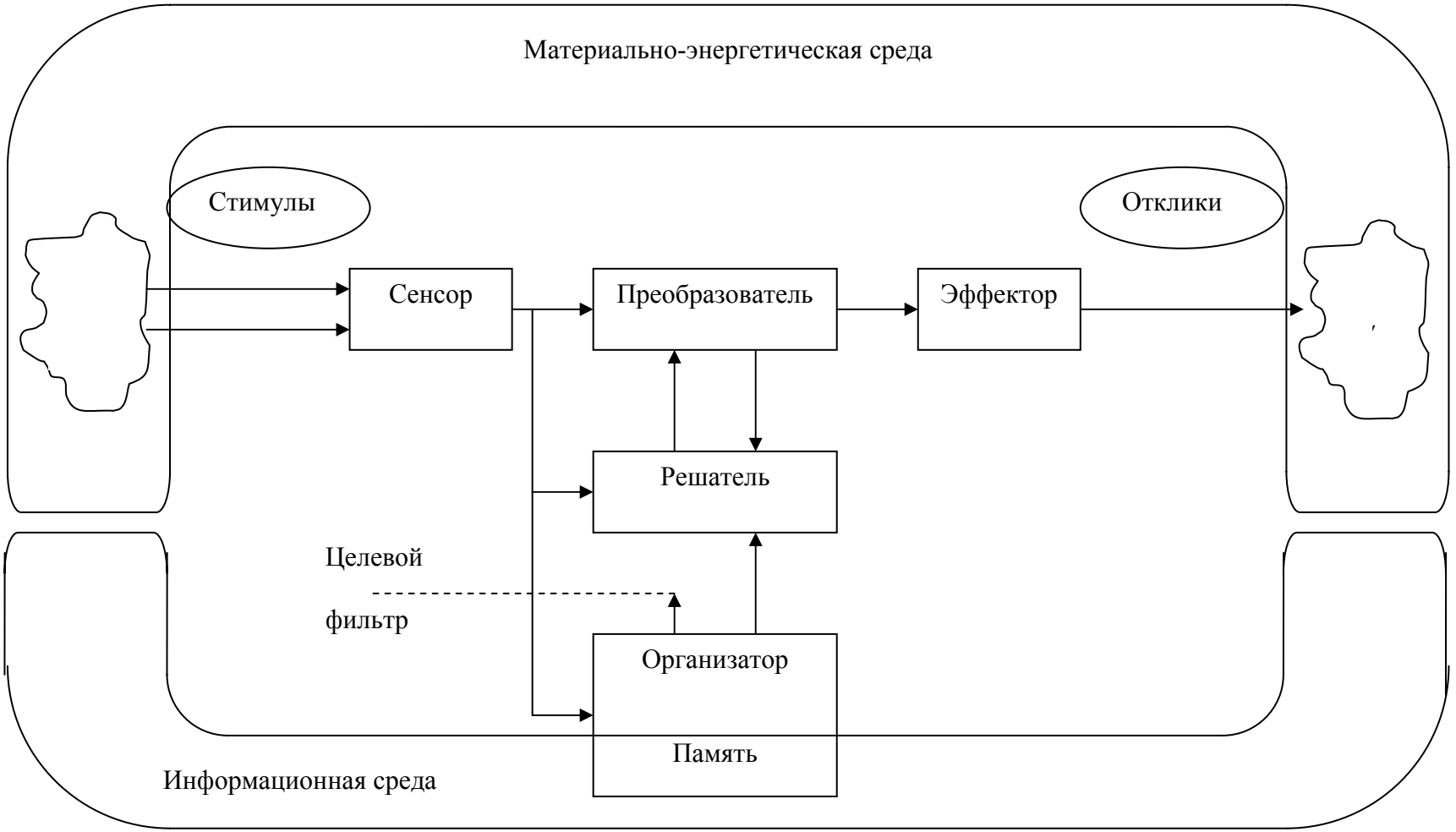


Рис. 2.3. Информация в системах управления

Здесь система - это часть реального мира (среды), реализующая целенаправленное взаимодействие с объектами ближней среды, и не совпадающая с ней. Результатом взаимодействия является изменение ближней среды.

К *ближней среде* относится та часть среды, которая существенно воздействует на систему и/или подвергается существенному воздействию со стороны системы. Средой могут быть и другие системы, в том числе и более высокого уровня. С этой точки зрения вход трактуется как причина, а выход - как следствие взаимодействия.

Состоянием системы будем называть множество существенных свойств, которыми обладает система, а *структурой* – ту их часть, которая остается в системе неизменной при изменении ее состояния.

В системе выделяются входной, преобразующий и выходной компоненты. Для систем, реализующих целенаправленное поведение, преобразующий компонент включает (или зависит от) компонента управления, который обеспечивает посредством управляющих сигналов изменение свойств преобразователя, т.е. характер преобразования объектов среды.

Входной компонент - это обобщенное средство восприятия среды, будь то активное ее воздействие на систему или, например, активное измерение системой параметров, характеризующих состояние среды.

Аналогично, *выходной компонент* - это обобщенное средство воздействия на среду или изменения соотношения с ней (в том числе, например, изменение самой системы по отношению к среде).

Преобразователь осуществляет изменение объекта, поступающего через входной компонент. Таким объектом может быть элемент среды, либо изменяемый компонент самой системы. Можно сказать, что первый случай соответствует системам, целью которых является преобразование среды или сохранение себя в условиях изменяющейся среды, а второй – саморазвивающимся системам, конечной целью которых является построение новой системы или новых взаимосвязей со средой.

Решатель осуществляет по фиксированным правилам (например, на уровне соответствий) выбор преобразования (например, через установку значений соответствующих параметров). Т.е., решатель осуществляет *регулирование* преобразования на уровне оперативного управления, которое реализуется в виде отдельной реакции на отдельное текущее воздействие. В том случае, когда решатель реализует принцип «обратной связи», т.е. вырабатывает сигналы управления на основе измерения параметров изменяемых объектов (как результата взаимодействия с системой), система может реализовать задачу «самосохранения».

Системы, имеющие первый контур управления, могут быть отнесены к саморегулирующимся и представляют собой класс детерминированных систем с управлением, не предвосхищающим будущие состояния, т.е. реагирующими только на *случившиеся входные воздействия*, а не на их будущие значения. Для рассматриваемых реальных систем ма-

териальность преобразования предопределяет единственность процесса с предопределенным разнообразием состояний. А в том случае, если для процесса выполняется некоторый принцип «самосохранения», управление может рассматриваться как задача оптимального управления (например, как максимизация эффекта при минимуме затрат).

Такого рода замкнутость и предопределенность условий предполагает наличие субъекта управления, имеющего целевые установки и выбирающего или определяющего пределы функционирования не только «решателя-преобразователя», но и всей системы в целом. Компонент, обеспечивающий выполнение таких функций, представлен в схеме *организатором*, собственно и осуществляющим выбор цели, а также определение методов и принципов ее достижения. Для этого (второго) контура управления характерны сигналы, представленные преимущественно в информационной форме. Кроме того, поскольку задачи прогнозирования поведения, обеспечивающего достижение цели, базируются на знании причинно-следственных связей, организатор должен обладать свойством *памяти* для хранения и выборочного использования данных о состояниях среды и «опыта» системы.

Рассматриваемая обобщенная модель объекта деятельности как управляемой системы, взаимодействующей со средой на уровне обмена веществом и энергией, уже включает в явной форме информационные компоненты, что собственно и обеспечивает саму возможность управления. Причем для систем человеко-машинных и социальных выделяют [Урсул1973] две формы информации[13]: материальную и идеальную. Под материальной информацией здесь понимается отражение разнообразия вне сознания человека и независимо от него. Отражение же в сознании человека есть идеальная информация, которая в свою очередь может материализоваться в виде разнообразия управляющих воздействий или знания, фиксирующего это разнообразие в документальной форме. То есть, информация изначально не создается системой, а только отображается. Однако при наличии второго контура возникает необходимость (и целесообразность) создания новой информации, отражающей сам процесс управления, например, в виде результатов, принципов и методов. Особенностью такой информации является *языковая форма* ее представления, что и обеспечивает возможности коммуникации в науке и обществе.

Выбор целей, принципов и методов решения представляет собой реализацию стратегии управления как способа распределения ресурсов в течение всего периода взаимодействия со средой и по существу является «…процессом отбора из всех возможностей системы одной или нескольких близких возможностей, которые и реализуются в действительности или, по крайней мере, осознаются как оптимальные» [Разумовский1983]. Здесь «оптимальность» представляется как своеобразный

[13] Здесь информация - это отраженное (системой) разнообразие (среды), влияющее на поведение или изменение состояния системы.

принцип сохранения, указывающий, что никакой результат не может быть получен «из ничего». Кроме того, свойственное человеко-машинным системам научное мышление «..не ограничивается фиксацией эмпирических фактов, а также частных и специфических законов. Оно с необходимостью, в ходе развития теорий, выходит на уровень обобщающих специфических и неспецифических научных принципов, отражающих в своем содержании объективно общее и всеобщее... Принципы науки суть фиксация опыта, и доказуемы они только опытом. Поэтому сфера их действия, а также сфера действия построенных на их основе теорий имеет те же границы, что и научная эмпирия. Применение тех или иных принципов вне сферы этого опыта, в частности обобщение специфических научных принципов до уровня неспецифических, происходит на основе предположения и последующего обнаружения общности и единства формы новых классов явлений с углублением в сущность на базе новой эмпирии. Такая экстраполяция носит гипотетический характер, но опирается на принцип материального единства мира, единства пространственно-временных отношений. Конкретные способы такого обобщения многообразны» [Разумовский1983]. Таким образом, функция оснований любой теории – это, прежде всего, организация наличного знания с целью удобного получения следствий. Но, даже имея некоторый минимум твердо установленных принципов, трудно априори утверждать, что к нему нельзя еще добавить какой-нибудь новый принцип.

Схема, представленная на рис. 2.3, отражает наблюдаемое в природе и человеческой деятельности стремление так распределять задачи, чтобы была обеспечена физическая реализуемость решения и минимизирована полная работа. Это выражается в принципе выделения блоков, ориентируемых на реализацию одной функции, с однородными операционными (входными) объектами или, по крайней мере, имеющими одну форму существования[14]. Для блоков компонента управления этот принцип реализуется в стремлении минимизировать количество управляющих сигналов и количество отслеживаемых параметров и показателей (т.е., хорошо структурированная система – это «*один вход – один выход – один управляющий параметр*»). Это позволяет минимизировать потери точности (адекватности), которые привносят многокритериальные методы.

[14] В случае систем обработки однородных объектов большой размерности аналогичные рассуждения применимы скорее не к самой системе, а к процессу ее создания через приведение операционных объектов строительства к размерности, соответствующей операционным возможностям субъекта.

2.2.2. Информация в системах обработки и генерации знания

Рассмотрим обобщенную информационную схему деятельности в научной сфере (рис. 2.4), где обрабатываемые (преобразуемые) объекты (в частности и сама среда) имеют преимущественно абстрактную природу (идеальное, а не материальное происхождение). Имея ту же информационно-функциональную структуру, такие системы принципиально отличаются по характеру взаимоотношения с окружающей средой.

Рассмотренные выше системы имеют преимущественно цель самосохранения, вырабатывая отклик в ответ на изменение ближней среды. Цель вторых – преобразование среды или самой системы (самосовершенствование) для изменения (к лучшему) условий взаимодействия со средой. В этом смысле термин «научный поиск» наилучшим образом отражает активный характер системы (по отношению к среде). Для систем такого типа стимул – это неопределенность как свойство, характеризующее, но еще не выражающее проблемную ситуацию [Дорожкин1995]. Состояние неопределенности (взаимодействие внутреннего знания и внешнего по отношению к системе стимула) – симптом возможного противоречия, которое, если будет выявлено в процессе развития (исследования), формализуется в виде проблемы – состоявшегося, признанного противоречия, которое, тем не менее, еще не имеет решения и/или может иметь неоднозначные решения. Причем, решение может быть найдено рациональным способом (сведением к решаемым задачам – через уменьшение размерности), так и иррациональным, т.е. выходом за пределы существующего знания. В первом случае проблемной ситуации может быть сопоставлен ряд средств, обеспечивающих решение, причем для задач исходными данными являются как сама обрабатываемая (входная) информация, так и методы ее обработки. Именно здесь появляется возможность множественности решения проблемы. В зависимости от цели решение, например, может находиться путем элиминирования факторов и сведения к типовым случаям. То есть, в контексте понятия «цель», «задача имеет актуально достижимую цель, а проблема – потенциально достижимую» [Дорожкин1995].

Таким образом, решение проблемы (в том числе и постановка) – это отыскание форм и условий взаимодействия нового элемента знания (стимула, породившего неопределенность) с существующим знанием.

Для процесса постановки–решения задачи (как ситуации определенного выбора) характерно наличие следующих четырех компонентов:

- субъект, осуществляющий выбор;
- среда решения – принятые в качестве рабочих понятия, гипотезы, законы, парадигмы и т.п.;
- доступные средства решения и практических действий;
- возможные результаты.

Проблема выбора (как частного случая нахождения) решения задачи тесно связана с методологической проблемой выбора *критерия оценки* альтернатив, где существуют различные точки зрения и основания, такие, как простота, обоснование правдоподобности, синтаксические, семантические и другие критерии.

Существенной особенностью ситуации выбора является то, что «выбираемые основания - это еще и средство организации наличного знания субъекта с целью *удобного* получения следствий» [Разумовский1983], при этом, сам процесс включает следующие взаимосвязанные и взаимообуславливающие этапы:

– определение потенциально или актуально достижимой цели (путем систематизации фактов и гипотез);

– определение основных законов и принципов (как решения, так и применения решения);

– постановку задачи – определение известных и неизвестных параметров, определение методов решения и критериев оценки решения.

Причем взаимосвязь принципов, целей и задач при их выборе обуславливается требованием «конструктивности» (разрешимости задач и достижимости цели), что допускает их вариантность и комбинативность.

Компоненты взаимодействия со средой обеспечивают как выделение (в том числе активными методами) тех объектов предметной области, которые в совокупности своей или взаимосвязи порождают неопределенность – стимул (проблемную ситуацию), так и обратную связь (управление) – оценку полученного решения как результата включения (вписывания) вновь полученного элемента среды в существующую структуру.

Так же, как и в рассмотренном выше случае модели управления (рис. 2.3), здесь в компоненте преобразования присутствуют три уровня:

1) генерация актуального результата как реализация предопределенного метода преобразования текущего входа (фиксированного набора значимых параметров в пространстве возможных состояний);

2) управление реализацией преобразования (и обеспечение его эффективности) в соответствии с предопределенными принципами и критериями – оперативное управление процессом достижения цели;

3) выбор целей (порождение гипотез), а также формирование оснований (теорий, принципов, законов) и критериев оценки применимости и эффективности преобразования – целеполагание и стратегическое управление.

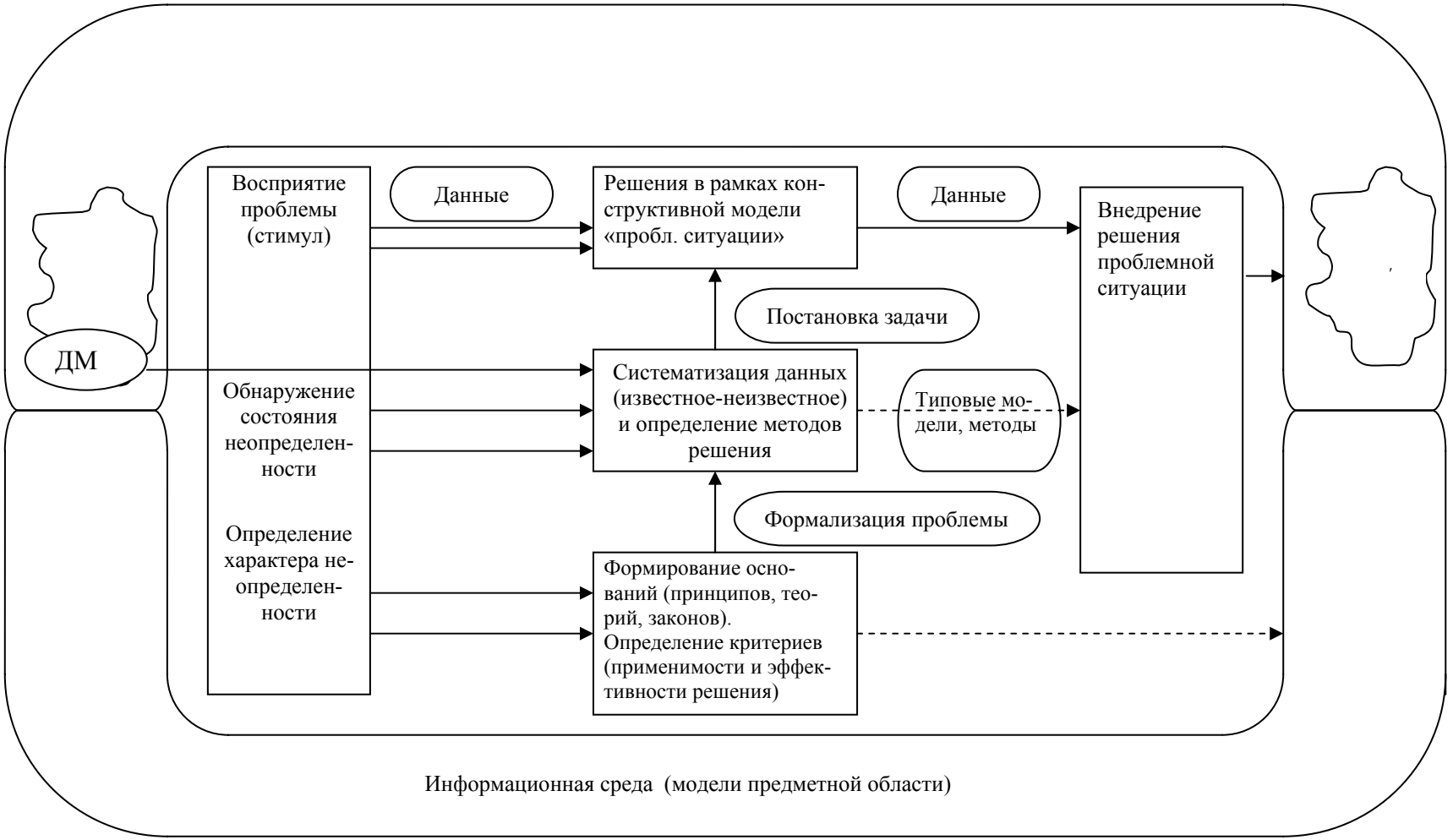


Рис. 2.4. Информация в системе воспроизводства знания

В общем случае, необходимость и достаточность двух контуров управления объясняется и с позиции принципа наименьшего действия, выражающемся в стремления к однородности объектов или преобразований:

– первый контур (оперативного управления) замыкает преобразование, изолируя его от изменяющегося во времени характера воздействия среды;

– второй контур (стратегического управления) – через выбор базиса управляет взаимодействием со средой (преобразуя состояния среды в сигналы), определяет степень и границы применимости выбранного решения, а также анализирует актуальные или потенциальные противоречия внедрения вновь получаемого объекта в структуру ранее существовавших.

С точки зрения пространственно-временной динамики второй контур обеспечивает построение предполагаемой траектории (плана), а первый контур реализует этот план для каждой дискретной точки пространства состояний.

И, поскольку план - это предварительно принятые, но еще не реализованные решения, достижение предполагаемой цели обязательно характеризуется неопределенностью, которая снимается либо реализацией решения (возможно ресурсоемкой), либо предварительной оценкой степени достижимости, получаемой моделированием или аналитическим путем. Снижение ресурсоемкости в этих случаях достигается за счет использования *информационных* ресурсов – теоретических или эмпирических знаний о значимости того или иного свойства или знаний об их устойчивой взаимозависимости в виде законов или закономерностей, обусловленных, например, природой объекта. Более того, если нет таких информационных ресурсов или среда не обладает свойством «памяти», то система может только *реагировать* на сигналы: ни «обучаться», ни «планировать» такая система не может.

Для рассматриваемого случая научного поиска новое знание (результат исследований) воплощается обычно в форме сообщения - документа, реализующего преобразование смысла в текст. Такая «материализация» «идеальных» знаний обеспечивает унифицированную форму отчуждения и за счет сравнительно низкой стоимости их тиражирования существенно расширяет сферу потенциальных потребителей. С другой стороны, как неизбежное противоречие, низкая стоимость публикации (по сравнению со стоимостью получения самого результата) приводит к колоссальному и все возрастающему объему публикаций, а унифицированность способа представления обуславливает внешнюю безликость сообщений. Кроме того, чтобы опубликованное сообщение стало *стимулом,* сенсор должен произвести и семантические преобразования: сообщение должно быть не только воспринято (выделено среди других),

но и понято (выделен смысл) и вписано в систему наличного знания (потенциально применено).

Соответственно, для обеспечения эффективности «узнавания» - первой фазы использования, сообщения должны иметь «сигнальные» признаки. Такие признаки могут быть сформированы, например, по схеме «род – видовое отличие», т. е. путем введения явной систематизации, что является достаточно естественным – научные знания всегда системны, т.к. создаются в рамках некоторой системы понятий соответствующей отрасли знаний.

Создаваемые сообщения (документы) сами в свою очередь образуют среду как воспроизводимую особую информационную ее часть, но как объекты они не могут быть отнесены к типу «активных» стимулов, что также обуславливает необходимость построения активных сенсоров – информационно-поисковых систем. Характерной особенностью этой среды является цикличность воспроизводства: объекты, ее образующие, являются результатами целенаправленной деятельности, и цель их создания – использование их самих для получения новых результатов (см. рис. 2.5). Это предопределяет «естественность» существования информационной деятельности [Попов1996], имеющей целью такое преобразование информационной среды, которое бы обеспечило ее полное и эффективное[15] использование в процессе воспроизводства. Причем, следует отметить, что такими преобразованиями могут быть либо непосредственное упорядочение сообщений, либо их «виртуальное» упорядочение – создание дополнительных (справочных) информационных сообщений, обеспечивающих альтернативный прямой вход во множество сообщений, ассоциируемых с решаемой проблемой.

С точки зрения управления информационные системы, как и ранее рассмотренные, также должны иметь двухконтурное управление.

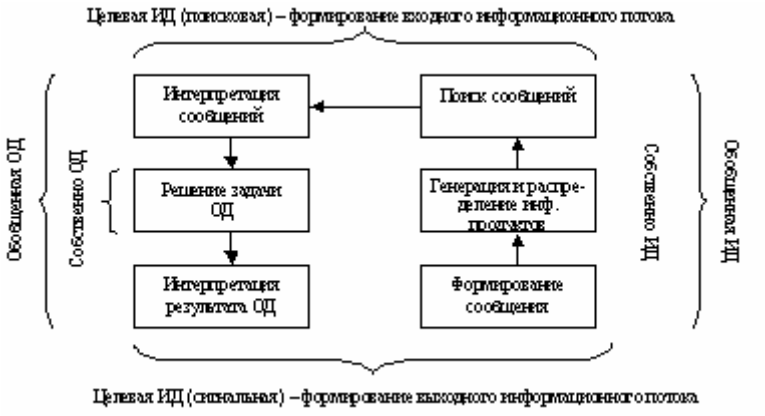


Рис.2.5. Основные элементы (фазы) деятельности в системе воспроизводства знаний (информации)

[15] Эффективность имеет двоякую природу: 1) такое использование информации чтобы были минимизированы затраты на основную деятельность или максимизирован результат; 2) такая организация самой информационной среды, чтобы была обеспечена сохранность и доступность любого объекта.

2.2.3. Характер информационных составляющих в системах управления

Как было показано ранее, основой любого управления является наличие информационной составляющей. В цепи управления, основной задачей которого является выполнение действий, организационно предшествующих самому преобразованию, именно информация выступает в качестве операбельного эквивалента преобразуемого объекта. Возможность абстрагироваться от фиксированного момента времени (извлечь объект из реального временного пространства) позволяет построить ряд сопоставимых объектов (образов, одинаковых с точки зрения управления), каждый из которых соответствует отдельному состоянию реального объекта в фиксированный момент времени, образуя тем самым траекторию во времени, что и предопределяет возможность прогнозирования будущих состояний уже реального объекта.

Здесь, с точки зрения как теории систем, так и этапности реализации решения (действие – оценка результата действия, используемая для корректировки повторно выполняемого действия) можно выделить следующие типы информационных компонентов (сообщений):

– фактографическую информацию, фиксирующую ситуацию (состояние взаимодействия системы со средой) как результат преобразования;

– систематическую информацию, определяющую прагматические условия (принципы, законы, критерии) применения преобразования и/или построения системы.

Причем, именно вторая составляющая отражает способность и возможность качественного развития системы (на основе изменения базовых системообразующих принципов, а не только экстенсивного роста за счет свойства комбинативности компонентов). И именно введение систематической составляющей обеспечивает снижение размерности задачи идентификации входных объектов и выделения ближней среды, позволяя абстрагироваться от конкретики ситуационной составляющей (фиксируя условия выделения самой системы) и представления ее в виде отдельной *структурной* информационной компоненты.

Для систем управления материальными объектами такая структурная составляющая представляется, например, опытом или сводами принципов и правил; для систем преобразования объектов в идеализированной форме (например, научные исследования) – это законы природы и научные принципы; для систем управления информационными потоками (информационной деятельности) – это метаинформация: тезаурусы, классификаторы и т.д.

Таким образом, информационная составляющая управления (например, сигнал обратной связи) каждого следующего уровня может

быть представлена как отражение фактической составляющей информационного компонента на системную.

Выделение ближней среды среди всех объектов по существу есть представление выделенного подмножества как системы – единого образования, целостного прежде всего по отношению к преобразующему компоненту. Т.е., с субъектно-объектной точки зрения, среда как объект преобразования - это система взаимосвязанных объектов, преобразующим субъектом которой в общем случае является совокупность взаимодействующих подсистем, в результате функционирования которых изменяются свойства объектов среды или отношения между ними.

Учитывая принцип наименьшего действия, определяющего требование однофакторности преобразования (и, соответственно, декомпозиции преобразующей системы на функциональные блоки с одним входом, одним выходом и одним управлением) можно говорить о подобии структур взаимодействующих систем – «среды» и «преобразователя», то есть, о существовании общего системообразующего основания системы, определяемой в [Урманцев1978] как

$\mathbf{S_i = \{M_i, A_i, R_i, Z_i\}}$, где $\mathbf{M_i}$ – множество объектов; $\mathbf{A_i}$ – множество системообразующих признаков объектов, относимых к системе; $\mathbf{R_i}$ – множество отношений; $\mathbf{Z_i}$ – системообразующий закон композиции системы.

Для объекта преобразования $\mathbf{S_i}$ – это, соответственно, множество характерных признаков, идентифицируемых преобразователем ($\mathbf{A_i}$); отношения связывающие объекты ($\mathbf{R_i}$); законы композиции системы, определяющие условия и границы возможности преобразования ($\mathbf{Z_i}$).

Для субъекта преобразования $\mathbf{S_i}$ – это, соответственно, множество базовых функций ($\mathbf{A_i}$); множество каналов и сигналов связи подсистем ($\mathbf{R_i}$); средства управления, обеспечивающие адаптируемость и развитие системы ($\mathbf{Z_i}$). Здесь набор базовых функций определяется свойствами объекта и характером преобразования, взаимосвязь подсистем – причинно-следственными взаимоотношениями объекта, а управляемость – целями этого преобразования, как этапа жизненного цикла субъекта.

На практике введение такого представления позволяет реализовать управление преобразованием: установление значений параметров преобразования до начала выполнения самого преобразования, т.е., определение поведения системы на основе сигналов, отражающих (с точностью, определяемой системообразующим основанием) состояние преобразуемой среды.

В этом контексте *информация* - это образ: результат отражения ближней среды (множество объектов и связей, удовлетворяющих требованиям системообразующего основания) как на входе, так и на выходе. Причем, физически образ реализуется как множество сигналов (сообщений), носителем которых являются объекты, не обязательно принадлежащие ближней среде.

Аналогично, управляющая информация - это модель (образ) преобразования, близкая к тем действиям, которые должны быть реализованы эффектором. Т.е., поведение системы определяется из условия, как будто система уже находится в целевом состоянии, которое реально может быть и не существует[16].

Таким образом, информационная модель любой системы, ориентированной на взаимодействие (со средой), необходимо включает компоненты двух типов[17] - множество сообщений, отражающих актуальное (текущее) состояния ближней среды (текущие значения параметров), и сообщения о системообразующем основании, выбранном для построения этой модели. Соответственно, управление реализуется через изменение значений параметров (регулирование, эволюционное развитие) или адаптивно через изменение компонентов системообразующего основания (революционное развитие, в том числе через изменение целей).

Информационные компоненты в свою очередь расширяют (вновь образуют) среду, что предполагает наличие носителя – «памяти», локальной - как свойства самой системы, или глобальной - как свойства среды, включая в том числе и компоненты системы. При этом носителями информации могут быть как искусственные специально создаваемые среды (например, бумажные или машиночитаемые носители), так и объекты естественной природы (в этом случае информация о предшествующих состояниях объекта может представляться, например, собственной *упорядоченной* структурой).

В предложенном изложении понятие информации рассматривается как объект или продукт процесса целенаправленного преобразования системой компонентов среды (или преобразование самой системы, как саморазвитие), имеющих как материальную, так и идеальную (информационную) природу.

В то же время, в целеустремленных системах (к которым относятся человеко-машинные информационные системы) проявляется двойственность природы информации. Здесь информация проявляется как движущая сила: именно она является инициирующей, организующей и управляющей составляющей процесса преобразования. В этом контексте информация - это собственно преобразование некоторого объекта (имеющего информационную природу, например, сообщения, передающего некоторый смысл) среды с целью изменения его состояния (смысла). Поскольку объекты такого рода имеют в основном «идеальную» природу, их состояние фиксируется другими объектами, имеющими материальную природу, например, документами.

Именно с этой точки зрения информация рассматривается «как преобразование, которому следует подвергнуть одно сообщение, чтобы полу-

---

[16] Информация является необходимым компонентом любого целевого преобразования, поскольку только в этом виде может быть обеспечено сохранение системообразующего основания преобразования (цели воздействия на среду), замещая собой уже не существующую в прежнем виде после выполнения преобразования среду.

[17] Аналогичная дихотомия типов информации вводится в [Криницкий1982], где информация подразделяется на осведомляющую и управляющую; [Абдеев1994] – оперативная и структурная; [Урсул1073] – кибернетическая (управляющая) и некибернетическая (структурная); [Бриллюэн1966] - свободная и связанная.

чить другое сообщение той же (смысловой) ассоциации» [Мазур1974]. Информация – это отражение в рамках принятого системообразующего основания процесса познания или, иначе, - *процесс преобразования состояния* наличного знания. Причем, исходное и конечное состояние (сообщения), фиксируемое традиционно в виде документов, в свою очередь является отражением основной информации в рамках познавательной парадигмы. Это представление является моделью передачи информации – *информирования*, реализуемого, например, в виде процесса обучения, научного исследования, а также и информационного обеспечения.

Реализация информационного обеспечения посредством информационных систем - это по существу есть преобразование информации, содержащейся в источнике (оригинал – отражение изменения наличного знания), в информацию приемника (образ – восприятие изменения состояния знания) через цепь промежуточных сообщений – *ассоциации сообщений*, связанных некоторыми информационными преобразованиями.

Используя основные понятия, предложенные в [Мазур1974], обобщенную структурно-функциональную модель ИС как системы, взаимодействующей с множеством сообщений внешней среды (цепочка *источник информации - информационная система – потребитель информации*), можно представить в виде схемы рис. 2.6.

Здесь информация $\mathbf{I_{X12}}$ цепи оригинала - ассоциация связанных по смыслу сообщений $\mathbf{\{X_1, X_2\}}$, преобразуется в продольных цепях в промежуточную информацию $\mathbf{I_{Y12}}$ и воспринимается приемником как результирующая информация (образ оригинала) $\mathbf{I_{Z12}}$ – ассоциация сообщений $\mathbf{\{Z_1, Z_2\}}$, в общем случае образующая, возможно, другой смысл. Соответственно, информирование – это преобразование информации, содержащейся в цепи оригиналов, в информацию цепи образов. Преобразование понимается как процесс, в результате которого одно сообщение превращается в другое, а ассоциация сообщений – это сообщения, связанные одним преобразованием.

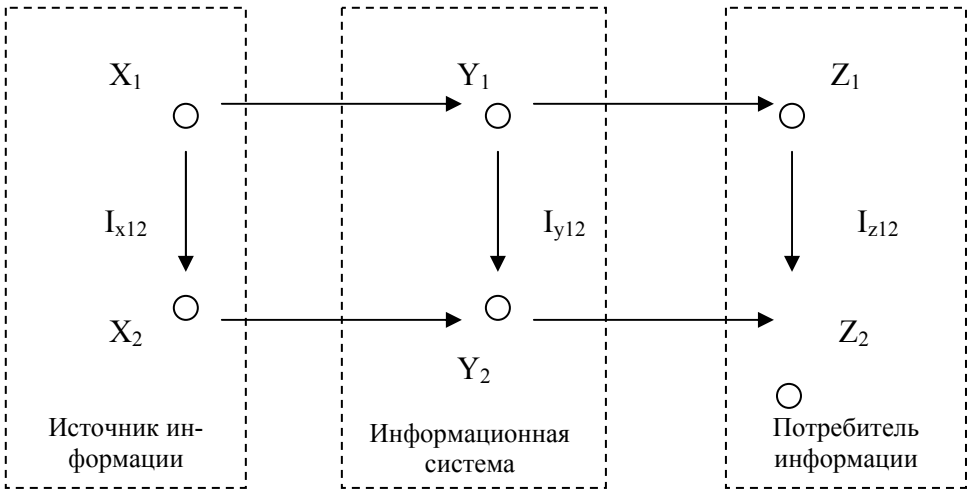


Рис. 2.6. Информационные преобразования в совокупной ИС

В схеме, представленной на рис. 2.6, различаются кодовые (продольные) и информационные (поперечные) преобразования. Кодовые преобразования имеют целью изменение формы представления сообщения без изменения его смыслового содержания. Например, *рукопись – статья – электронный документ*, или *авторское видение проблемы - публикация – образ, возникающий у читателя*. Информационные преобразования имеют аналитико-синтетический характер и связаны с изменением смыслового содержания. Например, *исходные данные – результат научного исследования – граф понятий,* или *документ – поисковый образ*.

Таким образом, процесс получения информации, как система информационных преобразований может быть представлена трехуровневой схемой (рис. 2.7).

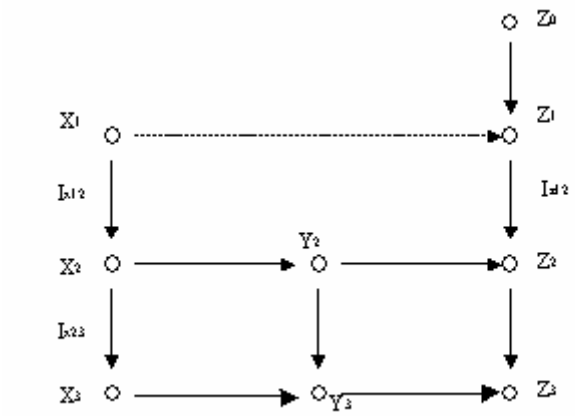


Рис. 2.7. ИС как система информационных преобразований

где:

X_1 – генерируемое источником (сознанием человека) новое знание на образном уровне;

X_2 – публикация – сообщение, подготовленное автором с учетом специфики предметной области издания;

X_3 – основные понятия и закономерности, использованные автором;

Y_2 – документ – опубликованные результаты исследования, оформленные в соответствии с требованиями издания;

Y_3 – поисковый образ документа;

Z_0 – наличное знание человека – приемника информации;

Z_1 – получаемое новое знание (образное представление), возникающее в сознании человека принимающего информацию;

Z_2 – восприятие содержания документа, в контексте понимания приемником специфики предметной области;

Z_3 – восприятие приемником поискового образа документа (отражение понятий ПОДа на системообразующую понятийную схему приемника).

Такая схема передачи знаний соответствует случаю параинформирования, когда образ знания $\mathbf{Z_1}$ будет получен не передачей по кодовой цепи (она реально отсутствует, так как нет носителя соответствующего уровню сознания), а будет восстановлен или вновь построен на основе образа $\mathbf{Z_2}$ и наличного знания приемника $\mathbf{Z_0}$.

С точки зрения представленного ранее рассмотрения информационных компонент в системах управления, можно сказать, что кодовые преобразования инвариантны относительно внешней среды, в то время как информационные – значимы.

2.3. Информационные технологии

Согласно определению, принятому в [Колин1995], информационная технология - 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. При этом достигается экономия затрат труда, энергии людских и материальных ресурсов, необходимых для реализации данного процесса.

Информационные технологии позволяют оптимизировать разнообразные информационные процессы, начиная от подготовки и издания печатной продукции и кончая информационным моделированием и прогнозированием глобальных процессов развития природы и общества.

В рамках класса задач, характерных для информационных систем, **технологией обработки информации** будем называть упорядоченную последовательность взаимосвязанных действий, выполняемых с момента восприятия информации до момента получения заданных результатов.

Технология обработки информации зависит от характера решаемых задач, используемых средств вычислительной техники, числа пользователей, систем контроля за процессом обработки информации и т.д. Технология, как некоторый процесс, всегда присутствует в любой предметной области, особенности которой, в свою очередь, оказывают существенное влияние на функции соответствующих технологий. Информационные технологии ориентированы на решение технических, экономических и управленческих задач, связанных с выполнением операций по сбору необходимой для решения этих задач информации, переработки ее по некоторым алгоритмам и выдачи результата лицу, принимающему решение, в удобной для него форме.

Обработка информации происходит в процессе реализации **технологического процесса**, определяемого предметной областью.

Информационные технологии, в отличие от производственных (как следствие свойств самого объекта – информации), обладают рядом относительно специфических функций, таких как: сбор, регистрация,

хранение, поиск, накопление, генерация, анализ, передача и распространение данных, информации и знаний. Информационная технология направлена на обработку и/или переработку “сырья” (в качестве которого выступают данные, информация) путем использования соответствующих “машин”, “механизмов” и “организационно-технологических приемов” (в качестве которых выступают аппаратные, программные, а также организационно-методические средства).

Таким образом, **информационную технологию** (ИТ) можно определить как систему методов, способов и средств сбора, регистрации, хранения, поиска, накопления, обработки, генерации, анализа, передачи и распространения данных, информации и знаний на основе применения средств вычислительной техники, программных средств и телекоммуникаций.

Сбор данных (информации) представляет собой процесс регистрации, фиксации, записи данных о событиях, объектах (реальных и абстрактных), связях, признаках и соответствующих действиях. При этом иногда выделяют в отдельные операции "сбор данных" и "сбор информации". Здесь сбор информации - это процесс идентификации и получения данных от различных источников, группирования полученных данных и представления их в форме, необходимой для ввода в ЭВМ.

Обработка данных включает в себя несколько взаимосвязанных операций, таких, как поиск, выборка, сортировка, слияние, проведение расчетов и т.д. Обработка данных представляет собой процесс управления данными, по возможности без учета смысла, заложенного в данные.

Обработка информации представляет собой переработку данных, отражающих информацию определенного типа (текстовой, цифровой, графической и др.) в процессе преобразования ее в информацию другого типа. Причем, зачастую и тип данных (как форма существования информации) определяется характером содержания, т.е., обработка информации – это преобразование данных с учетом их содержания.

Таким образом, говоря о технологии как процессе преобразования объектов (т.е., представленных в материальной форме), необходимо определить адекватный способ их идентификации. Это необходимо для их поиска - «узнавания» и выделения из множества других объектов окружающей среды.

Однако, с точки зрения рассмотренных в первой главе свойств информации, идентификация информации как объекта имеет двойственную природу: информация идентифицируется как целостный (неделимый) объект, имеющий как *форму* существования, так и *содержание*.

Это предполагает, что содержание представлено в форме композиции некоторых «атомарных» единиц информации. Например, текст, состоящий из слов.

Забегая вперед, отметим, что именно форма существования «единичного» неявно предопределяет разделение информационных систем (а точнее – баз данных, составляющих основу ИС) на фактографические и

документальные. Для первых «факт» - это самодостаточная единица информации, не теряющая смысл вне других фактов обрабатываемой коллекции. Для документальных систем значение слова как семантически значимой единицы, будет однозначно восприниматься только в контексте - в единстве с другими словами.

2.4. О реализации процесса поиска информации

Для рассмотрения особенностей реализации процессов поиска информации, как одной из основных операций ИТ, отметим тот простой факт, что поиск – это процесс, в ходе которого в той или иной последовательности производится соотнесение отыскиваемого с каждым[18] объектом, хранящимся в массиве. Причем, определяющими для понимания методов автоматизации поиска являются два следующих фактора: 1) сравниваются не сами объекты, а описания - так называемые «поисковые образы»; 2) сам процесс является сложным (составным и не одноактным) и обычно реализуется последовательностью операций. Первый фактор имеет коммуникативную природу, что обуславливает решение на уровне лингвистических средств. Второй – технологическую, когда оптимизация поиска может быть сведена к оптимизации структур данных и алгоритмов обработки.

Алгоритм поиска включает, по крайней мере (необходимо), следующие операции:

- выборку очередного объекта из массива для выполнения сравнения с запрашиваемым;
- сравнение выбранного объекта с образцом;
- принятие решения, соответствует ли объект образцу (определение степени соответствия и применение некоторого критерия для принятия решения на уровне двузначной логики «соответствует» / «не соответствует»);
- переход к выборке следующего объекта или завершение процесса поиска.

Различают два вида поиска информации – поиск целостного объекта (единицы хранения) и поиск по содержанию (точнее, некоторой части содержания – того, что не доставало пользователю в его практической деятельности). Здесь обязательно надо отметить условность такого деления. Определяющим (классификационным) признаком в этом случае является скорее не природа сравниваемых объектов, а то, что будет результатом процесса поиска в целом – документ (машинная запись как операционная единица в последующих процессах) или конкретная информация (которая будет сразу использована человеком, но, которая, естественно, представляет часть документа)[19].

---

[18] Естественно, что процесс поиска может быть завершен и раньше, например, если какой-либо очередной выбранный из массива объект будет полностью удовлетворять потребности, инициировавшей необходимость проведения самого поиска.

[19] Указанным двум видам поиска иногда ставят в соответствие способы поиска - *адресный* и *ассоциативный поиск.*

С точки зрения организации процедуры отбора этому соответствует два способа соотнесения затребованного с имеющимся в информационном массиве: 1) путем сопоставления запроса *непосредственно с содержанием объекта*, выбранного для сравнения; 2) опосредованно, когда запрос сопоставляется с образом, производным (вторичным) по отношению к самому объекту. С точки зрения представленного выше алгоритма разница между этими способами состоит в том, объект какой природы будет выбираться в цикле для сравнения – непосредственное содержание или некоторый идентификатор, отражающий содержание отчасти или в целом. Неявным, но, с точки зрения реализации алгоритма поиска – основным фактором здесь является форма (способ), определяющая порядок выборки: от порядка расположения объектов в массиве (например, в том «естественном» порядке, в каком они поступали для хранения), или в «искусственном» порядке, соответствующем, например, классификации предметной области. Но, поскольку и в том, и в другом случае мы имеем дело с перебором объектов, выбираемых из хранилища для сравнения, рациональность построения процедуры поиска будет определяться длиной перебора, что в свою очередь определяется как характеристиками хранимых объектов (в нашем примере – размерами документов), так характером запросов (в нашем примере – поиском по предмету или по шифру хранения документов). Соответственно, оптимизация процесса в первую очередь связывается с возможностью сокращения времени перебора, то есть - длины выбираемой последовательности.

В общем случае можно сказать, что 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 библиотеках[20].

В случае прямой организации массива документы размещаются в последовательности, не связываемой с порядком какой-либо классификации или алфавита, например, в порядке их поступления в хранилище. Но здесь надо отметить, что 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. Поиск по предмету при такой организации для больших массивов будет требовать достаточно много времени, так как для этого надо последовательно выбирать для сравнения с запросом все документы из хранилища, поскольку, не обратившись к документу, мы не можем судить о его содержании.

В инвертированном массиве документы могут быть, например, разбиты на подмножества, которые упорядочены в соответствии с некоторой классификацией и, что особенно важно, обозначены идентификаторами,

[20] 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 уровня автоматизации.

отражающими предметное содержание соответствующего класса. Более того, 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.

Целесообразность использования терминов «прямая» и «инвертированная» форма представления информации становится очевидной при рассмотрении «предельного» варианта организации инвертированного справочника, в котором в качестве индексов используются все без исключения слова документов, а в ссылку на документ, содержащий данное слово, включены данные о позиции этого слова в документе (например, номер главы, параграфа, предложения, позиции в предложении).

В этом случае избыточность данных может быть уменьшена за счет отказа от прямого массива документов. Но это приведет к дополнительным затратам пространства для хранения позиционных параметров и дополнительным действиям по сборке текста документов, которые необходимо выдавать пользователю, что, соответственно потребует больше времени.

При инвертированной организации 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 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 обычно на несколько порядков меньше длины документа, и, кроме того, индексы строго упорядочены, например, по лексико-графическому признаку.

Идентификация содержания с помощью индексов строится по принципам языковых систем (каждый индекс представляет то или иное множество характеристических признаков), что позволяет еще 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, то есть выбираются только те документы, которые описываются именно этим сочетанием. Кроме того, для индексирования содержания отдельного документа могут быть использованы разные лингвистические системы, то есть один документ может иметь несколько поисковых образов, отражающих его содержание в различных аспектах и с разной степенью детализации.

Использование технологии индексирования (и, соответственно, инвертированных форм представления информации) тем не менее, имеет ряд следующих принципиальных недостатков:

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

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

Тем не менее, автоматизация *поиска информации* основывается именно на технологии индексирования (как способа идентификации содержания) документов, поскольку документальные ИПС имеют следующие принципиально важные особенности [Солтон1979] построения и использования.

Во-первых, нужно помнить, что задачи в области документального поиска не сравнимы с другими задачами обработки текстов, такими, как автоматический перевод или поисковые процедуры типа вопрос-ответ (при которых даются прямые ответы на самые разные вопросы). 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 анализа, 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 анализа, но для среднего запроса наилучшими являются более простые методы индексирования.

Контрольные вопросы

1. Приведите примеры абстрактных систем.
2. Приведите примеры материальных систем.
3. Дайте определение понятия «информационная система»
4. Охарактеризуйте и классифицируйте информацию, как основной объект обработки в ИС.
5. Приведите классификацию ИС.
6. Охарактеризуйте основные компоненты ИС.
7. Перечислите и охарактеризуйте основные обеспечивающие подсистемы ИС.
8. Охарактеризуйте движение информации в системах управления и в системах воспроизводства знаний.
9. Определите понятие «информационная деятельность».
10. Дайте определение информационной технологии.
11. Перечислите основные операции процесса поиска информации.

3. Модели и структуры данных информационных систем

Рассматриваемые в контексте понятия «информационная система» элементы реального мира, информацию о которых мы сохраняем и обрабатываем, будем называть *объектами*. Объект может быть материальным (например, служащий, изделие или населенный пункт) и нематериальным (например, имя, понятие, абстрактная идея).

Набором объектов будем называть совокупность объектов, однородных с некоторой точки зрения (например, объектов *нашего* внимания, пусть даже и разнородных по своей внутренней природе).

Объект имеет различные свойства (например, цвет, вес, имя), которые важны для нас в то время, когда мы обращаемся к объекту (например, выбираем среди множества других) с какой-либо целью его использования. Причем свойства могут быть заданы как отдельными однозначно интерпретируемыми количественными показателями, так и словесными нечеткими описаниями, допускающими разную трактовку, зависящую, например, от точки зрения и наличных знаний воспринимающего субъекта.

Общим же фактором является то, что человек, работая с информацией, имеет дело с *абстракцией,* представляющей интересующий его фрагмент реального мира - той совокупностью *характеристических свойств (атрибутов)*, которые важны для решения его прикладной задачи. Абстрагирование – это способ *упрощения* совокупности фактов, относящихся к реальному объекту (по своей сути бесконечно сложному и разнообразному). 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 действий при решении.

Цель такого абстрагирования - построение конструктивного операбельного описания (рабочей модели), удобного в обработке, как для человека, так и для машины, позволяющего организовать эффективную обработку больших объемов информации, причем высокопроизводительной должна быть работа не только вычислительной системы, но и взаимодействующего с ней человека.

3.1. Семантика ИС, основанных на концепции баз данных

Как уже отмечалось, задачи информационных систем – это не только поддержка процессов планирования и управления, но и интеграция разработки и сопровождения основных и технологических объектов и процессов, диагностика, мониторинг, моделирование. Соответственно, задачи и назначение БД, как системы хранящей информацию обо всех составляющих – обеспечить информационную поддержку процессов жизненного цикла автоматизируемой системы.

Здесь база данных, как основная информационная компонента системы управления, – это отражение реальной предметной области, «действующая» информационная модель[21], которая, обеспечивая субъект информацией для принятия решения, позволяет в итоге управлять физическими объектами и процессами. Такая функциональная направленность (и, естественно, предполагающая достижение эффективности в первую очередь за счет использования именно БД) обуславливает и обратную зависимость: объекты, процессы и события ПрО выделяются таким образом, чтобы было возможно их представление в виде системы взаимосвязанных данных и процедур, удобных для их последующей (человеко-машинной!) обработки.

В каком-то смысле базу данных можно сравнить с сообщением о состоянии предметной области, воспринимаемым некоторым субъектом, задачей которого и является преобразование объектов этой ПрО, причем в своей деятельности субъект руководствуется информацией извлекаемой именно из этого «сообщения». Схема этого соотношения, приведенная на рис. 3.1, иллюстрирует еще и то, что система, преобразующая объект, принципиально является комплексной (состоящей, по крайней мере, из двух компонент, работающих с объектами разной природы: субъект преобразования взаимодействует преимущественно с материальными объектами, а БД – с информационными).

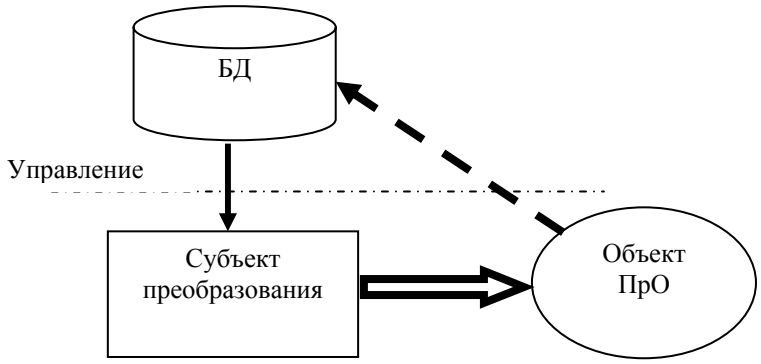


Рис. 3.1. Информационная модель преобразования

В общем случае, поскольку для сложных систем с многоуровневым представлением семантики, эффективность обработки достигается через специализированность представления объектов или процессов путем сведения представления множества обрабатываемых объектов к однородности природы и формы их представления, то для реализации эффективного межуровнего (межкомпонентного) взаимодействия (на каждом из которых объекты представлены в виде, наиболее адекватном функциональным средствам этого уровня или процесса) любая величина должна быть преобразована в соответствии с «контекстом» этого уровня для получения такого ее представления, которое будет «значимо» для

[21] Модель – лишь в том смысле, что она – представление, описание на уровне данных только некоторых аспектов, и только некоторой части реального мира, и поэтому не может быть тождественна реальным объектам. Но в тоже время БД и сама является частью реального мира.

воспринимающего уровня, т.е. может быть обработано средствами этого уровня. Здесь «контекст» - это декларативное или, иногда, процедурное определение способа использования элементарных составляющих величины для получения значения. Например, контекст - это порядок использования байтов при преобразовании вещественного числа, представленного в двоичной форме, в символьный формат.

Соотношение понятий *величина*, *контекст* и *значение* приведено на рис. 3.2. Здесь значение, получаемое на первом уровне (в первой подсистеме, процессе), на следующем рассматривается в свою очередь как величина, которая будет интерпретироваться в соответствии с контекстом своего уровн[22]

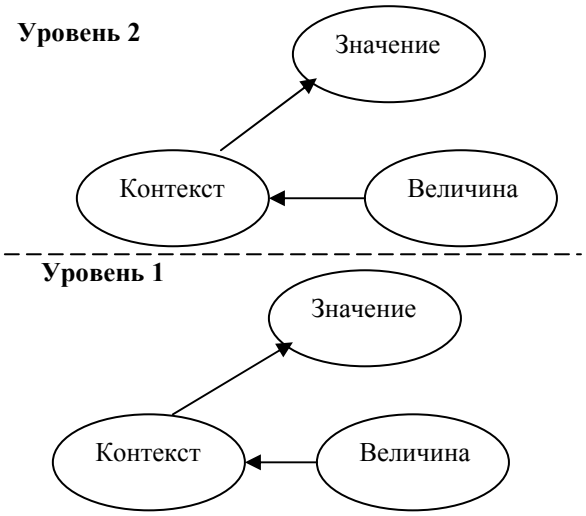


Рис. 3.2. Соотношение понятий «величина», «контекст» и «значение»

Таким образом, можно сказать, что значение в общем случае определяется парой *<контекст, величина>*. Причем, поскольку *контекст* и *величина* имеют разную природу, они должны быть представлены в вычислительной среде самостоятельными, скорее всего, разнотипными объектами.

Такое, хотя и упрощенное, представление БД как средства информационных коммуникаций, позволяет тем не менее увидеть взаимосвязь вида информации с формой ее представления и особенностью ее использования.

В этом смысле (с точки зрения способа представления и, соответственно, восприятия) в отдельный класс можно выделить *фактографическую информацию:* такое представление реально существующих событий и явлений, когда они могут быть описаны как *факты*, задаваемые парой *<имя, значение>*, где *имя* – знак, уникально определяющий (идентифицирующий) факт в заданной предметной области, и обычно не нуждающийся в явном определении или доопределении его существа; а

[22] Соотношение понятий «величина» и «значение» аналогично соотношению понятий «данные» и «информация».

значение – характеристика, задающая одно из множества возможных состояний.

Т.е., здесь факт (его значение) задается величиной, например, числовой для физически измеримых параметров, в том числе и логическими величинами «истина» / «ложь» для указания свершилось событие или нет[23].

Можно сказать, что особенностью фактографической информации является практическая очевидность (минимальная неопределенность, не требующая использования сложных или нечетких процедур) идентификации и интерпретации «факта», как его имени, так и состояния. То есть, в этом случае контекст в достаточной степени определяется однозначно понимаемым объявлением о назначении базы данных и таким именованием полей данных, когда в качестве имени используется общепринятое, не зависящее от прикладных задач, *имя свойства* (и таким образом определяются характеристические признаки). Именно такое состояние предопределяет для пользователя возможность адекватного восприятия содержания: способ интерпретации данных в этом случае практически не может быть неоднозначным, причем для пользователя *определение способа* происходит *неявно* (не требует от него явных действий для определения и использования контекста). Это, с одной стороны, позволяет свести представление предметной области к точной теоретико-множественной модели, а с другой – обуславливает возможность непосредственного использования данных в задачах обработки (на уровне прикладных программ) для генерации новой информации без участия субъекта (человека), внешнего по отношению к машинной среде, обеспечивающего определение и использование контекста.

Однако большинство задач, решаемых человеком, не могут быть сведены к «фактографическому» представлению и описываются (и, соответственно, представляются в машинной среде) средствами естественного или специализированного языков, оперирующих *лингвистическими переменными*, значение которых может зависеть не только от контекста предметной области, но также и от контекста ближайшего окружения – значения соседних переменных. Причем, появление нового смысла (факта) не обязательно приводит к появлению новой переменной: новый факт представляется с помощью уже существующих переменных. Например, словесные определения философских или географических понятий.

В отличие от ранее рассмотренного фактографического представления, для вербальной формы представления факта (выражениями языка с использованием лингвистических переменных) характерно то, что для задания *имени*, *значения* и *контекста* может использоваться единый способ и средства – лингвистические переменные одного и того же языка. Например, описание весовых свойств может быть представлено не-

[23] И, следует отметить, что такая форма в наибольшей степени соответствует машинным формам представления информации.

сколькими, но имеющих один смысл, вариантами предложений: «Чугунная заготовка весом 29 килограмм» или «Чугунная заготовка имеет свойство m = 29, где m – вес в килограммах».

Автоматическое приведение такого рода представлений к очевидной наилучшей для этого случая табличной форме, потребовало бы применения трудно реализуемых процедур морфологического и семантического анализа. Но, с другой стороны, выделение смысла (и генерация новой информации) обычно производится человеком, сознание которого (как среда преобразования) ориентировано именно на обработку лингвистических переменных.

Рассматривая процесс генерации новой информации (рис. 3.3), где в качестве источника исходных данных используются БД, нужно сказать, что отбор и обработка должны быть выделены в отдельные процессы, т.к. с точки зрения общей (суммарной) эффективности один из них (обычно поиск) должен быть опосредованным - оценка полезности найденной информации производится человеком в сознание человека - внешней по отношению к машине среде, работающей со слабоструктурированной информацией эффективнее машин.

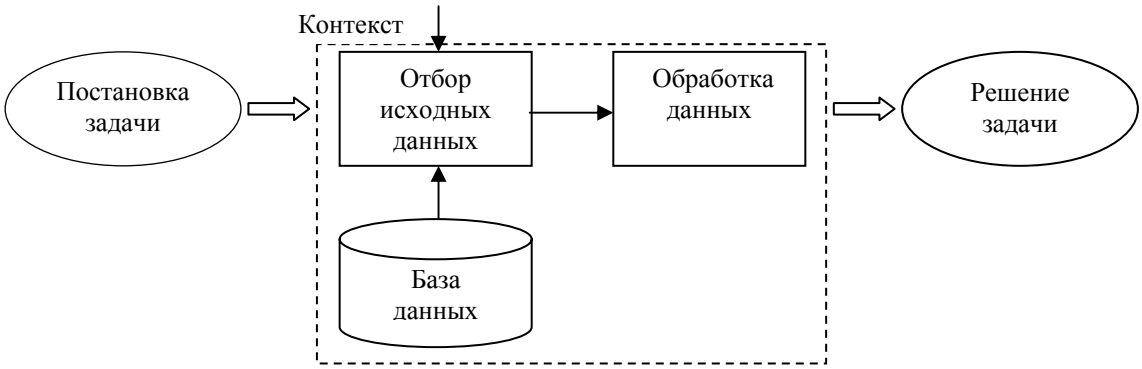


Рис. 3.3. Схема процесса автоматизированного решения задач

Случаи, когда информация представляется в форме не адекватной архитектуре Фон-Неймановских машин, могут быть обусловлены разными факторами. Рассмотрим следующие случаи.

1. Хорошо структурированная информация представляется в графическом или специальном формате. Например, структурные химические формулы, конструкторская документация и т.д. В этом случае для автоматической обработки требуются узко специализированные средства, что приводит к общей не унифицированности представления семантических элементов (например, графических примитивов) на уровне данных.

2. Информация точная по содержанию, но вариантно представляемая по форме. Например, описание в текстовом виде численно задаваемых параметров изделия. Лингвистические переменные в этом случае имеют точное значение, однако построение универсальной процедуры

автоматического выделения факта из текста трудоемко и потому нецелесообразно.

3. Слабоструктурированная информация, обычно представляемая в текстовой форме. Например учебная или научная публикация, где новые понятия строятся на основании ранее определенных. В этом случае значения лингвистических переменных могут принимать новые, ранее не определенные значения, которые определяются контекстом - ближним (словосочетания) или общим (темой сообщения).

Возвращаясь к процедуре поиска, как важнейшей составляющей использования ИС, еще раз отметим, что критерий отбора должен содержать не только величину (например, слово), но и контекст.

В реальных системах поиск документальной информации[24], представленной в текстовой форме, производится по вторичным документам – специально создаваемым поисковым образам точно идентифицирующим сам документ как единицу хранения, и приблизительно, в краткой форме путем *перечисления* основных понятий, отражающий смысловое содержание. Такой подход позволяет построить процедуры поиска на основе теоретико-множественной модели с точной логикой отбора по критерию наличия заданного сочетания терминов запроса в списке терминов поискового образа. Однако контекст использования терминов должен быть доопределен отдельно – либо во время поиска, например указанием тематической области, либо после отбора из базы – во время ознакомления человека с содержанием найденного.

Определение контекста предметной области, как будет рассмотрено далее, в целом осуществляется с помощью тезаурусов - терминологических систем, фиксирующих с помощью родовидовых и других отношений роль и семантику дескрипторов (выделенных терминов, которые используются для формирования поисковых образов документов).

3.2. Идентификация и поиск информации

В задачах обработки информации, и в первую очередь в алгоритмизации и программировании, атрибуты *именуют* (обозначают) и приписывают им *значения.*

При обработке информации мы, так или иначе, имеем дело с совокупностью объектов, *информацию о свойствах* каждого из которых надо сохранять (записывать) как *данные,* чтобы при решении задач их можно было найти и выполнить необходимые преобразования.

Таким образом, любое состояние объекта характеризуется совокупностью актуализированных атрибутов[25] (имеющих некоторое значе-

[24] Это соответствует третьему из вышеперечисленных случаев. Два первых мы не рассматриваем, т.к. в этих случаях используются специализированные системы.

[25] В общем случае объект может описываться совокупностью записей, относящихся к его составным частям или отражающих динамику изменения состояния.

ний в этот момент времени), которые фиксируются на некотором материальном носителе в виде *записи* – совокупности (*группы*) формализованных *элементов данных* (значений атрибутов, представленных в том или ином формате). Кроме того, в контексте задач хранения и поиска можно говорить, что значение атрибута *идентифицирует* объект: использование значения в качестве поискового признака позволяет реализовать простой критерий отбора по условию сравнения[26].

Также как и в реальном мире, отдельный объект всегда уникален (уже хотя бы потому, что мы *именно его* выделяем среди других). Соответственно, запись, содержащая данные о нем, также должна быть узнаваема однозначно (по крайней мере, в рамках предметной области), т.е. – иметь уникальный идентификатор, причем никакой другой объект не должен иметь такой же идентификатор. Поскольку идентификатор – суть значение элемента данных, в некоторых случаях для обеспечения уникальности требуется использовать более одного элемента. Например, для однозначной идентификации записей о дисциплинах учебного плана необходимо использовать элементы СЕМЕСТР и НАИМЕНОВАНИЕ ДИСЦИПЛИНЫ, так как одна дисциплина может быть прочитана в разных семестрах.

Предложенная выше схема представляет атрибутивный способ идентификации содержания объекта (рис. 3.4). Она является достаточно естественной для данных, имеющих фактографическую природу. Информацию, представляемую такого рода данными, называют *хорошо структурированной*.

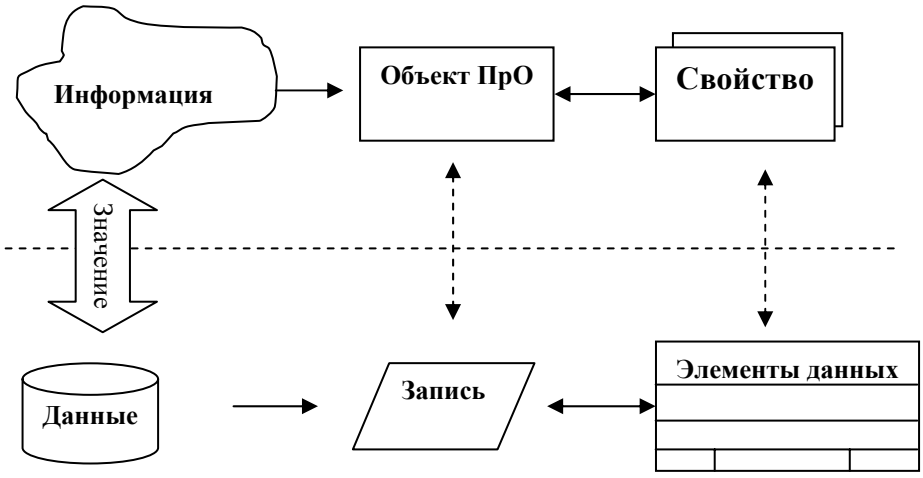


Рис. 3.4. Атрибутивный способ идентификации

[26] Следует отметить некоторые семантические проблемы идентификации через значение атрибута. Значение атрибута идентифицирует запись о **состоянии** объекта, и в случае изменения значения, например – табельного номера служащего, будет невозможно ответить на вопрос: идет ли речь о том же служащем, или о новом.

Здесь важно отметить, что структурированность относится не только к форме представления данных (формат, способ хранения), но и к способу интерпретации значения пользователем: значение параметра не только представлено в предопределенной форме, но и обычно сопровождается указанием размерности величины, что позволяет пользователю понимать ее смысл без дополнительных комментариев. Таким образом, фактографические данные предполагают возможность их *непосредственной* интерпретации.

Однако, как отмечалось ранее, атрибутивный способ практически не подходит для идентификации *слабо структурированной информации*, связанной с объектами, имеющими обычно *идеальную* (умозрительную) природу – категориями, понятиями, знаковыми системами. Такие объекты зачастую определяются опосредовано – через другие объекты, для чего используются естественные или искусственные языки (например, язык математики). Соответственно, для понимания смысла пользователю необходимо использовать соответствующие правила языка, и, более того, часто необходимо уже располагать некоторой информацией, позволяющей идентифицировать и связать получаемую информацию с наличным знанием. Т.е., процесс интерпретации такого рода данных имеет *опосредованный* характер и требует использования дополнительной информации, которая, в общем случае, не обязательно присутствует в формализованном виде в базе данных.

Таким образом, можно сказать, что основным отличием документальных ИС является опосредованный способ интерпретации данных, а не их организация.

Программисту или пользователю необходимо иметь возможность обращаться к отдельным, нужным ему записям (описаниям объектов) или отдельным элементам данных. В зависимости от уровня программного обеспечения прикладной программист может использовать следующие способы:

- задать машинный адрес данных и в соответствии с форматом записи прочитать значение. Это случай, когда «навигатором» должен быть программист;

- сообщить системе имя записи или элемента данных, которые он хочет получить, и, возможно, организацию набора данных. В этом случае система сама произведет выборку (по предыдущей схеме), но для этого она должна будет использовать вспомогательную информацию о структуре данных и организации набора. Такая информация по существу будет избыточной по отношению к объекту, однако общение с базой данных не будет требовать от пользователя знаний программиста и позволит переложить заботы о размещении данных на систему.

В качестве ключа, обеспечивающего доступ к записи, можно использовать идентификатор – отдельный элемент данных. Ключ, который идентифицирует запись единственным образом, называется *первичным (главным)*.

В том случае, когда ключ идентифицирует некоторую группу записей, имеющих определенное общее свойство, ключ называется *вторичным* (*альтернативным*). Набор данных может иметь несколько вторичных ключей, необходимость введения которых определяется практической необходимостью – оптимизацией процессов нахождения записей по соответствующему ключу.

Иногда в качестве идентификатора используют составной *сцепленный ключ* – несколько элементов данных, которые в совокупности, например, обеспечат уникальность идентификации каждой записи набора данных.

При этом ключ может храниться в составе записи или отдельно. Например, ключ для записей, имеющих неуникальные значения атрибутов, для устранения избыточности может храниться отдельно. На рис. 3.5 приведены два таких способа хранения ключей и атрибутов для набора простейшей структуры.

Ключ	Атрибут
112	а/м ВАЗ 2110
113	а/м ВАЗ 2121
441	а/м ГАЗ 3110
456	а/м ВАЗ 2110
457	а/м ВАЗ 2121
678	а/м ВАЗ 2101
789	а/м ГАЗ 3110
889	а/м ВАЗ 2110
998	а/м ГАЗ 3102

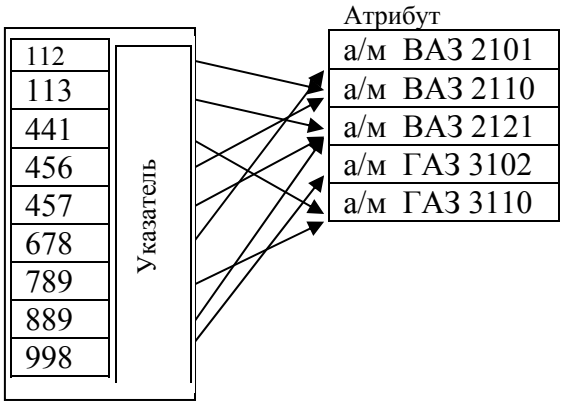


Рис. 3.5. Способы хранения ключа и атрибута

Введенное понятие ключа является логическим и его не следует путать с физической реализацией ключа – *индексом*, обеспечивающим доступ к записям, соответствующим отдельным значениям ключа.

Один из способов использования вторичного ключа в качестве входа - организация инвертированного списка, каждый вход которого содержит значение ключа вместе со списком идентификаторов соответствующих записей. Данные в индексе располагаются обычно в возрастающем порядке, поэтому алгоритм нахождения нужного значения довольно прост и эффективен. После нахождения значения запись локализуется по указателю физического расположения. Недостатком индекса является то, что он занимает дополнительное пространство и его надо обновлять каждый раз, когда удаляется, обновляется или добавляется

запись. На рис. 3.6 приведен инвертированный список для предыдущего примера.

а/м ВАЗ 2101	678
а/м ВАЗ 2110	112, 456, 889
а/м ВАЗ 2121	113, 457
а/м ГАЗ 3102	998
А/м ГАЗ 3110	441, 789

Рис. 3.6. Инвертированный список для ключа «Марка автомобиля»

В общем случае инвертированный список может быть построен для любого ключа, в том числе составного.

В контексте задач поиска можно сказать, что существуют два основных способа организации данных. Первый соответствует примеру, приведенному на рис. 3.5, и представляет прямую организацию массива. Второй способ является инверсией первого, он соответствует рис. 3.6. Прямая организация массива удобна для поиска по условию «Каковы свойства указанного объекта?», а инвертированная – для поиска по условию «Какие объекты обладают указанным свойством?».

В [Мартин] 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 имеет атрибут A в наборе?
6). **?(?) = V** Какие атрибуты объектов набора имеют значение равное V?

Здесь в запросах типов 2, 3, 6 вместо оператора равенства может быть использован другой оператор сравнения (*больше, меньше, не равно* или другие).

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(ям) производится по указателю, выбираемому из списка по значению вторичного ключа. Ответом в этих случаях будет *значение* атрибута или идентификатора. Запросы типа 3 имеют ответом *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.

Следует отметить, что в контексте обработки запросов 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 ускорения процесса поиска в системе используют два набора индексов: *индекс экземпляров* (значений ключей) и *индекс данных* (инвертированный список). С помощью индекса экземпляров можно найти в файле 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*.

3.3. Представление предметной области и модели данных

Если бы назначением базы данных было только хранение и поиск данных в массивах записей, то структура системы и самой базы была бы простой. Причина сложности в том, что практически любой объект характеризуется не только параметрами-величинами, но и взаимосвязями частей или состояний. Есть различия и в характере взаимосвязей между объектами предметной области: одни объекты могут использоваться только как характеристики остальных объектов, другие – независимы и имеют самостоятельное значение (рис. 3.7).

Кроме того, сам по себе отдельный элемент данных (его значение) ничего не представляет. Он приобретает смысл только тогда, когда связан с атрибутом (природой значения, что позволит интерпретировать значение) и другими элементами данных.

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 пред-

ставленные 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 будет задаваться (описываться) *удобными для пользователя* средствами в не адекватной человеку жесткой вычислительной среде с двоичной логикой и числовым представлением информации.

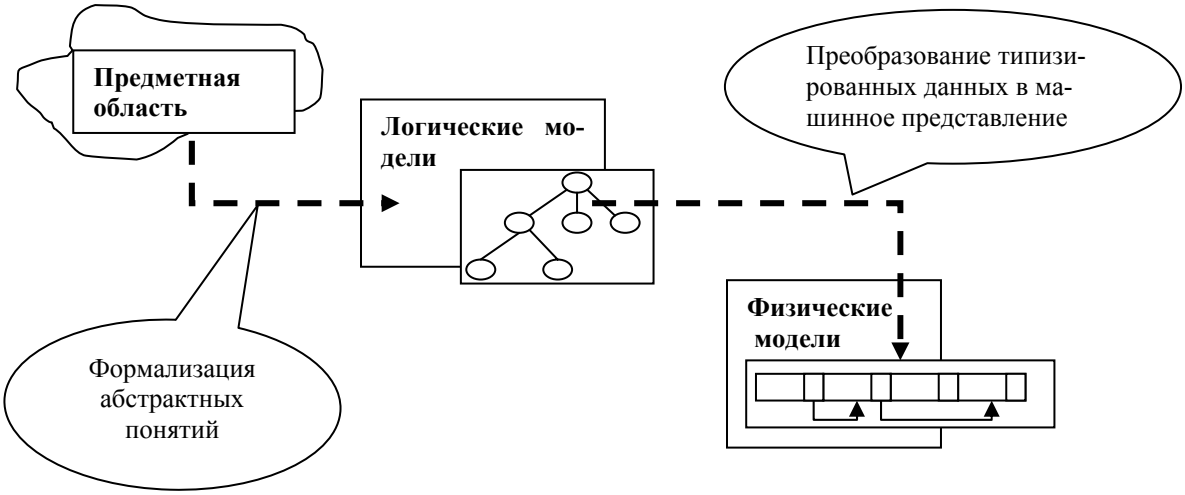


Рис. 3.7. Этапы преобразования представлений ПрО

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, свойственным языкам программирования: 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 (или система автоматизации проектирования БД), используя такие *типовые абстрактные понятия* (как например числа, множества, последовательности, агрегаты), определяет соответствующие информационные структуры. Система управления данными, используя *двоичные формы представления типизированных* данных, обеспечивает эффективные процедуры хранения и обработки данных.

Именно введение *промежуточного* уровня абстракции позволяет иметь раздельное описание логического и физического представления, освобождает конечного пользователя от необходимости беспокоиться о деталях внутримашинного представления и обработки, поскольку он может быть уверен, что программистом выбрана наиболее эффективная форма для данной ситуации. Однако эффективность здесь имеет определенные пределы. Чем ближе система абстракций к особенностям вычислительной среды, тем выше эффективность выполнения программы, но вынужденная «специализация» абстракций увеличивает вероятность того, что они станут неподходящими для некоторых других применений.

Модель данных должна, так или иначе, дать основу для описания данных и манипулирования данными, а также дать средства анализа и синтеза структур данных.

3.4. Структура информации и структура данных

При любом методе отображения предметной области в машинных базах данных 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 представления сложных композиционных объектов и связей реальной 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.;
- *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. Именно успешность структурирования данных определяет сложность процедур их обработки [Вирт]. Например, при функциональной обработке массива необходимо обращаться к отдельным элементам, в то время как в операциях присваивания или при записи массива в файл поэлементное обращении приведет к увеличению размера текста программы, а в ряде случаев - к увеличению времени выполнения.;
- *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 (с целью определения основных и вспомогательных функциональ-

ных массивов, а также совокупности унифицированных процедур манипулирования данными). 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. Именно это позволяет формализовать на общей понятийной основе взаимосвязь представлений информации в разных средах: обеспечить контролируемое сведение бесконечного разнообразия объектов и видов взаимосвязей реального мира к жестко детерминированному описанию – совокупности двоичных данных и машинно-ориентированных алгоритмов их обработки.

Отметим, что выделение трех указанных видов структур имеет, в некотором смысле, принципиальный характер. Структура определяет алгоритм выборки отдельных элементов данных, но в то же время она отражает и особенности «технологии» организации и обработки информации, свойственные человеку в его повседневной деятельности.

Физически понятию *структура* соответствует *запись данных*. Запись – это упорядоченная в соответствии с характером взаимосвязей совокупность *полей* (элементов) данных, размещаемых в памяти в соответствии с их *типом*[27]. Поле представляет собой минимальную *адресуемую* (идентифицируемую) часть памяти - единицу данных, на которую можно ссылаться при обращении к данным. Структура данных - здесь способ отображения значений в памяти: размер области и порядок ее выделения (который и определит характер процедуры адресации/выборки).

Таким образом, информационная база состоит из двух компонент:

1) коллекции записей собственно данных;

2) описания этих данных — метаданных.

Данные отделены от описаний, но в то же время данные не могут использоваться без обращения к соответствующим описаниям. Такая конструкция базы данных обеспечивает возможность того, что данные могут использоваться (т.е., представляться) по-разному. С одной сторо-

[27] Память, отводимая для хранения значения элемента данных (*поле данных*), должна выбираться в соответствии с диапазоном значений, которые может иметь этот элемент. Поскольку для выполнения операции *присвоения значения* элементу данных (установление соответствующих битов в «0» или «1») необходимо сначала выделить память, для чего используются две схемы– статическая и динамическая. Для первой характерно выделение памяти до того, как реально появляются значения (обычно на этапе трансляции программы); для второй – в тот момент, когда программа во время исполнения получает конкретное значение. Кроме того, характер данных (тип данных) определяет способ представления и, соответственно, некоторое множество стандартных операций (*примитивов*).

ны, разные прикладные задачи требуют разных наборов данных, в совокупности обеспечивающих функциональную полноту информации, а с другой – они должны быть различны для различных категорий субъектов (разработчиков или пользователей).

3.5. Организация данных в документальных информационных системах

Организация данных и механизмы поиска в базах данных документальных информационных систем, построены на тех же принципах, что и фактографические системы. Однако в физической реализации есть и существенные отличия, которые обусловлены в первую очередь информационной природой элементов данных:

1. Запись базы данных – документ, который задается как набор в общем случае *необязательных* полей, для каждого из которых определены имя и тип. Допустимы большинство стандартных типов (так называемые «форматные» поля, задающие числовые, символьные и другие величины), а также текстовые. Текстовые поля имеют переменную длину и композиционную структуру, не имеющую прямых аналогов среди стандартных типов языков программирования: текстовое поле состоит из параграфов; параграф - из предложений; предложение - из слов. При этом идентифицируемым (адресуемым атомарным) элементом данных с точки зрения хранения будет *поле*, а с точки зрения поиска (атомарным семантически значимым) – *слово*. Вследствие этого поисковые структуры строятся в виде инвертированных файлов.
2. Семантическая природа текстовых полей, представляющих смысл в основном на естественном языке, определяет необходимость учитывать важнейшие свойства используемых терминов: синонимию, полисемию, омонимию, контекстную обусловленность смысла отдельного слова и возможность выразить один смысл многими способами. Вследствие этого поисковые *индексы* могут быть отличны от соответствующих словоформ поля.

3.5.1. Организация данных в документальной информационно-поисковой системе STAIRS

На рис. 3.8 приведена примерная схема организации данных для представления и поиска информации диалоговой системы поиска документов STAIRS (Storage and Information Retrieval System), разработанной фирмой IBM в 70-х годах. Отметим, что такая структура характерна и для большинства современных АИПС.

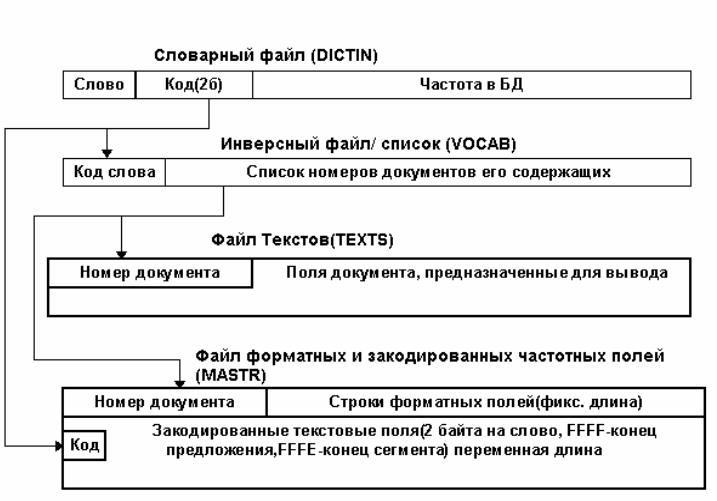


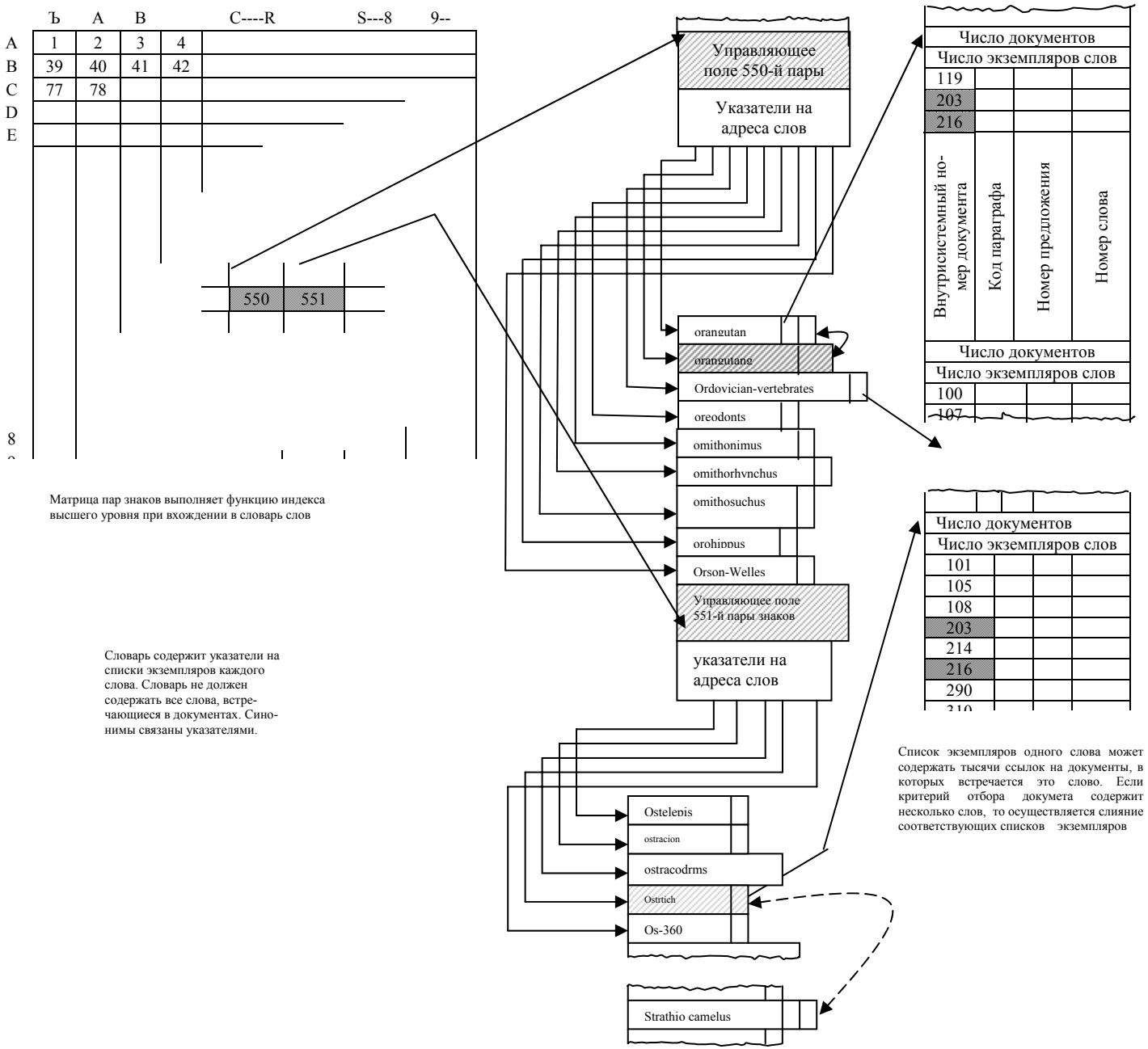
Рис. 3.8. Организация данных в документальной АИПС STAIRS

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

- 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 двухбайтовых кодов слов, находящихся в тексте данного документа. При необходимости, в соответствующих местах находятся разделители сегментов и/или предложений. Файл используется при форматном поиске и при наличии в запросах конструкций *SENT, SEGM, CTX.*

На рис. 3.9 представлен *словарь слов,* в котором содержится перечень терминов, встречающихся в документах. Ввиду значительных размеров словаря его организация должна предусматривать 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. Первые два знака слов, содержащихся в словаре, отсутствуют, но они показаны на рисунке, чтобы облегчить понимание структуры файла.

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


Ъ A B C----R S---8 9--
A 1 2 3 4
B 39 40 41 42
C 77 78
D
E
8
550 551
Управляющее поле 550-й пары
Указатели на адреса слов
Число документов
Число экземпляров слов
119
203
216
Внутрисистемный номер документа
Код параграфа
Номер предложения
Номер слова
Число документов
Число экземпляров слов
100
107
orangutan
orangutang
Ordovician-vertebrates
oreodonts
omithonimus
omithorhynchus
omithosuchus
orohippus
Orson-Welles
Управляющее поле 551-й пары знаков
указатели на адреса слов
Число документов
Число экземпляров слов
101
105
108
203
214
216
290
310
Ostelepis
ostracion
ostracodrms
Ostrich
Os-360
Strathio camelus
Матрица пар знаков выполняет функцию индекса высшего уровня при вхождении в словарь слов
Словарь содержит указатели на списки экземпляров каждого слова. Словарь не должен содержать все слова, встречающиеся в документах. Синонимы связаны указателями.
Список экземпляров одного слова может содержать тысячи ссылок на документы, в которых встречается это слово. Если критерий отбора документа содержит несколько слов, то осуществляется слияние соответствующих списков экземпляров

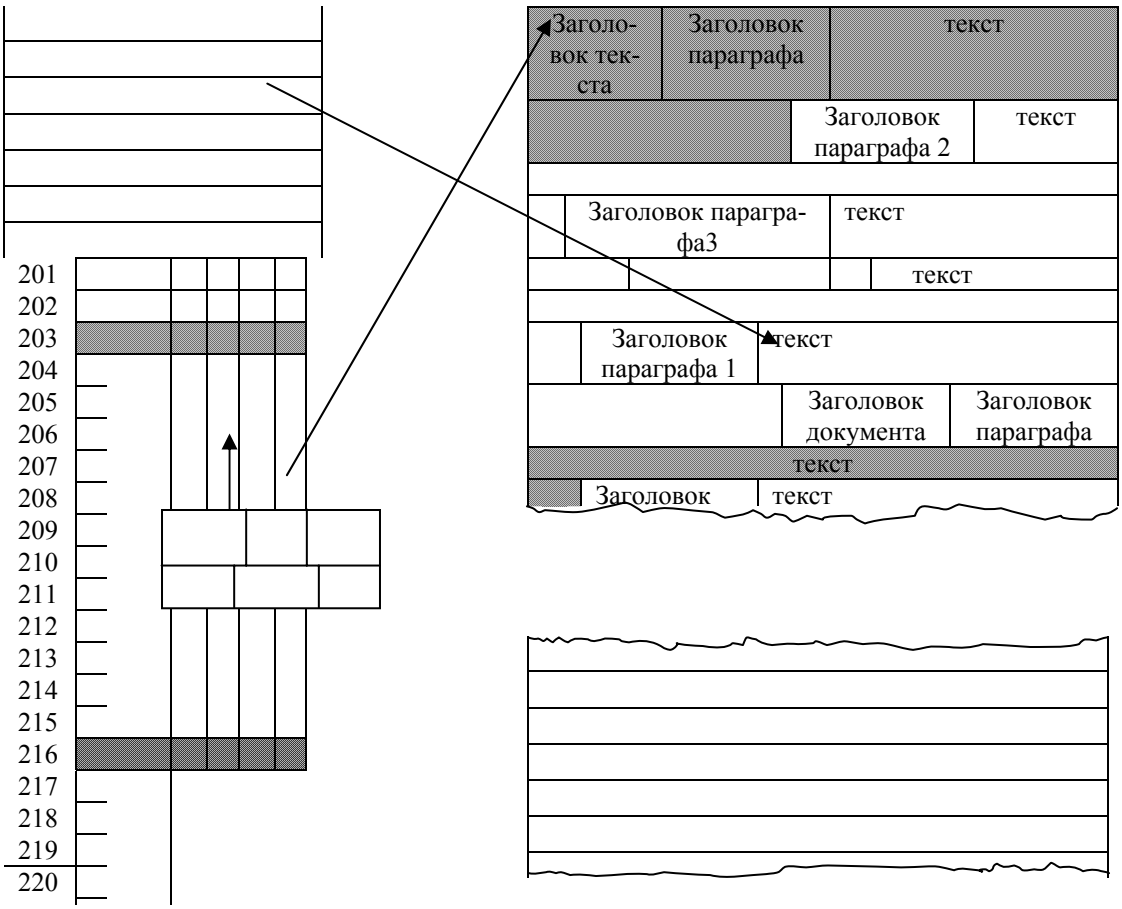


Рис. 3.10. Организация поисковых файлов документов АИПС STAIRS

Каждому слову на рис. 3.9 поставлен в соответствие указатель на *списки экземпляров,* являющихся перечнем документов, в которых встречается данное слово. Каждый список экземпляров содержит заголовок, из которого можно узнать число экземпляров слова во всем файле документов, а также число документов, в которых это слово встречается.

Система присваивает каждому документу уникальный номер. Этот номер является внутрисистемным и не связан с номерами, по которым пользователь может получить данный документ где-нибудь вне системы. В списке экземпляров, соответствующем какому-либо слову, содержатся внутрисистемные номера всех документов, в которых оно встречается. Поисковый критерий может включать требование *поиска всех документов, содержащих одновременно два специфических слова.* Например, можно осуществлять поиск документа, в котором содержится как слово ORANGUTANG, так и слово OSTRICH. В этом случае система находит множество документов, содержащих первое слово, а затем множество документов, содержащих второе слово, и путем их пересечения определяет множество документов, содержащих как первое, так и второе слово.

На рис. 3.10 показан *файл документов,* каждому из которых система сама присваивает внутренний порядковый номер. Документы состоят

из параграфов и текстов, причем тексты также пронумерованы. Каждому параграфу присвоен специальный код, определяющий его тип (например, заголовок, автор, аннотация и т. д.).

Внутрисистемный номер документа является ключом к *индексу документов.* Этот индекс содержит адреса соответствующих документов в памяти. В принципе *можно* хранить эти адресные указатели непосредственно в списке экземпляров, но это нецелесообразно, так как объем памяти, необходимый для хранения адреса, больше объема памяти, необходимого для хранения номера документа. Индекс документов содержит не только адреса, а также некоторые вспомогательные сведения о документах. К этим сведениям относятся внешний номер документа, признак удаления документа, указывающий, какие параграфы документа (или документ в целом) исключены из файла, а также уровень секретности.

В состав документов могут входить *параграфы различных типов,* поэтому пользователь может потребовать, чтобы заданное слово содержалось в названии документа, аннотации, введении или каком-либо конкретном параграфе. В критерии отбора можно указывать автора, место издания документа и дату издания. Независимо от содержания критерия отбора поиск документа осуществляется на уровне списка экземпляров без необходимости входа в файл документов.

3.5.2. Организация данных в документальной АИПС IRBIS

Логическая структура. База данных документальной ИПС IRBIS - это именованная совокупность массива документов и структурированных справочников, обеспечивающих эффективность поиска. Логическая структура БД документальной ИПС IRBIS представлена на рис. 3.11.

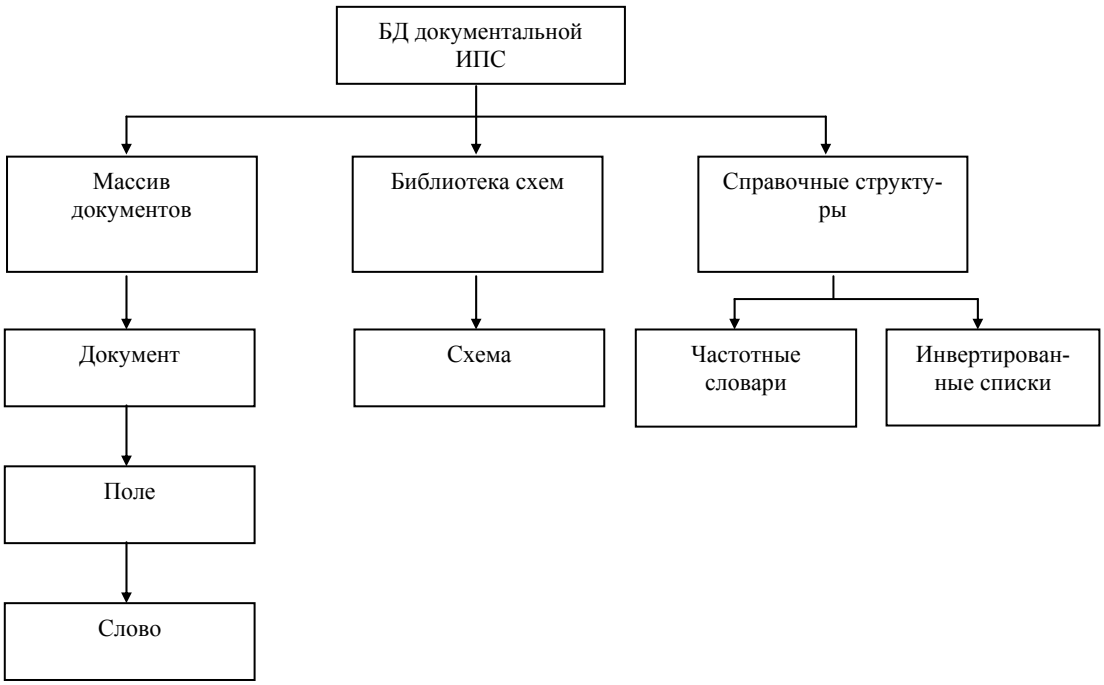


Рис. 3.11. Логическая структура БД документальной ИПС IRBIS

Документ базы данных как структурированная форма представления информации в общем случае определяется своим уникальным (в массиве документов базы данных) идентификатором и составом полей.

Поле как часть документа представляет собой однозначно идентифицируемый в информационном массиве фрагмент, для которого определены тип, имя и характер обработки.

Слово как фрагмент поля, выделяемый по некоторым формальным (заданным в схеме представления документа) правилам, является единицей информации в операциях поиска.

Схема базы данных (документа) определяет логическую связь именования, физического размещения и наполнения полей, образующих документ, а также стратегию поиска[28]. Особенностью этой реализации является логическая независимость схемы. Для одной базы данных может быть определено несколько разных схем, причем они в принципе равноправны, и в то же время одна и та же схема может быть использована для определения документов в нескольких базах данных.

[28] При этом представление всех или отдельных документов БД может быть доопределено контекстно средствами разметки. Использование встраиваемых определений структуры позволяет ввести «самоопределяемые» форматы представления документов. Это обеспечивает практически неограниченную гибкость при организации хранения коллекций разнородных документов, хотя и создает проблемы согласованного использования материала (из-за возможности различной интерпретации определений).

В схеме документ определяется совокупностью описаний отдельных полей, для каждого из которых задается:

- идентификация материала в базе данных, обеспечивающая пользователю доступ средствами документального поиска;
- представление материала при вводе и выводе (формат и длина поля, размещение и оформление материала при отображении и т.д.);
- спецификация стратегии документального поиска (прямое сканирование записей или использование инвертированных поисковых справочников). Для полей, специфицированных как ключевые, т.е. имеющих поисковые справочники, дополнительно определяются правила формирования дескрипторов (заданием списков символов-разделителей слов и списков стоп-слов).

Логически непрерывный массив документов базы данных в общем случае размещается в нескольких физических файлах, данные в которых связаны через указатель логического следования.

Справочник базы данных размещается отдельно от массива документов и имеет специализированную структуру (частотный словарь, алфавитный индекс и инвертированные списки). Поисковые справочники являются производным по отношению к массиву документов.

Физическая структура. В ИПС IRBIS используется следующая иерархия понятий, связанных с физической организацией информации.

База данных — некоторый объем файлового физического пространства для размещения данных, принадлежащих одной логической базе.

Файлы БД. Каждая база данных состоит не менее чем из двух типов файлов –файлов данных и файлов инвертированных структур. Отдельный файл может принадлежать только одной базе данных.

Экстент. Пространство для хранения данных в БД выделяется блоками (*экстентами*) по 8 следующих друг за другом страниц размером 8К. Экстент является единицей выделения пространства.

Страница. Файлы делятся на страницы размером по 8 Кбайт каждая. Логический номер страницы складывается из номера файла и номера страницы в файле (в простейшем случае логический номер равен номеру страницы в файле). В рамках БД файлы нумеруются, начиная с 1, и так же нумеруются страницы в рамках файла.

Словарные инвертированные структуры БД хранятся в отдельной области и представлены тремя типами страниц:

- индексные страницы;
- страницы текстового представления словарных структур;
- страницы инвертированных списков.

Страницы. На странице всегда хранится однородная информация. Все страницы имеют *заголовок*, в котором хранится общая информация, используемая ядром ИПС для работы со страницами всех типов:

- тип страницы;

- идентификатор (номер) страницы;
- идентификатор (номер) следующей страницы;
- идентификатор (номер) предыдущей страницы;
- число вхождений, размещенных на странице;
- длина фиксированной части вхождения.

Распределение пространства после заголовка зависит от типа страницы.

Индексные страницы. Индексные страницы содержат указатели на страницы текстового представления словарных структур. Каждая индексная страница содержит подзаголовок, хранящий следующую информацию:

- число вхождений (указателей), размещенных на странице;
- номер первой страницы инвертированных списков для множества страниц текстового представления словарных структур, описываемых индексной страницей.

За подзаголовком следуют указатели фиксированной длины, идентифицирующие отдельные страницы текстового представления словарных структур. В составе указателя следующая информация:

- метка сегмента (для представления общего словаря в виде объединения непересекающихся подмножеств);
- буква (символ), с которой начинается первое слово на странице;
- идентификатор (номер страницы).

Страницы текстового представления словарных структур. После фиксированного заголовка на странице следует подзаголовок, представляющий хранящийся на текущей странице фрагмент общего словаря. В состав подзаголовка входят:

- метка сегмента (для представления общего словаря в виде объединения непересекающихся подмножеств);
- номер первой страницы инвертированных списков;
- количество страниц инвертированных списков (для всех словарных структур, размещенных на текущей странице);
- размер свободного пространства;
- начало первого слова на странице (первые 4 буквы);
- начало последнего слова на странице (первые 4 буквы).

За подзаголовком следует карта размещения словарных структур, где для каждого отдельного вхождения фиксируются:

- длина слова (текстового выражения словарной структуры);
- кол-во документов (или длина инвертированного списка для словарной структуры);
- идентификатор страницы инвертированных списков, содержащей инвертированный список словарной структуры (по крайней мере, его начало);

- смещение начала инвертированного списка от начала списка страницы в целом.

Отдельный элемент карты размещения словарных структур располагается на пространстве страницы параллельно с самим текстовым выражением словарной структуры (длина текстового выражения при этом не должна превышать размера страницы за вычетом заголовочных областей). Размещение текстового выражения осуществляется по правилам лексикографической упорядоченности в рамках помеченного подмножества общего словаря и физически реализуется на странице снизу вверх. Тем самым свободное пространство на странице всегда представляет собой непрерывную область.

Страницы инвертированных списков. После фиксированного заголовка на странице следует подзаголовок, представляющий фрагмент инвертированных списков для некоторого подмножества словарных структур общего словаря. В состав подзаголовка входят:

- метка сегмента (для представления общего словаря в виде объединения непересекающихся подмножеств);
- номер первой страницы текстового представления словарных структур (для текущей страницы инвертированных списков);
- кол-во страниц текстового представления словарных структур (которым соответствует текущая страница инвертированных списков);
- размер свободного пространства.

За подзаголовком размещаются идентификаторы (физические номера) документов инвертированного списка. Под каждый номер отводится область фиксированного размера (этот размер указывается в поле «длина фиксированной части вхождения» общего заголовка страницы).

3.6. Уровневая модель представления информации в полнотекстовых БД

Как отмечалось ранее, важнейшими особенностями информационных систем, основанных на концепциях баз данных являются:

- многоаспектность представления[29] информации и, в том числе основанная на наиболее часто используемом принципе разделения физического и логического представлений;
- многократность и многофункциональность обработки (использования для решения различных прикладных задач) однажды сохраненной (созданной) информации и, как следствие, необходимость обработки запросов, заранее (при создании БД) не предусмотренных.

[29] Понятие представления здесь и далее рассматривается как форма существования информационных объектов, зависящая от особенностей среды и характера использования информации.

Однако для документальных систем важен и другой фактор, влияющий на эффективность работы человека с информацией - это форма хранения и представления (оформление) документа. Это особенно заметно при работе с объемными полнотекстовыми документами, когда от выбора формата (DOC, PDF, HTML и т.д.) зависит возможность дальнейшей обработки.

Для определения способов представления информации в полнотекстовых БД рассмотрим соотношение основных функциональных аспектов (базовых типов операций) АИС. Схема, представленная на рис. 3.12, отражает «перпендикулярность» функциональных направлений «поиск-выдача» и «ввод-запись-выборка», которые в организационно-функциональных решениях обычно выделяются в виде самостоятельных блоков. Здесь следует отметить, что в отличие от направления «поиск-выдача», подсистемы подготовки и ввода информации обычно организационно локализуются и включают в себя средства обработки ограниченного разнообразия, а функции хранения и выборки данных реализуются в одной среде – однажды выбранной СУБД.

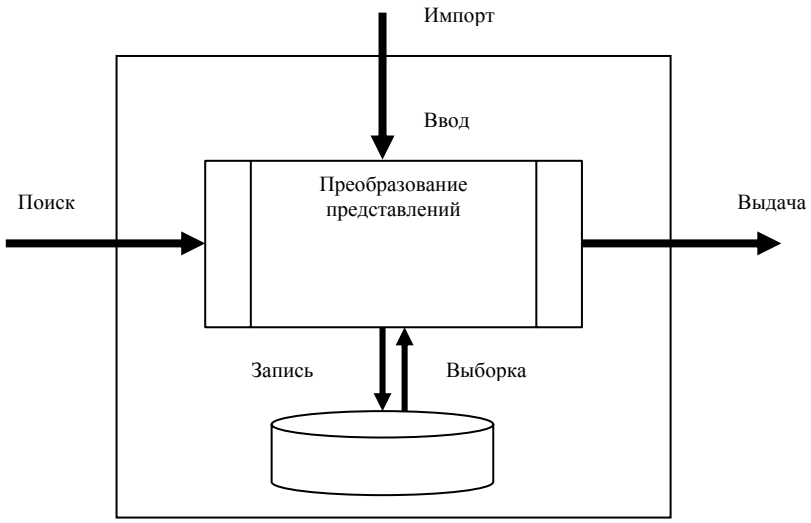


Рис. 3.12. Функциональные аспекты преобразований представлений информации

По каждому из этих функциональных направлений физические и логические представления различны, и в самом общем случае для каждого взаимодействия с системой требуются отдельные (в том числе и специализированные) представления информации: формы ввода запроса отличаются от форм представления содержания найденных по этому запросу документов; формы документов, используемые на этапе ввода данных в систему, отличаются от формы хранения введенных данных.

Соответственно, система должна обеспечить согласованную обработку представлений, используемых взаимодействующими сторонами,

т.е. форма поискового интерфейса должна согласовываться со структурой хранения и форматом выдачи документов.

И, как в случае уровневых схем теории баз данных, внешние представления должны определять семантику информационных объектов – свойства, атрибуты и характер последующего использования, а внутренние – типы данных и способы их идентификации.

Реально базы данных полнотекстовых документов не могут быть не гетерогенными, причем, в отличие от фактографических, полнотекстовым базам свойственна еще и разнородность физических представлений: документы, в силу разных причин, скорее всего, будут храниться в том формате, в каком они были получены. То есть, «разнородность» надо рассматривать в следующих аспектах:

1. Разнородность может проявляться как на уровне семантики (способов интерпретации величин), так и на структурно-форматном уровне (различных наборов и типов полей, образующих документ);
2. Разнородность может быть свойственна всем документам БД или отдельным. Т.е. «типизация» может быть определена либо для некоторого идентифицируемого подмножества документов, либо своя структура должна быть определена для каждого документа (например, с хранением схемы вместе с содержанием документа);
3. Для реализации разнородных БД может использоваться декларативный или процедурный способ определения структуры документа.

3.6.1. Преобразование представлений

Представление данных во многом определяет способы доступа к ним, возможности по преобразованию данных в другой формат с минимальными потерями содержания, а также оказывает значительное влияние на способы поиска и передачи данных. Логико-семантическая модель документальной информации должна учесть как «физический» аспект (многоуровневую вложенность разнородных компонентов), так и логику (семантику) использования информации - поиск документов по их вербальным и структурно-графическим компонентам, а также обеспечение навигации по документу.

Для полнотекстовых баз данных, для которых характерна разнородность как на уровне семантики, так и на физическом уровне согласованность представлений может быть обеспечена связыванием информационной и метаинформационной компонент.

В основу реализации процессов «декомпозиции-синтеза» данных[30] положена трехуровневая система следующих базовых информационных компонент:

- *элемент данных* – величина, представляющая в машинной форме логическую (семантически значимую) единицу информации. Обычно

[30] Сюда относятся и «стандартные» операции разбора и загрузки документа как взаимосвязанной совокупности элементов данных, а также выборка и верстка функционально ориентированных документов, как совокупности семантически упорядоченных полей.

представлена в вычислительной среде целостным физическим объектом и идентифицируется именем;

- *поле данных* – группа (последовательность) элементов данных, объединенных по какому-либо функциональному или семантическому признаку. Обычно представляет логически целостный объект, обеспечивающий полноту передачи контекстно-однородной информации;

- *документ* – структура, связывающая разнородные поля данных в соответствии с контекстом (или технологией) использования информации. Обеспечивает возможность адекватного восприятия содержания в целом: точность интерпретации значений полей, эффективность восприятия и понимания которых вне системы (обычно, человеком) обусловлена специфицируемой *структурой документа* - упорядоченной последовательностью соответствующим образом оформленного материала полей (верстkой документа).

Такой подход, отражающий в первую очередь семантику использования информации в сфере основной деятельности, имеет в своей основе логику, подобную логике управления данными: документ является упорядоченной совокупностью элементов данных, которая формируется в соответствии со схемой - *определением структуры*, задаваемой статически или динамически. Причем, в том случае, когда документальная система реализуется в среде универсальной СУБД (например, реляционной), наибольшая гибкость представления данных достигается при двухуровневой схеме определения структуры документа: поля определяются как композиция элементов данных средствами языка СУБД, а документ определяется как композиция полей средствами, внешними по отношению к СУБД (это могут быть средства языка программирования прикладной программы или генератора отчетов).

Однако в практике создания документальных БД оптимальность такого подхода далеко не очевидна. Действительно, можно выделить несколько критериев оптимизации, практически не связанных друг с другом. К таким критериям можно отнести количество элементов во внутрисистемной и внешних структурах документа, время или иные ресурсы, затраченные на преобразование документа из внешнего во внутрисистемное представление и обратно.

Выбор оптимального варианта модели в [Сысойкина2003] сводится к многокритериальной задаче теории принятия решений <T, A, K, X, F, G, D>,

где T – постановка задачи;

A – множество альтернатив;

K – множество критериев;

X – множество шкал оценок критериев;

F – отображение множества допустимых решений в множество предпочтений эксперта;

G – система предпочтений эксперта;

D – решающее правило.

В рамках поставленной задачи альтернативами будем считать варианты цепочек преобразований документа, получаемого из внешних источников, во внутрисистемное представление, и заканчивая генерируемыми документам, формируемыми системой на выходе. В самом полном варианте цепочка преобразований - операций декомпозиции-синтеза включает:

- разбор документа – выделение полей данных и элементов оформления;
- выделение элементов данных и преобразование их в формат внутренней схемы для загрузки в базу данных;
- преобразование выбираемых из базы элементов данных в формат внешней схемы;
- формирование документа – композиция полей данных и элементов оформления.

В качестве значимых выделим следующие семь критериев:

– число типов входных документов,

– число типов выходных документов,

– количество элементов данных во внутрисистемной структуре,

– количество элементов входной структуры, сопоставимых с элементами внутрисистемной структуры,

– количество элементов выходной структуры, сопоставимых с элементами внутрисистемной структуры,

– ресурсы, затраченные на преобразование входного документа к внутрисистемному представлению,

– ресурсы, затраченные на преобразование документа из внутрисистемного представления в выходной формат.

При этом первые пять критериев, определяющие эффективность системы с точки зрения управления документов, требуют качественной оценки, и должны принимать максимальные значения. Последние критерии, определяющие стоимость реализации конкретной альтернативы, могут быть явно выражены количественными характеристиками, причем их значение должно стремиться к минимуму. Они явно выражаются через функции преобразования документа путем суммированием затрат, произведенных на преобразование представления и структуры на каждом этапе.

При вычислении значений критериев времени используем следующее допущение. Время занесения одного элемента в память системы будем вычислять исходя из того, что в среднем операция считывания/записи в оперативной памяти, по крайней мере, на порядок быстрее операции чтения/записи на жесткий диск.

Так как мы имеем дело со слабоструктурированной моделью, в которой присутствуют и количественные, и качественные критерии, а число

альтернатив заранее известно, то можно использовать метод анализа иерархии, предложенный в [Саати1989].

Функции полезности каждой из альтернатив организации процессов преобразований были построены в предположении, что максимальное количество элементов не более числа элементов данных в форматах описания полнотекстовых документов и не превышает 250. Сравнительный анализ полученных зависимостей показал, что при числе элементов более 10 наиболее эффективным вариантом организации преобразований является полная цепочка и, соответственно, двухуровневая система определения структуры документа.

3.6.2. Структура полнотекстовой БД

Исходя из ранее приведенных положений, для хранения полнотекстовых документов используется двухуровневая схема представления.

В качестве базового средства представления полнотекстовых документов принята объектная модель (DOM). При этом, способы идентификации информационных элементов хранимых документов могут быть как контекстными (например, на основе XML), так и «декларативными» (традиционными для БД схемами). Принятие в качестве базовой DOM-модели позволяет использовать уже достаточно разнообразные языковые и программные средства, а также компоненты преобразования форматов, в том числе HTML, XML, RTF и т.д. С другой стороны, использование поэлементного представления информации, свойственного базам данных, обеспечивает гибкость обработки и эффективность поиска.

Отдельная запись БД соответствует физическому документу, который может содержать несколько логических документов. Структура записи определяется физической схемой БД и включает метаинформационную и информационную составляющие.

Метаинформационный компонент содержит помимо обязательных идентификационных несколько необязательно явно указываемых значений:

- формат физического документа (текст, XML, и т.д.);
- сведения о логической схеме документа (ссылка на схему);
- сведения о поисковых индексах.

Информационный компонент содержит материал документа и может быть представлен по-разному - в зависимости от возможностей используемой СУБД (например, двоичным полем, объектом, связанной записью, агрегатом полей и т.д.).

Логический документ может быть идентифицируемой частью физического документа, или композицией логических документов (динамический или «виртуальный» документ)

Структура документа может быть определена схемой документа отдельно от документа, или контекстно – поля документа и его структу-

ра могут быть определены, например, средствами XML внутри самого документа.

Исходя из требования запись-ориентированного доступа, определяемого архитектурой современных СУБД, можно определить два способа хранения полнотекстовых документов в базе данных.

Во-первых, можно выделить все значения узлов XML-документа и вносить их в базу поэлементно в соответствии со схемой БД, то есть значению XML-заголовка будет соответствовать имя поля, определенное в схеме БД. Однако в этом случае для документов, имеющих иерархическую структуру, возникают сложности при выдаче документа в первоначальном виде. Также проблемы могут возникать с сопоставлением имен полей и XML-заголовков (тэгов элементов).

Второй вариант – это хранение документа в виде «интегрального» массива, в данном случае в виде полноценного XML-документа. При этом возможность поэлементной выборки и обработки информации обеспечивается ассоциированной схемой базы данных, связывающей идентификацию полей в БД с элементами данных, определяемых XML-средствами.

Таким образом, в том случае, когда для хранения полнотекстовой информации используются базы данных, структура документов может быть определена двумя путями[31]:

1) так же как и для фактографических БД, заданием схемы – последовательности именованных типизированных полей данных;

2) контекстным определением – использованием специализированных языков разметки (например, HTML или XML), задающим индивидуальные особенности представления материала каждого документа.

Использование встраиваемых определений структуры позволяет ввести «самоопределяемые» форматы представления документов. Это обеспечивает практически неограниченную гибкость при организации хранения коллекций разнородных документов, однако создает проблемы семантические проблемы согласованного использования материала (из-за возможности различной интерпретации определений), что в свою очередь требует создания доступного всем пользователям репозитария метаинформации – описаний природы и способов представления информации.

[31] Для реляционной СУБД MS SQL Server 2000 реализован импорт/экспорт документов, представленных в XML-формате, в том числе с использованием схем сопоставления, определяющих соотношение элементов XDR-схем таблицам, а атрибутов – столбцам.

Контрольные вопросы

1. В чем различие между структурированной и слабоструктурированной информацией.
2. Определите понятие лингвистической переменной.
3. Приведите способ идентификации информации, ориентированный на ее поиск.
4. Дайте определение понятия «База данных».
5. Перечислите преимущества и недостатки использования баз данных.
6. Определите основные функции и назначение СУБД.
7. Назовите отличительные особенности использования баз данных в ИС.
8. Перечислите основные требования, предъявляемые к базам данных.
9. Определите назначение и организацию инвертированного списка.
10. Перечислите типы простых запросов.
11. Определите соотношение понятий «структура информации», «структура данных», «структура записи».

4. Модели поиска и оценки эффективности

Принятие решения о создании или вводе в эксплуатацию любой конкретной системы зависит не только от ответа на вопрос, способна ли она функционально обеспечить все поставленные перед ней задачи, но и от того, насколько эффективно она будет работать. При этом принято различать *экономическую* и *техническую (операционную) эффективность*.

Экономическая эффективность системы определяется денежными (или иногда – временными) затратами, необходимыми для выполнения данного набора задач.

Техническая эффективность информационной системы определяется ее способностью обеспечить потребителям требуемый уровень информационного обслуживания.

При комплексной оценке системы должны рассматриваться оба критерия, т.к. жизнеспособность системы в равной степени зависит и от качества выполнения тех или иных операций, и от стоимости их выполнения.

4.1. Оценка экономической и технической эффективности

По сравнению со средствами анализа технической (операционной) эффективности средства анализа экономической эффективности не так хорошо развиты. Это во многом объясняется тем, что получение точных данных о преимуществах усовершенствованных информационных служб в большинстве случаев невозможно. Кроме того, при определении затрат на информационные системы почти неизбежно сталкиваются с несравнимыми ситуациями, так как различия в затратах между такими системами, как автоматизированная и традиционная, не обязательно точно отражают значение каждой из систем. Автоматизированная система может, например, использоваться помимо задач информационного поиска и для решения других задач, или она может в отличие от традиционной системы функционировать 24 часа в сутки. Таким образом, в оценке экономической эффективности используется большое количество скрытых факторов, которые могут мешать конкретному анализу и привести к ненадежным или бессмысленным результатам.

И тем не менее вопрос анализа затрат необходимо рассматривать, так как маловероятно, чтобы информационные системы разрабатывались или внедрялись без каких-либо попыток оценить их потенциальную экономическую эффективность.

4.1.1. Экономическая эффективность

Принято различать анализ *экономической эффективности затрат* и анализ соотношения *затраты-выигрыш.* В первом случае требуется найти самые недорогие методы выполнения заданного набора операций или получить максимальные результаты при данных затратах. Во втором случае требуется систематическое сравнение стоимости выполнения отдельных операций и выигрыша, получаемого в результате их выполнения.

В стоимость системы обычно включаются *первоначальные затраты,* необходимые на разработку, испытания и оценку; *операционные затраты,* которые являются переменными и зависят от решаемых задач, участвующего персонала и количества требуемого оборудования; и, наконец, *фиксированные затраты* на аренду, налоги и другие стандартные расходы. Выигрыш, получаемый от усовершенствованной информационной системы, может быть связан либо с уменьшением затрат, либо с увеличением экономической эффективности, но чаще всего при замене ручных операций автоматизированными снижение затрат трудно документировать. Преимущества сложных информационных систем могут тогда состоять в улучшении возможности принимать решения, увеличении экономической эффективности, стимулировании исследовательских возможностей и тому подобном, причем значение всех этих факторов обычно невозможно точно учесть. Анализ эффективности затрат должен основываться на следующих положениях:

1. Должны быть четко определены цели системы.
2. Для достижения целей должны быть предусмотрены альтернативы.
3. Должна быть определена стоимость реализации альтернатив.
4. Должна быть создана модель, связывающая стоимость реализации альтернатив с целями, которые должны быть достигнуты.
5. Необходимо провести ранжирование альтернатив путем оценки для каждого случая затрат и ожидаемой эффективности.

Для случая информационного поиска, когда задан объем работы (количество документов, объем и стоимость документов, среднее число запросов и т. д.), основные альтернативы и выбор вариантов относятся к операциям индексирования и ввода документов, а также к процессам поиска и вывода информации.

Вообще говоря, любой критерий качества, например данный уровень точности, может достигаться многими различными способами, каждый из которых требует своего уровня затрат: так, точность можно повысить использованием высокоспецифичного языка индексирования, при этом необходима высокая квалификация индексаторов и большие затраты на индексирование; или же индексирование может проводиться менее тщательно, но результаты до передачи их потребителям долж-

ны просматриваться опытными экспертами, снижая тем самым стоимость индексирования, но увеличивая время поиска; и, наконец, можно переложить бремя на пользователей, предоставив им возможность самим вести поиск, уточняя формулировки запросов в надежде получить лучшие результаты.

Во многих случаях имеется возможность получить количественную информацию, связывающую различные системные альтернативы с эффективностью или качеством результатов, выдаваемых системой. В качестве примеров можно привести следующие зависимости:

- Зависимость между охватом массива и ожидаемым количеством выдач: значительный процент всех выдач приходится на очень небольшую долю документов массива, поэтому стоимость добавления к массиву большого числа менее продуктивных документов трудно оправдать с точки зрения улучшения результатов работы системы.

- Зависимость между временем индексирования и эффективностью поиска. Существует прямая связь между временем и полнотой индексирования и соответствующей ожидаемой полнотой поиска; к сожалению, при больших значениях полноты поиска потребляемые ресурсы для индексирования увеличиваются намного быстрее, чем полнота поиска, поэтому когда время или полнота индексирования превышают некоторый предел, вступает в действие закон уменьшающейся отдачи.

- Специфичность языка индексирования и баланс между полнотой и точностью. Обычно создание более специфичного языка индексирования обходится дороже и он дает более высокую точность, но он же может быть причиной ухудшения полноты; очевидно, что желаемый уровень точности и, следовательно, значение специфичности языка зависят от размера массива. При этом для очень больших массивов высокая точность является определяющей.

- Зависимость сложности средств автоматизации от ограничений в процессе обработки. Вообще говоря, разнообразные средства обработки позволяют обеспечить больший выбор выходных продуктов, например, упорядоченную выдачу документов. Однако более сложные средства обработки дороже по стоимости и по эксплуатации.

Даже если различные альтернативы можно надежно оценить количественно, может оказаться трудным принять оперативное решение, так как большие фиксированные затраты, связанные с реализацией, бывает нелегко возместить путем введения оплаты за предоставляемые услуги. До тех пор, пока не будет достигнуто согласие о ценности и выигрыше, которые дают информационные системы, анализ затрат не дает полного ответа, требуемого для принятия решений.

4.1.2. Техническая эффективность

В вопросе исследования технической эффективности информационных систем можно различить две точки зрения. Первая — это точка зрения пользователей, вторая — точка зрения администраторов системы. Однако в любом случае рассматривается основная функция ИПС – выдача информации в ответ на поисковый запрос.

Вопросы эффективности, интересующие администраторов, очень близки к следующим вопросам пользователей:

- Удовлетворяет ли система основным требованиям пользователей?
- Каковы основные причины невыдачи релевантных документов?
- Каковы основные причины выдачи нерелевантных документов?

Кроме того, администраторов интересуют расходы и доходы системы (то же интересует и пользователей, когда им приходится платить за обслуживание).

Из многих возможных показателей оценки технической эффективности, интересующих пользователей, основными принято считать следующие шесть [Солтон1979]:

1. *Полнота поиска,* т. е. способность системы выдавать все релевантные документы.
2. *Точность поиска,* т. е. способность системы отфильтровывать все нерелевантные документы.
3. *Усилия* (интеллектуальные или физические), затрачиваемые пользователями на формулирование запросов и просмотр выдаваемой информации.
4. *Время* с момента поступления запроса в систему до выдачи ответа.
5. *Форма представления* выдачи (что определяет дальнейшие возможности использования выданных материалов пользователями).
6. Полнота информационного массива в целом, т. е. степень *охвата* всех релевантных документов, которые могут быть интересны пользователям.

Упомянутые показатели оценки эффективности отражают и определяются эффективностью технических, лингвистических и организационных решений, заложенных в основу конкретной ИПС. Обобщенная схема взаимосвязи показателей эффективности и основных компонентов системы, как определяющих их факторов, приведена на рис. 4.1.

Показатели эффективности (за исключением полноты и точности) сравнительно легко измерить:

- затраты труда пользователей можно выразить через время, необходимое для формулирования запроса, диалога с системой и просмотра выданной системой информации;
- время реакции системы можно измерить непосредственно;
- форму представления выдачи можно оценить в процентном отношении к полному тексту.

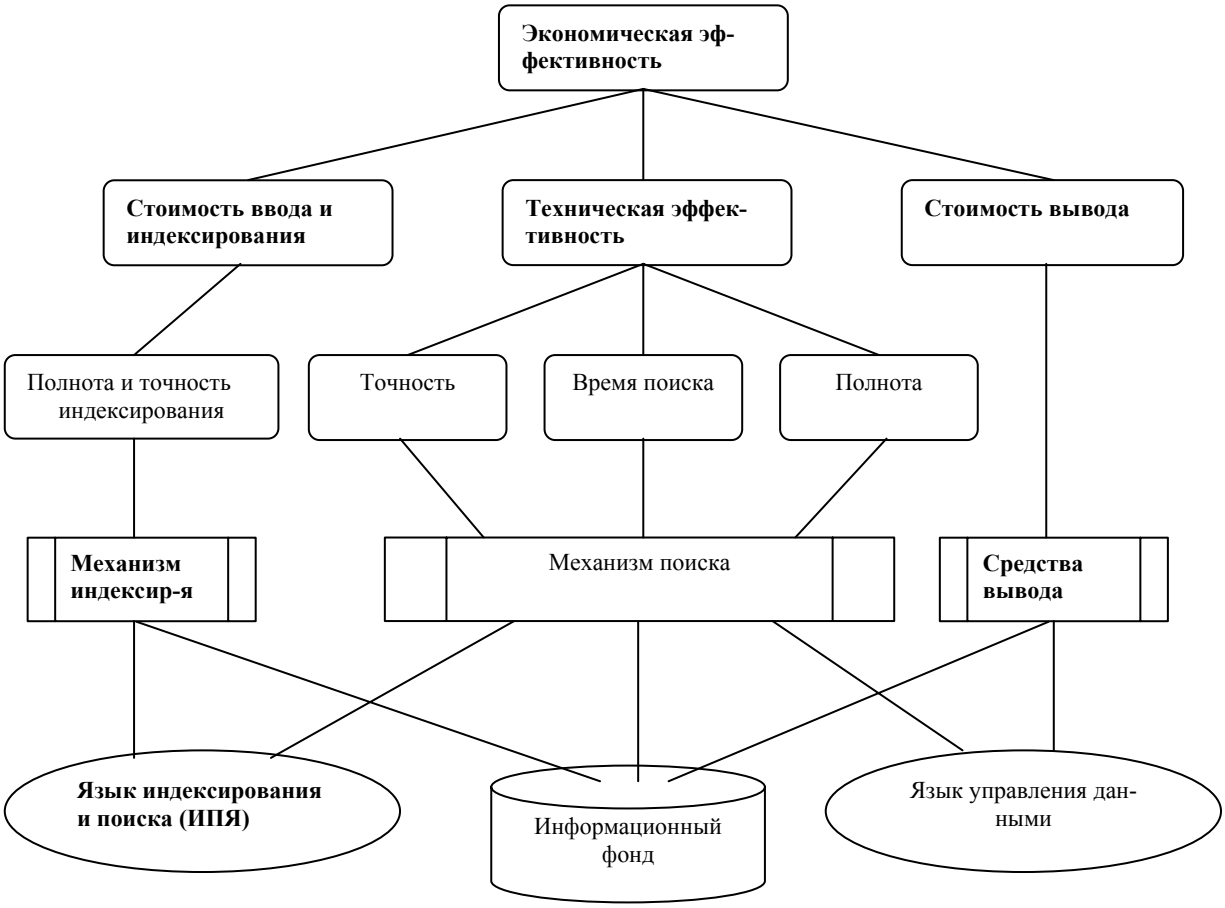


Рис. 4.1. Взаимосвязь показателей эффективности и компонентов ИПС

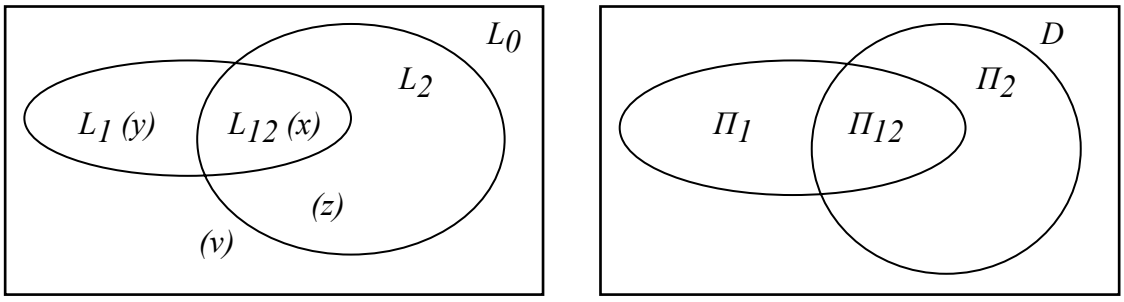
Определенные трудности может представить оценка степени охвата релевантных поисковой потребности документов, т.к. чаще всего заранее неизвестно общее количество материалов, потенциально представляющих интерес в данной предметной области.

Вычисление же мер полноты и точности является наиболее трудным как принципиально, так и практически.

4.2. Математические модели оценки технической эффективности

Для качественной оценки технической эффективности используется подход, предполагающий имитацию основных функций системы с помощью математической модели (с дальнейшей выработкой количественных критериев для оценки работы системы).

Для построения формальной модели критериев эффективности приведем диаграммы Эйлера-Венна возможных отношений между множествами терминов и/или документов информационного массива (рис.4.2).



а). б).

Рис. 4.2. Диаграммы Эйлера-Венна отношений множеств терминов/документов.
(а) - множества документов, (б) – множества терминов.

Здесь L_1 и L_2 - множества документов, L_{12} - их пересечение, L_0 . – множество документов информационного массива.

$П_1$ и $П_2$ - множества терминов (информационные профили соответствующих множеств документов, т.е. все значимые термины, хотя бы один раз встречающиеся во множестве документов), $П_{12}$ – пересечение информационных профилей, D – универсальный словарь.

Данные множества могут трактоваться следующим образом:

L_1 и L_2 - множества документов, связанных по общему термину;

$П_1$ и $П_2$ - списки терминов каждого из двух документов (термины, хотя бы раз встречающиеся в документах потока; или встречающиеся чаще чем некоторый порог f_{min}, или имеющие частоты, лежащие в интервале, задаваемом как f_{min} - f_{max}).

Рассмотрим случай, когда L_1 и L_2 - множества документов, связанных по общему термину. Выберем два произвольных термина T, t входящих в какие-либо документы из L_0.

Пусть L_1 - множество документов, содержащих термин T, L_2 - множество документов, содержащих термин t, тогда (Рис. *.1а):

$x = |L_{12}| = |L_1 \cap L_2|$ - количество документов, содержащих оба термина (T и t);

$y = |L_1 \setminus L_2|$ - количество документов, содержащих термин *T*, но не содержащих термин *t;*

$z = |L_2 \setminus L_1|$ - количество документов, содержащих термин *t*, но не содержащих *T*;

$v = |L_0 \setminus (L_1 \cup L_2)|$ - количество документов, не содержащих ни одного из терминов *T* и *t*.

$x + y + z + v = |L_0| = n_0$

Проблема оценки эффективности информационного поиска формально сходна с задачей сопоставления множеств документов и множеств терминов. Для измерения эффективности системы используются разностные меры множеств истинно релевантных (с точки зрения конечного потребителя) L^u и выданных (формально, с точки зрения системы, релевантных) L^c документов.

Рассмотрим так называемые *первичные координаты* описания выхода ИПС, представляющие соотношение выданных и не выданных множеств документов (диаграмму Эйлера-Венна <*L*> и таблицу сопряженности <*a,b,c,d*>).

Диаграмма <*L*> представляет соотношение множеств L_0 – всего информационного потока, L^u – множества истинно релевантных документов (т.е. документов, отвечающих информационной потребности пользователя) и L^c – множество документов, выданных системой в ответ на поисковый запрос (рис.4.3). Соотношение этих множеств и количественные оценки меры их близости могут характеризовать эффективность поискового механизма системы.

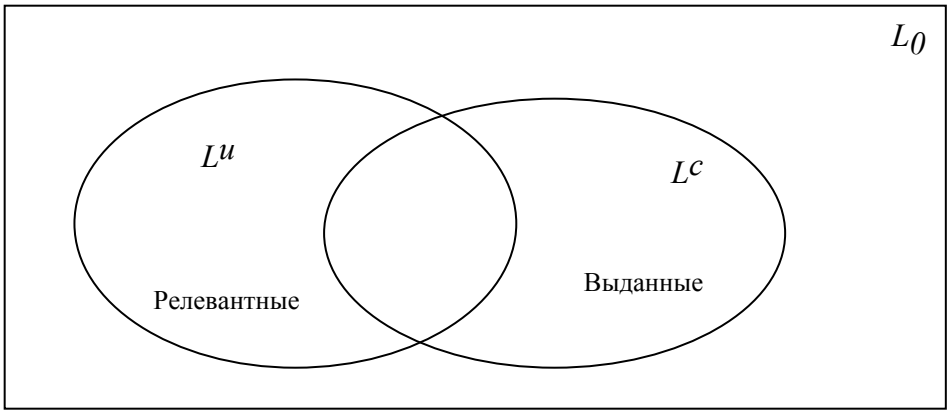


Рис. 4.3. Диаграмма <L>

Таблица сопряженности <*a,b,c,d*> отображает количественное соотношение выданных системой множеств релевантных (с точки зрения потребителя) и нерелевантных документов и не выданных множеств релевантных и нерелевантных документов (табл. 4.1).

Таблица 4.1.

Таблица сопряженности выдачи и релевантности

	Релевантные	Нерелевантные
Выданные (формально релевантные)	***a***	***b***
Не выданные	***c***	***d***

Очевидна следующая взаимосвязь представленных координат:

- число выданных релевантных документов: $a = x = \left|L^{и} \cap L^{c}\right|$;
- общее число релевантных документов: $a + c = x_0 = \left|L^{и}\right|$;
- количество выданных документов: $a + b = n = \left|L^{c}\right|$;
- общее число документов в L_0 : $a + b + c + d = n_0 = \left|L_0\right|$;
- число выданных нерелевантных документов: $b = n - x = \left|L^{c} \setminus L^{и}\right|$;
- число не выданных релевантных документов: $c = x_0 - x = \left|L^{и} \setminus L^{c}\right|$;
- число не выданных документов: $c + d = n_0 - n = \left|L_0 \setminus L^{c}\right|$;
- число нерелевантных документов: $b + d = n_0 - x_0 = \left|L_0 \setminus L^{и}\right|$;
- число не выданных нерелевантных документов:

$$d = n_0 - x_0 - (n - x) = \left|L_0 \setminus (L^{и} \cup L^{c})\right|$$

С приведенными первичными координатами связаны *частные критерии* оценки:

- *полнота* (доля выданных релевантных документов по сравнению с их общим количеством в информационном массиве):

$$r = \frac{a}{a + c} = \frac{x}{x_0} = \frac{\left|L^{и} \cap L^{c}\right|}{\left|L^{и}\right|}; \qquad (4.1)$$

- *точность* (доля релевантных документов среди выданных):

$$p = \frac{a}{a + b} = \frac{x}{n} = \frac{\left|L^{и} \cap L^{c}\right|}{\left|L^{c}\right|}; \qquad (4.2)$$

- *специфичность* (доля не выданных документов по сравнению с не выданными и выданными нерелевантными):

$$\sigma = \frac{d}{b + d} = 1 - \frac{n - x}{n_0 - x_0} = \frac{\left|L_0 \setminus (L^{и} \cup L^{c}\right|}{\left|L_0 \setminus L^{и}\right|}; \qquad (4.3)$$

- *общность* (или точность массива L_0), характеризует качество комплектования поискового массива (доля релевантных документов в информационном массиве):

$$p_0 = \frac{a + c}{a + b + c + d} = \frac{x_0}{n_0} = \frac{\left|L^{\text{Ц}}\right|}{\left|L_0\right|}. \qquad (4.4)$$

Каждая из переменных (4.1) - (4.4) изменяется в пределах от 0 до 1. Этот перечень может быть дополнен показателем относительного объема выдачи:

$$\nu = \frac{a+b}{a+b+c+d} = \frac{n}{n_0} = \frac{\left|L^c\right|}{\left|L_0\right|} \qquad (4.5)$$

Значение показателей технической эффективности во многом зависит от характеристик поисковых механизмов – моделей и методов отбора документов, а также возможностей используемого лингвистического обеспечения.

4.3. Модели механизмов информационного поиска в документальных БД

Модели поиска в диалоговой АИПС должны быть ориентированы на то, что реальная информационная потребность не удовлетворяется одним множеством документов, найденных по единственному запросу, а требует проведения серии отдельных поисков и выделения нужных фрагментов информации на каждой стадии развития запроса. Такие модели должны учитывать следующие факторы:

- поисковые запросы являются не статичными, а развивающимися (в том числе и с изменением представлений пользователя о предмете и задачах поиска);
- пользователь отбирает информацию итеративно, по частям, а не всю сразу в ответ на единственный запрос;
- пользователю доступны разнообразные поисковые методы, включая не только поиск по дескрипторам поискового запроса, но и, например, поиск документов по сходству;
- пользователь для работы с лексикой предметной области может применять широкий круг вспомогательных средств – тезаурусы, отраслевые рубрикаторы, словари и т.п.

Отвечающие этим требованиям модели будут намного ближе к реальному поведению человека, чем традиционная модель информационного поиска, требующая формулировки одного, пусть даже и точного, запроса, и может лучше управлять проектированием эффективных интерфейсов.

Для определения требований к поисковым механизмам рассмотрим АИПС как средство отыскания пользователем (субъектом поиска) решения находящейся в сфере его основной деятельности задачи P_i, ассоциируемой с системой понятий C_i , путем поиска документов, содержащих описание искомого решения. В этом случае процесс непосредственного решения задачи заменяется процессом поиска решения или методов его построения, полученных и *опубликованных* ранее. То есть, как показано на рис. 4.4, для получения решения задачи P_i , представляемой системой понятий C_i, необходимо найти множество документов D_i, используя в качестве поисковых (характеристических) признаков множество терминов T_i , представляющих понятия C_i.

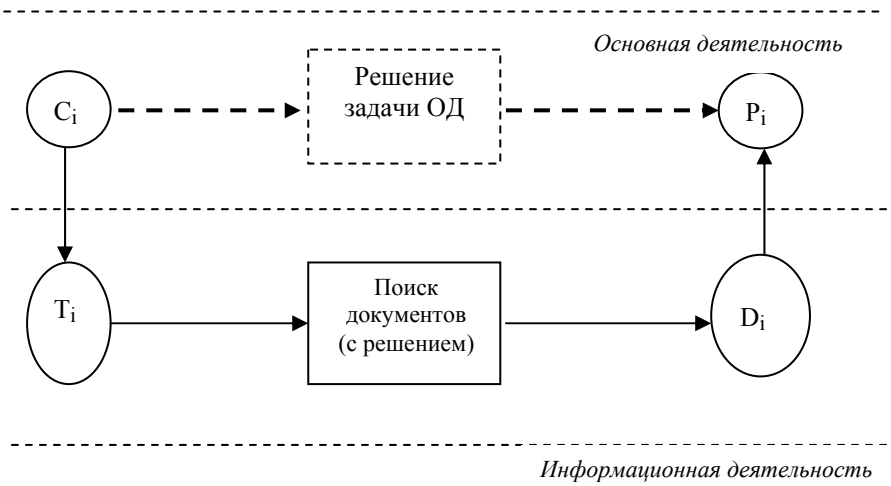


Рис. 4.4. Информационный поиск в процессе основной деятельности

Очевидно, что чем более структурирована и систематизирована предметная область основной деятельности (и система понятий как основа структуризации – отображение C_i<–>P_i), чем более устойчива терминологическая система (однозначность именования понятий и их композиций – отображение C_i<–>T_i), чем более «проработана» предметная область (полнота представления результатов ОД в пространстве документов – отображение D_i<–>P_i), тем более детерминированным должен быть механизм отбора S, реализующий отображение T_i<–>D_i . При использовании для индексирования нормализованной лексики (ключевых слов, обеспечивающих однозначность именования понятий) поиск эффективно реализуется на основе жесткой булевой логики.

«Нечеткость» приведенных отображений (конечно при требовании обязательного нахождения решения задачи) означает, что неточность любого из соответствий должна быть компенсирована увеличением полноты выдачи за счет уменьшения точности. Это может быть обеспечено следующими путями: 1) обогащением выражения запроса, 2) использованием менее жесткого механизма отбора, 3) использованием многоэтапных итеративных процедур поиска, обеспечивающих последова-

тельное расширение терминологического и документального пространства, например, по технологии обратной связи по релевантности. Обобщенная схема итеративного процесса поиска приведена на рис. 4.5.

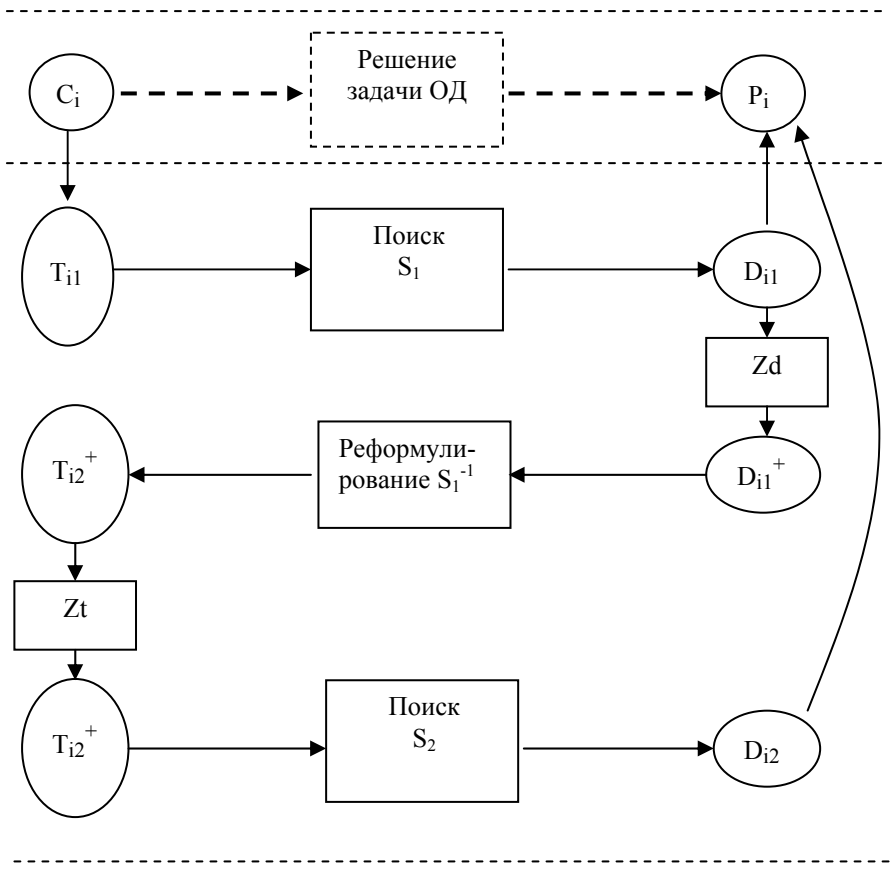


Рис. 4.5. Схема итеративного процесса поиска

Такая схема, обеспечивающая выполнение требования сопоставимости и оцениваемости результатов, включает в себя два типа обратной связи: внешнюю, отражающую оценку пользователя (выделение истинно релевантных документов Zd – отображение $D_i <-> D_i^+$), а также внутреннюю, учитывающую статистические особенности использования терминов в конкретной базе данных (выделение информативных для данной предметной области терминов – процесс S_1^{-1} , реализующий отображение $D_i^+ <-> T_i^+$). Отметим, что на схеме приведен и другой тип внешней обратной связи Zt, реализующий на уровне терминологии предметной области выделение информативных терминов – процедуру отображения $T_i <-> T_i^+$.

Результаты различных исследований, посвященных анализу методов и оценке эффективности поиска в интерактивных БД, позволяют сделать следующие выводы:

- стратегии, используемые при поиске информации, весьма разнообразны, и их анализ не позволяет однозначно выделить наиболее оптимальную;

- для повышения эффективности поиска поисковый аппарат должен включать развитые возможности как собственно поиска (использование контекстного расстояния, учет грамматических характеристик текстов и т. п.), так и отображения информации (форматы выдачи, удобные средства листания и пр.);
- полезно включение в состав лингвистического обеспечения ИС не только традиционных тезаурусов и рубрикаторов, но и дополнительных структур, являющихся результатом статистической обработки словарей БД.

Таким образом, для повышения эффективности поисковых процессов ИПС должна обеспечивать различные поисковые стратегии, позволяющие не только постоянно модифицировать исходный запрос, но и получать результаты, основываясь на статистической значимости лексики релевантных документов и на критериях, позволяющих искать «похожие» документы.

Далее будет рассмотрена совокупность моделей механизмов информационного поиска, в которых в качестве запроса используются как множества дескрипторов, так и множества документов (соответственно, модель механизма поиска по совпадению терминов, модели механизмов поиска по сходству), а также логические конструкции, построенные над этими множествами (модель механизма поиска по логическому выражению, модель механизма поиска с использованием контекстных операций, модель механизма поиска документов-аналогов).

Каждый из механизмов поиска предназначен для определенных типов БД, находится в соответствии с требованиями запросов и обладает уникальными достоинствами. В ИПС же особенно важно обеспечить возможность использования различных механизмов поиска (а также их комбинаций) для реализации всех типов поисковых задач.

4.3.1. Матрица «термин-документ»

В соответствии с [Попов1996] используем понятие универсального словаря D (прообразом которого может быть, например, тезаурус, рубрикатор, УДК), содержащего множество лексических единиц всего потока документов. Таким образом,

$l_i \subset D$ для всех i,

где l_i – совокупность лексических единиц некоторого документа (сообщения), который является элементом некоторого потока L:

$L = \{l_1, ... l_i, ... l_n\}, l \in L$

Аналогично универсальному словарю вводится понятие универсального потока (массива) L_0 (прообразы - поисковый массив ИПС, отраслевой справочно-информационный фонд, массив библиотеки), подмножеством которого являются все документы:

$L_0 = \{l_1, ...l_i, ...l_n\}, l_i \in L_0$ для всех i, причем $|L_0| = n_0$

где n_o - мощность множества L_0.

Линейное представление теоретико-множественного образа документа:

$$l_k = \begin{pmatrix} b_{1k} \\ \vdots \\ b_{ik} \\ \vdots \\ b_{Dk} \end{pmatrix}, \text{ где } b_{ik} = \begin{cases} 1, \text{если i - й т ермин входит в k - й документ} \\ 0, \text{если не входит} \end{cases}$$

Универсальный массив в линейном представлении есть матрица размерности $D \times n_o$:

$$L_0 = \begin{pmatrix} b_{11} b_{12} \cdots b_{1n_0} \\ \cdots\cdots\cdots\cdots \\ b_{i1} b_{i2} \cdots b_{in_0} \\ \cdots\cdots\cdots\cdots \\ b_{D1} b_{D2} \cdots b_{Dn_0} \end{pmatrix} \qquad (4.6)$$

Подобные матрицы известны под названием матрицы *«термин-документ»*. Каждый столбец матрицы соответствует документу и описывает множество терминов, содержащихся в нем. Таким образом, столбец матрицы характеризует поисковый образ документа (ПОД).

Строка матрицы соответствует отдельному термину и является перечнем документов, содержащих данный термин. Сумма элементов строки представляет собой частотную характеристику термина (F_i), присутствующую обычно в частотном словаре информационного массива:

$$F_i = \sum_{k=1}^{n_0} b_{ik}$$

Формализуем понятие *механизма поиска* как преобразователя ПОЗа, представленного с помощью матрицы L_0, в бинарный вектор результата Q (размерности n_0) и рассмотрим математическую интерпретацию основных поисковых механизмов.

4.3.2. Модель механизма поиска по совпадению терминов

При поиске по условию совпадения терминов в паре запрос-документ задается требование полного и/или частичного совпадение терминов (ключевых слов) для отбора документов, содержащих эти ключевые слова [Озкарахан1989]. Условие частичного совпадения можно задать, используя в терминах поискового образа так называемый несущественный символ – символ маскирования (обычно для этого используется знаки «*», «?» и «%»). Такие символы могут начинать тер-

мин, заканчивать его или находиться в середине, причем их может быть фиксированное или переменное число.

Формирование ПОЗа – это выбор из матрицы L_0 строк, соответствующих терминам, указанным в запросе. При этом, если некоторый термин не найден в словаре D, ему ставится в соответствие строка, состоящая из одних нулей (нулевая строка). Таким образом, для k терминов получаем подматрицу запроса (L_q), в которой отдельные строки могут быть нулевыми:

$$L_q = \begin{pmatrix} \cdots\cdots\cdots\cdots \\ b_{i_1 1} b_{i_1 2} \cdots b_{i_1 n_0} \\ \cdots\cdots\cdots\cdots \\ b_{i_2 1} b_{i_2 2} \cdots b_{i_2 n_0} \\ \cdots\cdots\cdots\cdots \\ b_{i_k 1} b_{i_k 2} \cdots b_{i_k n_0} \\ \cdots\cdots\cdots\cdots \end{pmatrix}$$

Построим результирующий вектор запроса:

$$Q = \left(\sum_{l=1}^{k} b_{i_l 1} \sum_{l=1}^{k} b_{i_l 2} \cdots \sum_{l=1}^{k} b_{i_l n_0} \right) \qquad (4.7)$$

Окончательный поисковый результат далее может быть сформирован по двум правилам: документ считается формально релевантным запросу, если содержит все k терминов, или документ считается формально релевантным запросу, если содержит хотя бы часть (один, два, три и т.д.) из k терминов.

При реализации первого правила получаем:

$$Q_k = (q_1 q_2 \cdots q_{n_0}), \text{ где } q_i = \begin{cases} 1, \text{если} \sum_{l=1}^{k} b_{i_l i} = k \\ 0 \text{ - в противном случае} \end{cases}$$

Для реализации второго правила зададим порог m, определяющий минимальное количество терминов (из k терминов запроса, $m \le k$), необходимое для отнесения документа к множеству формально релевантных запросу:

$$Q_k = (q_1 q_2 \cdots q_{n_0}), \text{ где } q_i = \begin{cases} 1, \text{если} \sum_{l=1}^{k} b_{i_l i} \ge m \\ 0 \text{ - в противном случае} \end{cases}$$

4.3.3. Модель механизма поиска по логическому выражению

Логическое выражение поискового условия – это синтаксическая конструкция языка, задающая порядок и способ вычисления величины, принимающей значение «0» или «1» («истина» или «ложь»).

В соответствии с правилами выражение представляет собой последовательность *операндов*, соединенных друг с другом знаками *опе-*

раций. Некоторые фрагменты выражения могут быть заключены в круглые скобки.

Нотация Бэкуса для такого выражения следующая:

<Выражение> ::=
<Операнд>|<Выражение><Операция><Операнд>|
<Операнд><Операция><Выражение> |
(<Выражение>)<Операция><Операнд>|
<Операнд><Операция>(<Выражение>)

В качестве *операнда* в поисковом выражении обычно выступают термины (дескрипторы), а в качестве *операции* – одна из логических операций AND (И), OR (ИЛИ), XOR (ИСКЛЮЧАЮЩЕЕ ИЛИ) и NOT (НЕ).

Первый этап вычисления логического выражения может состоять в построении двоичного дерева операций. Исходя из того, что все логические операции (кроме операции НЕ, которая, по существу, представляет собой инверсию исходного значения) являются бинарными, можно представить любое логическое выражение запроса в виде несбалансированного двоичного дерева, прохождение по которому снизу вверх приводит к получению результата.

В узлах такого дерева (рис.4.4), включая корневую вершину, расположены логические операции (o_i), а листья (конечные узлы) представляют собой строки матрицы L_0, соответствующие терминам запроса ($t_i = \left(b_{ij}, j = \overline{1, n_0}\right)$).

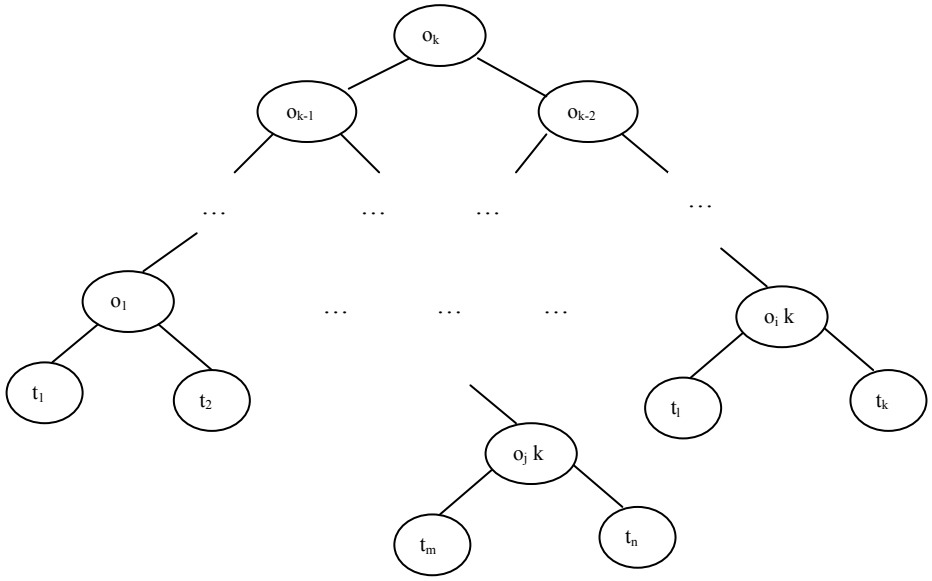


Рис. 4.4. Дерево структурных единиц запроса

Например, логическому выражению: $t_1 \wedge t_2 \vee t_3 \wedge t_4$, где t_i – термины запроса, соответствует двоичное дерево, приведенное на рис. 4.5.

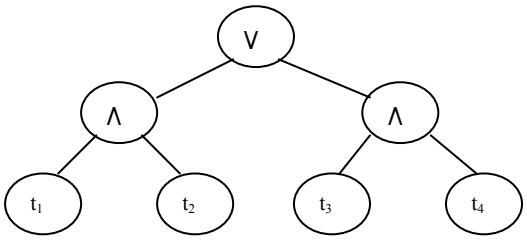


Рис. 4.5. Дерево логического выражения $t_1 \wedge t_2 \vee t_3 \wedge t_4$

Будем далее называть *операндом запроса* отдельно вычисляемое выражение, соответствующее поддереву запроса.

Рассмотрим расширенную матрицу «термин-документ» L_0', строки которой могут представлять собой не только показатели встречаемости терминов в документах информационного массива, но и результирующие векторы запросов (Q_i)

$$L_0{}' = \begin{pmatrix} b'_{11}b'_{12}\ldots b'_{1n_0} \\ b'_{21}b'_{22}\ldots b'_{2n_0} \\ \ldots\ldots\ldots\ldots \\ b'_{D'1}b'_{D'2}\ldots b'_{D'n_0} \end{pmatrix}, \text{ где } D' = D + K, \qquad (4.8)$$

K – количество включенных в матрицу результирующих векторов запросов,

$$\text{а } b'_{ij} = \begin{cases} b_{ij}, \text{ если строка принадлежит матрице } L_0 \\ q_{ij}, \text{ если строка представляет собой результат запроса} \end{cases}$$

Далее, поставим в соответствие каждой логической операции правило ее выполнения с использованием расширенной матрицы:

$$b'_i o_k b'_m = \left(b'_{ij} o_k b'_{mj}, j = \overline{1, n_0}\right), \qquad (4.9)$$

где o_k из множества бинарных логических операций:

$$o_k \in O, O = \{o_1, o_2, ..., o_s\} \qquad (4.10)$$

Для унарной операции NOT это правило реализуется следующим образом:

$$\neg b'_i = \left(\neg b'_{ij}, j = \overline{1, n_0}\right) \qquad (4.11)$$

Тогда алгоритм разрешения двоичного дерева поискового запроса состоит в последовательном выполнении снизу вверх логических операций и в пополнении на каждом шаге матрицы L_0 очередной строкой-результатом.

Условием выполнения *k*-той операции служит наличие в матрице L_0' строк, соответствующих правому и левому операнду. После выполнения *k*-той операции формируется результирующий вектор $q_k = b'_i o_k b'_m$, который становится $(D'+1)$-й строкой матрицы.

*Модель механизма поиска с использованием контекстных операций***.** При контекстном поиске указываются структурная единица (абзац, предложение) и/или расстояние между поисковыми терминами, которым должен удовлетворять документ. То есть, кроме определения взаимосвязи между терминами в том смысле, что они должны встречаться в какой-либо логической комбинации, используются и другие аспекты, свойственные естественным языкам, например, взаимное расположение терминов.

В этом случае поисковый алгоритм может рассматриваться как последовательное выполнение двух задач:

1) замена в поисковом запросе контекстных операций на логическую операцию AND, построение дерева запроса и выполнение алгоритма, описанного выше;

2) реализация на полученном в результате множестве документов контекстных операций путем непосредственного сканирования документов и вычисления координат терминов.

4.3.4. Модели механизмов поиска по сходству

Работа ИПС основана на использовании дескрипторов, которые лишь приблизительно описывают тематическое содержание документов и запросов. Поэтому обычно выдача в ответ на тематический запрос не бывает полной и точной.

Кроме того, широко распространенный поиск с использованием булевой логики имеет ряд недостатков. Наиболее существенный из них - плохая усваиваемость многими пользователями семантики булевых операторов и синтаксиса выражений. Поэтому ИПС, основное назначение которых – поиск релевантной информации с использованием тематических дескрипторов, для повышения эффективности поисковых процессов предлагают технологию обратной связи, типичная реализация которой, например, следующая [Robertson1986]:

- пользователь формирует список терминов для поиска, в результате которого выдается набор документов;
- документы в выдаче упорядочиваются в соответствии с некоторым алгоритмом взвешивания и ранжирования;
- пользователь просматривает выдачу, отмечая релевантные документы;
- после окончания просмотра система автоматически изменяет веса терминов и ранги документов в соответствии с информацией обратной связи.

Интеграция указанных процессов не совсем удобна, поскольку пользователю приходится инициировать порой довольно большое число булевых запросов, поэтому более технологично применять процедуры

автоматического или полуавтоматического расширения выражения запроса путем добавления терминов из релевантных документов.

Как видно, основная нагрузка при этом приходится на анализ весовых коэффициентов или различных мер близости терминов и документов.

В реальной практике информационно-поисковых систем, однако, основывать алгоритмы обратной связи на вычислении мер или функций сходства не всегда технологично, т.к. расчеты иногда требуют больших вычислительных и информационных ресурсов.

Другой путь – автоматическое расширение *выдачи,* т.е. предложение системой пользователю документов, похожих на ранее выбранные им и отмеченные как релевантные. Сходство в данном случае можно интерпретировать как расстояние между двумя информационными потоками: чем более похожи эти потоки, тем они ближе. Отличие такого способа заключается в том, что управление полностью осуществляется системой, а не пользователем.

Далее рассмотрим модели, которые позволяют реализовать механизм поиска документов по сходству.

4.3.4.1. Модель механизма поиска документов-аналогов

Аналогами некоторого документа назовем такие документы информационного массива, которые имеют заданное количество общих терминов с исходным, т.е. функция «похожести» реализуется простой количественной оценкой документов с точки зрения наличия в них терминов из некоторого подмножества словаря.

Источником для поиска аналогов служит некоторый отдельный документ информационного массива. Задача функции - отыскать «похожие» на него документы.

Выделим в матрице L_0 столбец $l_k = \left(b_{ik}, i = \overline{1, D}\right)$, соответствующий ПОДу рассматриваемого документа, и построим подматрицу L_{Doc}, оставив в матрице L_0 только те строки, в которых $b_{ik} \neq 0$. Далее, по матрице L_{Doc} строится результирующий вектор запроса на поиск аналогов (Q_{Doc}) и, аналогично п.2.1, может быть получен поисковый результат с учетом (или без) некоторого заданного порога «близости» (m).

В случае, когда универсальный словарь представляет собой набор отдельных словарей D_i, построенных по лексике отдельных структурных единиц документов (например, полей), процедура поиска аналогов может быть усложнена заданием различных пороговых значений для структурных единиц и построением логического выражения над множеством критериев отбора, связывающих поле и соответствующее пороговое значение. Например, поиск библиографий-аналогов может быть сформулирован следующим образом: найти документы, где в библиографии встречается хотя бы одна из фамилий авторов исходного доку-

мента, и, по крайней мере, две тематические рубрики, общие с исходным документом.

Рассмотрим реализацию процедуры поиска аналогов для случая:

$$D = \sum_{i=1}^{n} D_i \ , \ L_{D_i} = \begin{pmatrix} b_{11}^i b_{12}^i \dots b_{1n_0}^i \\ b_{21}^i b_{22}^i \dots b_{2n_0}^i \\ \dots\dots\dots\dots \\ b_{D_i 1}^i b_{D_i 2}^i \dots b_{D_i n_0}^i \end{pmatrix}$$

Тогда ПОД заданного документа представляет собой объединение ПОДов, построенных для различных структурных единиц:

$$l_k = \bigcup_{i=1}^{n} l_k^{D_i}, \ \ l_k^{D_i} = \left(b_{lk}, l = \overline{1, D_i}\right),$$

а подматрица аналогов - соединение подматриц:

$$L_{Doc} = \begin{pmatrix} L_{Doc}^{D_1} \\ L_{Doc}^{D_2} \\ \vdots \\ L_{Doc}^{D_n} \end{pmatrix}.$$

Построим матрицу результирующих векторов $Q_{Doc} = \left(q_{ij}, i = \overline{1,n}, j = \overline{1,n_0}\right)$, где каждая строка представляет собой результирующий вектор одной из подматриц с учетом заданного порога близости:

$$Q_i = \left(q_1^i q_2^i \cdots q_{n_0}^i\right), \text{где } q_l^i = \begin{cases} 1, \text{если} \sum_{l=1}^{D_i} b_{lj}^i \geq m_i \\ 0 \text{ - в противном случае} \end{cases}$$

Используя теперь матрицу Q_{Doc} вместо матрицы L_0 в модели поиска по логическому выражению, можно выполнять процедуры построения дерева запроса с последующим вычислением результата.

4.3.4.2. Модель механизма эвристического поиска

Эвристический поиск работает по принципу отыскания документов, «похожих» на усредненный «тематический» образ некоторого множества релевантных документов, указанных пользователем, и реализуется следующей последовательностью шагов:

Шаг 1. Построение словника по массиву релевантных документов.

Результатом этого шага является подматрица L_{Rel} матрицы L_0, построенная путем выбора столбцов, характеризующих заданные пользователем документы:

$L_{Rel} = \left(b_{ij_k}, i = \overline{1,D}, k = \overline{1,n}, 1 \leq j_k \leq n_0\right)$, n – количество документов, отмеченных пользователем как релевантные.

Шаг 2. Оценка терминов словника и построение Поискового Образа Темы (ПОТ).

Результатом оценивания должно быть выделение только тех терминов, которые могут быть включены в ПОТ. Желательно, чтобы в основе формальной оценки лежали частотные характеристики, которые могут быть получены из матриц L_0 и L_{Rel}:

$F_i = \sum_{j=1}^{n_0} b_{ij}$ (или i-тый элемент главной диагонали матрицы $L_0 \times L_0^{\ T}$),

$F_{i\,Rel} = \sum_{j=j_1, j_2, \ldots, j_k} b_{ij}$ (или i-тый элемент вектора $L_0 \times Q_{Rel}$),

где F_i – частота термина в информационном массиве, F_{iRel} – частота термина в множестве релевантных документов, Q_{Rel} – вектор релевантных документов (строка расширенной матрицы L_0').

Например, для оценки степени соответствия термина ПОТ может быть использована мера точности термина - отношение частоты термина в множестве релевантных документов к частоте термина в информационном массиве, а в качестве порога для отбора в ПОТ – относительный коэффициент C_R, вычисляемый в зависимости от эвристического параметра n_S, характеризующего количество ожидаемых документов (т.е. максимальное количество документов результата поиска). С другой стороны, эвристический параметр характеризует минимальную (ненулевую) точность термина, возможную в ожидаемой выдаче:

$$C_R = \frac{1}{n_s}.$$

Тем самым, в ПОТ отбираются термины, для которых выполняется неравенство:

$$\frac{F_{i\,Rel}}{F_i} \geq C_R \qquad (4.12)$$

Шаг 3. Построение матрицы «термин-документ» для функции поиска аналогов.

На этом шаге из матрицы L_{Rel} должны быть удалены строки, для которых не выполняется неравенство (4.12). В результате получаем матрицу $L_{ПОТ}$:

$L_{ПОТ} = \left(b_{ij_k}, i = \overline{1,M}, k = \overline{1,n}\right)$, где M – количество терминов в ПОТ, определяющее порог «близости» для следующего шага.

Шаг 4. Выполнение функции поиска аналогов с пороговым значением M.

По матрице $L_{ПОТ}$ строится результирующий вектор запроса на отбор документов-аналогов ($Q_{ПОТ}$) и формируется поисковый результат с учетом порога близости M. Если число документов полученного результата меньше, чем заданное в системе n_S, то пороговое значение M уменьшается на 1, и повторяется процедура поиска аналогов с новым

пороговым значением. Таким образом, на каждой *i*-ой итерации пороговое значение равно *M–i.*

Цикл заканчивается в одном из двух случаев: либо после выполнения очередной итерации число документов результата стало равно или превысило значение n_S , либо пороговое значение стало равно 0.

4.3.4.3. Модель механизма поиска с использованием обратной связи по релевантности терминов

Обратная связь по релевантности на уровне отдельных терминов должна обеспечить пользователю возможность целенаправленно изменять поисковый запрос путем повышения роли одних и понижения роли других терминов, не вникая в тонкости составления запроса, определяемые особенностями документального массива и ИПС. При этом процесс поиска обычно разбивается на последовательность несложных шагов, ведущих к поставленной цели.

В рамках модели (в соответствии с которой определяется обратная связь) существуют различные стратегии изменения весовых коэффициентов терминов, предлагаемых системой для расширения запроса, на основании информации о релевантности/нерелевантности выданных документов.

Рассмотрим диалоговую модель механизма поиска по обратной связи, предлагаемую в ИПС IRBIS.

Диалоговая модель поиска «по обратной связи» отличается от модели эвристического поиска тем, что после выполнения системой очередного шага пользователю предоставляется возможность управлять дальнейшим процессом формирования результата, т.е. последовательность шагов в диалоговой модели дискретна и реализуется (с точки зрения продолжительности) в зависимости от предпочтений пользователя.

Шаг 1. Построение и ранжирование словника релевантных документов.

Результатом этого шага является вектор $W = \left(w_i, i = \overline{1,k}\right)$, где k – количество терминов релевантных документов, а w_i, соответственно, значение весового коэффициента для *i*-го термина, удовлетворяющее неравенству $w_i \geq w_{i+1}$.

Расчеты весовых коэффициентов могут основываться на различных мерах близости и на этом шаге не влияют на количество выдаваемых пользователю терминов (пользователь в данном случае получает оценку всех терминов релевантных документов, которые находятся в частотном словаре, т.е. в ПОТ попадают все термины без исключения).

По завершении первого шага система передает управление пользователю, который самостоятельно (основываясь на выданных ему значениях весовых коэффициентов и упорядоченности терминов релевант-

ных документов) отмечает термины, способные улучшить поисковый запрос.

Отмеченные термины пользователь далее может самостоятельно добавить в поисковый запрос (для реализации моделей поиска по совпадению терминов или по логическому выражению) или инициировать второй шаг поиска по обратной связи.

Шаг 2. Формирование матрицы поисковых результатов.

Термины, отобранные пользователем на предыдущем шаге, рассматриваются как исходные для проведения поиска по совпадению терминов. Модель этого механизма поиска реализуется в данном случае построением подматрицы запроса (L_q), в которой отдельные строки могут быть нулевыми.

Рассмотрим теперь подматрицу L_q как исходную для проведения процедуры поиска аналогов и последовательно для каждого ненулевого столбца построим вектор Q_i – результат поиска аналогов с максимальным порогом близости (максимальный порог близости задается количеством единиц в столбце, а контекст результата задается перечислением самих терминов). Полученные векторы рассмотрим как строки матрицы поисковых результатов:

$$Q_{Theme} = \left(q_{ij}, i = \overline{1,n}, j = \overline{1,n_0}\right),$$

где n – количество ненулевых столбцов подматрицы L_q .

Отметим, что каждая строка сформированной таким образом матрицы снабжается контекстом – перечислением конкретных терминов, присутствующих в документах конкретного результата. Удалив из матрицы строки с одинаковым контекстом, получим кластеризованное пространство документов, где каждый кластер задается не только количеством терминов запроса, но и составом самих терминов.

Матрица поисковых результатов Q_{Theme} дает возможность обеспечить доступ к каждому отдельному результату для его просмотра и последовательного формирования нового множества релевантных документов.

4.4. Пример использования различных поисковых механизмов и оценка эффективности результатов

Рассмотрим примеры применения различных поисковых механизмов и проведем оценку эффективности поисковых стратегий на материале БД ВИНИТИ РАН «Информатика» с использованием ИПС IRBIS.

Сформулируем поисковые запросы для отбора документов, отнесенных к рубрике Рубрикатора ВИНИТИ 201.23.17.03 «Структура массивов. Формирование массивов и баз данных».

Формализуем поисковую задачу следующим образом: пусть необходимо отобрать документы этой рубрики, используя поисковые образы документов, представленные полем ключевых слов. В этом случае мно-

жество истинно релевантных документов определяется как множество документов, заиндексированных экспертами данной рубрикой (1469 документов), а множество выданных документов формируется как результат отбора в соответствии с логическим выражением, операндами которого служат ключевые слова.

Использование механизма поиска по совпадению терминов. Для реализации стратегии поиска по совпадению терминов зададим поисковый запрос, включив в него термины «массивы», «базы данных», «формирование», «структура».

Поиск по совпадению терминов с пороговым значением 2 (т.е. поиск документов, имеющих не менее 2-х общих терминов с запросом) дает в результате 102 документа, из которых истинно релевантных – 33. В табл. 4.2 представлены результаты вычисления полноты и точности.

Таблица 4.2.

Стратегия	**Полнота**	**Точность**
Поиск по совпадению терминов (порог 2)	0,02246	0,32353
Поиск по совпадению терминов (порог 1)	0,52553	0,15028
Поиск по логическому выражению	0,02246	0,34375
Поиск по совпадению терминов с маскированием (порог 2)	0,03744	0,36667
Поиск по совпадению терминов с маскированием (порог 1)	0,5488	0,14054
Поиск по логическому выражению с маскированием	0,03744	0,39287

Поиск по совпадению терминов с пороговым значением 1 (т.е. поиск документов, имеющих хотя бы один общий термин с запросом) дает в результате 5137 документа, из которых истинно релевантных – 772. Легко заметить, что увеличение полноты поиска при этом влечет за собой понижение показателя точности.

Использование механизма поиска по логическому выражению. Стратегия поиска с использованием булевой логики предполагает построение как можно более точного выражения запроса с применением лексики предметной области.

Ориентируясь на ту же лексику, что и в предыдущем случае, сформулируем выражение запроса с использованием логических операторов И и ИЛИ:

(KW:массивы ИЛИ KW:'базы данных') И (KW:формирование ИЛИ KW:структура)

Поиск по логическому выражению в результате дал 96 документов, 33 из которых оказались релевантными (те же, что и при поиске по совпадению терминов с пороговым значением 2). При том же значении полноты, что и в случае использования предыдущей стратегии, был получен лучший показатель точности.

В приведенных примерах была использована нормализованная лексика, в точности совпадающая с лексикой названия рубрики. Улучшать показатели эффективности в данном случае можно, только путем модификации логического выражения.

Рассмотрим далее результаты обеих стратегий в случае использования аппарата маскирования терминов запроса.

Предложение запроса для стратегии поиска по совпадению терминов имеет следующий вид:

KW:(массив* и баз*данн* и формирован* и структур*)

Поиск по совпадению терминов с пороговым значением 2 в данном случае дает в результате 150 документов (55 релевантных), а с пороговым значением 1 – 5735 документов (806 релевантных).

Использование маскирования в стратегии поиска по логическому выражению привело к следующему результату: всего найдено 140 документов, из них 55 релевантных. Из табл.4.2 видно, что маскирование терминов повышает показатели и полноты, и точности.

Применение стратегий, основанных на вводе терминов поискового запроса, во многом зависит от полноты и точности отдельных терминов в рамках информационного массива и не может вывести пользователя за пределы используемой в запросе лексики. Рассмотрим применение стратегий, основанных на поиске по некоторым формальным признакам документов, «похожих» на уже найденные релевантные.

Использование механизма поиска документов-аналогов. При просмотре релевантных документов функция поиска документов-аналогов может вывести пользователя на новую лексику – показать множество терминов, которое можно использовать для поиска новых релевантных документов. Например, инициировав поиск аналогов для документа, представленного на рис. 4.6, найдем релевантные документы, которые не могли быть найдены при использовании вербальных стратегий, т.к. не содержат терминов исходного запроса. Тем самым, механизмы поиска «похожих» документов позволяют не только увеличить показатель полноты поиска, но и выводят пользователя на новый неиспользованный в запросах пласт лексики предметной области.

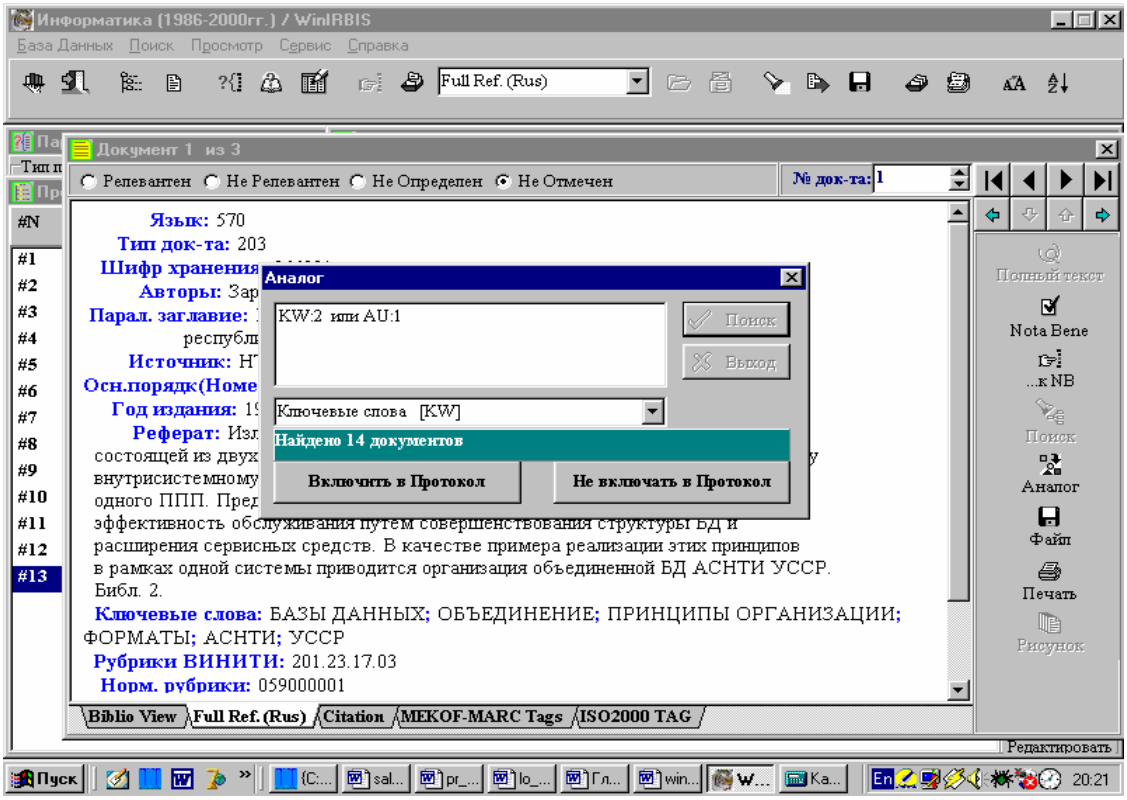


Рис.4.6. Поиск документов-аналогов

4.5. Информационно-поисковый язык документальной ИПС

Повышение эффективности поиска обычно связывают с совершенствованием поискового аппарата, основу которого составляет лингвистическое обеспечение, и, в частности, с ИПЯ, а также с созданием интерфейсных средств, обеспечивающих адекватность их использования в соответствии с характером объекта поиска и когнитивным состоянием потребителя.

Язык, как унифицированное средство выражения семантики сообщений и вопросов (контекста), является основным компонентом для понимания процесса поиска информации, поэтому перспективным представляется подход к информационному поиску, сконцентрированный и сосредоточенный на реальных пользователях с их реальными информационными потребностями и, как следствие, на индивидуализации представлений семантических объектов. Один из наиболее важных аспектов решения этих задач – построение языка как системы представления, позволяющей адекватно описывать и идентифицировать как потоки и массивы информации, так и отдельные документы или запросы.

Информационно-поисковый запрос с точки зрения документальной ИПС (на примере ИПС IRBIS[32]) представляет собой совокупность отдельных предложений запроса, в общем случае синтаксически и се-

[32] Рассматриваемый ИПЯ является расширенной реализацией типового языка запросов, свойственного большинству современных документальных ИПС.

мантически не связанных между собой. Однако, само понятие «Запрос» предполагает объединенную общей тематикой последовательность поисковых действий, направленных на получение обобщенного результата, что позволяет разрешать ссылки на результаты отдельных предложений в рамках текущего запроса, объединять поисковые результаты, выделять общее множество релевантных документов и т.п.

4.5.1. Предложение запроса

Структурной единицей Запроса в рассматриваемом ИПЯ является *Предложение запроса*.

Синтаксис *Предложения запроса* в нотациях Бэкуса-Наура следующий:

<Предложение запроса> ::= <Условие поиска> |

<Предложение запроса><Логическая операция><Предложение запроса>|

(<Предложение запроса><Логическая операция><Предложение запроса>)

<Логическая операция> ::= И | AND | ИЛИ | OR | , | НЕ | NOT | ^

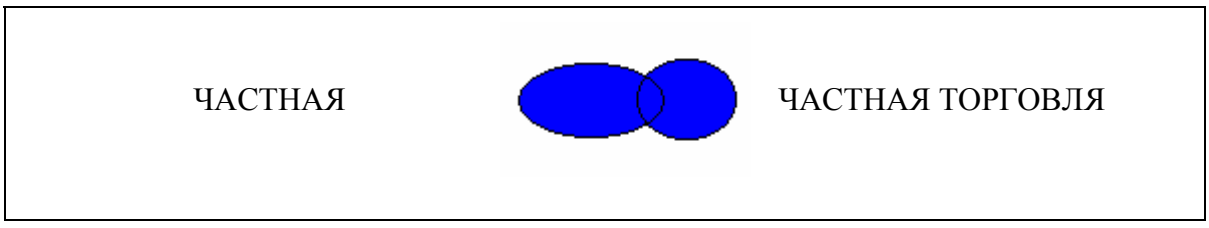
Предложение запроса в общем случае состоит из произвольного числа *Условий поиска*, связанных логическими операциями И (AND, «пробел»), ИЛИ (OR, «,») и НЕ (NOT, «^»). Внутри предложения допускается использование скобок, задающих дополнительные приоритеты выполнения условий поиска.

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

- **OR (ИЛИ)** – например:

KW:**(‘ЧАСТНАЯ СОБСТВЕННОСТЬ’ OR ‘ЧАСТНАЯ ТОРГОВЛЯ’)**

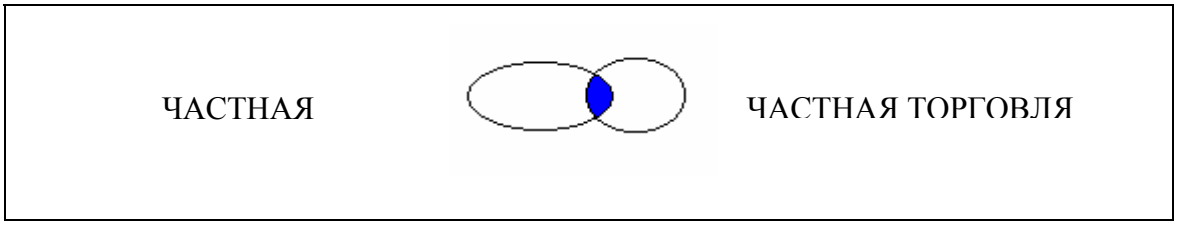
означает, что в результаты поиска включаются все документы, в которых в поле **KW** встречаются термины (словосочетания) «ЧАСТНАЯ СОБСТВЕННОСТЬ» или «ЧАСТНАЯ ТОРГОВЛЯ» или оба вместе:



- **AND (И)** –например:

KW:(‘**ЧАСТНАЯ СОБСТВЕННОСТЬ’ AND ‘ЧАСТНАЯ ТОРГОВЛЯ’)**

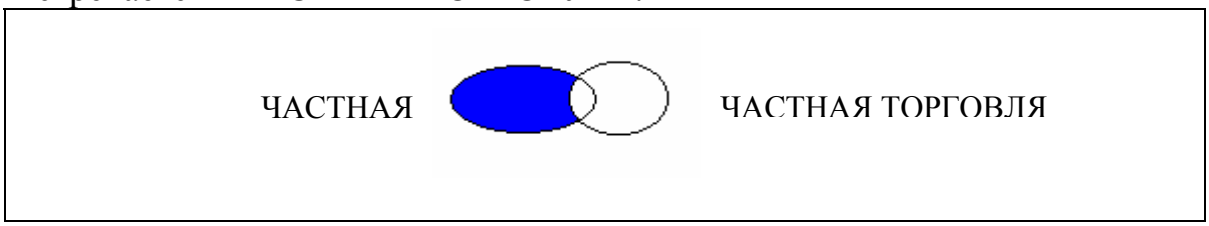
означает, что в результаты поиска включаются только те документы, в которых в поле **KW** встречаются оба термина «ЧАСТНАЯ СОБСТВЕННОСТЬ» и «ЧАСТНАЯ ТОРГОВЛЯ».



- **NOT (НЕ)** – например:

KW:(‘**ЧАСТНАЯ СОБСТВЕННОСТЬ’ NOT ‘ЧАСТНАЯ ТОРГОВЛЯ’)**

означает, что в результаты поиска включаются документы, в которых в поле **KW** встречается термин «ЧАСТНАЯ СОБСТВЕННОСТЬ» и не встречается «ЧАСТНАЯ ТОРГОВЛЯ».



4.5.2. Условие 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>

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

Выражение условия – набор терминов (поисковых дескрипторов), 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 предложение запроса состоит из имени поля, оператора вхождения и одного дескриптора, например:

KW : РОССИЯ

Область поиска. Область поиска задается именами структурных единиц документа – полей.

<Область поиска> ::= <Имя поля>|

(<Область поиска> <Логическая операция> <Область поиска>)

Из нотации видно, что допускается использование логических операций 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.

Отличительная особенность представляемого ИПЯ – возможность формирования логического выражения как в правой, так и в левой части условия поиска.

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 терминами запроса используются следующие операторы критерия (вхождения, сравнения):

<Оператор критерия> ::= : | = | EQ | <> | NE | > | GT | >= | GE | < | LT | <= | LE

- **”:”** (условие вхождения) – позволяет найти документы, которые содержат в указанной области поиска результат вычисления выражения условия;

- **”=”** (условие "равно", или **EQ**) - позволяет найти документы, для которых указанная область поиска равна результату вычисления выражения условия;

- **”<>”** (условие "не равно", или **NE**) - позволяет найти документы, которые не содержат в указанной области поиска результат вычисления выражения условия;

- **”>”** (условие "больше", или **GT**) - позволяет найти документы, которые содержат в указанной области поиска значения больше, чем результат вычисления выражения условия;

- **”>=”** (условие "больше или равно", или GE) - позволяет найти документы, которые содержат в указанной области поиска значения больше или равные результату вычисления выражения условия;

- **”<”** (условие "меньше", или **LT**) - позволяет найти документы, которые содержат в указанной области поиска значения меньше, чем результат вычисления выражения условия;

- **”<=”** (условие "меньше или равно", или **LE**) - позволяет найти документы, которые содержат в указанной области поиска значения меньше или равные результату вычисления выражения условия.

Выражение условия. **Синтаксис выражения условия в ИПЯ 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 и помещены в круглые скобки.

Контекстные операторы – это оператор расстояния (NEAR[N]), оператор расстояния со строгим следованием (CTX[N]), оператор предложения (SENT) и 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 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 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 SENT позволяет найти документы, в заданной области поиска которых поисковые дескрипторы находятся в одном предложении. Выражение условия имеет вид:

<дескриптор1>* SENT *<дескриптор2>

В ИПЯ ИПС IRBIS включен оператор пересечения полей CON[N], служащий для отбора документов, в заданных полях которых имеется не менее N одинаковых поисковых терминов (N может принимать значения от 1 до 255). Контекстный оператор пересечения полей позволяет использовать в выражении условия имена полей (выступающие в данном случае в роли дескрипторов), содержимое которых сравнивается на предмет отыскания общих терминов.

Выражение условия запроса имеет вид:

<имя поля1>* CON[N] *<имя поля2>

4.5.3. Синтаксис и семантика использования дескрипторов

Для поиска информации средствами ИПЯ поисковые дескрипторы могут быть заданы (включены в запрос) одним из следующих способов:

- выбор из частотного словаря;
- ввод с клавиатуры;
- отметка ключевых слов в тексте документа;
- выбор терминов из специализированных словарных, рубрикационных или тезаурусных структур.

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

Маскирование. ИПЯ разрешает употребление символов маскирования двух видов:

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

Символы маскирования могут использоваться вместо любого символа дескриптора, и их количество внутри дескриптора неограниченно.

Параметризированные символы маскирования произвольного количества символов (например, «*(N)»), означают, что в дескрипторе на месте символа маскирования может стоять произвольная последовательность длиной не более чем N символов (где N - от 0 до 255).

Нормализация. *Для расширения возможностей дескрипторного языка на этапе сопоставления ПОД и ПОЗ может быть использован аппарат нормализации дескрипторов. Например, в основу реализации аппарата автоматической нормализации в ИПЯ системы IRBIS положен повышающий полноту поиска модифицированный алгоритм нормализации ОСНОВА-2 [Захаров1997], построенный на следующих правилах:*

1. Три первые буквы дескриптора остаются без изменения.
2. Все следующие гласные буквы заменяются символом маскирования произвольного числа рядом стоящих букв.
3. Конечные буквы *в, г, м, х* в дескрипторе заменяются символом маскирования произвольного числа рядом стоящих букв.
4. В конце дескриптора проставляется символ маскирования произвольного числа рядом стоящих букв (если после всех преобразований конечный символ дескриптора не является символом маскирования).

Лингвистическое обоснование такой замены заключается в том, что смыслоразличительная роль согласных во много раз больше, чем гласных. Начальная часть слова включается в новый дескриптор без изменения, т.к. информативность первых трех букв в слове велика. Согласные *в, г, м, х* могут попадать в дескриптор из окончаний существительных и прилагательных, поэтому исключение этих согласных из дескриптора и замена их символом маскирования ведет к отсечению окончаний.

Нормализованный таким образом дескриптор позволяет обеспечить более полный дескрипторный поиск с использованием только лишь частотного словаря БД.

Рассмотрим, например, запрос, который на естественном языке представляет собой предложение: «Частотный анализ терминов словаря». Такой запрос в системе (с применением правил нормализации) автоматически преобразуется в следующий ПОЗ:

част$тн$ AND анал$з$ AND терм$н$ AND слов$р$

Нормализованный таким образом ПОЗ обеспечивает поиск по логическому выражению с разрешением символов маскирования:

част$тн$ = частотность, частотности, частотный, частотные, частотных, частотного, частотной;

анал$з$ = анализ, анализа, анализе, анализу, анализируется, анализируются;

терм$н$ = термин, термина, термину, термином, термины, терминов, терминах, терминология, терминологии, терминологию, терминологические, терминологическим, терминологических, терминологической, терминологический, терминосистем, терминологичности;

слов$р$ = словарь, словаря, словаре, словарем, словарей, словарные, словарными, словарных, словарного, словоформа, словоформе, словоформы, словоформ, словарные, словарно-грамматический, словоупотреблений.

4.5.4. 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.

Контрольные вопросы

1. Охарактеризуйте оценки эффективности информационного поиска.
2. Охарактеризуйте взаимосвязь показателей эффективности и компонентов ИС.
3. Определите понятие универсального информационного потока.
4. Дайте определение первичных координат описания выхода ИПС.
5. Охарактеризуйте матрицу «термин-документ» и ее свойства.
6. Дайте формальное определение понятия «механизм поиска».
7. Охарактеризуйте сходства и отличия механизма поиска по совпадению терминов и механизма поиска документов-аналогов.
8. Перечислите механизмы поиска документов по сходству.
9. Определите понятие расширенного логического выражения, операции и операнда.
10. Охарактеризуйте возможную семантику использования дескрипторов в дескрипторных ИПЯ документальных ИПС.

5. Лингвистическое обеспечение ИС

5.1. Роль и логика языковых средств поиска документальной информации

Информационно-поисковыми языками (ИПЯ) называются искусственные языки, специально сконструированные для выражения (формулировки) основного смыслового содержания документов и запросов с целью последующего их сопоставления. ИПЯ обеспечивают компактную, строго алгоритмизированную и удобную для ЭВМ запись наиболее существенных сторон содержания документов и запросов [Аветисян1981].

Целесообразно привести замечание, касающееся особенностей использования наиболее распространенных дескрипторных и иерархически организованных языков. Язык, построенный на основе классификации (УДК, рубрикаторы и др.), позволяет потребителю легко найти свое место в информационной среде, как бы причислив себя к классу других потребителей. Дескрипторный же язык дает потребителю средство индивидуализироваться, отбирать документы по существенным для него признакам. Таким образом, дескрипторный язык может выступать как дополнение к классификационному[33].

ИПЯ дескрипторного типа отличаются друг от друга различным уровнем использования средств выражения парадигматического и синтагматического аспектов языка. Именно учет семантических связей при информационном поиске позволяет находить не только те документы, которые непосредственно образуют определенную тему, но и те, которые с ней связаны только частично. При этом разграничение видов семантической связи помогает группировать документы, включаемые в выдачу, по характеру их близости теме запроса. Например, документы, имеющие *семантическую общность* с запросом, непременно (в случае безусловной общности) или возможно (в случае гипотетической общности) содержат сведения по теме запроса. Документы, имеющие *семантическую близость* с запросом, непременно (в случае безусловной близости) или возможно (в случае гипотетической близости) содержат сведения по темам, соприкасающимся с темой запроса.

Фактически язык выступает как модель, отражающая состояние и связи объектов и явлений реального мира. С этой точки зрения парадигматика языка отражает наличие тех или иных отношений, имеющих место на уровне означаемых и/или означающих [Скороходько1974]. Например, отношение синонимии (тождество означаемых при различных означающих) является парадигматическим отношением означаемых, то-

[33] Следует, однако, отметить, что с точки зрения типологии любой классификационный язык может быть отнесен к дескрипторному, если под дескриптором понимать признак (имя признака в пространстве предметно-тематических признаков), т.е. система имен признаков функционирует как система дескрипторов.

гда как отношение омонимии (тождество означающих при различных означаемых) является парадигматическим отношением означающих. Парадигматику обычно связывают с лексико-семантическим потенциалом языка.

Парадигматические отношения означающих сравнительно легко поддаются систематизации, тогда как систематизация парадигматических отношений означаемых является гораздо более сложно задачей, связанной с моделированием элементов интеллектуальной деятельности человека: формированием в сознании человека тех или иных субъективных отражений объективно существующих связей между различными предметами и явлениями окружающего мира. Кроме самих предметов и явлений, «инициаторами» таких связей (ассоциаций) являются также языковые представления этих предметов и явлений. Ассоциации, сформировавшиеся в сознании различных индивидуумов, в процессе познания уточняются, модифицируются; случайные связи фильтруются, исчезают, а связи, более адекватные объективно существующим, закрепляются, с тем, чтобы потом уступить место новым связям, еще более адекватным. Если учесть влияние фактора субъективности на динамику процесса возникновения и развития таких ассоциаций, то станет ясно, насколько сложна задача их корректной систематизации в рамках тех или иных предметных областей. Тем не менее, исходя из практической необходимости, во многих случаях целесообразно выделять некоторые виды ассоциаций, которые сравнительно устойчивы и поддаются систематизации. К числу таковых принадлежат отношения подчиненности и соподчиненности, т.е. родовидовые отношения. Именно надситуативный характер таких ассоциаций в значительной степени и определяет целесообразность их фиксации в соответствующих словарях.

Синтагматика обусловлена наличием в каждом конкретном языке определенных правил построения означающих. Синтагматика естественных языков отражает структуру их грамматического строя. Если парадигматику связывают с потенциальными возможностями языка, то синтагматика отражает динамику конкретно сложившихся контекстных ситуаций. В поисковых задачах можно считать, что парадигматические отношения есть взаимосвязь между замещаемыми словами в тексте, а синтагматические – между сочетаемыми словами. А с точки зрения информативности отношений можно оказать, что парадигматика содержит в основном метаинформацию (отражающую структуру языка), в то время как синтагматические отношения (посредством синтаксиса) детально выражают основной смысл текста, выделяющий этот документ среди других, в том числе тех, которые построены в значительной степени с использованием той же лексики.

Особенностью применения ИПЯ является то, что преобразование текста (высказывания) в поисковый образ (ПО) путем выделения существенных аспектов обеспечивает переход от смысла в тексте к комплексу понятий. При этом применение синтаксиса в поисковом образе не

обеспечивает обратного преобразования комплекса понятий в текст, а только видоизменяет сами понятия.

Более того, на практике наблюдается стремление уменьшить количество терминов, включаемых в поисковый образ. Для того, чтобы это компенсировать хотя бы частично, устойчивые для данной предметной области отношения включают в информационно-поисковый тезаурус (ИПТ), который помимо априорной фактической информации об описываемых объектах и отношениях между ними содержит психолингвистические особенности специального языка, а также систему взаимосвязанных понятий соответствующей области знаний.

ИПТ в современном состоянии содержит сравнительно ограниченный запас априорной информации о предметной области. Однако составление даже отраслевых словарей и тезаурусов вручную занимает несколько лет, причем за это время многое меняется и в проблематике и в лексике отрасли. Еще большим «отставанием» характеризуются политематические тезаурусы. Для ликвидации такого отставания автоматизируется ряд этапов составления и ведения словарей и тезаурусов, где весьма эффективными оказались статистические методы выявления ассоциативных связей. В основе таких методов лежит схема выявления ассоциативных связей по результатам статистического анализа совместной встречаемости тех или иных терминов в рамках одних и тех же документов.

Недостатком такого подхода является отсутствие учета конкретной поисковой ситуации «запрос - поисковый массив». Для поисковых образов, лишенных исходного линейного порядка, аффиксов и окончаний, можно построить множество различных контекстов, среди которых в общем случае могут быть контексты, содержащие как совпадающие, так и не совпадающие по содержанию с исходным контекстом. Причем, с увеличением глубины индексирования (т. е. числа дескрипторов, входящих в поисковые образы документов) быстро растет количество ложных сочетаний, которые можно составить из дескрипторов одного и того же поискового образа. Однако в задачах информационного поиска глубина индексирования поискового образа документа уже не будет иметь исключительного значения, так как, сколько бы ни было дескрипторов в ПОДе, вопрос признания данного документа релевантным решается на основании только тех дескрипторов, которые совпали с дескрипторами поискового предписания. Каждая новая лексическая единица, входящая в поисковое предписание, исключает многие ложные сочетания, которые семантически уже не могут сосуществовать с этой лексической единицей, что объясняется наличием в языке лексико-семантических корреляционных связей.

5.2. Состав и структура лингвистического обеспечения

Рассмотрим лингвистическое обеспечение ИС как совокупность языковых средств, позволяющих представить информационную составляющую ИС на различных этапах внутрисистемного взаимодействия и взаимодействия с пользователем.

Такое определение предполагает, соответственно, выделение двух аспектов рассмотрения: выражение смыслового содержания информационной составляющей ИС и выражение информационной потребности пользователя.

Основным средством описания информационной базы и информационной потребности служат *информационно-поисковые языки*, относящиеся к классу искусственных языков. Помимо таких строго формализованных с точки зрения семантики и синтаксиса средств, в качестве дополнительных широко применяются *терминологические структуры* различного назначения, имеющие как линейную, так и нелинейную (иерархическую, сетевую) организацию.

Состав лингвистического обеспечения (ЛО) информационных систем может быть представлен следующей схемой (рис. 5.1):

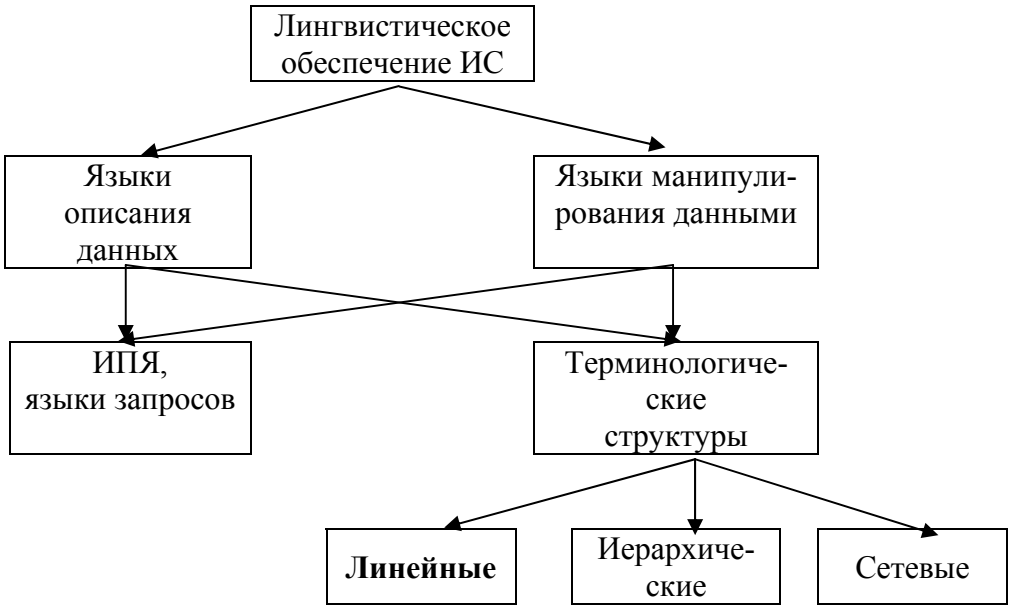


Рис. 5.1. Состав ЛО информационных систем

Предложенная схема не претендует на роль классификации, а представляет собой скорее иллюстрацию, отображающую эволюцию и особенности применения языковых средств в ИС: ИПЯ как средства выражения смыслового содержания документов и информационной потребности пользователя, SQL как попытки обобщения средств управления данными, терминологических структур как моделей данных, с одной стороны, и понятийных систем, выражаемых средствами искусственного языка с естественной лексикой – с другой.

Далее в этой главе приведем характеристики выделенных компонентов ЛО ИС.

5.2.1. Основные понятия лингвистического обеспечения

Для изучения принципов и методов построения и использования компонентов ЛО предварительно определим базовые понятия.

Язык — это знаковая система любой физической природы, выполняющая познавательную и коммуникативную функцию в процессе человеческой деятельности [Михайлов1968].

Искусственный язык (в отличие от *естественного,* представляющего собой средство общения и выражения мысли и неоднозначного по своей природе) – это специализированный язык, основное назначение которого состоит в устранении многозначности слов естественного языка и всего того, что характеризует эмоции и отношение к различным предметам. В искусственном языке должны выражаться лишь объективные характеристики предметов, их связей и соотношений [Горский1962].

Многозначность слов естественного языка, обеспечивающая «богатейшие возможности для вполне однозначного выражения тончайших оттенков мысли (в процессе речи эту многозначность устраняет контекст)» [Михайлов1968] и служащая в некотором роде показателем развитости языка, становится препятствием в случае использования в системе хранения и обработки информации. В связи с этим в ИС применяются искусственные языки, специально сконструированные для формулировки основного смыслового содержания информационной базы и информационной потребности с целью последующего их сопоставления. К таким языкам в первую очередь относятся *информационно-поисковые языки,* обеспечивающие компактную, строго алгоритмизированную запись содержания документов и запросов в ИПС.

Внутренняя структура языка характеризуется следующими составляющими — лексикой, грамматикой и словообразованием.

Лексика (или словарный состав) — это вся совокупность слов, входящих в состав языка.

Грамматика – это система способов и средств построения слов и предложений в рассматриваемом языке. Грамматика состоит из *морфологии* и *синтаксиса*.

Морфология — это 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.

Синтаксис — это совокупность действующих в языке способов и средств построения предложений.

Словообразование определяется как совокупность способов и средств образования слов на базе уже существующих.

Приведем данное в [Успенский1959] определение абстрактного языка, как формальной семантической системы: «Абстрактный язык, или формальная семантическая система, состоит из списка элементарных символов (знаков), правил образования (устанавливающих, какие комбинации знаков допускаются), правил преобразования (устанавливающих, какие допускаются преобразования выражений с целью получения логического вывода) и правил интерпретации (устанавливающих, какой смысл надлежит приписывать выражениям, составленным по правилам образования)».

Если из числа составных частей абстрактного языка исключить правила, которые устанавливают допустимые в нем преобразования с целью получения логического вывода, то ИПЯ можно определить как специализированную семантическую систему, состоящую из алфавита, правил образования (грамматики) и правил интерпретации (семантики).

Алфавит - это любая конечная совокупность знаков (букв, цифр и т.п.), используемых в ИПЯ. При построении ИПЯ выбор алфавита определяется не только соображениями практического удобства, но и внутренней структурой самого ИПЯ, а также предполагаемыми средствами технической реализации ИПС.

Выделяют *морфологические* и *синтаксические* правила образования (построения) *терминов* – слов языка. Морфологические правила определяют процедуру построения терминов ИПЯ из его морфем, а синтаксические — процедуру построения предложений (фраз) из этих терминов.

Синтаксические правила — обязательный элемент любого ИПЯ. В некоторых ИПЯ для соединения терминов в предложения (фразы) применяются специальные лексические средства.

Последний элемент ИПЯ, если его рассматривать как специализированный абстрактный язык, — это *правила интерпретации*, т. е. правила перевода терминов и предложений (фраз) ИПЯ на соответствующий естественный язык. Эти правила задаются, например, в виде двуязычных словарей, в которых каждому термину (лексической единице) ИПЯ ставится в соответствие определенное слово или выражение естественного языка, и наоборот. В такой словарь включаются также все символы, применяемые в данном ИПЯ для соединения терминов в предложения (фразы). Кроме того, правила интерпретации для ИПЯ, как и правила построения, формулируются на естественном языке в специальных инструкциях, методиках и т. д.

Словоупотребление в языке определяется двумя факторами: возможностью выбора слова, наиболее точно отображающего тот или иной предмет или явление объективной действительности, и возможностью сочетания этого слова с другими словами.

Предметы и явления объективной действительности связаны друг с другом определенными отношениями, которые существуют независимо от их языковой интерпретации. Эти отношения обусловлены предметно-логическими, а не собственно языковыми (семантическими) факторами и относятся к категории внеязыковых связей. Благодаря таким связям слова на основе того или иного семантического признака объединяются в лексико-семантические группы, которые называются *парадигмами.*

Рассмотрим, например, парадигму «грибы». К ней относятся слова: подберезовик, подосиновик, мухомор, сыроежка и т.п.. Внутри такой лексико-семантической группы можно выделить парадигмы «съедобные грибы», «условно съедобные грибы», «ядовитые грибы», «пластинчатые грибы», «трубчатые грибы» и т.п.. Причем, если парадигмы «съедобные грибы» и «ядовитые грибы» не пересекаются, то в парадигмах «съедобные грибы» (рыжик, млечник, подберезовик, подосиновик, сыроежка, боровик) и «пластинчатые грибы» (рыжик, млечник, сыроежка, мухомор, поганка) встречаются общие слова.

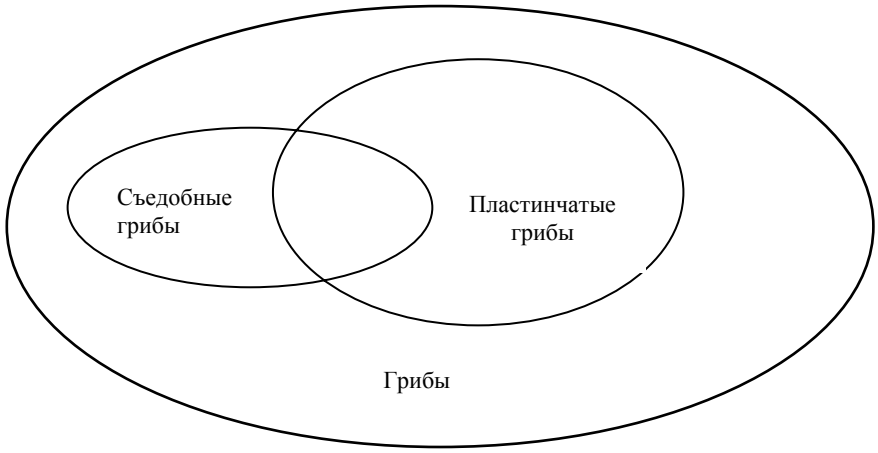


Рис. 5.2. Лексико-семантические парадигмы

Из приведенного примера следует, что парадигматические отношения в лексике не только многоступенчаты, но и многомерны (неоднолинейны). Одно и то же слово может быть одновременно членом нескольких лексико-семантических парадигм, в которых слова противопоставлены друг другу по какому-то определенному семантическому признаку (рис. 5.2).

Таким образом, *парадигматические отношения* предопределяют отбор слов для какого-либо сообщения, но сами остаются за его пределами.

Парадигматические отношения (аналитические отношения, базисные отношения, ассоциативные отношения) – логические отношения, существующие между лексическими единицами ИПЯ, независимо от их контекста.

Другой тип отношений между словами — *синтагматические отношения,* в которые слова вступают в пределах конкретного сообщения, фразы.

Синтагматические отношения (текстуальные отношения, синтетические отношения, синтаксические отношения) – отношения между лексическими единицами ИПЯ, выражающие действительные логические связи между соответствующими понятиями в тексте сообщения.

Типология ИПЯ. Рассмотрим типологию ИПЯ по способности к выражению смыслового содержания документов, как структурных единиц информационной базы ИС. Опираясь на лексику, грамматику и синтаксис, выделим два основных типа ИПЯ:

- языки классификационного типа;
- языки дескрипторного типа.

Классификация, как средство описания содержания документа, представляет собой процесс соотнесения содержания документов с понятиями, зафиксированными в заранее составленных систематических схемах. Основная цель классификации – приписать каждый документ классу, или, иначе – приписать каждому документу имя класса, формируя тем самым множества сообщений для обработки и поиска.

Языки дескрипторного типа поддерживают процесс *индексирования*, который заключается в формировании описания документа как совокупности дескрипторов, выбираемых из заранее созданных словарей понятий, либо из текстов документов.

На рис. 5.3 приведена типология методов описания содержания документов.

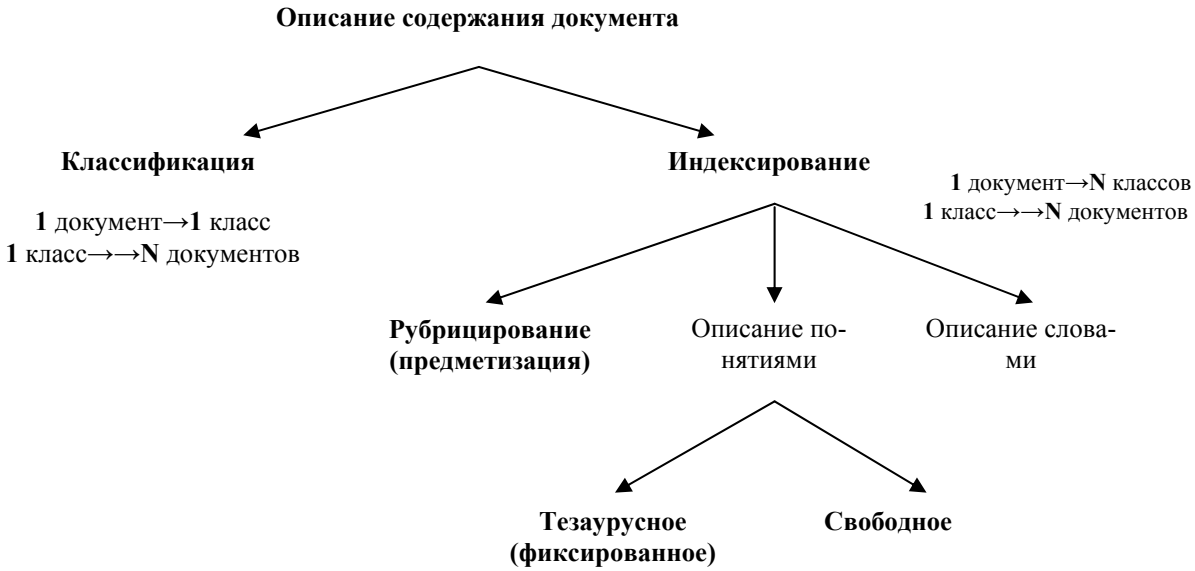


Рис 5.3. Типология методов описания содержания документов

5.3. Классификации

Классификации использовались для систематизации книг и других документов по их содержанию уже в глубокой древности. Например, есть сведения о том, что в Ниневийской библиотеке ассирийского царя Ашурбашшала (одна из древнейших библиотек - VII в. до н. э.) — клинописные глиняные плитки систематизировались в соответствии с их содержанием («история», «закон», «переписка» и т. д.) ».

В основе любой классификации лежит принцип деления объектов. *Классификацией* называется распределение объектов по классам на основании общего признака, причем распределение производится с таким расчетом, чтобы каждый класс относительно других классов занимал в получившейся системе точно определенное место. Признак, по которому производится деление, называется *основанием деления.*

Классификация должна подчиняться следующим формально-логическим правилам:

1) очередной шаг классификации проводится только по одному основанию;

2) получаемые в результате деления подклассы не должны пересекаться;

3) деление на подклассы должно быть соразмерным.

Пусть K — некоторый класс, который на основе некоторого признака разделен на подклассы $k_1\ k_2,\ k_3,\ \ldots,\ k_n$. Тогда сформулированные выше правила в обобщенном виде можно записать следующим образом:

$k_1 \cup k_2 \cup k_3 \cup \ldots \cup k_n = K$

$k_1 \cap k_2 \cap k_3 \cap \ldots \cap k_n = 0$

$k_i \cap k_j = 0 \quad \forall i \neq j, i = \overline{1,n}, j = \overline{1,n}$

Основные виды отношений между подразделениями классификации — это *иерархия (подчинение)* и *соподчинение.*

Иерархия — это отношение, при котором один класс является подклассом другого, более широкого: $k_i \subset K, i = \overline{1,n}$.

Отношением *соподчинения* связаны классы ($k_1,\ k_2,\ k_3, \ldots,\ k_n$), которые являются подклассами одного, более широкого класса ($k_1 \cup k_2 \cup \ldots \cup k_n = K$).

Различают два вида иерархии: сильную и слабую. При сильной иерархии каждый подкласс имеет один, и только один, непосредственно предшествующий ему класс, при слабой — более одного класса.

Классификации, классы которых связаны только отношениями сильной иерархии и соподчинения, будем далее называть *иерархическими классификациями.*

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.

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

- *Естественная классификация* — классификация, в основание которой кладутся существенные для выделяемых классов признаки.
- *Вспомогательная классификация* — классификация, в основание которой кладутся несущественные для выделяемых классов признаки.

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

- библиотечно-библиографические, специально предназначенные для систематизации книг и других документов;
- классификации изобретений, служащие для распределения объектов промышленной собственности, заявленных или признанных изобретениями;
- классификации наук, призванные систематизировать научную информацию.

5.3.1. Библиотечно-библиографические классификации

Основной целью естественных классификаций 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, так как только благодаря этому можно:

1) помещать предмет (книгу, статью, реферат и т. п.) в соответствующее место схемы или раскрывать содержание этого предмета через его положение в классификационной схеме;
2) быстро находить предмет (книгу, статью, реферат и т. п.), если известно, какое место закреплено за ним в схеме;
3) проводить группировку таких предметов, которые с точки зрения потребителя предпочтительно иметь в одном месте.

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 располо-

жения 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*.

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.

Различают три 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, философской или политической подготовкой.

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

В [Nitecki1989] отмечается, 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 когнитивой функций.

Основной функцией библиотечно-библиографической классификации является когнитивная. Это не просто система обозначений, а ментальная структура, выражаемая через обозначения.

Библиографическая функция классификации заключается в предоставлении систематического интеллектуального доступа к информации, содержащейся в библиографических материалах, который обычно осуществляется с помощью систематического каталога.

Существуют два подхода к библиотечно-библиографической классификации. Приверженцы *реалистического подхода* считают, 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. [Svenonius1989]

5.3.1.1. Иерархические библиотечно-библиографические классификации

В настоящее время наиболее распространенными библиотечно-библиографическими классификациями являются ББК, Десятичная классификация Дьюи, Библиографическая классификация Блисса и классификация Библиотеки конгресса США.

К иерархической библиотечно-библиографической классификации предъявляются следующие основные требования:

- для любого объекта в классификации должен быть предусмотрен один, и только один, исчерпывающий класс;
- классификация должна обеспечивать информационный поиск по любому сочетанию признаков.

Для удовлетворения этих требований в иерархической классификации необходимо иметь отдельные исчерпывающие классы для всех возможных объектов, т. е. перечислить их в классификационных таблицах. Поэтому такие классификации называются *перечислительными.*

Однако общее количество объектов классификации с развитием науки и техники непрерывно возрастает, поэтому ни одна классификация не дает возможности заранее перечислить все предметы. Кроме того, общие принципы построения иерархических классификаций таковы, что они не позволяют иметь в классификации один, и только один исчерпывающий класс для каждого объекта. Рассмотрим справедливость этих утверждений с формально-логической и с содержательной точек зрения.

С целью формально-логического анализа иерархических классификаций построим графическую модель иерархической классификации (рис.5.4).

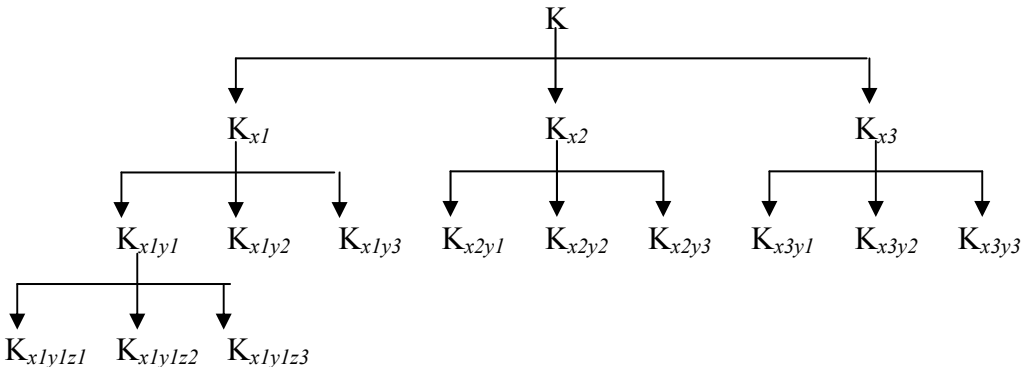


Рис. 5.4. Граф иерархической классификации

Пусть имеется класс объектов К и множество признаков (оснований деления) {*x, y, z*}. Каждый из этих признаков делит класс на три подкласса. Обозначим подклассы индексами, составленными из идентификатора признака и номера подкласса. Рассмотрим систему идентификации классов для примера из [Михайлов1968].

В качестве исходного класса К взят класс «гражданские самолеты», а в качестве оснований деления — соответственно следующие признаки: «целевое назначение» (*x*) «число моторов» (*y*) и «расположение крыла» (*z*). Применение основания деления *x* к исходному классу дает нам подклассы «пассажирские самолеты» *(*K_{x1}*),* «транспортные самолеты» *(*K_{x2}*)* и «специальные самолеты» *(*K_{x3}*)*. Применение признака *y* соответственно к классам K_{x1}, K_{x2} *и* K_{x3} дает нам подклассы «одномоторные пассажирские самолеты» *(*K_{x1y1}*),* «двухмоторные пассажирские самолеты» (K_{x1y2}), «четырехмоторные пассажирские самолеты» *(*K_{x1y3}*)* и т. д. Наконец, применение признака *z* соответственно к классам K_{x1y1}, K_{x1y2}, K_{x1y3}, K_{x2y1}, K_{x2y2}, K_{x2y3}, K_{x3y1}, K_{x3y2} и K_{x3y3} порождает подклассы «одномоторные пассажирские самолеты с низким расположением крыла» *(*K_{x1y1z1}*),* «одномоторные пассажирские самолеты со средним расположением крыла» *(*K_{x1y1z2}), «одномоторные пассажирские самолеты с высоким расположением крыла» (K_{x1y1z3}) и т. д.

При построении этой классификации деление, как того требуют формально-логические правила, должно проводиться только по одному основанию. Если бы на первом шаге деления были применены одновременно два признака — например, «целевое назначение» и «число моторов», то на одном уровне иерархии получились бы пересекающиеся классы «пассажирские самолеты» и «одномоторные самолеты» и т. д. Тогда документы об одномоторных пассажирских самолетах пришлось бы помещать не в один, а в два класса.

Описанная классификация, безусловно, не является идеальной. При ее построении, например, не был применен признак «тип движителя» (воздушный винт, реактивный поток). Классификация не является также и исчерпывающей: в число подклассов, образованных путем при-

менения признака «число моторов», не входят подклассы «трехмоторные самолеты» и «шестимоторные самолеты». Из-за несоблюдения формально-логических правил построения классификации в ней не найдется класса, к которому можно было бы отнести документ о трехмоторных транспортных самолетах или о реактивных пассажирских самолетах. Следовательно, при построении иерархической классификации соблюдение установленных формально-логических правил имеет исключительно большое значение.

Но даже безусловное соблюдение установленных формально-логических правил при построении иерархической классификации не устраняет ее главного недостатка - невозможность проведения группировки документов и информационного поиска по любому сочетанию характеристик.

Одно из основных требований, предъявляемых к библиотечно-библиографической классификации, состоит в том, чтобы она обеспечивала сбор документов по определенному предмету в одном, и только в одном, классе. Но это требование не выполняется даже в такой иерархической классификации, которая безупречна в формально-логическом отношении. Для построения иерархической классификации используется определенный ряд признаков (оснований деления). Число этих признаков обычно бывает невелико, и они применяются только в одной последовательности. Такая процедура позволяет построить классы предметов, обладающих лишь определенными сочетаниями признаков, выбранных в качестве оснований деления.

Если классификацию, которая приведена на рис. 5.4, использовать для информационного поиска по любому сочетанию признаков из множества оснований деления данной классификации, то придется собирать документы из нескольких классов в том случае, когда сочетание признаков отлично от исходного. Например, документы о гражданских самолетах с высоким расположением крыла в такой классификации распределены по 9 классам, документы о двухмоторных гражданских самолетах — по 3 классам и т. д.

Таким образом, для обеспечения возможности быстрого поиска документов по любому сочетанию признаков, взятых из некоторого данного множества (т.е. для предотвращения рассеяния таких документов по разным классам), необходимо построить отдельные классы для всех возможных сочетаний этих признаков. Эффективность этого метода зависит от числа классов самого нижнего уровня иерархии (обозначим это число через *N*), которое можно построить путем применения в разной последовательности исходного множества признаков. Очевидно, что чем больше *N,* тем выше трудоемкость процедуры классифицирования документов.

Проведем оценку трудоемкости этой процедуры путем вычисления числа иерархий, которое можно построить на данном множестве признаков.

Пусть C - множество признаков, используемых в качестве оснований деления:

$C = \{c_1, c_2, ..., c_n\}$, где n — число элементов множества.

$f(c_i)$ — число классов, порождаемых каждым признаком, взятым из C.

Тогда $N = \prod_{i=1}^{n} f(c_i)$ - число классов нижнего уровня одной иерархии.

При построении классификации можно выбрать ту или иную последовательность применения признаков. Каждая отдельная последовательность признаков порождает свою иерархию классов. Число возможных перестановок классификационных признаков равно $n!$. Отсюда следует, что общее число классов нижнего уровня всех возможных иерархий:

$$N = n! \cdot \prod_{i=1}^{n} f(c_i)$$

.

На самом деле, в основе любой иерархической классификации, создаваемой по формально-логическим правилам, должно лежать отношение «род-вид», которое фиксирует последовательность применения признаков – оснований деления.

При построении же реальных иерархических библиотечно-библиографических классификаций в качестве оснований деления используются не только видообразующие, но и другие признаки, имеющие характер отношения «от целого к части». Нарушение одного из важных принципов построения иерархических классификаций лишает процесс создания классификаций объективной формально-логической основы и делает его полуинтуитивным процессом, зависящим от конкретных практических потребностей, от опыта составителей схемы и от многих других факторов. В результате этого структура иерархической классификации усложняется, в ней появляются пересекающиеся классы и т. д.

Очевидно, что если специфика классифицируемых предметов такова, что классы в иерархических системах располагаются в определенном естественном порядке (например, для географической классификации естественным будет следующий порядок классов: часть света — страна — республика, штат, земля — область, край, воеводство, графство, департамент и т. д.), то такая классификация служит достаточно эффективным средством информационного поиска.

С другой стороны, имеется не меньше случаев, когда никакой естественной последовательности расположения классов в иерархической классификации не существует. Например, при создании иерархической схемы классификации грибов по их признакам (форма шляпки, вид ножки, пищевая ценность, цвет мякоти и т. д.) можно использовать любое

сочетание признаков для расположения классов в этой схеме, причем число таких вариантов, как было показано выше, может быть очень большим. Кроме того, приемлемость всех возможных вариантов иерархии фактически лишает иерархическую классификацию такого важного свойства, как обозначение родовидовых отношений между классами. Поэтому применение иерархической классификации в таких условиях становится неоправданным.

Рассмотрим далее содержательный аспект иерархической классификации как семантической системы, предназначенной для 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), давая одну серию подклассов. Классификационное дерево, получающееся в результате такого деления, является линейным и отличается большой жесткостью.

В то же время 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 библиотечно-библиографическая классификация непременно должна учитывать процессы анализа и синтеза в развитии 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.

В любой же иерархической классификации отдельные науки раз и навсегда разорваны разветвлениями жесткого классификационного дерева.

Особенно трудно включать в иерархическую классификацию новые межотраслевые предметы и области исследований, когда отсутствует даже общепринятая терминология для обозначения этих предметов. Однако, особенно в начальные периоды исследовательских работ в новых областях информационные запросы наиболее многочисленны, и быстрое удовлетворение этих запросов имеет особенно большое значение.

Таким образом, требуется периодическое изменение общей структуры той или иной реальной иерархической классификации, но на разработку новых таблиц иерархических классификаций уходит много времени и труда, и новый вариант иерархической классификации устаревает раньше, чем удается завершить работу по переклассифицированию документов. Поэтому выбирается путь соответствующей модификации уже существующих таблиц иерархической классификации, что придает ей все более условный характер.

Иерархические классификации обычно строятся на базе классификации наук с ее делением на отдельные отрасли. Но имеется множество предметов, 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. Поэтому в иерархической классификации создается несколько классов для одного и того же предмета, изучаемого разными науками или рассматриваемого в разных аспектах. Таким образом, невозможность иметь в иерархической классификации отдельные исчерпывающие классы для любого объекта делает такую классификацию малопригодной для информационного поиска по любым, заранее не предусмотренным сочетаниям признаков, а также по единичным признакам и признакам межотраслевого характера.

Суммируя вышеизложенное, можно сказать, что основные недостатки иерархических классификаций как семантической системы, предназначенной для поиска документов, состоят в следующем [Михайлов1968]:

1. Невозможность многоаспектного описания документов.
2. Непригодность для информационного поиска по любому сочетанию признаков, а также по узкопредметным и межотраслевым запросам[34].
3. Недостаточная глубина деления, из-за чего в классификации могут отсутствовать классы для специфических предметов.
4. Невозможность обеспечения взаимоисключаемости классов (явление «синонимии»).
5. Недостаточная эффективность механизма пересмотра классификации. Это означает, что в ней отсутствуют подразделения для новых объектов до тех пор, пока не будет опубликовано новое издание таблиц этой классификации. Таким образом, иерархическая классификация всегда отстает от достигнутого уровня развития науки и техники.

Тем не менее, наряду с перечисленными недостатками, иерархические классификации имеют и важные достоинства, благодаря которым различные классификации в течение многих веков используются в качестве основного средства для тематического поиска литературы. Ниже перечислены основные достоинства иерархических классификаций:

1. Классификация — один из привычных методов, которыми пользуются люди для определения объектов.
2. Иерархическая классификация пригодна для каталогов и указателей, имеющих любую физическую форму.
3. Для составления и использования систематического каталога или указателя не требуется никаких специальных средств и устройств.

[34] Строго говоря, непригодность иерархических классификаций для поиска документов по единичным и межотраслевым предметам не есть их недостаток, а одно из внутренних свойств.

4. На описание документов по иерархической классификации, как правило, расходуется меньше квалифицированного труда, чем на описание по фасетным и алфавитно-предметным классификациям.

5. Для шифровки классов в иерархической классификации обычно применяются арабские цифры и буквы латинского алфавита, имеющие широкое распространение. Это превращает ту или иную иерархическую классификацию в специализированный международный язык, понятный специалистам разных стран.

Наиболее известными иерархическими классификациями на сегодняшний день являются Десятичная классификация Дьюи, Библиографическая классификация Блисса, Классификация Библиотеки конгресса США, ББК.

Рассмотрим далее более подробно состав и структуру отечественной библиотечно-библиографической классификации (ББК).

ББК (отечественная библиотечно-библиографическая классификация).

В настоящее время ББК используется в отечественных библиотеках по гуманитарным наукам, публичных и детских библиотеках. Первое издание ББК вышло в 25 выпусках и 30 книгах (1961-1968 гг.). В дальнейшем разрабатывались и публиковались издания ББК для научных библиотек, сокращенные таблицы, однотомники для массовых, детских и школьных библиотек, таблицы для областных библиотек и специализированные таблицы и каталоги.

ББК имеет синтетическую структуру, позволяющую многоаспектно отражать содержание произведений печати. Пять уровней обобщения ББК, позволяющие установить единство между аналогичными по статусу, но различными по содержанию элементами универсальной классификации – это:

- основные таблицы,
- планы расположения,
- общие типовые деления,
- специальные типовые деления,
- территориальные типовые деления.

Типовое построение схемы ББК способствует единству структуры, сокращает объем таблиц, делает их обозримыми, мнемоничными. Единство структуры ББК выражается четырьмя способами:

- согласование структуры типологически сходных наук;
- подразделение одних отделов по аналогии с другими;
- подразделение с помощью знака отношения;
- таблицы типовых делений.

Эти способы также делают возможным дальнейшее совершенствование ББК. [Сукиасян1998, Сухманева1987]

Дальнейшее совершенствование классификации предполагает два возможных пути: изменения в рамках существующей структуры и модернизация самой структуры.

5.3.1.2. Алфавитно-предметные классификации

Алфавитно-предметная классификация представляет собой ИПЯ, основной словарный состав которого представлен упорядоченным по алфавиту множеством слов, словосочетаний и фраз естественного языка, обозначающих предметы какой-либо отрасли науки или практической деятельности. В [Шамурин1958] дано следующее определение алфавитно-предметной классификации: «Предметная классификация — классификация, в которой каждому предмету или вопросу отводится только один индекс, собирающий всю литературу в нем, независимо от разреза (аспекта), в котором предмет или вопрос в произведениях печати рассматривался».

Для рассмотрения структуры и основных свойств алфавитно-предметных классификаций введем следующие определения [Михайлов1968]:

Предметный заголовок — это слово, словосочетание или фраза естественного языка, используемая для обозначения предмета всего документа, части документа (текста) или информационного запроса.

Предметный словник — это упорядоченное по алфавиту множество предметных заголовков.

Предметная, рубрика — это совокупность предметного заголовка с описанием или адресом хранения хотя бы одного документа или с указанием тех мест в тексте документа, основная тема которых обозначается этим предметным заголовком.

Алфавитно-предметные классификации предназначены для построения указателей и каталогов, используемых главным образом для узкопредметного поиска. В таких указателях и каталогах под предметным заголовком даются сведения (адресные шифры или библиографические описания) по возможности о всех документах, предмет которых обозначен данным заголовком.

Алфавитно-предметные классификации могут быть двух типов: *алфавитно-систематические* и *словарные*.

В алфавитно-систематической классификации предметный словник состоит из названий предметов, которые подчинены названиям классов таких предметов. Число уровней иерархии в таком предметном словнике определяется практическими требованиями, предъявляемыми непосредственно к данной классификации. Предметные заголовки расположены в алфавитном порядке.

В словарной классификации предметный словник состоит только из расположенных в алфавитном порядке названий самих предметов. Идеальной моделью такой классификации можно считать именной указатель или каталог.

В словарной алфавитно-предметной классификации основными структурными элементами словаря считают предметный заголовок, предметный подзаголовок и предметную запись.

Предметный подзаголовок — это слово, словосочетание или фраза естественного языка, которая обозначает аспект рассмотрения предмета, указанного в предметном заголовке. Подзаголовок может также обозначать подкласс предметов, которые входят в более широкий класс, обозначенный предметным заголовком. Подзаголовки обычно включаются в предметный словник. Порядок расположения подзаголовков в пределах заголовка — алфавитный.

Предметная запись — это слово, словосочетание или фраза естественного языка, уточняющая и конкретизирующая подзаголовок. Как и подзаголовок, она может также обозначать в пределах заголовка подкласс предметов третьего уровня иерархии. Предметные записи обычно не включаются в словник. Иногда в качестве предметных записей используются заглавия документов или их библиографические описания. Последовательность расположения предметных записей в пределах подзаголовка может быть как алфавитной, так и любой другой.

Такое деление структурных элементов словарного состава алфавитно-предметных классификаций на заголовки, подзаголовки и предметные записи весьма условно. Действительно, если подзаголовок обозначает нулевой (пустой) класс, то такой подзаголовок может быть опущен. Наоборот, если в обозначаемом им классе содержится слишком много членов, то этот подзаголовок целесообразно сделать заголовком. Не во всех алфавитно-предметных классификациях применяются предметные записи. Поэтому различие между заголовками, подзаголовками и предметными записями обычно не проводится, и структурной единицей алфавитно-предметной классификации считается предметный заголовок, который может состоять из нескольких сегментов, а каждый сегмент — из одного и более слов (терминов).

Для примера рассмотрим фрагмент двухуровневого алфавитно-предметного классификатора товаров:

АБРИКОСЫ
- консервированные в сахаре
- консервированные для кратковременного хранения
- консервированные иным способом
- косточки и ядра
- паста
- свежие
- сушеные

АБСОРБЕРЫ
- газовые (скрубберы)
- холодильных установок
АБСОРБЦИОМЕТРЫ
АВАНТЮРИН
АВИЦИДЫ
АВОКАДО
АВТОБЕТОНОМЕШАЛКИ
АВТОВЫСТАВКИ
АВТОГРЕЙДЕРЫ
АВТОКЛАВЫ
- бланшировочные
- для вулканизации резины
- для вытапливания сала
- для консервирования пищи
- для омыливания жиров
- для приготовления и перемешивания смесей какао с одновременным нагревом
- для приготовления пищи (картофеля и др.), кормов
АВТОКРАНЫ
АВТОМАСТЕРСКИЕ
АВТОМАТЫ
- для размена денег
- оружие для ведения непрерывного огня
- токарные
- торговые, включающие подогревающие или холодильные устройства для продуктов и расфасованных напитков
- торговые, включающие подогревательные или холодильные устройства
- торговые для сигарет
- торговые прочие
- торговые типа парфюмерных пульверизаторов

Таким образом, одно из важнейших требований, предъявляемых к алфавитно-предметной классификации, заключается в том, чтобы обеспечить формулирование предметов документов и поисковых предписаний в одних и тех же терминах, т. е. единообразное употребление предметных заголовков не только предметизатором, но и потребителями, осуществляющими информационный поиск. Очевидно, что чем выше степень формализации предметного словника, тем больше такое единообразие и тем эффективнее действует алфавитно-предметный каталог или указатель.

5.3.1.3. Фасетные классификации

В основе *аналитико-синтетических* или *фасетных классификаций.* лежит *фасетный анализ,* идея которого принадлежит индийскому библиографу Ш. Р. Ран-ганатану.

Сущность фасетного анализа состоит в выделении в рассматриваемой предметной области категорий признаков классификации и в описании этих категорий множеством терминов. При этом категории называются *фасетами,* а каждый термин фасета называется *фокусом.*

Фасетный анализ проходит в несколько этапов:

- на основе анализа отрасли науки или техники составляется перечень основных категорий объектов, применяемых в данной отрасли;
- из изучаемых документов выписываются все существенные термины, относящиеся к данной отрасли науки или техники, которые группируются по фасетам, т. е. объединяются в соответствующие классы;
- фасеты делятся на субфасеты, субсубфасеты и т. д. (если необходима большая детализация);
- устанавливается полезная, но фиксированная последовательность расположения терминов внутри фасетов и фасетов — в схеме классификации;
- вводится обозначение соответствующими шифрами фасетов и терминов, входящих в эти фасеты;
- устанавливается фиксированная последовательность расположения фасетов при описании документа, которая называется *фасетной формулой.*

Процедура описания документов по фасетной классификации осуществляется следующим образом. Сначала на естественном языке формулируется основное смысловое содержание документа. Затем этот смысл выражается в терминах фасетной классификации, т. е. при помощи цепочки фокусов, взятых из фасетов и расположенных в фиксированном порядке. Часто вместо фокусов используются их шифры. Такая процедура позволяет создавать классы для документов, тематика которых выражается сочетанием нескольких разноаспектных характеристик.

Отметим, что отсюда следует весьма существенное различие между фасетными и иерархическими классификациями. При составлении иерархических классификаций в некотором систематическом порядке дается перечень всех предметных комплексов, которые уже описаны в литературе. При построении же фасетной классификации дается некоторое множество элементарных термов («строительных кирпичей»), из которых можно построить большое число наименований тематических классов.

Фасетная классификация не пользуется только готовыми классами. Названия классов (а следовательно, и сами классы) строятся на базе

разных сочетаний фокусов фасетной формулы, при этом ненужные фасеты пропускаются.

Таким образом, по сравнению с классификациями перечислительного типа фасетные классификации значительно облегчают многоаспектное описание документов. Представим рассмотренный ранее пример классификации гражданских самолетов с помощью совокупности фасетов:

«Типы гражданских самолетов»
x1 Пассажирские самолеты
x2 Транспортные самолеты
x3 Специальные самолеты
«Ч и с л о моторов»
y1 Одномоторные
y2 Двухмоторные
y3 Четырехмоторные
«Расположение крыла»
z1 Высокое расположение крыла
z2 Среднее расположение крыла
z3 Низкое расположение крыла

Нетрудно понять, что если расположить термины первого фасета на одной горизонтали, а затем приписать к каждому из этих терминов поочередно все термины второго фасета и после этого повторить описанную процедуру, используя термины третьего фасета, то мы получим иерархическую классификацию, приведенную на рис. 5.3. Таким образом, число всех возможных классов фасетной классификации во всяком случае не меньше числа иерархий, которые можно построить для эквивалентной ей иерархической классификации. Однако в такой классификации можно построить классы *x2z2* – «Транспортные самолеты со средним расположением крыла», *y1* – «Одномоторные гражданские самолеты» и т.п.

Фасетные классификации обладают рядом существенных преимуществ перед иерархическими классификациями. Основные из этих преимуществ состоят в следующем:

1. Возможность проведения многоаспектного описания документов.

2. Возможность собирать в одном месте все аспекты какого-либо предмета или темы. В общих иерархических схемах эти аспекты могут быть рассеяны по разным подразделениям (явление «синонимии»), причем отсутствуют правила отыскания таких подразделений.

3. Большая глубина деления понятий.

4. Возможность включать новые термины значительно проще и быстрее, чем в иерархические классификации.

Важнейший же недостаток фасетной классификации по сравнению с перечислительными классификациями — это большая трудоемкость построения фасетной классификации.

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

Ранганатан предложил пять основных фасетов, обозначаемых латинскими буквами: P (Personality, Индивидуальность), M (Matter, Материя), E (Energy, Энергия), S (Space, Место), T (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 ряда 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). УДК создавалась в основном еще до того, как был разработан фасетный принцип. Поэтому в УДК этот принцип получил лишь частичное воплощение и она является классификацией *полуфасетного* типа.

Все классы УДК сгруппированы в шесть фасетов, из которых два являются *фасетами формы* и четыре — *фасетами содержания* (табл. 5.1)

Таблица 5.1.

Фасеты формы и содержания в УДК (с их стандартным символическим обозначением)

№ пп	Тип фасета	Значение фасета	Стандартное символическое обозначение (нотация) фасета
1	Фасет формы	Язык документа	=
2	Фасет формы	Форма документа	(0...)
3	Фасет содержания	Место	(1/9)
4	Фасет содержания	Народность	(=...)
5	Фасет содержания	Время	"..."
6	Фасет содержания	Общий предмет	Отсутствует

Фасет содержания «Общий предмет» имеет десять основных подразделений, которые называются *главными классами.* Остальные фасеты УДК получили название *вспомогательных таблиц.*

Вспомогательные таблицы содержат перечни основных разделов УДК и таблицы вспомогательных фасетов классификации, а также других синтаксических средств, позволяющих комбинировать табличные классы в составе сложных тематических индексов, детально описывающих содержание классифицируемых материаловОсновными подразделениями 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

Общие определители языков обозначают язык документа, содержание которого обозначается основным индексом УДК.

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 удовлетворительно упорядочить выдачу.

В индексе УДК определители языка обычно располагаются на последнем месте. Однако они могут располагаться в середине и даже в начале составного индекса, если есть потребность располагать документы

по порядку языков, а не по их содержанию. При необходимости определитель языка может быть отделен от последующего индекса УДК двоеточием. Например:

663.4(493)(075)=112.5 Пивоваренная промышленность Бельгии. Учебник на фламандском языке

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. Определители формы 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, фотокопия, 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

Общие определители места Общие определители места служат для обозначения географического положения (или другого пространственного аспекта) объекта, классифицируемого по УДК, путем присоединения определителя к индексу основного ряда. 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.

Основной, наиболее часто используемой частью таблиц является раздел (4/9) «Страны и местности современного мира!. Этот раздел подразделяется в первую очередь по материкам (частям света): (4) «Европа», (5) «Азия», ..., (99) «Антарктика. Антарктида». Главный принцип, по которому построен раздел - современное политическое устройство

мира и политико-административное деление стран. Общие определители места, как правило, располагаются в соответствии с географическим положением стран и их областей. Большинство стран объединено по крупным частям материков.

При изменении политического статуса территорий (например, при образовании новых государств) Общие определители места обычно не менялись. Поэтому получилось, что в ряде случаев под 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.

При систематизации документов географический аспект часто играет определенную роль. Чтобы выразить связь рассматриваемого вопроса с указанным в документе местом, к основному индексу УДК присоединяют соответствующий определитель места, Например

621.311(410) Электростанции Великобритании

631.4(44) Почвы Франции

69(571.53/.62) Строительство в зоне БАМ

Общие определители народов (рас, этнических групп и национальностей). Общие определители народов (этнические определители) обозначают национальный или этнический аспект предмета, 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)

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

Определители времени обычно добавляются после основного индекса.

Примеры:

53"196" Физика 1960-х гг.

54"196" Химия 1960-х гг.

Среди вспомогательных таблиц находятся так называемые знаки соединения индексов УДК – дополнительное семантическое средство построения сложных индексов. В табл. 5.2 приведены характеристики этих знаков и примеры использования.

Таблица 5.2.

Знак	Назначение	Примеры применения
+ присоединение	Применяется тогда, когда содержание документа не может быть выражено одним индексом. В этом случае два (или более) индекса соединяются знаком.	622+669 Горное дело и металлургия (7+8) Северная и Южная Америка
/ распространение	Применяется тогда, когда содержание документа можно выразить с помощью нескольких индексов, следующих друг за другом в десятичном ряду. В этом случае первый и последний индексы соединяются знаком.	592/599 Систематическая зоология (вместо 582+593+...+599) 669.2/.8 Металлургия цветных металлов (вместо 669.2+669.3+...+669.8)
: двоеточие	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 двоеточия 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
:: двойное двоеточие	Применяется для закрепления определенной последовательности двух или более элементов в составном индексе	575::576.3 Цитогенетика 77.044::355 Военные фотосъемки

В правилах построения индексов УДК основное внимание уделено последовательности размещения характеристик в индексе документа, а также порядку расстановки таких индексов в систематическом указателе (каталоге).

В УДК принята определенная последовательность присоединения индексов специальных и общих определителей к основному индексу. Таким образом, в УДК тоже 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 предмет-

ный индекс УДК состоит из нескольких простых индексов, соединяемых друг с другом знаками присоединения, распространения, объединения или отношения, то эти простые индексы рекомендуется 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 в УДК являются:

1) недостаточная 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 1 приведена таблица одного из основных делений общепредметного фасета УДК «004 Информационные технологии. Вычислительная техника. Теория, технология и применения вычислительных машин и систем».

Примеры других фасетных классификаций и методик их построения.

В [Harris1987] представлена классификация The Dickens House Classification (DHC), предназначенная для описания совокупности фактов, связанных с изучением жизни и творчества Чарльза Диккенса. Классификация разработана Домом Ч. Диккенса в Лондоне в сотрудничестве с рядом других организаций.

DHC - фасетная (аналитико-синтетическая) классификация и принадлежит к числу немногих классификаций, специально разработанных для сферы гуманитарного знания. Она включает 5 основных разделов:

- общий раздел (сведения о языке, времени и месте фактов, связанных с Ч. Диккенсом),
- основы изучения творчества Ч. Диккенса;
- библиография о Ч. Диккенсе;
- изучение и оценка творчества Ч. Диккенса;
- произведения Ч. Диккенса.

В [Doucet1989] представлена система ESAR, используемая для классификации и анализа игровых принадлежностей. Она представляет собой оригинальную 6-фасетную схему. Фасеты обозначают этапы развития детей, соответствующие формам игры и основным поведенческим характеристикам - познавательным, инструментальным, социальным, языковым и эмоциональным. Схема позволяет классифицировать и анализировать игровые принадлежности путем выделения навыков, требуемых для каждой игры, и распознавания специфического вклада анализируемой игры с психологической точки зрения.

В [Брежнева1995] описана попытка разработки классификации информационных продуктов и услуг. Предлагается фасетная, многоаспектная классификация, цель которой - выявление признаков, необходимых и достаточных для описания свойств информационных продуктов, отражающих их возможности в плане удовлетворения потребностей потребителей, гарантирующих качество и приемлемую себестоимость подготовки. Считается, что использование фасетного подхода должно способствовать однозначности понимания существа процессов информационного обслуживания всеми специалистами, вовлеченными в эту сферу, нормализации технологии подготовки продуктов и услуг, обоснованному ценообразованию.

В [Полонский1989] описывается построение классификации исследований на основе фасетного метода. Выделяются четыре независимых фасета, характеризующие исследования с точки зрения теоретической и практической направленности: задачи, результаты, адрес исследования, вид документа. Для каждого из фасетов выделен ряд терминов, отражающих соответствующие признаки исследований. Эксперты определяют фасетные формулы (фиксированные в определенной последовательности сочетания признаков), которые соотносят с типом исследования: фундаментальное, прикладное, разработка или промежуточные типы. Полученная таким образом базовая модель служит эталоном для сравнения с любой конкретной работой в соответствующей области. Методика может быть реализована на ЭВМ.

5.3.2. Классификации изобретений

Наиболее известными и используемыми на сегодняшний день являются Международная классификация изобретений (МКИ) и Национальная классификация изобретений США (НКИ). Рассмотрим основные принципы, положенные в основу классификаций изобретений, на примере МКИ.

МКИ 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.

Тот факт, что МКИ по организационному построению является компромиссным решением между классификацией США, использующей в качестве основного функциональный признак, и германской классификацией изобретений, основанной на отраслевом делении, обусловил существование нескольких входов в отдельные понятия. С точки зрения индексирования это удобно, т.к. дает возможность индексировать основной признак по отраслевому принципу, не определяя функциональных особенностей. Однако поиск или индексирование близких понятий, отражающих предметы изобретения, построенные на одном принципе, но используемые в различных отраслях, связаны с известными трудностями. Исследования показали, что даже в наиболее благоприятных случаях, когда тема поиска дословно совпадает с текстом рубрики МКИ, последняя содержит лишь 50-70% общего количества релевантных ей документов. Это связано с тем, что предмет изобретения, воплощенный в способе или устройстве, может иметь разные существенные признаки.

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). Однако, во избежание громоздкости и многословности текстов, отсылки помещаются только там, где это является крайне необходимо.

Создатели МКИ стремились построить систему, обладающую возможностью расширения и максимальной логичностью деления понятий на классы. Классификация представляет собой специальную, имеющую технико-прикладной характер, линейную систему иерархического типа, предназначенную для ручного индексирования. Согласно с основными принципами применения Международной классификации изобретений, указывающими, 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;

Тогда изобретение может рассматриваться как некоторый объект, взятый относительно среды, т.е. формальное описание изобретения содержит описание объекта изобретения и некоторого предикатора (аспекта), отражающего его разновидность, особенность. В качестве таких предикаторов наиболее часто используют:

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

Индекс МКИ имеет несколько уровней иерархии, основными из которых являются:

- раздел, обозначаемой буквой латинского алфавита (от А до Н);
- класс, обозначаемый двумя цифрами (от 01 до 99);
- подкласс, обозначаемый буквой латинского алфавита (от А до Я);
- группа, обозначаемая одной или двумя цифрами (отделяемая графически от класса пробелом);
- подгруппа, обозначаемая двумя-тремя цифрами (отделяемая от группы знаком «/»)

Иерархия понятий (начиная с уровня группы) указывается в графической (печатной) форме количеством точек, предшествующих коду подгруппы.

Рассмотрим логику построения МКИ. Как уже отмечалось, МКИ является атерминологической системой. Будем считать каждую рубрику, допустимую в системе, отдельной лексической единицей (ЛЕ). Логически такая ЛЕ соответствует некоторому понятию, множеством значений которого являются некоторые реальные объекты (материалы, процессы и т.д.), объединенные в класс толерантности по некоторому общему для них признаку. С другой стороны ЛЕ можно рассматривать как состоящую из двух частей: *темы* – то, *о чем* говорится (предмет изобретения), и *ремы* – то, *что* говорится - некоторая предикация или атрибуция темы.

Ремы могут быть упорядочены (сгруппированы) по аспектам предикации. Каждая рема в свою очередь может стать темой при дальнейшем делении понятия.

Тема, определенная на уровне подкласса, имеет некоторое множество рем, определенных на уровне группы. Следует отметить, что эта двухуровневая система не является в общем случае классификационной схемой, т.к. не всегда выполняется требование деления понятия по единственному общему признаку, выбранному для данного уровня. Для устранения этого недостатка вводится дополнительный уровень «аспекта» (между уровнями группы и подклассов).

Иерархическая систематизация уровня подгрупп имеет тот же недостаток. Однако, классификационный признак, соответствующий уровню "аспекта", содержится в неявном виде в определениях содержания соответствующих рубрик (находящихся на одинаковом уровне, т.е. имеющих одинаковое число предшествующих точек).

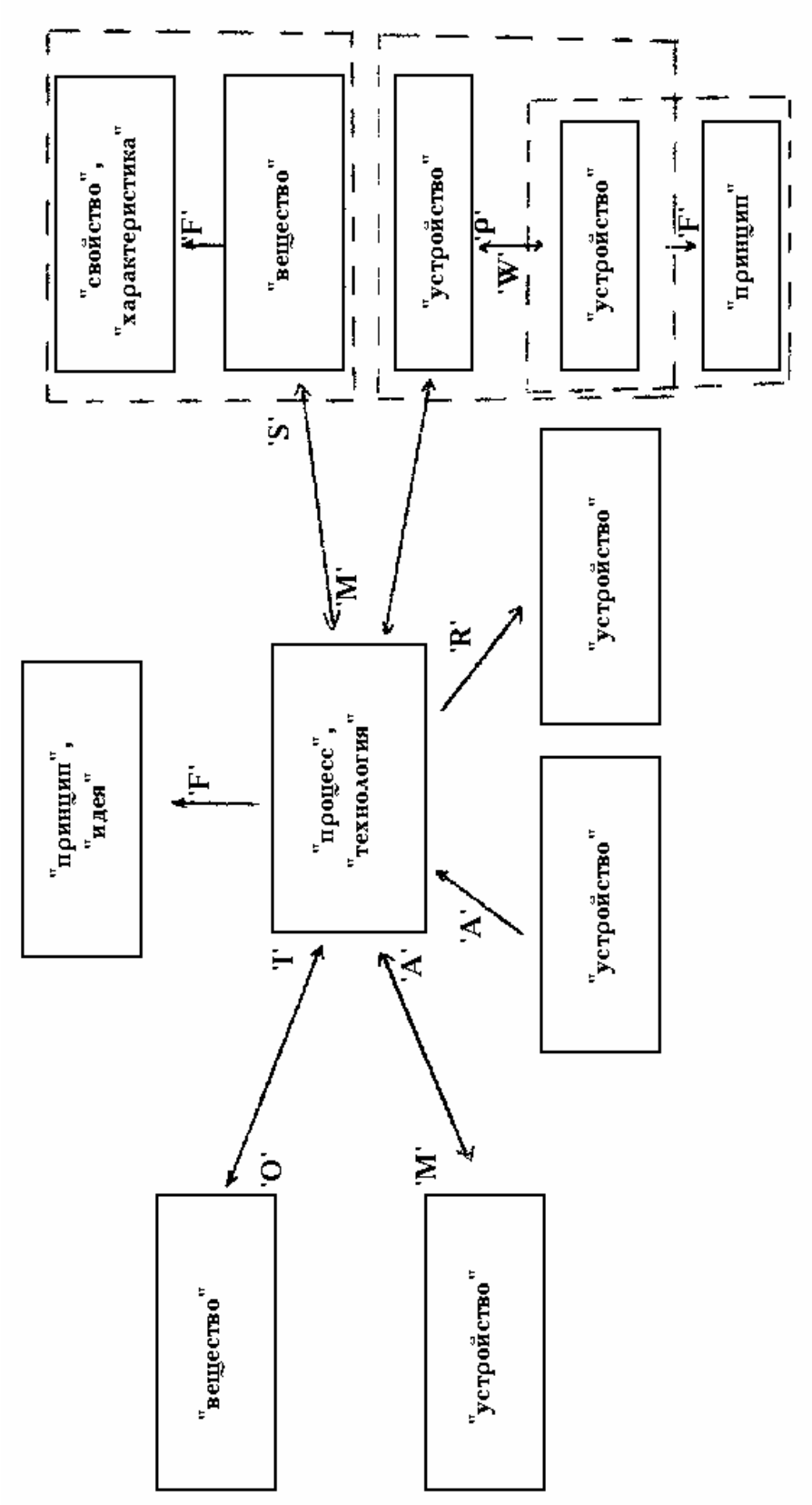


Рис. 5.5. Типы отношений между лексическими единицами МКИ

Возможные типы отношений между ЛЕ можно систематизировать, руководствуясь следующей обобщенной схемой (рис. 5.5).

В кавычках приведено возможное организационное выражение (обозначение) в классификации.

На схеме не указаны отношения типа "часть-целое" и "целое-часть", занимающие значительное место в МКИ. Эти типы отношений выражаются признаками "имеет в составе", "конструктивные особенности", "конструктивные элементы".

Встречаются также производные (от приведенных) типы отношений: "совместное использование", "предметы, используемые при построении, получении" и т.д. Полный перечень используемых отношений и их обозначение приведены в табл. 5.3.

Таблица 5.3.

№	Код	Содержание
1	**B**	Родовое понятие
2	**N**	Видовое понятие
3	**C**	Соподчиненные понятия
4	**U**	"Используй" понятие
5	**I**	Объект воздействия, использование
6	**S**	Субъект воздействия, средство
7	**O**	Результат, выход
8	**A**	Функциональное назначение
9	**R**	"Реализуется в"
10	**F**	"Основан на", "характерезуется"
11	**M**	Способ получения, технология
12	**W**	Целое
13	**P**	Часть
14	**Z**	Ассоциативное понятие

Основными достоинствами МКИ являются следующие:

- весь поток патентной информации индексируется в соответствии с МКИ;
- МКИ отражает 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.

В то же время необходимо отметить, что использование 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 (т.е. материалов, способов, устройств, являющихся предметами изобретения). Выбор именно такого типа лексических единиц обеспечивает единообразную трактовку содержания рубрик, независимо от особенностей употребления специальной научно-технической терминологии для различных научных школ и языков. Эта атерминологичность частично компенсируется введением в рубрику дополнительных видовых понятий при построении классификации за счет построения алфавитно-предметных указателей в классификации изобретений.

Среди недостатков построения МКИ отмечают также нелогичность выделения составных частей, внесение отдельных разделов в другие, некорректности отнесения классов к разделам, подгрупп к группам, смешение целого и части, отсутствие четко обозначенных отношений.

5.3.3. Отраслевые классификационные системы

Среди наиболее используемых в настоящее время отраслевых классификаций научной информации необходимо отметить Государственный рубрикатор научно-технической информации (ГРНТИ) и номенклатуру специальностей Высшей аттестационной комиссии (ВАК РФ).

Государственный рубрикатор НТИ и локальные (или отраслевые) рубрикаторы, построенные на его основе, используются при формировании всех видов информационных изданий. Систематизация документов в издании осуществляется в соответствии с последовательностью расположения рубрик, описывающих тематику издания.

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

Рубрикатор имеет три уровня иерархии. Коды рубрик состоят из пар арабских цифр, разделенных точкой.

Рубрикам (в качестве справочной информации) поставлены в соответствие индексы Универсальной десятичной классификации (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.[35]

[35] 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.

В основные функции Рубрикатора входит:

- определение тематического охвата информационных служб, систем, банков и баз данных;
- формирование информационных массивов с целью последующего обмена;
- систематизация материалов в информационных изданиях;
- индексирование документов и поиска их по тематическим рубрикам;
- переадресация запросов в информационных сетях;
- выполнение нормативной функции при разработке и совершенствовании локальных рубрикаторов;
- выполнение функции языка-посредника между другими классификационными системами (УДК, ББК и др.), используемыми в автоматизированных информационных системах.

Реализация этих функций позволяет:

- минимизировать дублирование при обработке документов и запросов;
- повысить эффективность информационного обслуживания в традиционном и автоматизированном режимах;
- минимизировать затраты при формировании информационных массивов;
- унифицировать структуру локальных рубрикаторов и обеспечить их совместимость в рамках сети;
- обеспечить совместимость тематического описания информационных ресурсов при работах по межгосударственному сотрудничеству в рамках СНГ;
- унифицировать методы ведения локальных рубрикаторов;
- упорядочить тематическую структуру информационных систем и баз данных;
- вести статистический анализ информационных массивов и потоков и т. д.

Принципы разработки Рубрикатора. При разработке Рубрикатора соблюдались следующие основные принципы и требования.

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

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 Общие вопросы*	73.*01 Общие вопросы*
59.*01.17 Международное сотрудничество*	73 .*01.17 Международное сотрудничество*
59.*01.29 Информационная деятельность*	73.*01.29 Информационная деятельность*
59.*01.79 Кадры*	73.*01.79 Кадры*

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)

Наряду с иерархической классификационной структурой в Рубрикаторе с помощью ссылок 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

Система нотаций строится на единых принципах и обеспечивает введение новых рубрик без изменения имеющихся кодов.

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.

Структура Рубрикатора. Рубрикатор имеет 3 уровня иерархии, при этом весь универсум знаний условно разделен на 4 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.

Аппарат ссылок и примечаний при рубрике отражает наличие полиерархических связей между рубриками, способствует уточнению и разграничению содержания рубрик, а также адекватному отнесению документов к релевантным рубрикам при индексировании.

Используются следующие виды ссылок:

Смотри ("см.") - содержит наименование понятия или темы, не находящей отражения в данной рубрике, и одновременно указывает ее место в рубрикаторе, например:

19.21 Массовая коммуникация. Социология массовой коммуникации

см. 04.51.54 Социология информации и коммуникации

Отсылка от ("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

Ост. от 66.15.17 Валка леса

Смотри также ("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

Эквивалентно ("Экв.") - используется для указания случаев тождественного наименования двух рубрик в разных разделах рубрикатора и также является симметричной, например:

15.81.43 Военная психология 78.21.15 Военная психология

Экв. 78.21.15 Экв. 15.81.43

При рубрике может быть дополнительная помета в виде текста ("Примечание"), которая 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 также приписаны индексы УДК и Номенклатуры 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 используетя 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 ВАК рубрикам сопоставлены коды по Номенклатуре специальностей научных работников Высшей аттестационной комиссии России. Так же, как и при сопоставлении с УДК, рубрикам приписывается один, либо несколько кодов специальностей по Номенклатуре 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

Локальный рубрикатор строится как выборка произвольного множества рубрик Государственного рубрикатора НТИ с возможным дальнейшим развитием на глубину ниже третьего уровня. Глубина классификации разрабатываемого рубрикатора не ограничивается и определяется информационной потребностью органа НТИ. При этом разные рубрики могут быть развиты на разную глубину. Отобранные рубрики Рубрикатора переносятся в локальный рубрикатор без изменения кода и наименования. Рубрики в локальном рубрикаторе располагаются также в порядке возрастания кодов.

При многоаспектном индексировании каждое понятие и аспект должны быть выражены кодами рубрик Рубрикатора. При этом для адекватного и полного отражения содержания рекомендуется использовать коды рубрик более низкого уровня обобщения (глубоких уровней иерархии). Использование рубрик первого уровня для индексирования документов не допускается.

Для отражения в поисковом образе политематического многоаспектного содержания документа может быть использовано "блочное" индексирование. Блок отражает только одну тему документа. В зависимости от количества выделенных при анализе содержания тем поисковый образ может состоять из одного или нескольких блоков.

Каждый блок включает основной код рубрики, отражающий главную тему документа, и дополнительные коды рубрик, которые отражают аспекты рассмотрения темы. Основной код ставится на первое место. Минимальный блок состоит из одного кода. Если блок содержит более одного кода, то они отделяются запятой, например:

Документ "Борьба с вредителями при выращивании зерновых культур".

Для индексирования используются рубрики: 68.35.29, 68.37.29

где:

68.35.29 Зерновые культуры

68.37.29 Вредители сельскохозяйственных растений и борьба с ними.

В Приложении 2 приведен раздел Рубрикатора ВИНИТИ РАН «20 Информатика».

5.4. Дескрипторные информационно-поисковые языки

Выше было показано, что ни одна из классификаций, не обеспечивает возможности поиска документов по любому, заранее не заданному сочетанию признаков (характеристик). Именно поэтому с начала 50-х годов начал распространяться метод *координатного индексирования* документов и основанные на этом методе *дескрипторные ИПЯ.*

5.4.1. Координатное индексирование

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.

Назывные слова не обозначают предмет, а выделяют его путем указаний. К категории назывных слов относятся также имена собственные. В [Кацнельсон1965] выделяются следующие признаки назывных слов:

– *надситуативность*, т. е. выделение определенного объекта независимо от того, присутствует он в данный момент в чувственной ситуации или нет;

– *неэгоцентричность* (выбор слова не зависит от говорящего лица, его местоположения и времени высказывания);

– *объективность* (опора на некие релевантные признаки предмета);

– *семантическая устойчивость и контрастность* (в каждом акте употребления они сохраняют некое инвариантное ядро, определенный минимум признаков, необходимых для распознавания предмета).

Кроме назывных в качестве ключевых слов могут выступать также соответствующие численные характеристики, хронологические данные, диапазоны температур, давлений и т. д. Ниже приведен пример координатного индексирования реферата, помещенного в РЖ «Информатика» за 2003 г.

Управление библиотечным и информационным центром. Trosow Samuel E., Libr. Quart., 2000, № 70, 153-155 Рецензируемая книга (R. D. Stueart, B. B. Moran. Library and information center management, 5th ed. Englewood, Colo.: Libraries Unlimited Inc., 1998, xxv, 509 p.), вышедшая уже в 5-м издании, 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. При подготовке данного издания авторы решили сохранить классическую структуру, включающую планирование, организацию, подбор персонала, определение направлений, координацию и финансирование. Они считают такой подход продуктивным и не признают постоянной критики данной структуры со стороны современных теоретиков по организации. 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.	библиотеки информационные центры управление функции руководителей рецензии США

Приведенный справа от реферата алфавитный список ключевых слов образует так называемый *поисковый образ*. Практический опыт показывает, что для координатного индексирования одного документа обычно бывает достаточно 6-12 ключевых слов.

Таким образом, *координатное индексирование* — это способ выражения основного смыслового содержания документа или информационной потребности в виде определенной совокупности ключевых слов. Координатное индексирование называют также *методом координации понятий, коррелятивным индексированием, унитерм-индексированием, ассоциативным индексированием, комбинаторным индексированием и т. д.*

При «чистом» координатном индексировании ключевые слова в поисковых образах никак не связаны одно с другим и функционируют самостоятельно. Для отыскания документов, отвечающих на какой-либо информационный запрос, необходимо выполнить определенные логические операции над классами, которые обозначены ключевыми словами поисковых образов документов. В простейшем случае, когда поисковое предписание сформулировано в виде логического произведения (конъюнкции) некоторого множества ключевых слов, документ считается отвечающим на информационный запрос и подлежит выдаче, если в поисковом образе этого документа одновременно содержатся все ключевые слова поискового предписания.

Рассмотрим некоторые примеры «чистого» координатного индексирования, показывающие основные случаи недостаточности такого подхода для обеспечения высокого качества информационного поиска[36].

1. *Ложная координация.* Предмет информационного запроса — «Информационные системы в экономике». Поисковое предписание сформулировано так: *информационные системы, экономика.* В ответ на такое поисковое предписание документальная ИПС выдаст как релевантный запросу документ «Информационная система по экономическому и социальному планированию», так и не отвечающий запросу документ «Экономика информационных систем», т.к. оба поисковых образа содержат ключевые слова *информационные системы* и *экономика.* В данном случае недостаточно использовать в запросе только координатную связь между ключевыми словами в поисковом образе документа.

2. *Неполная координация.* Предмет информационного запроса — «Вклад поставщика в разработку электронных каталогов». В данном случае поисковое предписание может быть сформулировано так: *поставщики, электронные каталоги, разработка.* В результате поиска по такому запросу документальная ИПС выдаст нерелевантный документ «Вклад пользователя в разработку электронных каталогов: точка зрения поставщика», т.к. поисковый образ документа содержит ключевые слова *поставщики, пользователи, электронные каталоги, разработка*. Причина выдачи нерелевантного документа заключается в том, что для формулировки поискового предписания были использованы ключевые слова, достаточные для выражения предмета информационного запроса, но недостаточные для выражения предмета документа.

3. *Синонимия, полисемия и омонимия ключевых слов.* Предположим, что предмет информационного запроса — «Применение анкетирования при изучении информационных запросов пользователей». Поисковое предписание сформулировано следующим образом: *информационная потребность, пользователи, анкетирование.* Документальная ИПС при этом не выдаст в ответ на такой запрос явно релевантный документ «Изучение информационных потребностей пользователей Ка-

[36] Эти примеры приводятся в целях иллюстрации, и поэтому им намеренно придан тривиальный характер.

надской информационной системы по полярным исследованиям», при индексировании которого были использованы ключевые слова *информационные системы, информационная потребность, пользователи, анкетный опрос*. Причина невыдачи релевантного документа заключается в том, что вместо ключевого слова *анкетирование* в поисковом образе документа был использован его синоним *анкетный опрос*.

Случай полисемии иллюстрируется, например, использованием в информационном запросе ключевого слова *базы данных*. В ответ на такой запрос (если не было дополнительного уточнения) будут выданы документы о полнотекстовых, реферативных, библиографических, фактографических базах данных, хотя реально документы по каком-либо из этих видов баз данных могут не соответствовать информационному запросу.

4. *Необозначенностъ родо-видовых связей между ключевыми словами.* Пусть предмет информационного запроса — «Библиотечное дело в Европе». Тогда поисковое предписание будет сформулировано следующим образом: *библиотечное дело, Европа.* В ответ на такое поисковое предписание не будет, например, выдан соответствующий информационному запросу документ «Школы, библиотеки и новая политическая система Венгрии», т.к. в его поисковом образе присутствуют ключевые слова *библиотечное дело* и *Венгрия*.

5. *Ложные синтагматические связи.* Предмет информационного запроса — «Передача электроэнергии из Шотландии в Англию», поисковое предписание: *передача, электроэнергия, Шотландия, Англия.* Документальная ИПС выдает документ «Передача электроэнергии из Англии в Шотландию», хотя этот документ не соответствует информационному запросу. В данном случае ИПС выдала нерелевантный документ, хотя имело место точное совпадение поискового предписания с поисковым образом документа. Отсюда следует, что для предотвращения выдачи нерелевантных документов необходимо, чтобы ключевые слова поисковых образов документов и ключевые слова поисковых предписаний можно было связывать более сильными синтагматическими связями, чем простая координация.

5.4.2. Семантическая сила дескрипторных ИПЯ

Приведенные примеры показывают, что для существенного повышения качества информационного поиска, основанного на применении координатного индексирования, необходимо:

1) устранение синонимии, полисемии и омонимии ключевых слов, используемых в качестве лексических единиц ИПЯ;

2) построение специальных словарей, таблиц или схем, в которых бы эксплицитно были выражены наиболее существенные парадигматические связи между ключевыми словами;

3) разработка для дескрипторного ИПЯ такого синтаксиса, который бы позволял использовать при построении поисковых образов документов и поисковых предписаний не только простую координацию дескрипторов, но и более сильные синтагматические связи.

Первая задача относится к области семантики, т. е. к аспекту отношений слов к предметам и явлениям, которые они обозначают, а вторая задача — к области отношений между предметами и явлениями, обозначаемыми словами. Совокупность методов и средств, применяемых для решения этих двух задач, называется *контролем за словарным составом ИПЯ*. Благодаря введению такого контроля обеспечивается использование одинаковых ключевых слов для координатного индексирования одинаковых по смысловому содержанию документов и информационных запросов.

Контроль за ключевыми словами, которые используются для координатного индексирования документов и информационных запросов, может иметь разные степени. При *нулевом* или минимальном контроле 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. Такой подход применяется в реальных ИПС при координатном индексировании документов. Индексирование информационных запросов в этом случае должно проводиться весьма тщательно и с избыточностью, необходимой для нейтрализации отрицательных явлений, которые порождаются отсутствием словарного контроля при индексировании документов.

При *полном* контроле за словарным составом ИПЯ разрешено использовать для координатного индексирования документов и информационных запросов лишь *дескрипторы*, т. е. такие ключевые слова, которые содержатся в некотором нормативном списке (например, в тезаурусе). В таком списке или словаре полностью устранены синонимия, поли-

семия и омонимия ключевых слов, а также обозначены определенные парадигматические связи между ними.

На самом деле, на разных этапах работы ИПС - при вводе документов в систему (т. е. при индексировании документов) или при формулировании поисковых предписаний – требуется свой словарный контроль. Еще Г. П. Лун высказывался за минимальный словарный контроль при вводе документов в ИПС. Он писал: «Чрезмерное редактирование явно увеличивает вероятность влияния сегодняшних интересов, опыта и точек зрения. Вследствие этого полезность системы уменьшится, если изменятся задачи и интересы. Поэтому кажется очевидным, что чем меньше информация систематизируется и сокращается при вводе, тем больше она будет поддаваться динамической интерпретации при ее выводе» [Luhn1957].

В целях дальнейшего рассмотрения основных методов и средств, применяемых для контроля за словарным составом дескрипторного языка, а также для выражения синтагматических отношений между дескрипторами в поисковых образах документов и в поисковых предписаниях, уточним термины «дескриптор» и «дескрипторный язык».

Дескрипторы — это предназначенные для координатного индексирования документов и информационных запросов нормативные ключевые слова, которые по определенным правилам отобраны из основного словарного состава того или иного естественного языка и у которых искусственно (при помощи соответствующих отсылок и помет) устранены синонимия, полисемия и омонимия.

Дескрипторным языком называется специальный ИПЯ, словарный состав которого состоит из дескрипторов, а грамматика — по крайней мере, из способа построения поисковых образов документов и поисковых предписаний путем координации соответствующих дескрипторов.

В [Михайлов1968] приводится история возникновения идеи координатного индексирования. В сообщении, сделанном на конференции по проблемам применения классификаций для информационного поиска (Великобритания, 1957 г.), Б. Виккери отмечал, что координатное индексирование (с инвертированной организацией картотеки), по-видимому, применялось еще в Шумере около трех тыс. лет назад, о чем свидетельствуют найденные глиняные таблички, каждая из которых отведена для какого-либо симптома болезни. На табличке перечислялись названия всех болезней, для которых характерен этот симптом. Очевидно, что если по обнаруженным у больного симптомам подобрать соответствующие таблички, а затем определить, название какой болезни содержится на каждой из них, то это и будет вероятное заболевание, которым страдает данный больной. Эти таблички можно считать прообразом современных диагностических машин.

Принцип координатного индексирования, по-видимому, был известен и американскому орнитологу Г. Тейлору, который в 1915 г. получил патент на суперпозиционные (просветные) перфокарты. В геологии

этот принцип (в сочетании с суперпозиционными перфокартами) применяется с 1920 г. в системах для идентификации минералов. В 1923 г. француз А. Либё предложил использовать координатное индексирование (также в сочетании с суперпозиционными перфокартами) для поиска в картотеке персонала лиц, обладающих заданным сочетанием характеристик.

Первым (или, по крайней мере, одним из первых), кто применил метод координатного индексирования для информационного поиска, был У. Баттен, когда в 1939 г. создавал для английского концерна «Империал кемикл индастриз, лтд» поисковую систему для патентов. Эта система была реализована на суперпозиционных перфокартах. У. Баттен пишет, что идея координатного индексирования у него возникла тогда, «когда он тщательно изучал два списка патентов, один из которых относился к предмету *A,* а другой — к предмету *B,* с целью выбрать из них патенты, относящиеся к сочетанию предметов *A* и *B»*.

В 1947 г. американский математик К. Муэрс разработал и запатентовал систему механизированного поиска документов, основанную на использовании идеи координатного индексирования. Созданную им ИПС К. Муэрс назвал Системой *зато-кодирования* (по имени организованной и возглавляемой им консультативной фирмы «Затор ко.»). Для реализации этой ИПС были созданы особые карты с вырезками вдоль краев, названные «зато-картами», а также специальный селектор. Запись характеристик на таких картах проводилась по суперпозиционному коду (с использованием случайных чисел), который был разработан К. Муэрсом и назван «зато-кодированием».

Необходимо отметить, что именно К. Муэрс в 1950 г. ввел для обозначения лексических единиц ИПЯ, применяемого в его поисковой системе, термин «дескриптор». Им были введены также термины «информационный поиск» (information retrieval), «информационно-поисковая система» (information retrieval system), «информационно-поисковый язык» (retrieval language), «дескрипторный словарь» (descriptor dictionary), «поисковый образ» (tally) и другие термины, которые в настоящее время широко используются в области информационного поиска.

Исключительно большой вклад в теоретическое обоснование, развитие и пропаганду идей координатного индексирования внес американский логик М. Таубе, который в 1951 г. разработал так называемую *систему унитермов,* получившую широкое практическое применение. Основное отличие системы унитермов М. Таубе от системы зато-кодирования К. Мауэрса заключается в том, что в первой мы имеем дело со словами, выражающими понятия, а во второй — с понятиями, выраженными словами. Обе эти системы послужили основой для разработки современных дескрипторных языков.

Унитерм (uniterm) М. Таубе — это ключевое слово (как правило, простое), которое может быть снабжено соответствующей отсылкой или

пояснительной пометой, устраняющей его синонимию, полисемию и омонимию. Унитермы не имеют никаких помет, которые бы обозначали парадигматические связи между ними. В качестве унитермов могут использоваться имена собственные, географические и фирменные названия и т. д.

Для сравнительной оценки семантической силы разных ИПЯ рассмотрим особенности словарного состава (лексики) таких языков.

Существуют ИПЯ, в которых словосочетания и фразы, выражающие сложные понятия (т. е. состоящие из двух и более простых понятий), заданы в словаре наряду со словами, выражающими простые понятия. Это означает, что в таких словосочетаниях или фразах образующие их слова связаны координатной (соподчинительной) или какой-либо иной связью до процесса индексирования. ИПЯ такого типа называются *предкоординированными* (pre-coordinate). Словарный состав предкоординированных языков в какой-то мере напоминает двуязычные разговорники, в которых заранее составлены наиболее употребительные, с точки зрения составителей, фразы. К предкоординированным языкам относятся, например, как перечислительные, так и аналитико-синтетические классификации.

Когда применяется предкоординированный язык, то для обозначения основного смыслового содержания документов можно использовать только отдельные, никак не связываемые между собой слова, словосочетания и фразы, взятые из словарного состава данного языка. Таким образом, проводится *классификация* документов, т. е. их отнесение к классам, обозначенным лексическими единицами этого языка. Семантическая сила предкоординированного языка определяется главным образом тем, насколько тщательно и исчерпывающе разработана его лексика. При составлении словаря для такого языка, которое проводится до его использования, недостаточно учесть лишь специфику информационных запросов прошлого и настоящего. Необходимо также предвидеть изменения информационных потребностей в будущем. Это очень трудная задача, удовлетворительное решение которой едва ли возможно. Таким образом, предкоординированным языкам органически присуща недостаточность словарного состава, а, следовательно, и небольшая семантическая сила.

Другой тип ИПЯ — это такие, в которых лексические единицы (термины, слова) объединяются в «предложения» (поисковые образы) лишь во время индексирования документов или даже в процессе их поиска. Такие ИПЯ называются *посткоординируемыми* (post-coordinate). До индексирования лексические единицы посткоординируемых языков не связаны никакими синтагматическими отношениями. Можно провести определенную аналогию между посткоординируемыми языками и алфавитом любого естественного языка. В естественном языке соединение букв в слова производится лишь в процессе письма, а до этого буквы никак не связаны одна с другой. Точно так же из сравнительно неболь-

шого числа лексических единиц посткоординируемого языка можно построить «предложение» (поисковый образ), выражающее практически любой смысл. Очевидно, что такие языки будут семантически более сильными, чем предкоординируемые. На рис. 5.6 приведена типология ИПЯ с точки зрения их семантической силы.

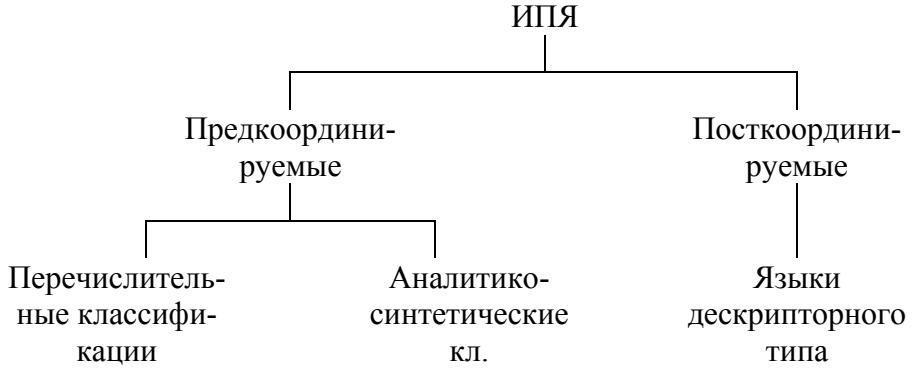


Рис. 5.6. типология ИПЯ с точки зрения семантической силы

Еще раз отметим, что главная особенность классификаций состоит в том, что их словарный состав задается в виде фиксированного списка слов, словосочетаний и фраз. При переводе текста, выражающего основное смысловое содержание документа или информационного запроса с естественного языка на предкоординируемый язык можно пользоваться только словами, словосочетаниями и фразами, содержащимися в фиксированном списке. Введение новых лексических единиц строго ограничено; оно возможно лишь до индексирования документов, т. е. при создании языка. Поэтому классификации не обладают такой «семантической силой», которая необходима для эффективного информационного поиска по запросам любого типа.

Соответственно, основные трудности информационного поиска возникают вследствие того, что процессы индексирования документов и их поиска разделены во времени, причем иногда весьма значительно. Если бы можно было заранее предвидеть все варианты информационных запросов, которые могут возникнуть у потребителей информации, то документы можно было бы адекватно заиндексировать, по-видимому, средствами любого ИПЯ, хотя для этого, возможно, пришлось бы в необходимой степени расширить основной словарный состав классификации. Но такое условие не выполнимо в принципе, т.к. специфика информационных запросов зависит от развития науки и техники, от изменяющихся научных интересов потребителей информации, их индивидуальных особенностей и т. д. и определяется факторами, которые фактически не поддаются учету и предвидению.

5.5. 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. 5.1, 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. 5.7. 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 отношений – отношения иерархии, ассоциативные отношения и отношения эквивалентности. Синтагматические связи показывают логические отношения между понятиями (рис. 5.7).

5.5.1. 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, аффиксы, суффиксы и 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 логической мышления, и поэтому при поиске информации необходимо полностью или частично исключить влияние аффиксов и окончаний слов естественных языков.

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 словарей аффиксов и окончаний заключается в автоматической проверке на наличие в словах естественных языков элементов, вошедших в наперед заданные (составленные экспертами-лингвистами) словари аффиксов и окончаний, и отсечении их, если они имеются.

От качества составления словарей аффиксов и окончаний 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 отсекается из всех слов естественно-языкового употребления в базе данных, независимо от того, является ли для конкретно рассматриваемого слова морфемой или 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 ВИНИТИ РАН «Информатика» за период с 1986-2000 гг., по следующим рубрикам Рубрикатора ВИНИТИ 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. 5.3).

Таблица 5.3.

Рубрика	Кол-во доктов	Общее количество терминов	Слов	Словосочетаний	% 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

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*).

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.

5.5.2. 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) [Белоозеров2001].

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. 5.8 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. 5.8. Фрагмент Рубрикатора ВИНИТИ

Словарь синонимов. Словарь синонимов, который для каждого входа словаря определяет одну или больше 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 выявить 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
    ДИСКРИМИНАЦИОННЫЙ 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
    ТАКСОНОМЕТРИЧЕСКИЙ АНАЛИЗ
    ФАСЕТНЫЙ АНАЛИЗ
    ЧИСЛЕННЫЙ АНАЛИЗ
    ЭКСПЕРИМЕНТАЛЬНЫЙ АНАЛИЗ

5.5.3. Терминологические структуры с сетевой организацией

Основными представителями сетевых терминологических структур являются *тезаурусы*.

Весь окружающий нас 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 мира, 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 Брутто 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» [Браславский1997].

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 [Браславский1997].

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 [Информационно-поисковый1987] 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.

Рассмотрим основные определения и виды связей, обозначаемых в тезаурусах, на примере информационно-поискового Тезауруса по информатике [Информационно-поисковый1987].

Лексические единицы тезауруса поделены на дескрипторы (графически выделенные написанием прописными буквами) и *ключевые слова* - недескрипторы (написание строчными буквами). 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 3000 лексических единиц, из которых примерно половину составляют дескрипторы. Общее количество словосочетаний составляет около 70%, а количество дескрипторов-словосочетаний – 58% от общего числа дескрипторов.

Лексические единицы в тезаурусе снабжены словарными статьями.

Словарная статья дескриптора состоит из собственно дескриптора (заглавного дескриптора) и списка дескрипторов и ключевых слов, связанных с заглавным дескриптором по смыслу.

Общеупотребительные аббревиатуры входят в тезаурус в качестве дескрипторов. Каждая из них снабжена расшифровкой, которая приводится в косых скобках строчными буквами. Эта расшифровка служит также ключевым словом, синонимичным дескриптору-аббревиатуре.

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

- заглавный дескриптор;
- ключевые слова, условно синонимичные заглавному дескриптору;
- вышестоящие дескрипторы;
- нижестоящие дескрипторы;
- дескрипторы, связанные с заглавным дескриптором одним из ассоциативных отношений.

Ключевые слова, входящие в класс условной эквивалентности, снабжены пометой «с» (синоним). Если дескриптор использовался в сочетании с другим дескриптором для замены ключевого слова (соответственно со ссылками «исп к» или «исп а»), то данное ключевое слово приводится после всех остальных ключевых слов со ссылкой «ср» (сравни).

Дескриптор, являющийся в иерархическом отношении вышестоящим по отношению к заглавному дескриптору, снабжается в словарной статье пометой «в» (вышестоящий). Этой ссылкой обозначается как родовой дескриптор, в объем понятия которого входит объем понятия заглавного дескриптора, так и дескриптор, обозначающий «целое» по отношению к понятию, выраженному заглавным дескриптором.

Дескрипторы, нижестоящие по отношению к заглавному дескриптору, вводятся в дескрипторную статью со ссылкой «н» (нижестоящий). Этой ссылкой обозначаются как видовые дескрипторы, так и дескрипторы, обозначающие компонент (часть) понятия, выраженного заглавным дескриптором. В словарную статью каждого дескриптора включаются вышестоящие и нижестоящие дескрипторы только одного – ближайшего к заглавному дескриптору – уровня иерархии.

Дескрипторы, связанные с заглавным дескриптором другими видами отношений, включены в дескрипторную статью со ссылкой «а» (ассоциация).

Приведем примеры полных словарных статей дескрипторов:

ИПЯ /информационно-поисковые языки /
с информационно-поисковые языки
поисковые языки
ср информационные языки
в ИСКУССТВЕННЫЕ ЯЗЫКИ
ЛИНГВИСТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ
н ДЕСКРИПТОРНЫЕ ИПЯ
КЛАССИФИКАЦИИ
а ВХОДНЫЕ ЯЗЫКИ
ВЫХОДНЫЕ ЯЗЫКИ
ЕСТЕСТВЕННЫЕ ЯЗЫКИ
ИНДЕКСАЦИОННЫЕ ТЕРМИНЫ
ИНДЕКСИРОВАНИЕ
ИНФОРМАЦИОННО-ЛОГИЧЕСКИЕ ЯЗЫКИ
ИНФОРМАЦИОННЫЙ ПОИСК
ИПС
ИПТ
ПАРАДИГМАТИЧЕСКИЕ ОТНОШЕНИЯ
ПОСТКООРДИНАЦИЯ
ПРЕДКООРДИНАЦИЯ
ФОРМАЛИЗОВАННЫЕ ЯЗЫКИ

АНАЛИТИКО-СИНТЕТИЧЕСКАЯ ПЕРЕРАБОТКА ДОКУМЕНТОВ (ПРОЦЕСС)
с аналитико-синтетическая обработка информации
аналитико-синтетическая переработка информации
аналитическая обработка информации
обработка документов
синтезирование информации
ср переработка информации
преобразование информации
в НИД
н АННОТИРОВАНИЕ
ИНДЕКСИРОВАНИЕ
ИНФОРМАЦИОННЫЙ АНАЛИЗ
КАТАЛОГИЗАЦИЯ
КЛАССИФИЦИРОВАНИЕ
ОБРАБОТКА ТЕКСТА
ПЕРЕВОД
ПРЕДМЕТИЗАЦИЯ
РЕФЕРИРОВАНИЕ
РЕЦЕНЗИРОВАНИЕ
РУБРИКАЦИЯ

а АНАЛИТИЧЕСКИЕ ОБЗОРЫ
АННОТАЦИИ
ДОКУМЕНТЫ
ОБЗОРЫ
ОБРАБОТКА ИНФОРМАЦИИ
РЕФЕРАТЫ

Для устранения неоднозначности или уточнения значения лексических единиц используются *релятоpы.* Релятор приводится в круглых скобках прописными буквами, если он относится к дескриптору, и строчными, если относится к ключевому слову. Например, при использовании релятора с ключевым словом:

классификация (процесс)
см КЛАССИФИЦИРОВАНИЕ

Словарная статья ключевого слова состоит из собственно ключевого слова, снабженного отсылкой «см» (смотри), и дескриптора, служащего его условным смысловым эквивалентом, например:

порождающие грамматики
см ГРАММАТИКА

В случаях замены ключевого слова сочетанием двух или более дескрипторов используется ссылка «исп к» (используй комбинацию), например:

документально-фактографические ИПС
исп к ДОКУМЕНТАЛЬНЫЕ ИПС; ФАКТОГРАФИЧЕСКИЕ ИПС

В случае неоднозначности ключевого слова и необходимости его замены одним из двух или более дескрипторов используется ссылка «исп а» (используй альтернативно):

индексация
исп а ИНДЕКСИРОВАНИЕ; ИНДЕКСЫ

В табл. 5.4 приведены возможные связи между дескрипторами и ключевыми словами в Тезауруса по информатике с указаниями типа отношения.

Таблица 5.4.

Обозначение	Название	Тип отношения
/…/	расшифровка	эквивалентность
(…)	релятор	категория
исп к	используй комбинацию	эквивалентность
исп а	используй альтернативно	эквивалентность
С	Синоним	эквивалентность
В	вышестоящий	иерархия
Н	нижестоящий	иерархия
А	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 [Кулик1977]:

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.

Приведем примеры терминов из Тезауруса по информатике:

- ключевое слово:

языковые системы
см ЯЗЫКИ

- дескриптор без связей:

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ
а АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ИНФОРМАЦИИ
КЛАССИФИКАЦИИ
КЛАССИФИЦИРОВАНИЕ

- родоначальный дескриптор:

ЯЗЫКИ
с подъязыки
языковые системы
н ЕСТЕСТВЕННЫЕ ЯЗЫКИ
ИСКУССТВЕННЫЕ ЯЗЫКИ
МЕТАЯЗЫКИ
а ЯЗЫКОВЫЕ СРЕДСТВА
ЯЗЫКОЗНАНИЕ

- видовой дескриптор первого уровня:

ЕСТЕСТВЕННЫЕ ЯЗЫКИ
с ЕЯ
в ЯЗЫКИ
а ВХОДНЫЕ ЯЗЫКИ

ВЫХОДНЫЕ ЯЗЫКИ
ИНОСТРАННЫЕ ЯЗЫКИ
ИПЯ
МОДЕЛИ ЯЗЫКА
ПИСЬМЕННОСТЬ
ЯЗЫКИ ЗАПРОСОВ
ЯЗЫКОВЫЕ БАРЬЕРЫ
ЯЗЫКОВЫЕ СРЕДСТВА

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-ом 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* те-

зауруса 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 :

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

В Тезаурусе по информатике лексические единицы распределены следующим образом: всего терминов – 3330, из них дескрипторов – 1647 (49,46%), ключевых слов – 1683 (50,54%).

В табл. 5.5 и на рис. 5.9 представлен словарный состав Тезауруса по информатике.

Таблица 5.5.

№ пп	Множество	Количество	% от общего количества
1	дескрипторы без связей	651	39,53
2	родоначальные дескрипторы	95	5,77
3	видовые дескрипторы I уровня	406	24,65
4	видовые дескрипторы II уровня	272	16,51
5	видовые дескрипторы низших уровней	233	14,15

На I-ом уровне специфичности в Тезаурусе по информатике находится 746 дескрипторов, или 45,29% всех дескрипторов тезауруса.

На II-ом уровне специфичности – 1152 дескриптора или 69.95% от общего количества дескрипторов.

Степень развития иерархической структуры тезауруса: $I = \frac{1647}{746} \approx 2{,}208$

Коэффициент синонимии: $S_n = \frac{1684}{3330} \approx 0{,}506$

Рис. 5.9. Словарный состав Тезауруса по информатике

***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 индексировании (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. [Тезаурус1988]

Основные проблемы, с которыми приходится сталкиваться при построении тезауруса, можно разделить на три класса: [David1993]

- 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-комногим», 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.

Для анализа использования лексики тезаурусов было проведено исследование динамики использования дескрипторов и ключевых слов информационно-поискового тезауруса по информатике [Информационно-поисковый1987] при индексировании базы данных ВИНИТИ РАН «Информатика». Результаты исследования представлены в Таблице 5.6 и на Рис. 5.10.

Таблица 5.6.

Анализ использования дескрипторов тезауруса

Год	Кол-во док-тов	Кол-во кл. сл. в поле KW	Кол-во дескрипторов тезауруса в словнике	Доля дескрипторов в поле KW
1981	5	18	11	0,611
1982	23	93	59	0,634
1983	42	147	111	0,755
1984	167	454	258	0,568
1985	713	1461	628	0,430
1986	4718	5242	1408	0,269
1987	6165	6398	1461	0,228
1988	6575	6518	1428	0,219
1989	7017	7000	1406	0,201
1990	6715	6805	1350	0,198
1991	5699	6515	1226	0,188
1992	4473	5977	1113	0,186
1993	3932	5218	1018	0,195
1994	4487	7436	1008	0,136
1995	4424	9220	975	0,106
1996	4418	9107	954	0,105
1997	3323	8445	871	0,103
1998	3838	9848	915	0,093
1999	3927	10300	862	0,084
2000	672	2727	401	0,151

Рис. 5.10. Распределение доли дескрипторов тезауруса по годам.

Исследования показали, что доля дескрипторов тезауруса, использованных при индексировании документов, существенно уменьшается с течением времени.

Контрольные вопросы

1. Охарактеризуйте состав лингвистического обеспечения документальных ИПС.
2. Определите понятие «искусственный язык».
3. Определите понятие «информационно-поисковый язык».
4. Дайте определение и приведите примеры парадигматических и синтагматических отношений между лексическими единицами ИПЯ.
5. Охарактеризуйте назначение и приведите типологию информационно-поисковых языков.
6. Дайте определение классификации.
7. Охарактеризуйте сходства и отличия перечислительных и аналитико-синтетических классификаций.
8. Охарактеризуйте свойства и приведите примеры перечислительных классификаций.
9. Охарактеризуйте свойства и приведите примеры аналитико-синтетических классификаций.
10. Дайте определение дескрипторного ИПЯ.
11. Охарактеризуйте метод координатного индексирования.
12. Перечислите и охарактеризуйте недостатки чистой координации.
13. Приведите типологию терминологических структур.
14. Приведите примеры использования линейных терминологических структур при индексировании и поиске.
15. Приведите примеры иерархических терминологических структур.
16. Дайте определение понятия «тезаурус».
17. Охарактеризуйте назначение и структуру информационно-поискового тезауруса.
18. Приведите примеры тезаурусов.

6. Поисковые задачи и технологии информационного поиска

В задачах информационного поиска качественно различают две составляющие: *концептуальную* и *технологическую*.

К концептуальным составляющим относятся, прежде всего, методы и средства представления собственно информации (знаний) и метаинформации, которые используются в качестве основы как для проектирования механизма поиска, так и для организации процессов взаимодействия пользователя с АИПС.

К технологическим составляющим относятся средства пользовательского интерфейса, алгоритмы индексирования и поиска, языки запросов, средства интеграции информации из различных источников и т.д.

Как отмечалось ранее (см. главу 2), принципиально важным фактором, определяющим направление развития современных информационных систем, является то, что взаимодействие пользователей с информационными ресурсами происходит в режиме «информационного самообслуживания», когда пользователь, по существу, уже не разделяет свою деятельность на информационную и основную[37].

Особенности технических решений при проектировании и эксплуатации автоматизированных информационных систем, ориентированных на информационную поддержку основной деятельности и интегрирующих такие специализированные функции, как поиск, обработка и организация информации, определяются двумя следующими, имеющими разную природу, факторами.

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

2. Поисковые средства и технологии, используемые для реализации информационных потребностей, определяются типом и состоянием решаемой пользователем задачи основной деятельности: соотношением его знания и незнания об исследуемом объекте. Кроме того, процесс взаимодействия пользователя с системой определяется уровнем знания пользователем содержания ресурса (полноты представления, достоверности источника и т.д.) и функциональных возможностей системы как

[37] Это особенно важно учитывать в задачах информационного обеспечения научных исследований, когда объект поиска не может быть четко определен заранее и когда цель поиска, сформулированная на начальной стадии работы, может измениться уже в процессе самого поиска, например, при ознакомлении с найденным документом. Причем факт изменения цели возможно даже не будет явно осознан исследователем, что в итоге может привести к неполному результату поиска.

инструмента. В целом эти факторы обычно сводятся к понятию «профессионализма» - информационного (подготовленный/неподготовленный пользователь) и предметного (профессионал/непрофессионал).

6.1. Динамика информации в системах основной и информационной деятельности

Взаимодействие пользователя с комплексом разнородных информационных ресурсов должно рассматриваться как процесс, зависящий от двух групп основных факторов. С одной стороны – это свойства информации и закономерности информационных преобразований в сфере основной деятельности (ОД), учитывающие специфику восприятия и переработки человеком как основной (целевой) информации, так и технологической, обеспечивающей условия его взаимодействия с информационной средой. С другой стороны, организация информационного пространства должна рассматриваться как задача такого управления ИР, при котором персональная АИС пользователя позволяла бы работать с ними как с единым ресурсом, что требует решения вопроса об идентификации ресурсов, а на уровне потребителя информации связано с проблемами разработки интерфейсов и средств доступа, обеспечивающих *персонификацию* представления информационных объектов.

Рассмотрим обобщенную схему воспроизводства информации, в основу которой положено предложенное в [Попов1996] представление совокупной информационной системы (генератор – потребитель информации), определяющее исследуемые объекты и процессы автоматизации в контексте взаимозависимости основной и собственно информационной деятельности (рис. 6.1).

С точки зрения задач управления потоками здесь можно различить две совокупности процессов: *формирование потока информации* (документов) в соответствии с заданными характеристиками (тематичность, полнота охвата и т.д.) и *распределение входных и выходных потоков* и их составляющих в соответствии с информационными потребностями[38]. И, если основная деятельность имеет дело с поиском и содержательной обработкой научной информации (т.е. сообщениями, описывающими некоторые свойства исследуемого объекта), то научно-информационная – это по возможности инвариантные относительно смысла преобразования текста в форму, приемлемую для автоматизированной идентификации, хранения и поиска.

[38] Отметим, что информационные потребности и запросы также можно рассматривать как поток информации - гипотетической или в какой-то части неактуализированной.

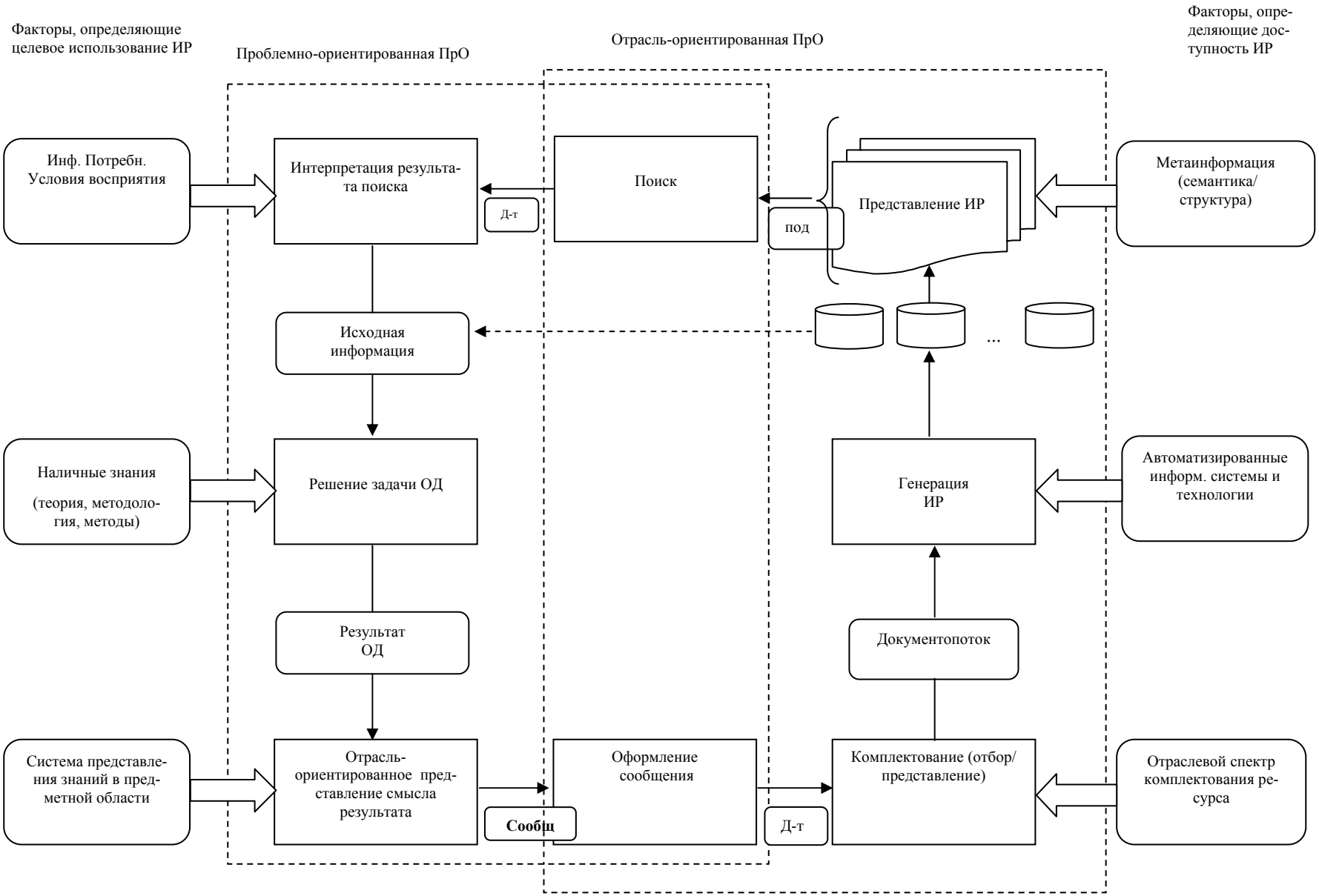


Рис. 6.1. Обобщенная схема воспроизводства информации

По характеру информации в совокупной системе (рис. 6.1) можно выделить три следующих *уровня преобразования* информационных объектов.

Первый уровень – это основная деятельность, где объектами являются предметы реального мира, а результатами – новое знание. Носителем информации этого уровня является человеческое сознание, для которого характерны системность организации и ассоциативность выборки, а коммуникационным объектом является сообщение – знание, адресно отраженное на систему понятий предполагаемого приемника - потребителя информации.

Второй уровень – создание общественно-полезной информации – одна из форм овеществления знаний через обобществление результатов в документальной форме. Средством представления знаний (коммуникаций) здесь является язык, а носителем – документ как функционально ориентированное сообщение, структурирующее информацию и идентифицирующее ее, например, путем выделения логических или физических частей - семантически однородных полей.

Третий уровень – собственно информационная деятельность – управление потоками информации для обеспечения основной деятельности. Работа с компактными по объему вторичными документами, позволяет совершенствовать процесс поиска нужных сообщений. Здесь информация (поисковый образ документа) – это хорошо структурированный материал, компактно и предметно отражающий содержание документа, а также обеспечивающий идентифицируемость документа в целом и на уровне отдельных элементов данных.

Для выявления характера взаимосвязи информационных объектов используем приведенное в гл. 1 определение понятия «информация» как отражения результата упорядочения и ограничения разнообразия описаний объектов ОД и их взаимосвязей (в реальном мире) в соответствии с требованиями, обуславливаемыми возможностями средств представления (языка описания). Отсюда следует, что использование абстракций различного порядка в итоге дает возможность (упрощая описание объекта одного семантического уровня за счет введения объектов другого уровня) представлять объекты с помощью конечного числа терминов. Соотношение и характер взаимосвязей информационных объектов, форм и средств их представления, рассматриваемых в контексте задач информационного обеспечения основной деятельности, приведены на рис.6.2.

Здесь преобразование форм представления информации является последовательным отражением содержания, а по существу - фильтрацией информации путем снижения разнообразия форм и аспектов представления смыслового содержания через вынесение части смысла в метаинформационную составляющую или простое отбрасывание.

Например, сообщение предполагает фиксацию (ограничение) предметной области; документ – фиксацию вариантов способа пред-

ставления через выделение семантически однородных полей и, соответственно, определение характера и способа их наполнения; поисковый образ фиксирует способы указания значения отдельного элемента (типа данных).

Соответственно, адекватность средств отражения информации (а в случае информационно-поисковых систем это средства лингвистического обеспечения) должна рассматриваться как с точки зрения возможности неискажающего преобразования самой информации в цепи генерации-потребления информационного ресурса, так и с точки зрения адекватности восприятия пользователем функциональных возможностей этих средств.

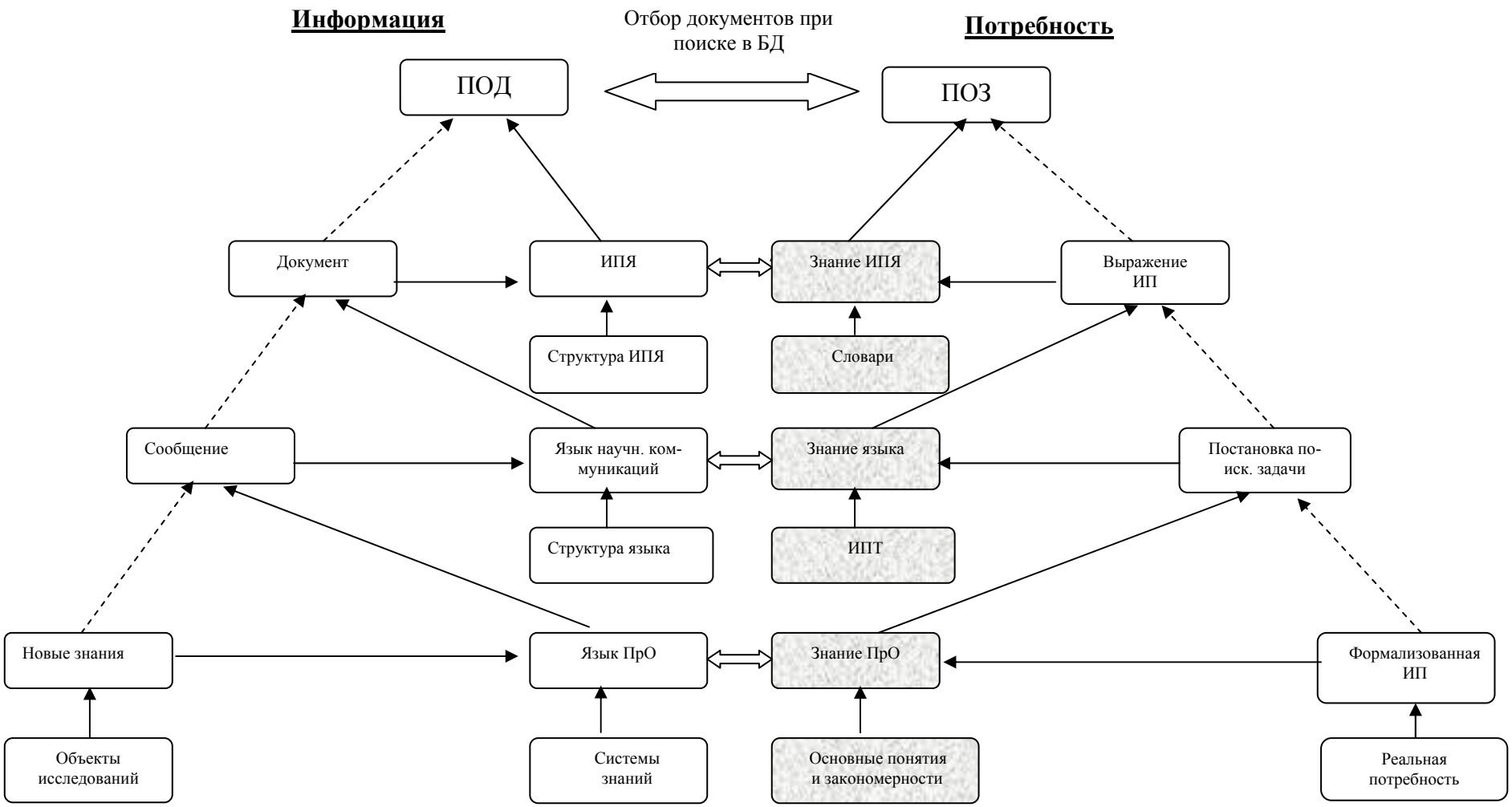


Рис. 6.2. Уровневая модель взаимосвязи информационных объектов

6.2. Поисковые задачи и виды информационного поиска

Поскольку автоматизированная система является всего лишь *инструментом, используемым человеком при поиске* (а не *интеллектуальным автоматом для поиска* информации - готовых решений задач основной деятельности), эффективность ее использования зависит от того, насколько хорошо человек знает природу объектов и свойства инструмента, посредством которого он работает с этими объектами.

Особенностью поискового процесса, рассматриваемого как взаимодействие двух систем представления знаний (пользователь – ИПС), является многоуровневость и, часто, неоднородность объектов в цепи информационных преобразований.

6.2.1. Типология поисковых задач

Операционными объектами, непосредственно участвующими во взаимодействии (сравнении потребности и документов в базе данных), являются поисковый образ документа (ПОД) и поисковый образ запроса (ПОЗ), соответствие которых устанавливается поисковым механизмом АИПС на формальном уровне (рис. 6.2). Установление же истинного соответствия предполагает соотнесение содержания на смысловом уровне: пользователь практически реконструирует возможное содержание по перечислению основных понятий и далее полученный образ соотносит с реальной потребностью. При этом адекватность образа действительному содержанию документа определяется не только качеством процесса свертки информации, но и уровнем знания субъектом средств отражения - концептуальной схемы предметной области и возможностей информационно-поискового языка.

В соответствии с характером задач 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. Поисковая модель (логическая идентификация объекта поиска) может быть представлена как поиск по логическому выражению 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 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 метода решения задачи. Поисковая модель в этом

случае - поиск по части известного понятия с использованием накопленных ранее результатов.

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. Логическая поисковая модель - поиск «похожих» документов, содержание которых ассоциируется с задачей пользователя.

6.2.2. Типология информационных потребностей

Характер информационных потребностей[39] в значительной степени определяется формой представления знания, которая в свою очередь зависит от среды – носителя информации. В контексте многоуровневой модели процессов преобразования информации (рис. 6.2) и в соответствии с терминологией [Tailor1968] определяют следующие типы информационной потребности пользователя (ИПП).

Реальная информационная потребность. Это потребность в информации, еще не вполне осознанная, но отражающая проблемную ситуацию пользователя, характерная для начальной стадии ОД.

Осознанная ИПП. В процессе понимания проблемной ситуации реальная ИПП преобразуется в *осознанную* ИПП*, представленную в виде вопроса или задачи,* которую далее пользователь выражает на привычном ему языке, формируя *запрос на естественном языке* и затем переводя его в *поисковый запрос,* представленный в терминах ИПЯ. Для *запроса* характерно то, что вопросы типа «как» и «почему» должны быть преобразованы в вопрос типа «ли», поскольку именно такая форма представления потребности является наиболее адекватной теоретико-множественой модели поиска. Отметим, что преобразование вопроса в запрос имеет по существу качественный характер.

Переход от реальной к осознанной ИПП тем сложнее, чем менее определена задача пользователя. Для поисковых задач проблемного типа этот переход наиболее труден, так как пользователь не представляет, *какая именно* информация нужна для решения его задачи.

---

[39] Информационная потребность рассматривается как дополнение, возможно – гипотетическое, известного знания.

6.2.3. Типология информационной неопределенности и виды информационного поиска

Особенности представления информации на разных уровнях человеко-машинной среды обусловливают различные типы неопределенности – семантическую, лингвистическую, метаинформационную (последняя относится как к семантике, так и к синтаксису представления информации).

В этом смысле процесс поиска можно определить как последовательность шагов, задачи которых - снятие перечисленных неопределенностей.

Семантическая неопределенность связана с формализацией запроса. Формируя запрос, пользователь явно или неявно синтезирует ту информацию, которая, возможно, есть в отыскиваемом тексте. Определяются понятия и связи между ними, т.е. происходит реконструкция пользователем гипотетического текста, предположительно совпадающего в известной части проблемы с возможно уже существующим текстом, и обозначение связи известного знания с выявленным неизвестным.

Лингвистическая (лексическая) неопределенность связана с формулировкой ПОЗа. Формулируя запрос, пользователь должен учитывать, что его представление об информативности термина необязательно совпадает с представлениями индексатора. Для ИПЯ дескрипторного типа это в значительной степени лексическая неопределенность.

Метаинформационная неопределенность связана с тем, что пользователь должен иметь адекватное представление о самой системе и способе представления информации в ней. Например, как и по каким полям проводить поиск.

Рассмотрим человеко-машинный поиск информации как процесс отыскания неизвестных (по крайней мере, для субъекта поиска) сведений (фактов, идей и т.д.), необходимых для получения знания, нового для данной предметной области (например, позволяющего вскрыть связи между фрагментами знания или найти примеры, опровергающие научные гипотезы).

Такой процесс характеризуется двойственностью целей человека. С одной стороны – это создание нового знания, включая этапы структуризации и формализации проблемы, нахождения или разработки методов решения. С другой стороны – это поиск сообщений и оценка полезности найденного.

Столь же значимой особенностью поисковой ситуации является опосредованность отбора информации, материально представленной в виде документов. Потенциально полезные документы (предположительно содержащие нужные сведения) выделяются из всего доступного множества через соотнесение поисковых образов (информационной потребности и содержания документа, выраженных средствами информационно-поискового языка).

Аналогичная опосредованность наблюдается в случае рассмотрения сред представления информации: смысловая обработка (соотнесение содержания сообщения с реальной, т.е. осознанной потребностью) происходит в сознании человека, а отбор документов, формально соответствующих потребности – в машинной среде с жесткой двоичной логикой. Причем такая схема установления соответствия (отражения) построена на сведении информационной потребности (как неопределенности и неизвестности) к перечислительной форме *гипотетически известного*, представляющей потребность гипотетическими документами. Этот прием обеспечивает однородность сопоставляемых поисковых образов и применим в том числе и к наиболее распространенным видам информационного поиска, например, библиографическому или отысканию публикаций об объектах, уже существующих, так или иначе известных субъекту.

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

1) неопределенности соотношения «известного/неизвестного» в предмете поиска;

2) неопределенности системы характеристических признаков для структуризации предмета поиска;

3) семантической неопределенности формулировки предмета поиска;

4) лексической неопределенности как фактора степени соответствия информационно-поискового языка естественнонаучному языку предметной области;

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

6) неопределенности интерпретации ПОЗов (субъективность и неполнота реконструирования пользователем смысла найденных документов).

Не являясь практически измеримыми величинами, эти параметры, тем не менее, позволяют обозначить характер изменения состояния сторон и структурировать процесс, выделяя компоненты не столько по функциональному, сколько по структурному принципу. 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.

В вопросно-ответной логике обнаружения нового знания [Белнап1981] одна из форм вопроса - это задание списка альтернатив ответа

и правила (алгоритмы) построения прямого ответа на основе этого списка. В этом контексте информационный поиск средствами неинтеллектуальной[40] АИПС – это только первая составляющая: нахождение сообщений, предположительно содержащих прямой ответ (или его составляющие, которые в дальнейшем могут быть объединены субъектом поиска), причем не обязательно альтернативных и не обязательно удовлетворяющих требованиям полноты и различения.

Резюмируя вышесказанное и учитывая, что особенностью поисковой ситуации является то, что пользователь за новым знанием обращается в массив уже известного знания (хотя, возможно, и противоречивого[41]), будем представлять поисковый запрос как гипотетический документ, описывающий реальный или предполагаемый или создаваемый объект. То есть, в этом контексте задача поиска может быть сформулирована следующим образом: *найти уже существующие документы, которые являются содержательным аналогом запрашиваемого гипотетического*[42].

Для случая свойственной информационно-поисковым системам атрибутивной модели представления смысла (в том числе и вопроса) объект задается набором характеристических признаков и связей. Внутренние связи определяют структуру самого объекта, а внешние – структуру взаимоотношений с другими объектами.

Тогда запрос, рассматриваемый на понятийном уровне – это структурно-логическое определение неизвестного (реальная информационная потребность) через известные характеристические признаки и связи, если предполагаемый аналог существует, или, в противном случае – через дополнение, т.е. характеристические признаки и связи объектов, частью которых является или с которыми связан искомый объект.

С другой стороны, поскольку запрос (его поисковый образ) является формально описанной моделью информационной потребности пользователя, то по смысловыражению ПОЗ и ПОД должны быть приведены в соответствие. Т.е. либо пользователь должен принимать концептуальную платформу (и знаковую систему) индексатора, либо, если ИПС является "интеллектуальной", она на основе запроса должна реконструировать проблему, решаемую пользователем. Однако при этом надо учитывать, что и формулировка пользователем своей информационной

---

[40] Интеллектуальной системой можно считать [Смирнов1981] информационную систему, обладающую следующими свойствами: способностью упорядочивать массив сведений по степени существенности; способностью извлекать из массива все возможные сведения как следствия, выводимые посредством логики; способностью к «рефлексии», т.е. оценке хранимых сведений; способностью формировать новые типы вопросов в ответ на получение новой информации из внешнего мира.

[41] Массив системы включает в том числе и документы, содержащие сведения, например, об исследованиях, которые могут быть неполными, непроверенными, а также взаимно противоречивыми.

[42] При этом следует учитывать «вариантность» такого рода гипотетических документов, являющуюся, в том числе, и следствием профессионального менталитета потребителя, например, отмеченную в [Cory1999, Ellis1993] склонность гуманитариев искать новые аналогии, а не причины. Учет этого фактора особенно важен при проектировании интерфейсов поисковых систем.

потребности в виде запроса, и оценка меры соответствия найденной информации реальной потребности носят субъективный характер.

Таким образом, с точки зрения структурной полноты определения поискового объекта[43], типологию поисковых задач можно представить в виде таблицы (табл. 6.1).

Таблица 6.1

Вид поиска	**Логическая модель объекта поиска**	**Логическая модель механизма поиска**
Предметный (атрибутивный) поиск	Объем понятия, задаваемого именем	Поиск по логическому выражению над именами понятий, задаваемыми терминами или их комбинацией (значениями определенного характеристического признака)
Тематический поиск	Определение нового понятия или понятийных связей, косвенно определенного объемом этого понятия	Поиск по части известного понятия (или связям), частично задаваемым комбинацией характеристических признаков, с использованием накопленных ранее результатов
Проблемный поиск	Документальное определение нового понятия или связей путем реконструкции образа по его части.	Поиск «похожих» документов, поиск с использованием технологии «обратной связи».

Рассмотрим типологию видов поиска с точки зрения семиотики как знаковой системы, которой свойственна неизоморфность отображения системы обозначающих (знаков) системе обозначаемых (объектов - денотатов). Рассматриваемая типология ассоциируется со следующими семиотическими ситуациями.

---

[43] Отметим, что в контексте представления объекта поиска как системы, выделение трех видов поиска соответствует существу приведенного ранее (глава 2) определения системы S_i как композиции множества первичных элементов M_i через задание набора системообразующих признаков A_i , системообразующих отношений R_i и системообразующего закона композиции Z_i.
В этом контексте предметный (атрибутивный) поиск – это нахождение объекта - системы $S_i = \{M_i\}$, по заданному (полностью определенному) системному его основанию $< A_i, R_i, Z_i >$. Тематический поиск - это нахождение подмножества систем $\{S_i , i=1,n\}$, для которых задано Z_i и одно из оснований A_i или R_i. Проблемный поиск – это разновидность тематического поиска с не единственным законом композиции Z_i .

Предметному поиску соответствует ситуация формирования (выбора) знака (знаковой конструкции), устраняющего неопределенность знаковой системы в контексте полноты и точности представления объекта, т.е. такого знака, который позволит эффективно выделить (отличить) объект из множества других при фиксированном (единственном) концепте, что иллюстрируется рисунком 6.3 а).

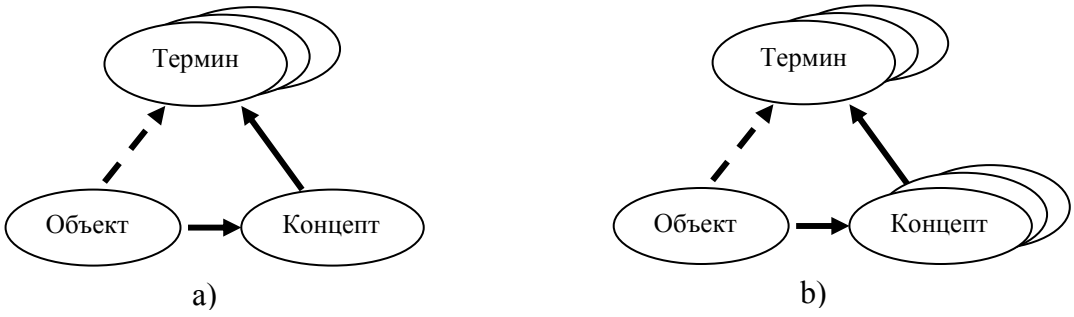


Рис. 6.3. Семиотические ситуации поиска

Для случая *тематического поиска* ситуация отличается тем, что мы имеем упорядоченное ограниченное множество концептов, позволяющих представлять объект в различных аспектах, что представлено на рис. 6.3 b).

Для случая *проблемного поиска* мы уже имеем неупорядоченное и не четко определенное множество концептов.

6.3. Компоненты и обобщенная схема информационного поиска

Функционирование современных ИПС основывается на двух предположениях: 1) документы, необходимые пользователю, объединены наличием некоторого характеристического признака или комбинации признаков; 2) пользователь способен указать эти признаки. Оба эти предположения на практике редко выполняются и можно говорить только о вероятности их выполнения. Поэтому, процесс поиска информации обычно представляет собой последовательность шагов, приводящих при посредстве системы к результату, качество которого будет иметь случайную природу и определяться многими факторами. При этом поведение пользователя, как организующее начало управления процессом поиска, определяется не только информационной потребностью, но и инструментальным разнообразием системы - технологиями и средствами, предоставляемыми системой.

Отметим, что понятия стратегия и технология поиска, средства и методы, модели и алгоритмы являются достаточно употребляемыми, однако разные авторы используют эту терминологию в разных контекстах и зачастую наделяют разным содержанием. Приведем несколько примеров.

В [Шкаренкова1987] стратегия поиска сведена к выбору критерия отбора, максимизирующего количество релевантных документов или минимизирующего количество выданных документов или обеспечивающего уникальность получаемой информации.

В [Bates1987] стратегия поиска определяется как общий план диалогового сеанса, а тактика - как путь дальнейшего перемещения в процессе поиска. В [Pejtersen1986] вводятся библиографическая, аналитическая и эмпирическая стратегии, которые выделяются по специфичности информационной потребности и требованиям к знаниям пользователя и ресурсам АИПС. Библиографическая стратегия ориентирована на поиск по заданным библиографическим характеристикам документа, аналитическая - на поиск, когда конкретные характеристики документа отсутствуют, однако требования к его информационному содержанию известны достаточно точно. Эмпирическая - на поиск, когда вводится информация о пользователе, которая используется для построения профиля его интересов. Профиль сопоставляется с хранящимися в системе профилями-прототипами (если профиль близок к прототипу, в качестве ответа на запрос предлагаются результаты поиска по запросу-прототипу). В [Димитрова1995, Карначук1986] определяются четыре основных класса стратегий информационного поиска: случайная (последующий вариант формулировки поискового предписания никак не связывается с предыдущими пробами), стратегия расширения, стратегия сужения, смешанная (композиционная) стратегия. В [Bates1995] стратегия связывается с выбором различных подвидов функции просмотра и печати результатов поиска с точки зрения снижения суммарных затрат пользователя.

В [Boughanem1999, Shaw1994, Spink1997] рассматриваются технологии отбора документов и формы задания условия отбора; определяется поиск с обратной связью по релевантности как одна из поисковых стратегий. В [Mohan1993] рассматриваются технологии поиска с использованием логических операторов и методов сходства документов по ближайшему окружению. В [Spink1997] как ключевая часть поиска рассматривается интерактивный отбор поисковых терминов. В [Kerr1990] информационный поиск рассматривается как способ ориентации в базе данных. В [Miyamoto] предложена логическая модель информационного поиска в ситуации неопределенности.

В [Rijsbergen1988] рассматриваются три вида методов поиска: 1) математические (вероятностный, векторного пространства и кластеризации); 2) лингвистические; 3) алгоритмические. В [Ingwersen1988] анализируются четыре метода информационного поиска (булева логика с точным совпадением, расширенная булева логика, вероятностный поиск, поиск по кластерам) в связи с характером информационной потребности. В случае уточнения (пользователю известны какие-то библиографические признаки нужного ему документа) наиболее эффективен поиск на точное совпадение терминов. При тематическом поиске (пользователь может сформулировать тематику своего запроса в адекватных понятиях

и терминах) рекомендуется использовать кластерные или вероятностные методы и расширенную булеву логику. В случае неопределенного поиска (пользователь хочет получить информацию о понятиях и отношениях в малознакомой ему тематической области) лучше применять кластерные методы.

Введем определения следующих понятий, важных с точки зрения моделирования и эксплуатации информационно-поисковых систем.

Методы поиска - это 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.

Механизмами поиска будем называть реализованные в системе модели и алгоритмы процесса формирования выдачи документов в ответ на поисковый запрос.

Средства поиска – это, с одной стороны, взаимозависимый комплекс ИПЯ и языков определения/управления данными, обеспечивающий структурные и семантические преобразования объектов обработки (документов, словарей, совокупностей результатов поиска), а с другой – это объекты пользовательского интерфейса как технологические решения, обеспечивающие управление последовательностью выбора операционных объектов конкретной АИПС.

Поисковые технологии – унифицированные (оптимизированные в рамках конкретной АИПС) последовательности эффективного использования в процессе взаимодействия пользователя с системой отдельных средств поиска для устойчивого получения конечного и, возможно, промежуточных результатов.

Стратегия поиска - общий план (концепция, предпочтение, предрасположенность, установка) поведения пользователя для выражения и удовлетворения информационной потребности, обусловленный как характером цели и типом поиска, так и системными «стратегическими» решениями - архитектурой БД, а также методами и средствами поиска конкретной АИПС. Выбор стратегии в общем случае является оптимизационной задачей, однако, на практике в значительной степени определяется искусством достижения компромисса между практическими потребностями и возможностями имеющихся средств. С точки зрения способа задания условия соответствия информационной потребности информационным ресурсам можно говорить о двух «чистых» стратегиях – «вербальной», являющейся аналогом функционального задания, и «кластерной» - отражающей особенности перечислительного способа.

Навигация как реализация процесса поиска по запросу в выбранной БД - это целенаправленная, определяемая стратегией, последовательность использования методов, средств и технологий конкретной АИПС для получения и оценки результата.

Средства навигации позволяют пользователю осуществлять управление процессом поиска. Такие средства предоставляются пользователю в виде *интерфейса*, позволяющего организовать более или менее эффективный процесс взаимодействия с базой данных. При этом «дружественность» интерфейса характеризуется не только эргономичностью и понятностью, но и вариантностью выбора операционных обьектов. Именно средства интерфейса должны «объяснить» существо системы, т.е. обеспечить основу и поддержку определения стратегии поведения, а также соотнесения каждого действия с обобщенной моделью процесса, явно задать содержание каждого объекта и обосновать переход от предположений к действиям по реализации процесса поиска.

6.3.1. Обобщенная схема информационного поиска

Процесс поиска информации представляет собой последовательность шагов, приводящих при посредстве системы к некоторому результату, и позволяющих оценить его полноту (может ли пользователь быть уверенным в том, что полученный результат – исчерпывающий и ничего более по искомой проблеме не содержит). Так как пользователь обычно не имеет исчерпывающих знаний об информационном содержании ресурса, в котором проводит поиск, то оценить адекватность выражения запроса, равно как и полноту получаемого результата, он может, основываясь лишь на внешних оценках или на промежуточных результатах и обобщениях, сопоставляя их, например, с предыдущими.

В контексте рис. 6.2 задача поискового процесса – построить согласованное множество моделей объекта поиска (т.е., осознанной, зафиксированной в сознании человека информационной потребности). Для информационной цепи формирования ПОЗа – это две имеющие лингвистическую природу модели, относящиеся к двум верхним уровням: 1) модель коммуникативная, являющаяся представлением ИП, ориентированным на передачу для соотнесения с аналогично представленными объектами, описанными в уже опубликованных документах; 2) модель поисковая, атрибутивно представляющая ИПП и ориентированная на точечное (теоретико-множественное) соотнесение с аналогично представленными поисковыми образами документов.

Согласованность здесь имеет двоякий смысл: очевидное (вертикальное) соответствие выражения поисковому образу документа (ПОД – это выражение, представленное средствами ИПЯ), и согласованность «горизонтальная» - степень соответствия представления пользователя реальным возможностям языка для выражения ИП.

Поскольку информационный поиск реализуется в машинной среде, в основе которой лежит двоичная логика, процесс любого отбора должен быть сведен к атрибутивной модели, т.е. отбору через задание имени и значения характеристического свойства (атрибута). Таким образом, только для случая предметного поиска отбор документов может

быть реализован по «одноактной» схеме – операцией проверки, есть ли в базе данных документы, имеющие заданное значение атрибута.

Для того чтобы пользователь имел возможность реально управлять процессом поиска (на основе объективных данных, позволяющих оценить эффективность выполняемых действий), необходимо произвести декомпозицию целостной с точки зрения конечной задачи пользователя запросно-ответной схемы процесса поиска. Такая функциональная декомпозиция должна в итоге обеспечить возможности для последовательного снятия неопределенностей всех типов, что в организационном плане выражается в выделении подпроцессов-процедур и соответствующих операционных объектов. С точки зрения целевого назначения ИС, то есть для процесса поиска в целом, мы имеем всего два типа *основных операционных объектов* - запрос и документ, которые представляют средствами языка некоторый семантически целостный фрагмент предметной области. Другие операционные объекты - *технологические* в рамках декомпозированного процесса – это самостоятельные семантически значимые объекты метаинформационного назначения, или объекты, производные от основных[44]. Назначение и природа технологических объектов – дать возможность локализовать и снять или зафиксировать неопределенность отдельного типа.

И запрос, и документ являются моделями, представляющими средствами языка отдельные части и аспекты некоторого целостного фрагмента предметной области.

Используя определение понятия «система» применительно к схеме «основная - информационная деятельность» (глава 2), получаем, что в контексте уровневой модели преобразования информационных объектов фрагменту ПрО можно поставить в соответствие два взаимосвязанных подмножества систем $\{S_i\}$ и $\{S_i^1\}$.

Подмножество систем S_i представляет предметную область на «вещном» уровне, где m_i, R_i – объекты и связи предметной области, выделяемые в рассмотрение в соответствии с целями ОД, задаваемыми Z_i, и характеристическими признаками A_i .

Подмножество систем S_i^1 , представляющее ПрО на информационном уровне, образуется отдельными сообщениями, где, соответственно, m_i, R_i – имена объектов и связей (уже включая отношения терминологической системы), а Z_i отражает еще и точку зрения автора сообщения.

Документ является *конкретной* (хотя и не единственной) формой выражения определенной проблемной ситуации (разрешение которой было предметом ОД, что и привело к появлению этого документа). ПОД представляет эту конкретику композицией в общем случае не уникальных характеристических признаков, выбираемых из множества призна-

[44] Отметим, что имеющие лингвистическую природу, как основные, так и технологические объекты могут быть представлены как в прямой, так и в инвертированной форме.

ков, свойственных и другим объектам, информация о которых хранится в базе данных.

Цель создания ПОДа – представить изначально уникальный смысл документа компактной композицией признаков (например, в случае дескрипторных ИПЯ - ключевых слов), по возможности, не увеличивая комбинативность порождаемых ими возможных смыслов.

Цель построения ПОЗа – сохраняя уникальность проблемной ситуации, увеличить комбинативность смыслов, порождаемых композицией поисковых признаков запроса, для того, чтобы максимально охватить аспекты представления объекта поиска.

Для человека идеальной коммуникативной формой представления реальной ИПП является вербальная, где характеристические признаки неизвестного (искомого) будут связаны с конкретным контекстом проблемной ситуации, то есть запрос фактически будет представлен как документ, содержащий высказывания, которые в гипотетической форме описывают предположительно существующие объекты[45].

Для задач ИС реальные объекты и связи отражаются в форме высказываний (описательно), которые воплощаются в виде логически связанных предложений документа. Принципиальным отличительным свойством высказываний при этом является изначальная контекстная определенность (хотя этот контекст, возможно, представлен только в сознании высказывающего). Отдельное высказывание, воспринимаемое как грамматическая форма (предложение), может порождать в сознании воспринимающего несколько смыслов, и исходный смысл высказывания будет воспринят только при условии передачи исходного контекста.

Именно потому, что воспринимающий использует ситуационную или собственную контекстную составляющую, человек извлекает из текста больше информации, чем явно выражено словами. В этом случае смысл слов на основе этого контекста может быть развернут до полных высказываний. Такое развертывание при восприятии текста человеком происходит неосознанно: человек, восстанавливая или генерируя смысл, не производит явного выделения лингвистических, логических и предметных компонентов в полученном сообщении – в сознании человека они неотделимы друг от друга и не представлены в «чистом» виде.

Необходимость выделения объектов появляется при частичной или полной автоматизации. Декомпозиция цельного процесса *поиска-восприятия-использования* информации необходима для распределения функций между подсистемами. Именно для перехода от предметного

[45] При этом содержание запроса частично или полностью может быть представлено реально существующими в БД документами (либо как решениями проблемной ситуации, либо как информационными сообщениями, имеющими какую-либо семантическую общность с ИПП). Это является основанием для утверждения, что форма и способ представления запроса принципиально имеет двойственную природу: при стремлении к совершенной вербальной (однородной и целостной) форме выражения запроса, в силу неизвестности, присущей реальной ИПП, часть или весь запрос может быть представлен в форме отдельных документов или их кластеров, что соответствует дискретной, фрагментарной «мозаичной» форме.

уровня (реальной потребности) к лингвистическому (запросу и его поисковому образу – выражению формализованной ИПП средствами языка) и, далее – к технологическому, необходимо явно выделить еще и логический уровень, где в качестве операндов используются понятия и структуры, что и обеспечивает переход к формированию формализованной ИПП (объекты, связи и контекст их представления) на уровне имен и лингвистических связей.

В случае процесса построения запроса с использованием слов естественного языка, которым свойственны синонимия и полисемия, для неискажающего смысл межуровневого перехода применяются метаинформационные, внешние по отношению к сообщению конструкции, позволяющие фиксировать (выбирать или указывать) контекст словоупотребления.

Полнота и точность передачи контекста при такой организации процесса зависят от вида запроса: для типовых по отношению к предметной области, и потому хорошо представленных метаинформационными средствами они могут быть высокими, а для запросов проблемных – низкими. При этом в процессе явно не передается ни контекст, ни *характер* неопределенности. Поэтому *первым шагом* должна быть *локализация* реальной ИПП – структурно-логическая декомпозиция ПрО, при которой для возможных аспектов рассмотрения объекта поиска выделяются характеристические понятия и связи. Это преимущественно не автоматизируемый процесс выделения структурных единиц мышления (понятий и высказываний), содержанием которых является предметная информация, а метаинформационная составляющая является априорной и обеспечивает семантику имен понятий, отражающую устоявшуюся прагматику предметной области, рассматриваемой в отдельных аспектах.

Следующий шаг – переход на лингвистический уровень – это нахождение множества имен понятий и, соответственно, терминов, образующих класс условной эквивалентности для каждого из исходных объектов в заданном аспекте. Другими словами, это формирование возможных вариантов грамматических (терминологических) конструкций, выражающих существо ИПП в каждом из возможных аспектов рассмотрения. Метаинформационный компонент обеспечивает семантику словоупотреблений, характерную для языка и отражающую практику индексирования. Этот шаг достаточно эффективно автоматизируется[46], однако вероятностный характер процесса построения классов условной эквива-

[46] Основой этого являются исследования и разработки в области теоретической и прикладной лингвистики – создание машинных словарей, средств и алгоритмов синтаксического анализа, систем машинного перевода. Эти достижения нашли широкое применение в сфере информационного поиска при создании АИПС с естественно-языковыми интерфейсами и, в частности, поисковых машин Internet. Отметим, однако, что средства такого рода не могут отражать семантику проблемной ситуации ИП пользователя уже хотя бы потому, что она отдельно не предусматривалась при их создании. Как следствие, в общем случае пользователь не может получить в одноактном процессе поиска полный ответ, используя только «интеллект» системы.

лентности предполагает обязательность последующей проверки состоятельности гипотезы – оценки пользователем адекватности терминов по степени их важности в контексте найденных по ним документов.

Несколько иная ситуация при *использовании* выдачи – содержания документов, найденных по запросу. Значение термина (как лингвистической переменной), его контекст как структурной единицы в рамках более крупных конструкций, таких как предложение или документ, определяется пользователем достаточно точно (хотя и субъективно) и обычно без явного использования метаинформации. Менее полно может быть определен смысл композиционных структурных единиц – предложения и сообщения в целом: в лучшем случае мы можем точно определить, дает содержание документа исчерпывающий ответ на практический вопрос или нет, однако мы не будем знать, исчерпывается ли этим весь смысл сообщения. При этом характер и контекст словоупотреблений может быть использован в качестве оценки эффективности запроса с точки зрения как семантики словоупотребления, так и семантики предметной области.

Соответственно, использование в запросе терминов, выбираемых из текста релевантных документов, является по существу реализацией схемы реформулирования запроса по обратной связи. По характеру контекста здесь мы имеем два типа обратной связи: внутреннюю – на лингвистическом уровне, и внешнюю – на уровне семантики предметной области.

Внутренняя обратная связь хорошо автоматизируется, поскольку отражает лингвистические особенности *использования* языка, адекватно представляемые статистическими характеристиками, построенными на основе частотных показателей БД. В том числе, как показано на рис. 6.4, дистрибутивно-статистический анализ лексики релевантных документов позволяет автоматически строить тематико-статистический спектр (ТСС) по теме запроса [Попов1973], используемый системой для ранжирования документов, а структурно-статистический анализ – строить мини-тезаурус[47] темы, который может быть использован не только для автоматического лексического расширения выражения запроса, но и в качестве отдельного технологического объекта. Такой мини-тезаурус является структурно-лингвистической моделью предметной области поиска, отражающей не только общепризнанные, но также и актуальные, характерные для проблемной ситуации особенности представления ИПП (в том числе, и может быть, в первую очередь – новизну подхода пользователя к решаемой им проблеме).

Существование мини-тезауруса в виде операционного объекта обеспечивает также и возможности построения автоматизированных технологий, реализующих внешнюю обратную связь. Мини-тезаурусы,

[47] Здесь под мини-тезаурусом понимается построенный в автоматизированном режиме словарь нормализованной лексики тематической области, не только упорядоченный по лексикографическому принципу, но и допускающий существование иерархических уровней.

получаемые (и используемые) в итеративном процессе поиска по теме запроса, образуют ряд объектов, что позволяет, используя дистрибутивно-статистические методы, количественно оценивать эффективность процесса.

В зависимости от способа выбора терминов и характера их использования для развития запроса можно выделить два типа процесса: модификация выражения запроса в том случае, когда запрос представлен в вербальной форме, и реформулировка запроса – если запрос представлен в кластерной форме. Этот фактор определяет конечность множества типов механизмов поиска.

Поскольку, как отмечалось ранее, принципиально есть две формы представления (выражения) ИП – вербальная и кластерная, каждая из которых отдельно не может быть исключительно полной и точной, механизмы поиска должны быть представлены двумя типами: у первых в качестве запроса используются терминологические *ИПЯ-выражения*, у вторых – документы.

Практически, в зависимости от предполагаемого разнообразия типов поисковых задач и типов интерфейсных технологических объектов, реализация АИПС может включать достаточно разнообразные механизмы поиска. Например, в АИПС IRBIS представлены следующие механизмы поиска. В случае формирования запроса на ИПЯ – это механизм поиска по совпадению терминов, когда поисковый запрос представляет собой множество терминов, присутствие хотя бы части которых обязательно в документе, или механизм поиска по логическому выражению, когда термины связываются логическими операциями, и для принятия решения о релевантности документа необходимо формировать результат вычисления логического выражения. Если же запрос представлен документами, то, в зависимости от типа интерфейсного объекта, мы имеем: в случае отдельного документа – поиск аналогов, если поиск выполняется автоматически, а если поисковые термины указываются в документе пользователем – поиск по совпадению терминов. Соответственно, если интерфейсный объект есть множество релевантных документов, то в случае, когда система формирует словник автоматически, мы имеем эвристический поиск, а если поисковые термины в словнике указываются пользователем – «поиск по контексту».

Обобщенная схема процесса поиска, поэтапно позволяющая локализовать неопределенности перечисленных ранее типов, приведена на рис. 6.4.

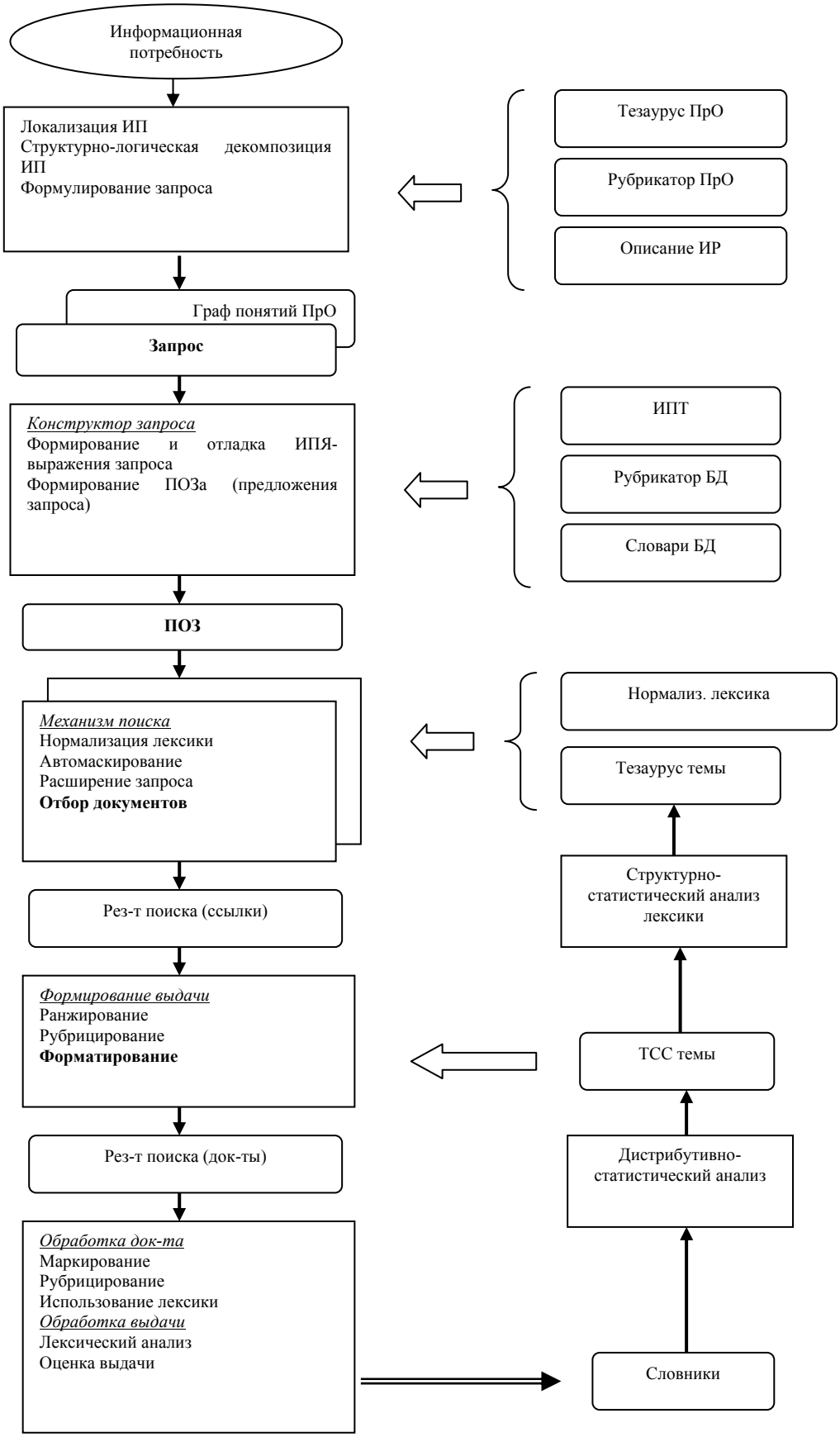


Рис. 6.4. Обобщенная схема информационного поиска

Как было отмечено ранее, задача информационного поиска относится к классу человеко-машинных. Уже на основании того факта, что образ информационной потребности имеет в качестве носителя сознание человека, и что *именно человек* производит сопоставление образа со смысловым содержанием отбираемых документов, а также оценивает адекватность используемых средств и объектов, можно сделать вывод, что система должна предоставлять интерактивный режим для организации гибкого процесса, эффективного в первую очередь с точки зрения человека.

Причем на уровне интерфейса такие технологические объекты и инструменты должны быть выделены среди средств поиска и работы с документами, что облегчит пользователю переключение с задачи своей основной деятельности (сбора информации для решения задачи) на информационную - оценку своих поисковых действий и состояний[48].

Представленный на рис. 6.4 итеративный человеко-машинный процесс информационного поиска в общем случае является интерактивным (где роль системы – пассивная информационно-технологическая поддержка) и включает следующие этапы:

1) определение темы запроса, ее локализация в предметной области и формализация на уровне понятий основной и смежных областей, а также идентификация ресурса. Здесь система предоставляет систематизированное описание предметной области, а также метаинформирование о тематике, наполнении, структуре и методах доступа к выбранному ресурсу;

2) формирование, а также структурное и лексическое адаптирование выражения запроса, где система предоставляет вспомогательные информационные объекты (словари, тезаурусы, шаблоны и т.д.);

3) отбор документов с помощью одного из механизмов поиска по критерию, адекватному степени неопределенности информационной потребности, где система предоставляет выбор механизма поиска или, например, автоматически с помощью лексикографических словарей и проблемно-ориентированных тезаурусов нормирует и расширяет лексику запроса;

4) формирование и управление выдачей найденных документов, где система обеспечивает масштабирование (форматирование) пространства представления выданных документов, а также сортировку и, возможно, рубрицирование или ранжирование по некоторому формаль-

---

[48] При этом активность системы по отношению к пользователю может реализоваться различными путями:

- непосредственным вмешательством в процесс через изменение параметров процедур, например изменением порога выдачи или стратегии поиска;
- построением прямых или косвенных оценок выдачи (показателей эффективности поиска);
- генерацией технологических объектов, являющихся дополнительными или альтернативными по отношению к тем, которые получены пользователем (например, построение словников при реформулировании запроса по обратной связи).

ному критерию соответствия, например, с использованием тематико-статистических распределений, характерных для проблемной области;

5) оценку результата поиска на уровне отдельного документа, где система обеспечивает возможность фиксировать значение степени соответствия запросу пользователя и непосредственное использование лексики документов для непосредственной модификации выражения запроса;

6) итоговую оценку результатов поиска на уровне всего запроса или отдельных предложений с точки зрения принятия решения о завершении поискового процесса (например, исчерпывающее удовлетворение информационной потребности, или несоответствие цели поиска), где система позволяет количественно оценивать динамику качества выдач и обеспечивает возможность выборочного обращения к результатам отдельных этапов процесса поиска или формирования проблемно-ориентированных словарей;

7) развитие запроса по технологии реформулирования по обратной связи по релевантности или использование каких-либо других ресурсов, например, ассоциированных баз данных вторичной или справочной информации, где роль системы – адекватное информирование о такого рода возможностях.

В следствие того, что объект поиска обычно не задан в виде образца, с которым можно соотнести найденный результат, а, с другой стороны, пользовательские ресурсы всегда ограничены, задача организации процесса поиска имеет оптимизационный характер – при временных ограничениях максимизировать показатели выдачи и получить максимальную (субъективную) уверенность в качестве поиска за счет предоставления пользователю в процессе диалога альтернативных направлений, а также количественные и качественные оценки их соответствия запросу.

Отметим еще раз, что здесь мы имеем два типа обратной связи[49]. 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 выделить значимые термины (ранжированием или кластеризацией по статистическим пока-

[49] Обращаясь к приведенной на рис. 2.7 схеме, отметим, что передача знаний посредством информационной системы, согласно [Мазур1974], соответствует случаю параинформирования. Такой подход позволяет определить условия, при которых информационно-поисковая система на основе вторичной информации будет обеспечивать выполнение требования $\mathbf{I_{Z12} = I_{X12}}$ (адекватность информирования).

В автоматизированных информационно-поисковых системах это может быть реализовано двойной реформулировкой запроса по технологии обратной связи по релевантности: построением на основе лексики истинно релевантных документов словников, термины которых в свою очередь размечаются с точки зрения их способности смысловыражения темы и, далее, используются в качестве выражения запроса для следующей итерации поиска.

зателям). Соответственно, для построения словников могут использоваться разные методы, что позволяет, в свою очередь, иметь разные стратегии реформулирования, реализуемые разными технологическими (интерфейсными) средствами.

Возможность совместного использования нескольких стратегий поиска позволяет реализовать процесс итерационного повышения эффективности поиска путем генерации новых ПОЗов, учитывающих как ситуационную (проблемную) ориентацию запроса, так и тематические свойства массива документов.

Контрольные вопросы

1. Определите основные информационные объекты и преобразования в схеме воспроизводства информации
2. Охарактеризуйте технологические составляющие информационного поиска.
3. Приведите типологию поисковых задач и примеры поисковых задач каждого типа.
4. Охарактеризуйте типы информационной неопределенности при поиске.
5. Определите условия установления соответствия информационной потребности и содержания документа БД.
6. Проведите сравнительный анализ понятий поисковая стратегия и поисковая навигация.
7. Охарактеризуйте основные этапы процесса информационного поиска
8. Перечислите основные и технологические объекты, используемые при поиске
9. Определите назначение «обратной связи» в процессе информационного поиска.
10. Перечислите информационные объекты, используемые для реализации технологии «обратной связи» в процессе информационного поиска.

7. Модели интерфейсов человеко-машинного информационного поиска

Особенности представления информации в документальных базах данных определяются их назначением – обеспечением эффективного (быстрого и исчерпывающего, прежде всего, по полноте) поиска нужных данных или, если таковые не обнаружены - сведений о документах, предположительно их содержащих.

Для достижения общности представления, минимально зависящей от точек зрения, особенностей изложения и восприятия информации, создаются каталоги и базы данных вторичной информации. При создании вторичного документа содержание первичного редуцируется до уровня перечисления основных понятий, в той или иной степени однозначно характеризующих его содержание, но в контексте именно той предметной области, для которой создается база данных. В свою очередь, для обозначения таких характеристических понятий используется ограниченная (нормализованная) лексика, снижающая влияние свойств синонимии и полисемии. Представление информационного содержимого конкретных документов в виде поисковых образов обеспечивает, с одной стороны, очень эффективную вычислительную процедуру (когда отбор производится по условию простого сопоставления термина запроса с термином документа), а с другой стороны – дает пользователю возможность получать достаточно полные и точные ответы на запросы, выражаемые упрощенным, “телеграфным” стилем. Такой подход построен на основе свойства концентрации информации и отражает существо уровневой модели, изложенных ранее (см. главу 1).

В ряде случаев, когда предметом проблемного поиска является что-то неизвестное (по крайней мере, пользователю), удовлетворительным результатом поиска скорее всего будет не отдельный документ, содержащий ответы на все вопросы, а некоторое множество документов, образующее для пользователя информационное пространство (понятий, фактов, идей и т.д.), достаточное для построения неизвестного до того решения, т.е. генерации новой информации. Таким образом, в случае проблемного поиска запрос - это не столько вопрос, предопределяющий ответ, сколько определение предметной области для поиска неизвестного через уже известное - документы и лексику.

С точки зрения внутримашинного представления информации семантическая тривиальность структур данных, реализующих документальные системы в фон-Неймановской архитектуре вычислительных машин, предопределяет, что развитие запроса и смысловая оценка результата поиска – исключительно прерогатива пользователя, а система - ассистент. И даже принимая во внимание определенные успехи в области разработок искусственного интеллекта, по крайней мере два следующих фактора не позволяют надеяться на скорое равноправие сторон:

– выражение запроса на естественном языке (даже хорошо формализованное) слишком лаконично для того, чтобы можно было бы выделить глубинную сущность проблемы, для решения которой должна быть найдена информация;

– в ряде случаев пользователь не может однозначно специфицировать (выразить наличными лингвистическими и понятийными средствами) информационную потребность, особенно если она связана с этапом постановки задачи.

И, как уже отмечалось в гл. 6, система является лишь инструментом, используемым человеком при поиске информации, поэтому эффективность ее использования зависит от того, насколько хорошо человек знает природу и свойства объектов и инструментов с которыми он работает.

В области автоматизации сложился ряд подходов, в рамках которых рассматриваются различные аспекты разработки эффективных средств обеспечения диалога «человек-компьютер», приемлемого для работы пользователей с разным уровнем подготовки и квалификации. Однако все они, так или иначе, связаны с необходимостью анализа субъективного «видения» пользователем особенности своей работы за компьютером.

Построение пользовательских интерфейсов, в основу которых положены сценарии с предвосхищающей схемой поведения системы, связано с необходимость анализа психологических особенностей человека (в первую очередь психологии познавательных процессов – ощущения, восприятия, внимания, мышления, памяти). Другим подходом к построению интерфейсов является использование теории деятельности, а также их совместное применение.

В контексте психологического подхода можно сказать, что для пользователей имеет значение лишь та информация, которая соответствует их пониманию недостающего знания. Т.е., полученная информация может быть оценена пользователем как релевантная только тогда, когда он уже обладает достаточными знаниями в некоторой области.

Таким образом, особенностью работы с информационными ресурсами является то, что деятельность пользователя складывается из двух составляющих – выполнения основной задачи (поиск документов, содержащих информацию, способствующую решению прагматической проблемы в сфере его основной деятельности), и явно или неявно осознаваемого освоения средств для взаимодействия с системой.

Рассматривая процесс коммуникации как взаимодействие двух сред представления информации, следует учитывать тот факт, что в каждой среде представление имеет свою специфику и то, что:

1) система всегда будет иметь дефицит понимания проблемы пользователя, поскольку разработчик вкладывает в систему свое представление о типе решаемой задачи;

2) разнообразие ситуаций, возникающих во время взаимодействия «человек – АИПС», достаточно велико, и создавать непрерывно адаптирующийся интерфейс не имеет смысла, так как в этом случае его «интеллект» должен быть сравним с человеческим;

3) методология оценки работы АИПС в целом традиционно сводится к количественным критериям оценки результатов выдачи, часто не учитывая реальной потребности пользователя. Кроме того, не существует строго определенных критериев, которые позволили бы адекватно оценить качество интерфейса как посредника.

Одним из возможных направлений в решении указанной проблемы является выделение *стереотипов «поискового» поведения* пользователей, на основе которых на уровне интерфейса могут быть созданы *стереотипы разрешения поисковых ситуаций,* что позволяет включать в систему такие технологические средства, которые бы обеспечили пользователю возможность реализовать в рамках интерфейса *свое видение проблемы*, построить модель проблемы в терминах и категориях, ему доступных.

7.1. Типология поисковых задач и план действий при взаимодействии пользователя с информационной системой

В соответствии с характером информационной потребности пользователя в гл. 6 были выделены три типа поиска: атрибутивный, тематический и проблемный.

Отметим, что поисковая задача ориентирует субъекта на будущее исполнение. При этом формулировка задачи на «технологическом языке» сопровождается неформальным толкованием на обыденном языке. Задача указывает цель, средства и пространство. Смысл задачи обычно определяется вербальной формулировкой.

Каждому типу поиска соответствует свой тип запроса, форма его выражения и, соответственно, характер результата. В первом случае необходимо найти все о некотором объекте - в результате пользователь получает некоторую выдачу, где нужные сведения присутствуют в явном виде. При решении задач, требующих использования поиска второго типа, строится предметная область проблемы, описываются объекты и связи и затем осуществляется поиск недостающих объектов и/или связей. В результате пользователь получает выдачу, где присутствует (или нет) понятие, которое может быть использовано для построения нового объекта, а не только для подтверждения факта его существования. Задачи, которые решаются с помощью поиска третьего типа, характеризуется тем, что тема запроса зачастую формируется или изменяется непосредственно в процессе поиска.

Стремясь понять задачу, субъект использует пространственные и временные образы, представления, схемы. Продумывание исполнения и

оценка сложности задачи дают план – схему действий. При формулировании целей появляются критерии, позволяющие ограничить разнообразие действий.

Характерно, что в деятельности пользователя-непрофессионала целеполагание представлено в сложной «развернутой» форме, тогда как у более опытного пользователя оно заменяется свернутым процессом понимания заданной цели. В [Hacker1985] выделяют три типа схемы действий в соответствии с уровнем их осознанности:

1) бессознательная программа, как последовательность простых действий;

2) полуосознанная схема действий;

3) осознанная схема действий, метаплан, набор стратегий и условия их использования.

Уровень осознанности может определяться в четырех аспектах:

- *понимание текущей ситуации* - система должна вести диалог и отслеживать, реагирует ли и каким образом пользователь на ее сообщения,

- *четко поставленная конечная цель* - заключение о ее наличии в случае поиска может быть сделано, например, на основе информации об изменении объема выдачи и анализа лексики релевантных документов;

- *последовательность действия для достижения цели* - система предоставляет различные способы поиска: в работе используются, как правило, их сочетания, при этом взаимодействие с системой неопытного пользователя представляет собой "хаотическое блуждание", опирающееся, например, на контекстно-зависимую систему подсказок;

- *условия выполнения или изменения последовательности действий* - система может отслеживать такие количественные характеристики выдачи, как полнота и точность и на основе их изменений судить об эффективности применяемых действий, о необходимости смены стратегии поиска.

Здесь, с одной стороны, осознанность деятельности свидетельствует о степени подготовленности пользователя. С другой стороны - можно предположить, что даже хорошо знакомый с поисковой средой человек начинает работу с системой, не имея четкого представления о том, что конкретно он хочет найти. Такая ситуация наблюдается, например, при проблемном поиске.

Следует также отметить и тот факт, что пространство действия не совпадает с пространством восприятия. Бессознательное исполнение без знания нельзя назвать действием. Действие развертывается на фоне уже знакомого мира, расширяя пространство восприятия: воспринимаемым становится то, что ранее не воспринималось - предметы, различия между ними, тонкости упорядочения. Действие — расширение пространства, но текущее, происходящее, живое [Стрелков2000]. Любая деятельность,

производимая человеком в реальной действительности для решения заданной задачи или достижения цели, одновременно разворачивается во внутреннем плане, то есть на уровне мышления. Но это “второе действие нельзя поместить в те же самые временные рамки и пространственные формы, т.к. действие реальное конечно, а действие на уровне мышления продолжается благодаря памяти" [Зинченко1996]. Эта особенность восприятия информации человеком играет особенно важную роль при тематическом поиске.

7.2. Типология информационных потребностей пользователя

Для пользователя, решающего проблему, может быть (и, как правило, бывает) затруднительно сразу сориентироваться в сложной среде поисковой системы. С другой стороны, система не может самостоятельно точно воспроизвести потребность пользователя. Поэтому в основе процесса взаимодействия лежит диалог, целью которого является последовательность шагов, направленных на решение проблемы ОД пользователя, для поэтапного снятия неопределенностей – семантической, языковой (лексической) и метаинформационной.

При этом характер информационных потребностей, рассматриваемых как дополнение, возможно гипотетическое, известного знания, определяется формой представления информации[50], что в свою очередь зависит от среды – носителя информации. Здесь можно выделить четыре следующих информационных уровня объектов (сред представления) (рис.6.2):

- уровень знания, где носителем является сознание человека, оперирующее образами;
- уровень сообщения – знания, отраженного на систему понятий (например, предполагаемого или реального приемника), и выраженного средствами языка;
- уровень документа, материализованной формы обобществления знаний, для которого характерна структурная и лингвистическая унифицированность представления сообщения (например, выделение семантически однородных полей и правил их заполнения), а также семантическая и адресная идентифицируемость сообщения в целом;
- уровень машинной записи, имеющей предопределенную структуру и тип (способ) представления информации (для ЭВМ – обрабатываемые данные), что обеспечивает ее идентифицируемость и атрибутирование на уровне отдельных полей[51].

---

[50] Например, в [Yoon1999] приводятся результаты предварительного изучения межличностного взаимодействия при поиске информации между пользователем и человеком, являющимся источником информации. Изучение показало, что при взаимодействии пользователи выражают свои информационные потребности ("неопределенность") преимущественно в терминах того, что они знают ("определенность"). Артикулированные определенность и неопределенность при взаимодействии могут быть классифицированы как высказывания, сконцентрированные или на теме (о чем говорит пользователь), или на комментарии (как эта тема связана с ситуацией или проблемой пользователя).

[51] Следует отметить, что в первую очередь именно жесткость требований формализации этого уровня, выражающаяся в типизации способа представления информации в зависимости от ее характера, приводит к выделению двух классов информационных систем – документальных и фактографических. Для первых смысловое значение и, соответственно, способ интерпретации задается в форме понятий и их комбинаций, порождающих новый смысл. Во втором случае для фактографической ин-

Аналогично (в контексте уровней представления информации) определяют несколько типов информационной потребности пользователя (ИПП).

Как было определено в гл. 6, прежде, чем быть адресованной системе, информационная потребность пользователя проходит четыре этапа формализации. Первоначально пользователь, находящийся в проблемной ситуации, имеет *реальную информационную потребность;* в процессе ее понимания ИПП преобразуется в *осознанную ИПП;* осознанную ИПП пользователь выражает на привычном ему языке, то есть, формируя *запрос на естественном языке,* который далее формулируется в терминах ИПЯ в виде *запроса к системе - ПОЗа*.

Понимание – переход от реальной ИПП к осознанной тем сложнее, чем менее определена задача пользователя. Для поисковых задач проблемного типа она наиболее трудна, так как пользователь не представляет, какая именно информация нужна для решения его задачи. Отмечаются следующие характерные ситуации при формализации ИПП на этапе ее осознания:

- «неполное представлении о мире», в связи с чем система не может определить, в какой момент субъект находится в «состоянии готовности» осуществлять целенаправленное взаимодействие с миром адекватно своей потребности;
- «аномальное состояние знания», когда пользователь не имеет достаточно определенного представления о том, что он хочет узнать, в том числе – незавершенное состояние знания и неопределенное состояние знания.

Выражение осознанной информационной потребности на естественном языке также порождает неоднозначность вследствие синонимии или "эффекта ярлыка", когда человек описывает проблему, используя понятия, вызванные ассоциативной связью с близкой предметной областью.

Формальное выражение информационной потребности пользователя средствами ИПЯ, то есть формирование ПОЗа, позволяет привести в соответствие с точки зрения смысловыражения то, что является объектом поиска, и то, что заиндексировано.

Опытный пользователь располагает многочисленными *сверхзнаками*: определенная «ситуация ... для него не конгломерат … изолированных признаков, которые рассматриваются по отдельности, а некоторая структура. Такие сверхзнаки - продукт опыта. Они редуцируют сложность – группа признаков порождает один сверхзнак. Следовательно, сложность системы определяется всегда по отношению к определенному действующему лицу с его набором сверхзнаков» [Дернер1997]. Способность предположить, как изменится ситуация при конкретном вмешательстве (использовании той или другой технологии), появляется в том случае, когда у пользователя

формации используется в основном жесткая «атрибутная» форма представления смысла, где атрибут помимо имени (аналогично словам - именам понятий) специфицируется также форма представления (вплоть до перечисления возможных значений) и способ интерпретации, например, через общее для всех записей указание размерности.

есть структурное знание о самой системе. "Общее множество предположений, относящихся к односторонним или многосторонним, простым или сложным связям переменных некоторой системы принято называть *моделью реальности*" [Дернер1997]. Оно может существовать эксплицитно, в осознанной форме, или имплицитно, когда субъект имеет лишь некоторые предположения (интуиция). При этом, эксплицитное, вербализованное знание необязательно является действенным. Оно может существовать как теоретическое знание, обладатель которого не сумеет его применить.

В процессе взаимодействия с поисковой системой пользователь снимает неопределенности разного вида: создавая модель реальности, формирует (осознанно или нет) множество предположений относительно работы системы. Обычно такая модель бывает и неполной, и неправильной (если только в качестве пользователя не выступает разработчик системы). Но, несмотря на то, что человеку разумно было бы учесть возможность изначальной некорректности и неполноты модели, ему обычно свойственно упорствовать в своем заблуждении.

7.3. Технология поиска и интерфейс АИПС

Самая общая схема алгоритма поиска представляет собой последовательность, представленную на рис. 7.1.

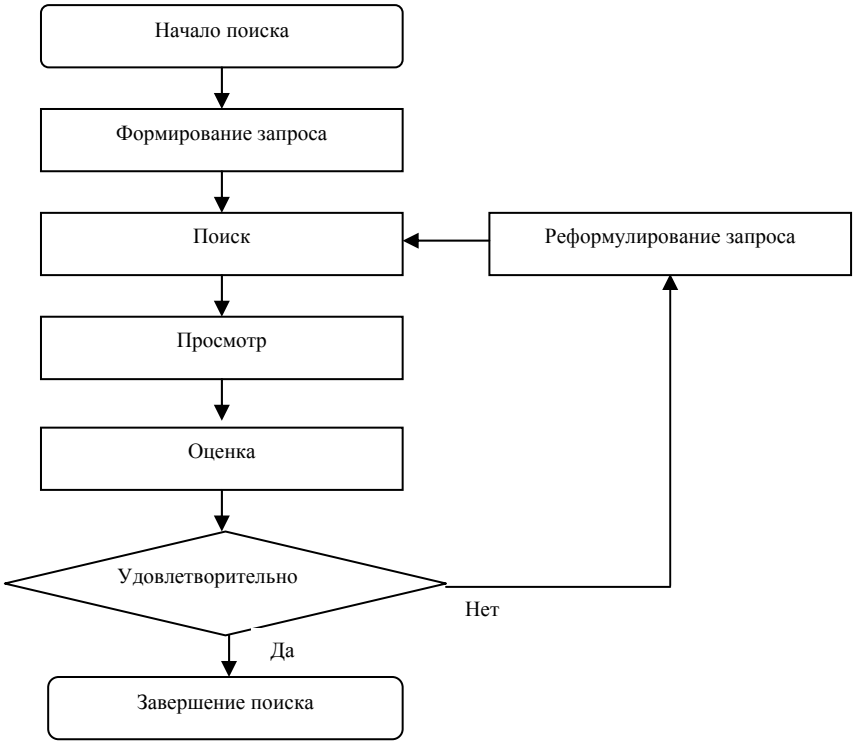


Рис. 7.1. Обобщенный алгоритм поиска

Выбор человеком конкретной стратегии и технологии решения задачи практически в любой области деятельности обусловлен его привычками и типом мышления. В процессе решения человек может переходить от одной технологии к другой, но выбор новой технологии опять же будет обусловлен его предшествующим опытом. Кроме того, при тематическом или проблемном поиске нужная информация не всегда содержится в явном виде, и для ее получения необходимо выполнять над полученными текстами документов некоторые логические операции, требующие привлечения дополнительных знаний о ПрО, непосредственно не содержащихся в тексте.

7.3.1. Творческий процесс и стереотипы мышления

Когда человек попадает в проблемную ситуацию, где стереотипы мышления и действий оказываются недостаточными, необходимо творчество - особая форма познавательной деятельности, которая приводит к новому видению проблемы. В процессе решения обнаруживается и, возможно, снимается противоречие между накопленным опытом и новизной условий задачи. Самостоятельное преодоление субъектом такого противоречия и становится в итоге творческим решением.

Процесс творческого мышления не может осуществляться без его самоорганизации, рефлексии, которая выступает в двух аспектах: 1) в открытии новых свойств предметного мира и средств его преобразования (интеллектуальная рефлексия); 2) в изменении представлений о самих себе и, следовательно, в преобразовании себя как целостной личности (личностная рефлексия). Поэтому организация и развитие творческого мышления возможны не иначе, как через формирование рефлексии, чему должна способствовать система, производя или, по крайней мере, стимулируя самообучение в среде с помощью средств интерфейса.

7.3.2. Интерфейс пользователя

Интерфейс пользователя - это совокупность правил, методов и программно-аппаратных средств, обеспечивающих взаимодействие пользователя с компьютером. Пользовательский интерфейс часто понимают только как внешний вид программы. В действительности интерфейс пользователя включает в себя все аспекты, которые оказывают влияние на взаимодействие пользователя с системой, и определяется такими факторами, как:

- набор задач пользователя, которые он решает с помощью системы;
- используемая системой метафора (например, рабочий стол и т.п.);
- элементы управления системой;
- навигация между структурными компонентами системы;
- визуальный дизайн системы.

Главная задача интерфейса – дать возможность пользователю эффективно работать с информацией без помощи специалиста-посредника.

7.4. Поведение пользователей при взаимодействии с АИС

7.4.1. Уровневая модель человеко-машинного взаимодействия

Определяя понятие человеческой деятельности как «целенаправленное, сознательное, спланированное поведение субъекта деятельности» [Cranach1980], выделяют следующие различия природы поведения человека и системы, построенной на базе вычислительных комплексов (см. табл. 7.1).

Таблица 7.1.

	Деятельность человека	**Поведение системы (компьютерной)**
Управляемость	Внутренняя мотивация, намерения	Определяется действиями Пользователя
Направленность	Может сменить направление неожиданно	Только последовательное от одного состояния к другому
Природа	Психологическая и физиологическая	Механическая и электронная
Характер деятельности	Активная, творческая, гибкая	Пассивная, исполнительная
Адаптация	Более быстрая, успешная	Медленная
Сознание	Индивидуализированное	Стандартизированное
Внимание, память	Отвлекаемость, забывчивость	Исполняемость, точность
Темп действия	Медленный	Высокоскоростной

Поведение систем, реализованных в компьютерной среде, обладает «блочной» гибкостью - в определенные моменты, предусмотренные при разработке, система способна изменять поведение, в том числе, например, ориентируясь на действия пользователя. Но в промежутках между такими состояниями «выбора и переключения» работа жестко детерминирована.

С точки зрения процесса взаимодействия когнитивных подсистем, поиск может быть представлен как серия коммуникативных актов между стратами (уровнями), направленных на решение основной задачи и взаимную адаптацию партнеров [Saracevic1997]. Страты партнера-человека в процессе взаимодействия претерпевают следующие изменения (рис. 7.2):

– на поверхностном уровне определяется характер процесса общения - применение различных тактик поиска, формулирование и рефор-

мулирование запроса, добавление, удаление терминов, работа с документами;

– аффективный уровень определяет эмоциональность - удовлетворенность или неудовлетворенность пользователя результатом;

– на когнитивном и ситуативном уровне происходят изменения, связанные с интерпретацией полученной информации, осмыслением и переработкой, а также изменение локальных задач, возникающих при достижении общей цели удовлетворения ИПП.

7.4.2. Когнитивные аспекты человеко-машинного взаимодействия

Рассмотрим более подробно составляющие когнитивной страты, как наиболее значимой для понимания особенностей восприятия и использования информации (рис. 7.2).

Первый уровень взаимодействия - это взаимодействие когнитивного уровня человека-оператора и поверхностного уровня компьютерной системы (интерфейса).

Второй уровень - взаимодействие когнитивного уровня человека-оператора и уровня процессов и задач.

Третий - взаимодействие пользователя с предоставляемой системой информацией; является самым сложным и практически не формализуемым уровнем общения.

Восприятие информации. Начальным этапом отбора и обработки информации является обнаружение и интерпретация сенсорных сигналов. Существует несколько классических теорий восприятия, которые рассматривают различные аспекты того, что в общем является единым процессом. С одной стороны, восприятие рассматривается как процесс переработки информации, прием которой обеспечивается сложными нейрофизиологическими механизмами. С другой стороны, восприятие может рассматриваться как процесс проверки и подтверждения гипотез.

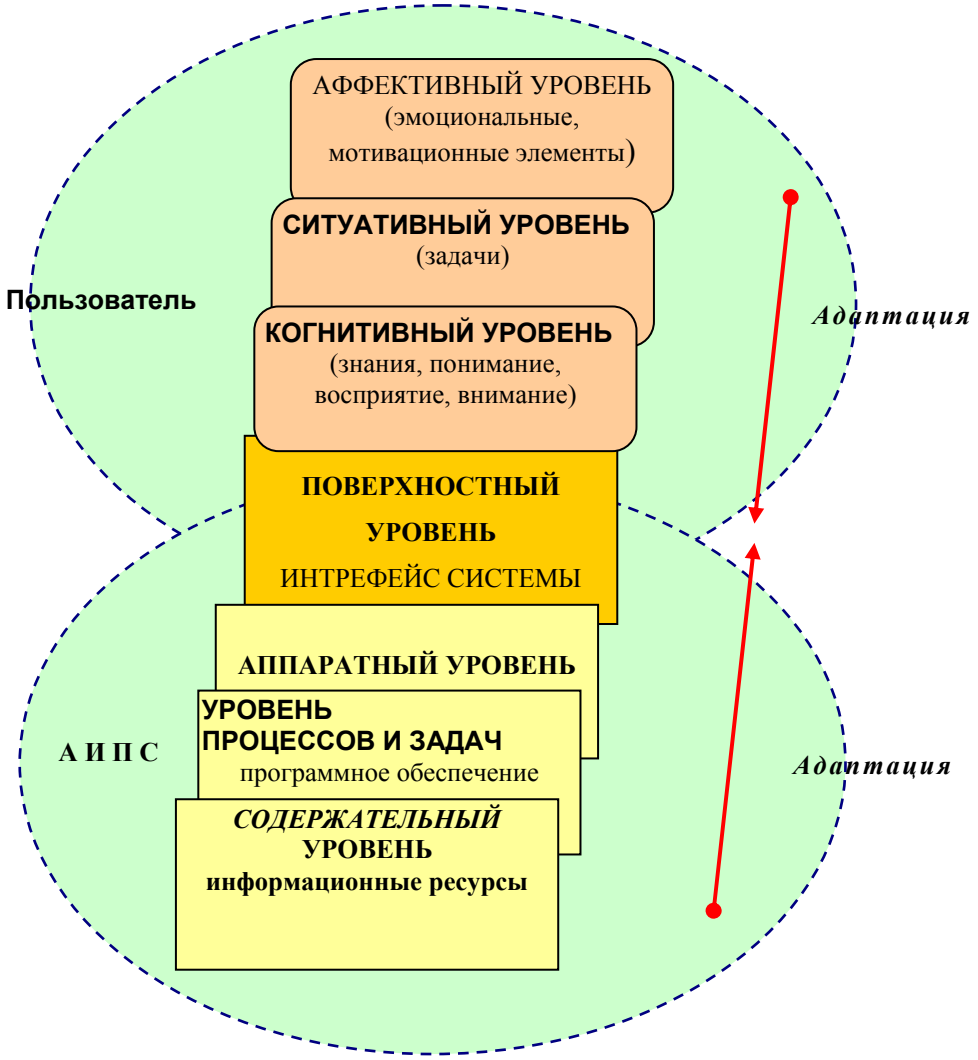


Рис. 7.2. Уровневая модель взаимодействия пользователя и системы

Подход, который объединил классические теории восприятия, был предложен в [Найссер1981], где восприятие рассматривается как квалификационная деятельность, отличающийся тем, что «ее влияние на окружающую среду ничтожно мало».

Восприятие происходит в процессе перцептивного цикла (рис. 7.3), важнейшим компонентом которого является «предвосхищающая» схема, подготавливающая человека к принятию строго определенной информации и таким образом управляющая активностью. Схема направляет когнитивную активность в поиске той информации, которую способна принять. В свою очередь, полученная в каждый момент времени информация способна изменить схему, модифицировать ее или сконструировать новую схему для принятия новой информации, в тот момент, когда она станет доступной.

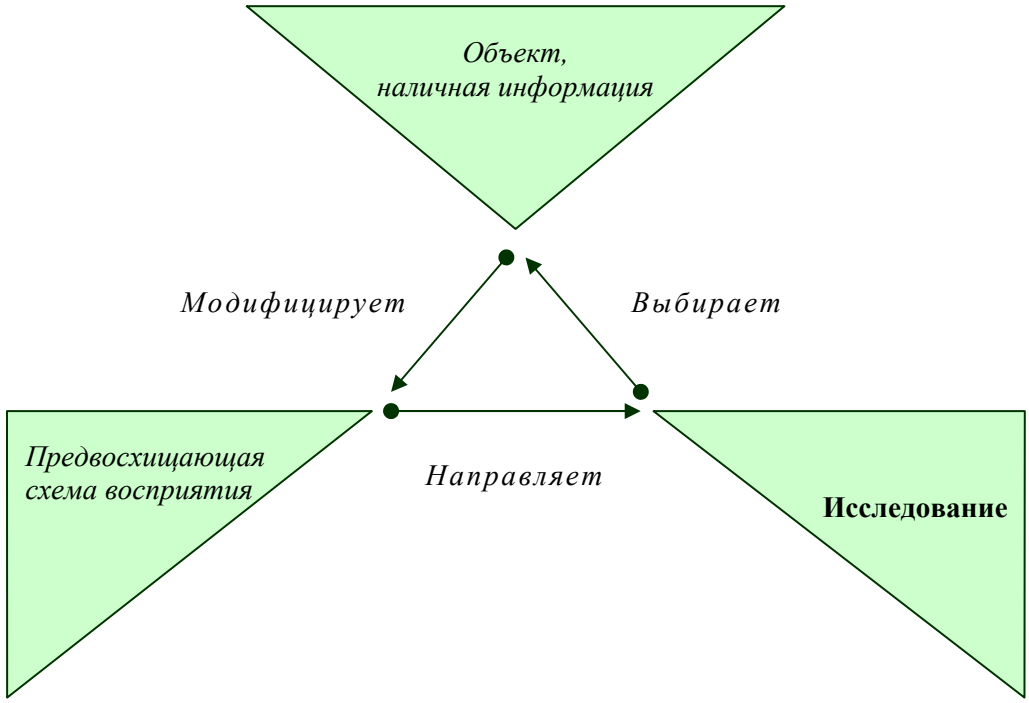


Рис. 7.3. Перцептивный цикл восприятия информации.

Понимание и мышление. Понимание (ориентировка), мышление (исполнение по правилам) и рефлексия (самоанализ) являются основными составляющими умственной деятельности. Понимание ситуаций, целей и задач является началом всей последующей деятельности. Новая информация может быть усвоена тогда, когда уже существуют когнитивные структуры и информация. И наоборот, недостаточное знание ограничивает понимание, так как человек должен разработать некоторую структуру знания об исследуемой проблеме, а также найти в этой структуре место для вновь полученного материала - тогда новая информация становится знанием [Солсо1996]. При таком подходе процесс понимания скорее является подтверждением гипотезы о том, каким представляется мир, чем просто усвоением некоторых новых фактов.

Мышление, как процесс манипулирования знаниями, направлено, а его результаты проявляются в поведении, которое решает некоторую проблему или направлено на ее решение. При этом человек использует определенные стереотипы действия, сформированные или заимствованные.

Значимую роль здесь играет фантазия, если рассматривать фантазию как переход от одного стереотипа к другому. При этом, чем больше стереотипов, тем выше уровень фантазии. В [Григорьев1999] выделяют два типа фантазии:

– опосредованная фантазия - переход в интерпретации рассматриваемого предмета к такому стереотипу, который является конкретным следствием только этой интерпретации;

– непосредственная фантазия - переход к такому стереотипу, который является конкретным следствием не только этой интерпретации.

В свою очередь, стереотип непосредственной фантазии может быть простым или сложным. Простой стереотип – являет собой относительно самостоятельное понятие, причинно-следственную связь или факт. Сложный стереотип – это набор простых стереотипов, логически связанных между собой.

Тип мышления определяется не только уровнем фантазии, но и способностью к логическим построениям, которая характеризуется тремя уровнями:

– уровень стереотипов - при решении задач человек использует в основном готовые стереотипы, построенные или заимствованные им ранее;

– уровень рассудка - человек активно подбирает или конструирует способы, которые бы обеспечили успех в достижении цели;

– уровень логики - в действиях пользователя обнаруживается стратегический подход, когда используемое решение есть следствие хорошо разработанной аргументации, целостно и многосторонне отражающей предмет анализа.

7.4.3. Типология и стереотипы поведения пользователей

Процесс взаимодействия пользователя и системы при поиске информации в АИПС может быть представлен множеством ситуаций, определяющихся состояниями пользователя и системы, а также их представлениями, как о своем собственном состоянии, так и о состояния системы.

Модель ситуации в представлении пользователя можно определить системой планов сознания, взаимосвязь которых обеспечивается рефлексивным планом сознания, который развертывается параллельно с бытийным слоем – слоем структур опыта [Стрелков2000]. В рефлексивном слое замыкаются связи между подготовкой, выполнением действия и анализом результатов. Можно предположить, что пользователи, обладающие способностью к рефлексии, будут работать более эффективно[52].

[52] К характеристикам когнитивного уровня относят показатель «локуса контроля», введенный американским психологом Дж. Роттером, который установил, что есть категория людей («экстерналы»), которые в своих действиях полагаются главным образом на самих себя, на собственные возможности, поэтому в поведении они ориентированы в основном на контроль за своими состояниями и действиями. Другая категория людей («интерналы») в своей деятельности больше полагается на внешние условия, считая, что именно от них главным образом зависит ее успех, поэтому свой контроль они локализуют в основном на внешних факторах, ситуациях [Котик1993]. При работе с АИПС пользователи-экстерналы проявляют

Пользователей можно условно разделить на четыре типа (см. Рис. 7.4):

Тип 1: доминирует бытийный уровень сознания: человек с трудом отказывается от выбранной тактики поведения, полагает себя правым, а в неуспехе склонен винить в первую очередь систему; не отличается настойчивостью, в случае неудачи быстро прекращает дальнейшие попытки, *тип мышления, как правило, стереотипный*;

Тип 2: также доминирует бытийный уровень сознания, причину неудачи человек прежде всего, видит в работе системы, но активно ищет альтернативные варианты действий, в случае неудачи попытается выразить свою прежнюю позицию другими средствами, *обладает восприимчивым или рассудочным типом мышления*;

Тип 3: доминирует рефлексивный уровень сознания, такой пользователь способен взглянуть на свою задачу с другой точки зрения, в случае неудачи вначале анализирует собственные действия в поисках ошибки, но для формирования дальнейших действий ожидает подсказки или помощи системы;

Тип 4: доминирует рефлексивный уровень сознания, но пользователь анализирует собственные действия, обладает многосторонним взглядом на свою проблему, а при неудаче - в результате самоанализа способен быстро изменить стратегию поиска. Таким пользователям в основном свойственен *конструктивный или нестандартный тип мышления*.

В процессе взаимодействия человек способен менять поведение в сторону более высокого типа, что объясняется не только обучением работе с системой, но и «научением», т.е. приобретением навыков применения типовых стратегий поиска, а также, возможно, развитием способности к самоанализу, хотя такой процесс происходит и менее интенсивно. Возможные направления изменения поведения пользователя показаны на Рис. 7.4.

Соответственно, интерфейсные средства системы должны помочь человеку перейти к четвертому типу поведения – осмысленному, эффективному творческому процессу созидания нового знания.

С одной стороны, – это унифицированные (так называемые, стандартные) средства, в том числе:

– готовые технологии поиска (для пользователей 1-го типа);

– система контекстно-зависимых подсказок, помощников (для пользователей 3-го типа).

большую активность, что выражается в готовности изменить стратегию поиска, в подборе возможных вариантов действий. Пользователи-интерналы наоборот будут ожидать «направляющих» действий системы.

С другой стороны, – это средства, побуждающие пользователя к осмыслению процесса поиска, в том числе:

– сообщения, поясняющие полученный результат, развернутые сообщения об ошибках, информация о возможных дальнейших действиях (для пользователей 1-го типа;

– «провоцирующие» действия системы, например, отображение дополнительных объектов или функциональных возможностей (для пользователей 2-го типа).

7.4.4. Типология поведения пользователя в различных деятельностных состояниях

У каждого человека набор поведенческих схем соотносится с определенным состоянием его сознания. Рассмотрим возможные состояния сознания пользователя при его взаимодействие с АИПС.

Пользователь обращается к АИПС с уже сформированным намерением. В процессе взаимодействия он выполняет действия, приводящие (или нет) к ожидаемому (или неожиданному) результату, воспринимает полученную информацию, оценивает ее и свое новое состояние и либо завершает свою работу с системой, либо продолжает работать дальше.

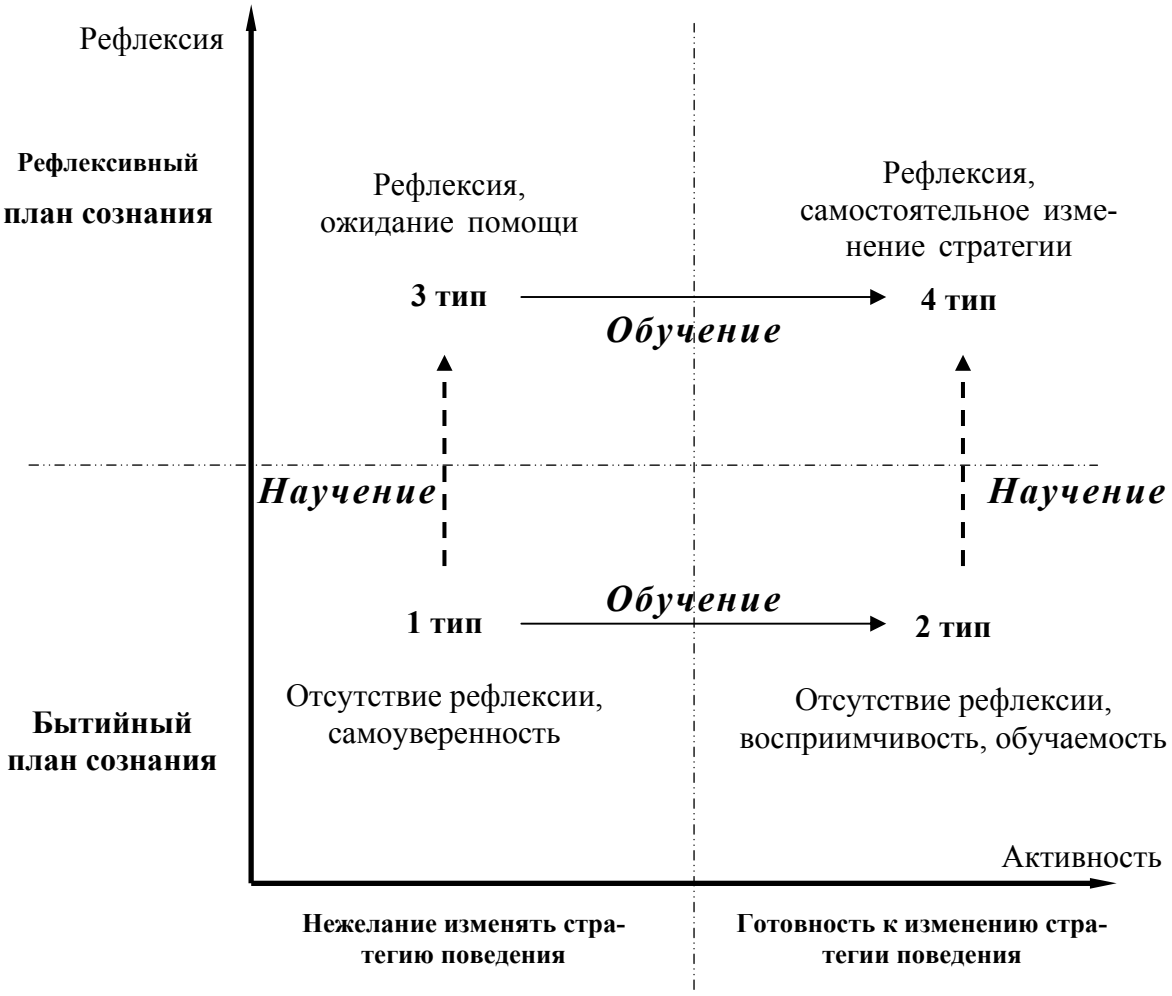


Рис. 7.4. Типы пользователей в зависимости от их активности и способности к рефлексии

Здесь можно выделить следующие деятельностные состояния пользователя:

1) формирование намерения - осознание информационной потребности, определение первоначального плана действий, формирование (может быть, неявного) образа предполагаемого результата;

2) собственно действие - выбор операционных объектов и операций исходя из общей схемы и текущей ситуации;

3) восприятие и осознание полученной в результате действия информации, как основной - результатов поиска, так и вспомогательной - системных сообщений, статических данных и т.п.). Изменение (или подтверждение) существующего знания;

4) оценка результата - изменения состояния знания пользователя, постановка новых задач, например, следующего этапа достижения цели, или, возможно, изменение целевого намерения.

Характеристика деятельностных состояний пользователя (воплощение поведения) как результат деятельности, характерной для разных страт сознания пользователя, представлен в таблице 7.2.

Таблица 7.2.

Деятельностное состояние	**Аффективная** страта сознания	**Когнитивная** страта сознания	**Ситуативная** страта сознания	**Поверхностная** страта сознания
формирование намерения	ожидание результатов взаимодействия	осознание ИП в существующей структуре знания	формирование плана поиска информации	конкретизация темы в условиях доступных ИР, формирование ПОЗа
Действие	эмоциональное напряжение в ожидании результата действия	понимание сути выбранного действия	выбор действия исходя из общей схемы решения и поисковой задачи	выбор стратегии поиска, редактирование запроса, начало поиска
Восприятие, осознание	удовлетворение или раздражение при получении результата, или ошибок	восприятие вспомогательной информации, осознание полученной информации, изменение структуры знания,	в случае ошибки изменение локальной подзадачи, выбор альтернативного действия	работа с документами - чтение, редактирование, реакция на системные сообщения
Оценка	удовлетворение / неудовлетворенность	оценка полученной информации; в случае неудачи - поиск причины	изменение стратегии поиска, изменение ИПП	оценка релевантности документа, ранжирование и сортировка выдачи, завершение работы

Фактически же поведение пользователя «формируется на основе множества стереотипов, часть из которых выработана человеком при решении других задач, а часть синтезируется в процессе работы» [Григорьев1999]. Действие пользователя есть результат выбора из нескольких вариантов, каждый из которых представляет собой стереотип или конструкцию из стереотипов той или иной степени сложности. Процесс выбора также подчиняется этой схеме: он вариантен и также основывается на стереотипах.

Соответственно, адаптивность блочно организованного поведения системы основывается на выделении и разработке модулей интерфейсных решений, адекватных предполагаемым стереотипам пользователей, а при работе (обслуживании реальных запросов) выражается в квалифицировании (на основании оцениваемых характерных признаков стереотипов) пользователя и предоставлении ему соответствующих интерфейсных средств.

Такого рода признаки по существу отражают мотивационную составляющую, представленную на аффективном уровне. В зависимости от степени соответствия *полученного* результата *ожидаемому*, человек принимает решение о направлении дальнейших действий. Возможны следующие ситуации, характеризующие состояние взаимодействия с системой:

- пользователь ожидает получить сразу готовый ответ,
- пользователь готов к длительному процессу поиска,
- пользователь заранее уверен, что ничего не получится.

В первом случае пользователи первого типа, скорее всего, прекратят работу с системой, пользователи второго и третьего типов попытаются либо изменить способ задания запроса, либо изменить сам запрос.

Во втором случае полученный после удачного запроса хороший результат провоцирует пользователя прекратить поиск, удовлетворившись найденными документами. Человек в принципе не может знать наполнение ресурса и не может быть уверен, что найденные документы исчерпывающе представляют проблему, однако, пользователям первого и второго типов такая мысль может и не прийти в голову. В таком случае работа с системой также будет быстро завершена.

Статистический анализ функциональной активности пользователей баз данных научной реферативной информации (по разнообразию функций, количеству операций в сессию, последовательности выполнения функций) показал, что поведение пользователя устойчиво коррелирует с поведенческим типом, определенным выше.

Первую группу в основном составляют пользователи 3-го типа: основной особенностью этой группы является стереотип длительного применения одной и той же непоисковой функции – просмотра словарей и документов, хотя общее число используемых функций достаточно велико.

Вторая группа – пользователи 4-го типа, использующие достаточно широкий спектр функций при сравнительно малом числе использованных операций. Поведение пользователей этой группы в основном свидетельствует о наличии определенной стратегии (и, как правило, успешной, так как практически все заканчивают работу, заказав электронные копии нескольких найденных документов).

Третья группа – пользователи, использующие мало функций и мало операций. Эту группу составляют пользователи всех типов, но в то же время для многих из них характерны ситуации, свойственные второму и третьему типу пользователей, например, «метания» между базами данных, когда после одного, двух запросов (обычно, слабо различающихся) пользователь переходит в другую БД и повторяет запрос там. Для третьей группы, большинство которой предположительно составляют пользователи второго типа, вероятность уточнения формулировки запроса значительно больше, чем вероятность выбора функции просмотра документов, причем с каждым последующим запросом доля первых увеличивается, тогда, как доля вторых уменьшается. Для второй группы, которая предположительно состоит в основном из пользователей четвертого типа, происходит обратный процесс.

С точки зрения стратегии развития запроса здесь можно выделить две типовые схемы:

1) поиск, при котором пользователь, получив вначале большое число документов (несколько сотен или тысяч), постепенно уточняет свой запрос и в итоге получает небольшую выдачу (несколько десятков), содержащую высокий процент релевантных документов;

2) поиск, при котором пользователь, получив большую выдачу, не торопится реформулировать запрос, а просматривает значительную часть выдачи и, возможно, также находит релевантные документы, с помощью лексики которых строит свою дальнейшую навигацию, например, применяет эвристический поиск или поиск аналогов.

7.5. Характеристика компонентов информационного поиска

Исходя из анализа логических моделей поиска, систематизированных в табл. 6.1, рассмотрим возможные способы представления (выражения) информационных потребностей

Типология и взаимосвязь информационных потребностей и средств поиска в зависимости от типа поисковой задачи приведена в табл. 7.2.

Таблица 7.2.

Вид поиска	**Характер ИПП**	**Задание предмета поиска (известное/неизвестное)**	**Состав ПОЗ**
Предметный (атрибутивный) поиск	Найти документы, содержащие информацию об известном объекте (факте) по известным значениям характеристических признаков	Объект полностью определен признаками и/или связями. Неизвестно значение отдельного признака или связи.	Термины индексирования документов, отнесенные к семантически определенным полям
Тематический поиск	Найти документы, содержащие информацию о гипотетическом объекте (или методе его построения) по предполагаемым характеристическим или ассоциированным признакам и связям	Объект задан частью признаков/связей; Полностью известна структура, как системное свойство, определяющее целостность. Неизвестны отдельные признаки или связи.	Термины документов и термины дополнительных поисковых структур – тематических рубрикаторов, тезаурусов и т.п., отнесенные к семантически определенным полям.
Проблемный поиск	Найти документы, позволяющие определить информационное пространство, подтверждающее/отрицающее возможность постановки и решения задачи; поиск нового качества в предметной области	Объект задан частью признаков/связей; структура, как системное свойство, определяющее целостность полностью не определена. Неизвестны отдельные группы признаки/связи или структура.	Документы, входящие в итеративно формируемое пользователем информационное пространство

Учитывая, что идеальный результат поиска должен удовлетворять требованиям единственности, полноты и непротиворечивости, получаем, что различные виды поиска определяют различные требования к функциональным возможностям системы.

Только для случая предметного поиска доказательство полноты и непротиворечивости является тривиальным: результат поиска непосредственно (собой) подтверждает или опровергает существование объекта, обладающего искомыми свойствами.

Результат тематического поиска в этом смысле неоднозначен (множественен) и, соответственно, требует *последующей* систематизации – еще одного процедурного шага для упорядочения полученного

множества объектов по значениям основания, которое явно не определено.

Проблемный поиск, соответственно, предполагает уже двухуровневую систематизацию.

Отметим, что такая дополнительная отдельная во времени обработка требует наличия в системе средств *идентификации* получаемых объектов (как отдельных элементов, так и их композиций), средств выборочного использования, а, кроме того, возможно, связывания с методами их получения.

7.5.1. Основные компоненты процессов и систем поиска документальной информации

Полнота результата зависит от степени идентичности соотносимых поисковых образов, т.е. либо гарантированной тождественности преобразования системой запроса в ПОЗ (и далее – в результат), либо, что более реально, предоставления пользователю не только средств преобразования, но и средств оценки (доказательства степени идентичности) этого преобразования.

В контексте технологической локализации процесса поиска в гл. 6 были выделены основные компоненты, которые могут быть отнесены к следующим стратам (средам и состояниям) преобразования запроса в результат:

– страта планирования, как предопределение пользователем стратегии поиска;

– страта организации и технологии поиска, как системные решения, воплотившие на стадии разработки и адаптации системы особенности поведения предполагаемых пользователей и характер хранимой информации;

– страта выполнения (навигация), как проведение реальным пользователем конкретного поиска в рамках выбранной им стратегии и средствами АИПС.

Рассмотрим далее основные понятия, относящиеся к поисковому процессу.

7.5.1.1. Стратегия поиска и классификация АИПС

Поведение пользователя, как организующее начало управления процессом поиска, мотивируется не только неопределенностью информационной потребности, но и разнообразием технологических объектов и средств, предоставляемых системой. Однако, пока существовал только один класс ИПС, реализующих диалог с одной активной стороной и позволяющих только получать выдачу на запрос (причем пользователя часто не интересовал механизм поиска), о стратегии можно было гово-

рить, имея в виду разве что исследовательские наклонности самого пользователя.

В контексте способов организации доступа к информации, представленной в документальной форме, и отдавая должное истории развития ИС, можно говорить о двух типах решений, воплощаемых в промышленных АИПС.

Первые - традиционные ИПС, берущие начало от библиотечных систем, информационный вход в которых реализуется через дополнительные (вторичные по отношению к текстам документов) справочные структуры различного типа.

Вторые - гипертекстовые ИС, в которых переход к потенциально полезному документу реализуется через контекстную ссылку, размещенную в тексте самого документа.

Относительная независимость развития этих двух направлений в значительной степени обуславливалась функциональным различием информационного продукта и техническими ограничениями среды хранения. Учитывая тождественность конечной задачи, т.е. обеспечение доступа к реально полезной и полной информации, следует отметить, что гипертекстовые системы использовались в основном для реализации справочных систем, базирующихся на проблемно-ориентированных коллекциях зачастую слабо структурированных полнотекстовых документов, а эффективность поиска документов или их фрагментов достигалась через более или менее полную систему контекстно-определенных ссылок. В свою очередь традиционные ИПС предназначались для обработки большого количества однородных (структурно-регулярных), чаще всего вторичных документов небольшой длины.

Можно сказать, что для этих двух типов систем принципиально различаются пути нахождения пользователем реально нужной информации.

Координатный принцип индексирования документов и использование в запросе терминов вне контекста предопределяет для ИПС необходимость последующей пользовательской оценки некоторого подмножества документов, отобранных системой по формальным признакам.

Гипертекстовые системы позволяют более целенаправленно (хотя лишь в рамках отдельного документа) управлять переходом к следующему документу за счет контекстной определенности ссылки. Соответственно, облегчается выработка решения о завершении поиска по критерию удовлетворения потребности или отсутствия новых релевантных документов в последующих выдачах. Однако, по крайней мере для класса задач поиска новой информации, более эффективным представляется метод координатного индексирования, базирующийся на свойстве комбинативности. Кроме того, линейность (одномерность) текста - носителя гипертекстовых ссылок предполагает однозначность навигации, предопределяя переход к единственному документу (может быть и наилучшему, но в жестко предопределенном контексте). Такого рода навигация в

большинстве случаев обеспечивает эффективность подачи информации, однако не дает представления о вариантности ситуации и возможных альтернативных или дополнительных аспектах.

В зависимости от формы представления информационной потребности (виду запроса) и учитывая историю развития информационных систем, можно выделить два вида "магистральных" направлений поисковых стратегий.

Большинство промышленных АИПС обеспечивают поддержку традиционной *вербальной стратегии*, отличительной чертой которой является обязательное построение завершенного, логически и синтаксически правильного выражения, посредством которого может быть получена выдача формально релевантных запросу документов.

Другим видом стратегии является *кластерная*, обобщающая понятие "документ" или "совокупность документов" до уровня запроса. Такой подход основывается на предположении, что документ, его фрагмент или группа документов может рассматриваться не только как результат поиска, но и как средство навигации, т.е. некоторый поисковый образ. Технологии, поддерживающие кластерные стратегии, в значительной мере позволяют сократить объем просматриваемой при поиске информации за счет определения на основе знаний пользователя групп документов для эффективной идентификации его потребностей.

В этом смысле сопоставление классических АИПС и гипертекстовых систем сводится не к состоянию конкурирования, а к взаимно не исключающей альтернативности.

С появлением технических возможностей реализации полнотекстовых ИС альтернативность выбора "стратегических" технологических решений практически перешла в плоскость экономических критериев о степени насыщения документов ссылками и определении пределов размеров словарей для индексирования всех или отдельных полей документов. Наиболее показательным примером является WWW-технология Internet, где массивы документов изначально создаются по гипертекстовой технологии, а в дальнейшем строятся индексы, реализующие "классические" технологии ИПС.

7.5.1.2. Методы поиска

Методы поиска, т.е. выделение подмножества документов, потенциально содержащих описание решения задачи ОД, являются отражением процесса нахождения решения и зависят от характера задачи и предметной области.

Если не рассматривать семантические проблемы смысловыражения, оптимизационная задача АИС - это минимизация совокупных временных затрат за счет снижения суммарного объема выдач, просматриваемых потребителем.

Рассматривая поиск как итеративный процесс, методы сокращения пространства перебора (просматриваемого подмножества) образуют по существу методологическую основу стратегии поиска и могут быть разделены на следующие классы:

– методы поиска в одном пространстве (обычно, тематическом);

– методы поиска в иерархически упорядоченном пространстве;

– методы поиска в альтернативных пространствах;

– методы поиска в динамическом (изменяющемся в процессе поиска[53]) пространстве.

Для случая документальных ретроспективных БД наиболее актуальными являются два первых случая, где в свою очередь можно выделить следующие подклассы:

– поиск методом уточнения / расширения области;

– поиск с использованием абстрактных пространств (динамически выделяемых в соответствии с некоторым фиксированным набором признаков);

– поиск с использованием метапространства (динамическое определение набора признаков для выделения подпространств), т.е. с переопределением метода поиска.

Учитывая опосредованность процесса извлечения информации из БД, можно сказать, что практически всегда процесс выполняется в два этапа (что соответствует и числу сторон - участников процесса). Первый этап - автоматизированный отбор документов по формальному критерию, в той или иной степени полно и точно соответствующих информационной потребности (предпочтительно более полно, хотя и менее точно), и второй - "ручной" отбор с непосредственным просмотром.

К методам поиска необходимо относить все функциональные решения - от методов сопоставления ПОЗов и отбора документов по некоторому критерию смыслового соответствия (КСС) до методов упорядочивания документов в выдаче, включая использование результатов поиска для реформулирования запроса.

Отбор документов по формальным критериям. Понятие метода[54] отбора документов по существу сводится к понятию критерия смы-

---

[53] Например, предлагаемый в [Borlund1997] метод основан на концепции "моделированной ситуации рабочей задачи" и привлечения к оценке эффективности результатов поиска реальных конечных пользователей. Метод базируется также на смешивании моделированных и реальных информационных потребностей и предполагает использование для оценки результатов как группы испытуемых, так и индивидуальных планов экспертной группы. При оценках различается тематическая и ситуационная релевантность. Учитывается также динамический характер информационных потребностей, которые у одного и того же потребителя могут со временем существенно изменяться.

[54] Например, в [Ingwersen1998] анализируются четыре метода информационного поиска (булева логика с точным совпадением, расширенная булева логика, вероятностный поиск, поиск по кластерам), рассматриваемые в связи с характером информационной потребности (уточнение - пользователю известны какие-то библиографические признаки нужного ему документа; тематический поиск - пользователь может сформулировать тематику своего запроса в адекватных понятиях и терминах; неопреде-

слового соответствия. По признаку использования мер близости (полному или частичному совпадению поисковых образов) методы можно разделить на две группы.

Методы первой группы в основном используются для реализации традиционного поиска по булевому выражению.

Второй группе соответствуют многочисленные реализации формальных моделей, таких как поиск, использующий векторные меры близости; весовой поиск и поиск по нечетким множествам; кластерный поиск; сетевые модели и модель распространяющихся активизаций (к последним также можно отнести модель детерминированного гипертекста).

Упомянутые методы достаточно хорошо проработаны и широко распространены для ИПЯ дескрипторного типа, когда смыслообразующей единицей является слово естественного (в том числе и нормализованного) или искусственного (например, классификационного) языка. То есть, сопоставляемыми в критерии объектами являются подмножества семантически связанных дескрипторов, в совокупности обозначающие характерные признаки объекта и таким образом позволяющие реконструировать его образ.

Несколько иные возможности предоставляют языки описания данных (и, возможно, некоторые ИПЯ), которые позволяют включать в качестве информационных компонентов такие поисковые средства, как явные или опосредованные ссылки на документы или их подмножества. Наиболее распространенными примерами являются гипертекстовые структуры, где можно выделить следующие структурно-технологические решения реализации ссылок:

– непосредственные ссылки - детерминированный гипертекст, связывающий в явной форме фрагмент одного документа с другим документом;

– опосредованные ссылки, устанавливаемые динамически, например, через некоторый нормализованный словарь или словарь поискового поля;

– динамически разрешаемые ссылки, определяемые поисковой функцией в контексте семантического поля документа или кластера документов.

Методы построения поискового образа запроса. Выбранные КСС в сочетании с языковыми средствами определяют способ представления информации в АИС и таким образом предопределяют методы построения ПОЗа. Однако значительное влияние имеет и характер цели

ленный поиск - пользователь хочет получить информацию о понятиях и отношениях в малознакомой ему тематической области. В случае уточнения наиболее эффективен поиск на точное совпадение. При неопределенном поиске лучше применять кластерные методы. Для тематического поиска рекомендуется использовать кластерные или вероятные методы (при наличии общей формулировки - в сочетании) и расширенную булеву логику.

(объекта) поиска. Можно было бы считать, что это влияние должно полностью учитываться на стадии выбора КСС и проектирования информационного и лингвистического обеспечения, однако, для такой системы, как документальная БД, можно говорить только о возможном преобладании (статистически) какого-то типа неопределенности. То есть, реализуемый метод построения ПОЗа должен обеспечивать более или менее эффективные способы построения запроса для достижения целей различного типа.

В соответствии с приведенной ранее типологией "чистых" стратегий поиска рассмотрим особенности методов построения запроса.

Стратегии, определенные как вербальные, ориентированы на структурные методы построения запроса, когда объект поиска достаточно хорошо определяется как известной, так и неизвестной стороной. Этому соответствуют поисковые ситуации, когда субъекту поиска известен аналог, или предмет поиска хорошо определяется в системе понятий предметной области.

В этом случае процесс построения запроса основывается на структуризации (декомпозиции объекта и синтезе понятий) объекта и смежных областей и дальнейшей редукции синтагматических связей между понятиями к отношению "совместной встречаемости".

Традиции естественнонаучной систематизации, основанные на построении древовидных структур, обуславливают "естественность" построения дерева понятий, хорошо определяемого логическими выражениями, где в качестве операндов используются наименования понятий, которые в свою очередь достаточно просто обозначаются логическиподобными выражениями в рамках конкретного ИПЯ.

Такой "естественный" и методически завершенный процесс реализуется следующими шагами (соответствующими уровням представления объекта):

– формирование понятийного образа с выделением основных и ассоциативных понятий (аналог, смежный, противоположный);

– построение дерева (сети) понятий;

– нормализация сети понятий и редуцирование связей до "совместной встречаемости";

– терминологическое (естественнонаучное) определение понятий;

– обозначение терминологических определений понятий средствами ИПЯ.

Кластерные стратегии тяготеют к построению "собирательного образа" некоторой части предметной области, границы которой либо предопределены при создании БД (чаще всего реализованы в виде систематических указателей к массивам), либо определяются динамически в процессе поиска.

В этом случае процесс построения запроса почти сливается с процессом построения выдачи, т.е. отсутствует явная граница - момент, до которого формируется запрос (например, поисковое предложение) и после которого выполняется отбор документов из базы данных по критерию соответствия. В качестве запроса могут выступать как отдельный документ, так и кластеры документов, а отбор производится "по подобию" - близости (схожести) формы или содержания.

В простейшем (но наиболее распространенном) случае основой для построения мер близости являются дискретные модели, опирающиеся на понятие множества (терминов и документов). Именно термины и документы в отдельности или их упорядоченные подмножества являются технологическими, интерфейсными объектами, обеспечивающими возможность управления поиском.

Управление процессом поиска, начинающегося с любого, может быть самого простейшего, вхождения в предметную область (например, через иерархический указатель), построено по принципу "обратной связи" [Сэлтон1973], когда каждая выдача системы (как термины, так и документы) оценивается на соответствие теме и используется в качестве запроса для следующей итерации отбора - так называемый метод реформулирования запроса по обратной связи.

Для "простых поисков" (и "простых" пользователей) кластерные стратегии эффективны чаще всего за счет простоты представления массива, предопределенного иерархиями классификаций, рубрикаций и указателей ПрО.

Для сложных поисков с высокой степенью неопределенности объекта поиска кластерные стратегии, не требующие предварительного построения точного выражения, позволяют выделить кластеры, содержащие нужные документы. Кроме того, эшелонирование выдач обеспечивает пользователя дополнительной информацией, позволяющей формулировать более объективные оценки качества поиска и, соответственно, иметь более обоснованные решения о завершении процесса поиска.

Методы обработки результатов поиска. По характеру преобразований (в контексте дальнейшего использования результатов обработки) методы обработки результатов поиска можно условно разделить на две группы:

- структурно-форматные преобразования;
- структурно-семантические преобразования (информационно-аналитические, логико-семантические).

Структурно-форматные преобразования обеспечивают необходимое разнообразие форм представления информации как на уровне отдельных документов, так и на уровне их совокупностей.

Такие преобразования реализуются:

– методами, обеспечивающими селективность выбора данных на уровне полей и агрегатов данных - внутри документа, и на уровне помеченных подмножеств документов - внутри совокупности наборов документов (например, выдач по отдельным запросам);

– методами представления (пространственного размещения) информации, включая методы "масштабирования" - возможностью иметь сокращенную форму, позволяющую свернуть, например, документ до одной строки, а выдачу - до одного экрана, обеспечивая локализацию области восприятия информации и создавая некоторую "образность";

– методами упорядочения как по структурным параметрам (например, значениям отдельных полей документов), так и по семантически значимым параметрам, например мерам близости в пространстве "термин-термин" или "документ-документ".

Структурно-семантические преобразования должны обеспечить пользователя (субъекта управления процессом поиска) информацией, необходимой для развития процесса поиска, например, путем реформулировки запроса по обратной связи или информацией для принятия обоснованного решения о достаточности полученного результата или прекращении процесса по причине несоответствия профиля базы данных теме запроса.

Соответственно, эти методы можно разделить на две группы:

– информационно-аналитические, т.е. обобщающие результаты, например построением распределений, статистически характеризующих выдачу или множество терминов. Такие количественные характеристики могут использоваться для принятия решения “в целом” (как внешняя независимая статистическая экспертиза);

– логико-семантические, т.е. обеспечивающие межуровневые преобразования информации (например, преобразования лексики документов некоторого семантически связного кластера в кластеры ранжированных словников).

Поисковые методы, таким образом, образуют интерфейсный слой, а при проектировании программного обеспечения АИПС - определяют требования к «внутрисистемным» методам организации и поиска данных.

7.5.2. Технологии поиска и обработки результатов

Так же, как и методы поиска, технологии, используемые в АИС, необходимо рассматривать, с одной стороны, в контексте этапности обработки запроса, и с другой стороны - как управляемый пользователем процесс получения и преобразования операционных объектов в среде конкретного программно-технологического комплекса. Соответственно, интерфейсные объекты как средства управления можно разделить по отношению к базе данных на *основные* и *технологические*.

Основными будем считать «необходимые» объекты – документы и запросы.

Технологическими – все те объекты, которые являются вспомогательными, обеспечивающими эффективность доступа как с точки зрения быстродействия, так и удобства. Такие объекты могут являться физической частью БД – словари, схемы, или временно создаваться, например, списки запросов, словники и т.д.

Технологии поиска. Технологии поиска (и как итог - получения выдачи) объединяют два процесса:

- процесс объявления (выражения, обозначения) пользователем информационной потребности;
- процесс построения системой информационного массива - множества документов, выдаваемых пользователю в ответ на поисковое требование.

Технология генерации выдачи полностью определяется архитектурой программного и информационного обеспечения конкретной системы.

В этом смысле выделяются два класса систем: *диалоговые* и *пакетные*. В первом случае технология ориентирована на работу в реальном масштабе времени, а условие отбора по одному запросу соотносится со всеми поисковыми образами документов (если база данных не имеет инвертированных массивов, являющихся избыточными по отношению к основному). При пакетной обработке запросов, предназначенной для использования в системе избирательного распределения информации, каждый поисковый образ документа (чаще всего при поступлении в систему) соотносится с поисковыми образами всех запросов.

Разнообразие технологий подготовки запроса, таким образом, в основном относится к диалоговым системам, обеспечивая (в той или иной степени) не только эффективность вхождения в базу, но и качество поиска в целом.

В контексте понятия “выражение запроса“, как главного (основного) операционного объекта можно определить три типа не взаимоисключающих технологии его спецификации[55]:

– непосредственного набора булевоподобного выражения;

– формирования булевоподобного выражения с помощью “конструктора запросов”, облегчающего использование словарей, имен полей и операторов структурно-логической связи;

– форматно-ориентированных форм генерации “запроса по образцу”.

Технологии обработки результатов поиска. Возможности АИПС по обработке результатов поиска можно рассматривать в следующих аспектах:

– управление формой представления документов и списков документов (оперативное изменение формата и наполнения, фрагментирование);

– управление последовательностью выдачи (сортировка, ранжирование, оперативные переходы с возвратами, “закладки”);

– локализация результата на уровне отдельного документа или совокупности (отметка степени соответствия информационной потребности);

– использование фрагмента документа, ссылки на документ или совокупность документов в последующих запросах, а также для развития процесса поиска;

– использование результатов поиска для оценки качества поиска.

Развитые средства обработки результатов поиска предопределяют возможность разработки средств и технологий автоматической или автоматизированной реформулировки запроса.

Методы ранжирования документов в выдаче основываются на использовании количественных мер, отражающих либо статистику встречаемости[56] и взаимосвязи терминов в документах, либо статистику взаимосвязи документа с другими документами в выборке или в предметной области

---

[55] Интерфейсные решения средств подготовки запроса будут рассмотрены в следующей главе.

[56] Например, в [Courtois1999] проведен сравнительный анализ методов ранжирования результатов поиска использующих три стратегии. Первая предполагает наивысшее ранжирование документов, содержащих все перечисленные ключевые слова или термины. Вторая предоставляет наивысший ранг документам, которые содержат все заданные термины в рамках одного предложения или фразы. Третья стратегия предусматривает наивысшее ранжирование тех документов, которые содержат заданные термины в своих заголовках, резюме или аннотациях. На основании проведенных экспериментов сделан вывод о том, что наиболее точное ранжирование (соответствующее пользовательским оценкам релевантности результатов) обеспечивает первая стратегия, а наименее точное - третья стратегия.

Технологии управления. Управление процессом поиска для диалоговых АИС определяется:

– разнообразием операционных объектов и средств их обработки, определяющих возможные методы получения результата и особенности его представления;

– интерфейсными решениями (зависящими от архитектуры программно-технического комплекса), определяющими степень гибкости сценария и его управляемость как пользователем так и системой.

Рассматривая поиск в контексте понятия «стратегия» и представляя его как динамический процесс с изменяющимися состояниями сторон диалога (пользователь - система), можно (хотя и очень условно) выделить три типа поисковых технологий:

– технологию “запрос-ответ”, как реализацию истинно вербальной стратегии;

– технологию “ накопления результата”, когда система позволяет не только использовать ссылки на результаты отдельных поисков, но и получать сам результат способом, отличным от запросно-ответного;

– технологию “распространяющейся активности”, позволяющую не только изменять способ получения результата, но и изменять цель - предмет поиска, обеспечивая как дифференцированное использование результатов, так и восстановление поисковой ситуации.

7.5.3. Поисковые интерфейсы

Пользовательский интерфейс – это «совокупность информационной модели проблемной области, средств и способов взаимодействия пользователя с информационной моделью, а также компонентов, обеспечивающих формирование информационной модели в процессе работы» [Гультяев2000].

Интерфейс информационной системы как пользовательское средство управления процессом поиска - это комплекс компонентов (операций и технологических объектов), возможно, объединенных в блоки по функциональному или какому-либо другому принципу и реализованный в виде системы команд, меню или сценариев.

С другой стороны (проектировщика), интерфейс - это целесообразная реализация совокупности операционных объектов и механизмов, обеспечивающих необходимый уровень избирательности вхождения в предметную область базы данных и должный уровень управляемости процессом для достижения его сходимости (или, по крайней мере, определенности условий его завершения). При наличии пользователей различных категорий система должна обладать средствами выбора пользовательского интерфейса, удовлетворяющего требованию необходимого и достаточного уменьшения сложности для пользователей различных категорий.

Интерфейс с точки зрения психологии восприятия должен «отражать не только отдельные объекты и их взаимосвязь в данный момент времени,

но и те изменения, которые происходили в предшествующие моменты, а также тенденцию дальнейших изменений» [Ломов1986]. Т.е., восприятие интерфейса пользователем включает элементы предвосхищения действий и состояний, обеспечивая возможность своевременности действий и их оптимизацию.

Управление системой (посредством интерфейса) предполагает наличие у пользователя представления не только об отдельных технологических объектах, но и о последовательности их использования. Такое представление может быть следствием опыта (накапливания восприятия типовых ситуаций) или появляться как следствие восприятия метафор[57], являющихся упрощенной моделью ситуации и использующих операционные объекты иной (общепонятной) природы.

Эффективное же управление возможно только при наличии у пользователя точной (адекватной) модели системы, построенной на основе вербально-логического понятийного мышления с агрегированием (абстракцией) и декомпозицией объектов на уровне понятий и использованием знаковых систем для фиксирования знаний и их передачи. В этом смысле выделяют [Попов1987] следующие виды знаний:

- стратегические (знания о цели и плане ее достижения);
- структурные (знания о систематизации объектов предметной области);
- поддерживающие (обеспечивающие восприятие или построение правил взаимодействия).

При этом именно средства интерфейса должны «объяснить» [Попов1987] существо системы, т.е. обеспечить основу и поддержку определения стратегии поведения, а также соотнесения каждого действия, направленного на получение результата, с обобщенной моделью процесса, явно задать содержание каждого объекта и обосновать переход от предпосылок к выводам. В общем случае такое объяснение должно быть информативным, пригодным для всех типов пользователей, быть многоуровневым и иметь ту или иную степени конкретности/абстрактности в зависимости от ситуации.

7.5.3.1. Организация интерфейсных объектов

Для того, чтобы дать пользователю возможность оценить полноту полученного результата (по сути, возможность управлять процессом поиска), необходимо предоставить специальные инструменты, позволяющие обращаться к ранее полученным объектам и результатам. На уровне интерфейса эти объекты должны быть отделены от средств поиска и работы с документами, чтобы пользователю проще было переключаться с задачи своей основной деятельности (сбора информации для решения задачи) на проблемы оценки своих действий и состояний.

---

[57] Понимание метафоры всегда ситуативно и субъективно и его невозможно предвидеть.

Для оценки динамики эффективности процесса поиска дистрибутивно-статистическими методами необходимо, чтобы все этапы технологии относились к одному (семантическому однородному) пространству объектов (т.е., задача не должна быть многокритериальной). Иначе говоря, результаты, получаемые на разных этапах и, соответственно, по разным поисковым образам, должны относиться к одному исходному (семантически замкнутому) запросу, который как предмет поиска представляет тематически отдельную реальную информационную потребность пользователя. То есть, в этом случае оценивается эффективность уже поискового образа запроса, а повышение эффективности поискового процесса основывается на последовательном повышении эффективности ПОЗа по отношению к предшествующему варианту (которая может определяться на основе, например, корреляционного анализа подмножеств $\{T_1, T_2, \ldots T_n\}$ и $\{ D_1, D_2, \ldots D_n \}$, введенных в гл. 4).

Для реальных запросов, которые практически являются многоаспектными и включают несколько подтем, общий результат будет получен последовательностью фактически самостоятельных (завершенных с точки зрения получения и *оценки* результатов) вышеописанных многоэтапных процессов поиска, каждый из которых должен быть выполнен для каждой подтемы и аспекта. То есть, как это представлено на рис. 7.5, каждому отдельному элементу тематически-аспектной декомпозиции запроса, представляющей информационную потребность как семантически значимый объект поиска на *логическом* уровне, соответствует отдельный *физический* процесс поиска и результат. При этом в реальных ИПС результаты поиска по отдельным этапам *последовательно* фиксируются в протоколе, позволяющем отобразить ход процесса и, возможно, на следующих этапах обратиться к ранее полученным результатам.

Однако изолированность объекта поиска и, соответственно, результатов, предопределенная требованием оцениваемости, на практике трудно достижима: множество документов, выданных при поиске по одному аспекту, обычно содержит документы, относящиеся и к другим аспектам. И, кроме того, в многоэтапном процессе развития запроса пользователь, получая значимый или просто интересный документ, но относящийся к другому аспекту, обычно переключает внимание именно на него и, соответственно, выходит за пределы тематически замкнутого пространства, что нарушает требование однородности и снижает эффективность поиска.

Это означает, что представление процесса поиска на физическом уровне (последовательность получения результата, зафиксированная в протоколе в виде интерфейсных объектов) не будет соответствовать последовательности на логическом уровне. Для обеспечения соответствия вводится промежуточный *интерфейсный уровень* представления процесса поиска. Объекты этого уровня (и характер их представления, например, упорядочение) структурно будут соответствовать логическому уровню, и каждый из них будет представлять (объединять) элементы

(ПОЗы, словники, результаты поиска), относящиеся к соответствующему предмету поиска, но физически полученные, возможно, на разных этапах.

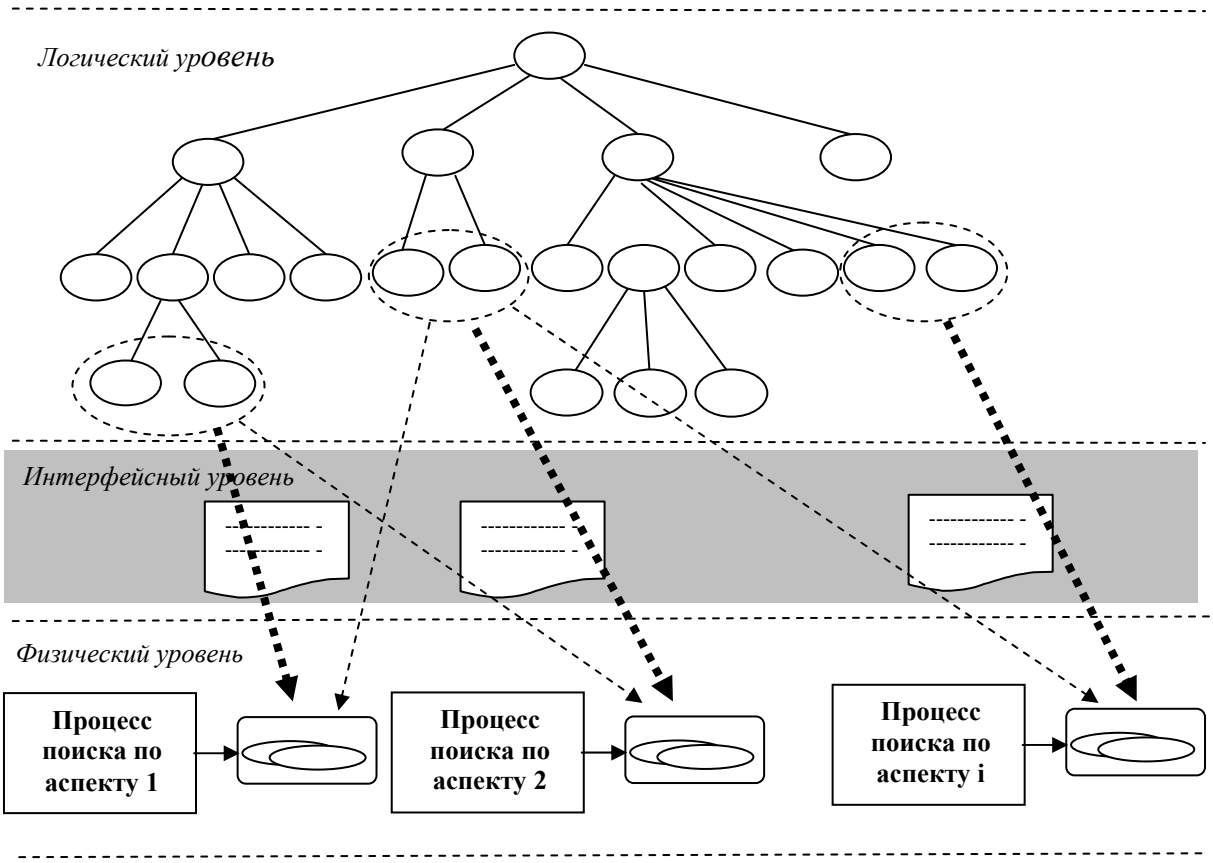


Рис. 7.5. Уровневая модель поискового процесса

Тем самым, на передний план выдвигается проблема организации взаимодействия пользователя с системой в процессе поиска.

И если для процесса в целом (с точки зрения конечного пользователя) мы имеем всего два типа *основных операционных объектов* – запрос (как пользовательское представление ИПП) и документ (как семантически целостный ответ или его часть, сформированный системой – отображение запроса в пространство документов), то с точки зрения организации процесса взаимодействия интерфейс системы должен иметь разнообразные объекты. При этом разнообразие типов объектов пользовательского интерфейса определяется «развитостью» технологических и процедурных возможностей системы. Для случая обобщенной схемы, технологически обеспечивающей снятие информационной неопределенности всех типов, такими объектами являются:

- тезаурусы, обеспечивающие ориентацию пользователя в предметной области;
- словари поисковой системы, используемые для формирования поискового выражения;
- тематические словники, представляющие информативную лексику предметной области.

Эти объекты, являясь технологически вспомогательными, используются на разных этапах поиска и обеспечивают возможность более или менее адекватного выражения информационной потребности пользователя. При этом для отражения индивидуальных особенностей ИПП, они, как интерфейсные объекты, не могут быть эффективно использованы, поскольку, в следствие усредненной природы, представляют ПрО в целом.

Для этого на промежуточном интерфейсном уровне можно использовать иерархически организованные структуры, отражающие пользовательское видение системы понятий предметной области. Причем, каждый такой объект представляет как общепринятое, так и индивидуальное видение ПрО. Интегральность такого представления достигается за счет того, что оно реализуется объектами как уровня ресурсов (подборками документов, ссылками на ассоциированные ресурсы и т.д.), так и уровня терминологии (тезаурусами, рубрикаторами, словниками)

При таком подходе информационная система может помимо стандартных и расширенных поисковых возможностей иметь средства систематизации информационных массивов, формирования и развития компонентов лингвистического обеспечения, а также оценки и анализа результатов поиска. Такими средствами, ориентирующими пользователя в предметной области в части терминологии, могут быть:

- автоматическое формирование наборов терминов для расширения запроса;
- терминологические таблицы, организованные в виде семантических сетей, каждому узлу в которых сопоставлен некоторый набор документов;
- средства автоматизированного ведения пользовательских минитезаурусов и тематических рубрикаторов и др.

7.5.3.2. Роль системы в процессе взаимодействия

Сложность разработки представления интерфейсных инструментов в целом заключается в том, что неизвестно, каким образом пользователи будут воспринимать объекты, систематизированные и сгруппированные разработчиком в соответствии с его, разработчиком, пониманием. Основой для прогнозирования возможных реакций и действий пользователя является выявление типичности поведения пользователей.

Процесс взаимодействия можно рассматривать на трех уровнях: глобальном, тематическом и отдельного шага[58].

Глобальная структура процесса взаимодействия зависит от цели ОД, состояния предметной области и будет определять стратегию поиска. Тематическая структура зависит от характера отдельных задач, их проработанности, политематичности и будет определять характер навигации – последовательности выполнения отдельных поисков информации по различным направлениям. Структура шага соответствует отдель-

[58] В контексте ОД этому соответствуют следующие этапы: определение цели, выделение задач и подзадач, решение задач.

ной технологии нахождения документов по отдельному запросу и будет определяться набором средств, доступных пользователю (их наличием в интерфейсе, а также готовностью пользователя к их применению) .

Развитые АИС могут иметь несколько интерфейсных решений. Применительно к случаю категорий пользователей можно рассматривать:

1) интерфейс конечного пользователя, обеспечивающий выбор объектов и методов из предлагаемого (чаще всего фиксированного) набора;

2) интерфейс системного администратора, обеспечивающий не только расширенный набор средств, но и позволяющий изменять или создавать новые интерфейсные объекты или сценарии.

По типу диалога (в зависимости от степени активности сторон) можно выделить три уровня системной активности:

1) диалог в режиме "запрос-ответ", когда в ответ на запрос (объект единственного типа) система формирует ответ, включающий объекты, возможно, другого, но также единственного типа;

2) режим информационно-советующих систем, когда помимо прямого ответа на запрос система формирует или обеспечивает доступ к справочной или оценочной информации;

3) режим симметричного (равноправного) диалога, т.е. с активной ролью системы.

При этом активность системы по отношению к пользователю может реализовываться различными путями:

– непосредственным вмешательством в процесс через изменение параметров процедур, например, изменением порога выдачи или сценария поиска;

– построением прямых или косвенных оценок параметров выдачи (показателей эффективности поиска);

– генерацией технологических объектов, являющихся дополнительными или альтернативными по отношению к тем, которые получены пользователем (например, построение словников при реформулировании запроса по обратной связи).

Это позволяет компенсировать односторонность представления пользователя о способе выражения (обозначения) предмета поиска и таким образом уйти от ситуации конкурирования (за "наилучший" запрос) различных способов выражения запроса к состоянию целенаправленного сочетания альтернатив.

Особенностью поисковых моделей человеко-машинного взаимодействия является различие принципов (основания) «системности». АИПС – кибернетическая система, целенаправленно созданная и имеющая детерминированную структуру, причем ее поведение (динамика)

оценивается с точки зрения устойчивости. Субъект поиска (система потребления информации) - саморазвивающаяся, синергетическая система, самоорганизация которой основана на противоречии, и часто приводит к структурной реорганизации, а целостность определяется свойствами, не сводящимися к сумме свойств составляющих ее элементов. Эта особенность определяет различие в уровне требований к адаптивным возможностям взаимодействующих систем и организации диалога. Профессиональному пользователю система должна предоставить набор средств, обеспечивающий адаптацию через настройку (целенаправленный выбор конфигурации и параметров самим пользователем). В других случаях система должна предлагать набор типовых стратегий и технологий поиска, выбор которых может производиться как пользователем, так и системой на основании диагностики поведения пользователя.

Таким образом, интерфейсные решения определяются:

– размерностью пространства отображения (одномерного - строка, двумерного - рабочий стол экрана, многомерного, использующего псевдотрехмерные изображения, цветовые и фоновые выделения);

– реактивностью системы, зависящей от средств обмена между терминалом пользователя и процессором, реализующим функции системы (режим удаленного терминала, полный диалог в режиме реального времени, режим пакетного обмена).

Роль системы на основных этапах поиска представлены в таблице 7.3.

Таблица 7.3.

Этап поиска	*Роль системы*
Определение (локализация и формализация) ***темы запроса*** и идентификация ресурса	Обеспечение метаинформирования о тематике, наполнении, структуре и методах доступа к выбранному ресурсу
Формирование, а также структурное и лексическое адаптирование ***выражения запроса***	Предоставление вспомогательных информационных объектов (словарей, тезаурусов, шаблонов и т.д.)
Отбор документов по критерию, по возможности адекватному степени неопределенности информационной потребности	Предоставление выбора механизма поиска или, например, автоматическое расширение лексики запроса
Формирование и управление выдачей найденных документов	Обеспечение масштабирования (форматирования) пространства представления выданных документов, а также сортировки и, возможно, ранжирования по некоторому формальному критерию соответствия
Оценка результатов поиска с точки зрения полноты удовлетворения информационной потребности (т.е. завершение поискового процесса) или их соответствия цели поиска и степени освоения информационного ресурса по теме запроса	Количественная оценка динамики выдач и обеспечение возможности выборочного обращения к результатам отдельных этапов процесса поиска
Развитие запроса по технологии «реформулирования по обратной связи по релевантности» или использование других ресурсов, например, ассоциированных баз данных вторичной или справочной информации	Адекватное информирование о возможностях развития запроса и средствах оценки результата

Интерфейс системы, являясь той операционной средой, которая может обеспечить решение двойственной задачи (получение результата и нахождение средств его получения), должен удовлетворять следующим требованиям:

1) Организация пространства по принципу однородности;
2) Предоставление информации для оценки результата;

3) Возможность отслеживания и изменения "траектории движения";

4) Возможность планирования «навигации» (подсказки к дальнейшим действиям);

5) Наличие информации для оценки степени завершенности процесса;

Соответственно, оптимизационная задача процесса поиска – при временных ограничениях максимизировать показатели выдачи и получить максимальную (субъективную) уверенность в качестве поиска, предоставляя пользователю в процессе диалога альтернативные направления, а также количественные и качественные оценки их соответствия запросу.

Обобщенная схема поиска, приведенная на рис. 7.6, представлена на примере АИПС IRBIS, которая имеет следующие основные интерфейсные блоки[59]:

- интерфейс формирования запроса,
- интерфейс поискового модуля,
- интерфейс обработки результата и развития поиска.

[59] Примеры реализации интерфейсов подготовки и развития запроса приведены в гл. 8.

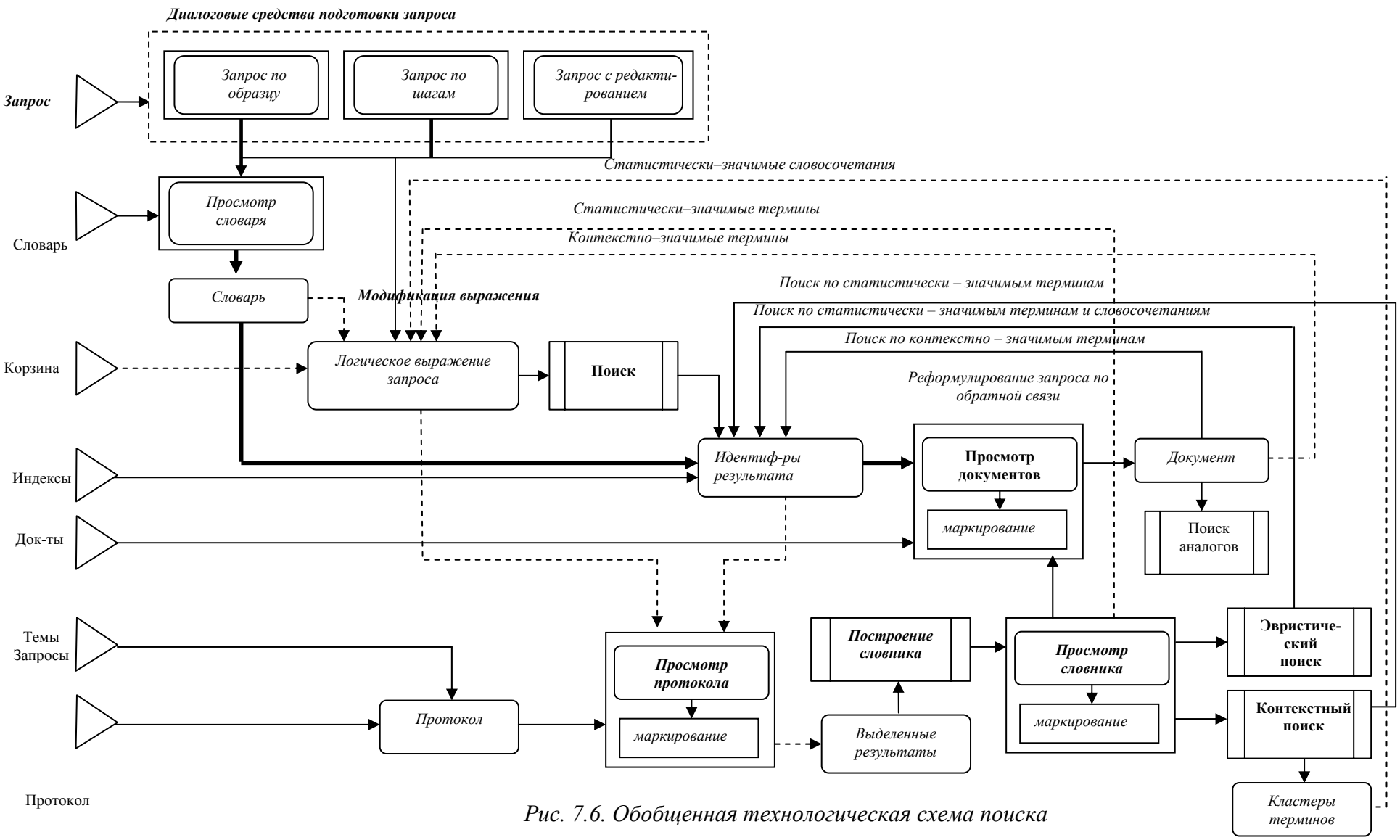


Рис. 7.6. Обобщенная технологическая схема поиска

Интерфейсы формирования запроса, особенно значимые при начальном входе в информационное пространство базы, подразделяются на «вербальные» и «кластерные».

Первые являются процедурно ориентированными средствами, обеспечивающими формирование на ИПЯ логического выражения той или иной сложности, что предполагает предварительную[60] структуризацию на семантическом уровне и лексическую адаптацию запроса. Представленные в системе процедурные интерфейсы – конструкторы запроса «по образцу», «по шагам» и редактор запроса, являются типовыми решениями с различным уровнем дружественности[61]. Эта группа является практически полной с точки зрения степеней свободы выбора операционных объектов для случая спецификации запроса в виде обобщенного логического выражения. Характер ограничений на уровне компонентов составного выражения отбора (логической связки выражений для отдельных областей поиска) отражен в таблице 7.4.

Таблица 7.4

<table>
<tr><th colspan="3">Компоненты выражения</th><th>Конструктор «по образцу»</th><th>Конструктор «по шагам»</th><th>Редактор запроса</th></tr>
<tr><td colspan="3">Ограничения на условие поиска</td><td>Фиксировано шаблоном. Оператор фиксирован (обычно И)</td><td>Линейное выражение без вложенности</td><td>Скобочное выражение любой сложности</td></tr>
<tr><td rowspan="2">Область поиска</td><td colspan="2">Имя поля</td><td>Фиксировано или ограниченный выбор</td><td>Выбор из списка</td><td>Скобочное выражение любой сложности</td></tr>
<tr><td colspan="2">Оператор</td><td>Нет или фиксирован (обычно ИЛИ)</td><td>Выбор из списка</td><td>Любой</td></tr>
<tr><td colspan="3">Оператор отбора</td><td>Фиксирован</td><td>Выбор из списка</td><td>Любой</td></tr>
<tr><td rowspan="3">Операнд-значение отбора</td><td rowspan="2">Операнд</td><td>Выраж.</td><td>Нет</td><td>Термин или предшествующее выражение</td><td>Скобочное выражение любой сложности</td></tr>
<tr><td>Термин</td><td>Выбор из списка или ввод</td><td>Выбор из списка</td><td>Скобочное выражение любой сложности с операторами усечения</td></tr>
<tr><td colspan="2">Оператор</td><td>фиксирован (обычно ИЛИ) или ввод</td><td>Выбор из списка</td><td>Любой</td></tr>
</table>

[60] Это противоречит практически и «идеологически» значимому требованию не заставлять человека обрабатывать вводимые данные вне системы [Коутс1990] и на практике приводит к низкой эффективности поиска.

[61] Здесь «дружественность» сводится исключительно к снижению уровня требований к подготовленности пользователя за счет снижения синтаксической сложности выражения.

Интерфейсы формирования запроса, основанные на «кластерных» методах, используют операционные объекты с разным уровнем отражения информационного содержания базы:

– на «атомарном» уровне - это в основном инвертированные формы (словари, индексы) или документы БД, представляемые в физической последовательности;

– на уровне тематическом - это структуризации с иерархической или линейной упорядоченностью через тематические рубрикации или типовые запросы;

– на уровне частных запросов - это коллекции выражений запросов и сохраняемые запросы (выражения и соответствующие результаты поиска).

Интерфейс поискового модуля обеспечивает выбор и управление механизмом отбора документов по сформированному условию поиска, включая:

– предобработку выражения: нормализацию или автоматическое усечение терминов, лексическое расширение выражения, связывание с предыдущими или уточняющими условиями;

– выбор метода отбора (критерия смыслового соответствия) – булев или нечеткий поиск, определение порога выдачи;

– постобработку результата – сортировку, ранжирование или кластеризацию найденных документов по содержанию или вычисляемым показателям.

Интерфейсные средства обработки результата и развития поиска использует два типа операционных объектов – отдельные документы и коллекции документов. Особенностью предложенной схемы является функциональное подобие интерфейсных блоков и то, что помимо функций обработки материала документов (печать, редактирование, сортировка, вывод в файл и т.д.), система предоставляет средства развития процесса поиска либо путем модификации выражения, либо путем реформулирования запроса по обратной связи по релевантности.

Модификация выражения запроса является методически достаточно простой операцией - выражение редактируется переносом в него статистически или контекстно значимых терминов из словников или документов.

Поиск путем реформулирования запроса на основе лексики документов, релевантность которых подтверждена пользователем, реализуется двумя типами поисковых технологий: непосредственным использованием лексики отдельного документа или на основе терминов, статистически взвешенных на множестве документов.

Приведенные кластерные технологии являются эвристическими в том смысле, что в результате процесса поиска формируется одна или не-

сколько документальных областей, в той или иной степени соответствующих потребности пользователя. Принципиальным моментом является то, что система готовит *альтернативы*, обеспечивая тем самым упорядоченность и идентифицируемость направлений дальнейшего перебора. Сравнительная характеристика видов поиска, реализующих кластерные технологии, приведена в табл 7.5.

Таблица 7.5.

Вид поиска	Стартовый объект	Лексическая основа для реформулировки запроса	Механизм поиска	Постобработка выдачи	Специализированные интерфейсные средства управления
Поиск по терминам, выделенным в документе	Отдельный просматриваемый документ	Термины, выделенные пользователем	Булев поиск по всем текстовым полям с автоматической нормализацией терминов	Нет	Нет
Поиск «аналогов»	Отдельный просматриваемый документ	Содержание просматриваемого документа	Нечеткий поиск: по условию частичного вхождения с указанным порогом	Ранжирование по суммарному числу вхождений поисковых терминов	Диалоговая панель «Поиск аналогов»
Эвристический поиск	Множество документов предложения запроса из протокола	Термины всех поисковых полей из документов, отмеченных как релевантные (словника не предъявляемого пользователю)	Поиск по статистически наиболее значимым кластерам терминов из словника	В соответствии с порядком ранжирования кластеров	Нет
Контекстный поиск	Множество документов предложения запроса из протокола	Отмеченные термины словника создаваемого системой из всех поисковых полей документов, отмеченных как релевантные	Поиск по кластерам терминов из словника, отмеченных как релевантные	В соответствии с порядком ранжирования кластеров	1. Словник; 2. Предложения протокола, содержащие результат поиска по каждому кластеру терминов

Контрольные вопросы

1. Перечислите типы информационной потребности пользователя и определите их связь уровнями информационных объектов.
2. Роль стереотипов в процессе организации и управления поиском.
3. Дайте сравнительную оценку характера деятельности человека и компьютерной системы.
4. Приведите основные процессы в уровневой модели взаимодействия пользователя и системы.
5. Дайте определение понятия «интерфейс пользователя».
6. Охарактеризуйте модели взаимодействия пользователя с АИПС и базой данных.
7. Охарактеризуйте влияние интерфейсных средств на адаптацию пользователя.
8. Приведите примеры диалоговых интерфейсных средств обучения пользователя работе с АИПС и базой данных.
9. Проведите сравнительный анализ вербальной и кластерной стратегий поиска.
10. Определите зависимость методов построения запроса и стратегий поиска.

8. Интерфейсные средства информационного поиска

Как отмечалось в главе 6, функционирование современных ИПС основывается на двух предположениях: 1) документы, необходимые пользователю, объединены наличием некоторых характеристических признаков; 2) пользователь способен указать эти признаки. Оба эти предположения на практике редко выполняются и можно говорить только о вероятности их выполнения. Поэтому, процесс поиска информации обычно представляет собой последовательность шагов пользователя, который, обращаясь к различным интерфейсным объектам, так или, иначе - формирует поисковый запрос, более или менее адекватно отражающий эти характеристические признаки.

С точки зрения «интеллектуальности» средств поиска и в зависимости от характера информации (и возможностей разработчика) в основу конкретной, соответственно, более или менее сложной АИПС может быть положена одна из следующих технологий поиска: литеральный поиск – поиск подстроки, происходящий без привлечения знаний о лексической, грамматической и семантической структуре обрабатываемого материала; поиск, в ходе которого используется лексико-грамматическая информация, то есть привлекаются лингвистические словари, программы морфологического анализа текста; семантический поиск, осуществляющийся на основании знания об отношениях между понятиями предметной области, выраженными средствами естественного языка.

В последнем случае носителями такого рода информации, в частности, являются тезаурусы, уже более трех десятилетий использующиеся для информационного поиска. Кроме того, огромную роль в организации диалога между пользователем и информационно-поисковой системой играют хотя и менее сложные, но разнообразные словарные структуры. Используя их, пользователь может развивать поиск, модифицируя запрос (выражение его информационной потребности) согласно особенностям представления объекта поиска средствами конкретной ИПС и базы данных.

Некоторые решения интерфейсных средств представления запросов будут рассмотрены ниже на примере АИС IRBIS.

8.1. Средства формирования запросов

Поисковые механизмы построены на основе ИПЯ, однако технология и средства формирования запроса, предоставляемые пользователю в виде *поисковых интерфейсов*, не должны требовать от него обязательного знания и навыков построения выражений алгебраического вида.

Поисковые интерфейсные средства условно можно разделить на два класса. Первый класс (сценарии типа «укажи и выбери») - это конструкторы запросов, которые позволяют, используя термины поисковых словарей или других поисковых структур (тезаурусов, рубрикаторов, словников), в режиме диалога построить выражение той или иной сложности, которое на следующем шаге (выполнения поиска) даст результат.

Второй класс – это средства, реализующие простейший сценарий типа «укажи и получи». В этом случае пользователь выделяет в отображаемом объекте (документе или множестве документов) значимые с его точки зрения элементы (термины в документе или словаре; документы в выборке или протоколе) и, используя механизмы поиска по сходству (поиск аналогов, эвристический поиск, поиск с использованием обратной связи), получает выдачу, минуя этап составления поискового выражения.

В основу формирования поискового запроса по технологии «укажи и выбери» в системе положено три различных подхода к построению выражений запросов разной степени сложности (ориентированных на разные степени подготовленности пользователя):

- *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 предопределенных списков. Причем, по умолчанию предполагается, что отдельное условие - это список терминов (синонимов), обычно выбираемых из словаря и обозначающих одно и тоже понятие.
- *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 путем непосредственного набора выражения запроса с возможностью обращения в произвольном порядке к словарям, спискам имен полей и т.д.

8.1.1. Формирование запроса «по образцу»

Режим поиска «по образцу» ориентирован на автоматизированное формирование достаточно простого логического выражения, объединяющего условия, относимые обычно к взаимно ограничивающим разнотипным полям документа.

Например, по условию:

PB:(M.) и KW:(SCIENTOMETRICS или БИБЛИОМЕТРИЯ) и DT:(1999 или 2000)

будут найдены библиографические описания документов, опубликованных в Москве в 1999 или 2000 гг., поисковые образы которых содержат понятие «библиометрия».

Средства построения запроса в конструкторе следующие (рис.8.1):

- панель просмотра словаря текущего поискового поля;
- панель шаблона, представляющего для каждой поисковой области окно выбора поискового поля и окно поискового условия для этого поля;
- панель инструментов.

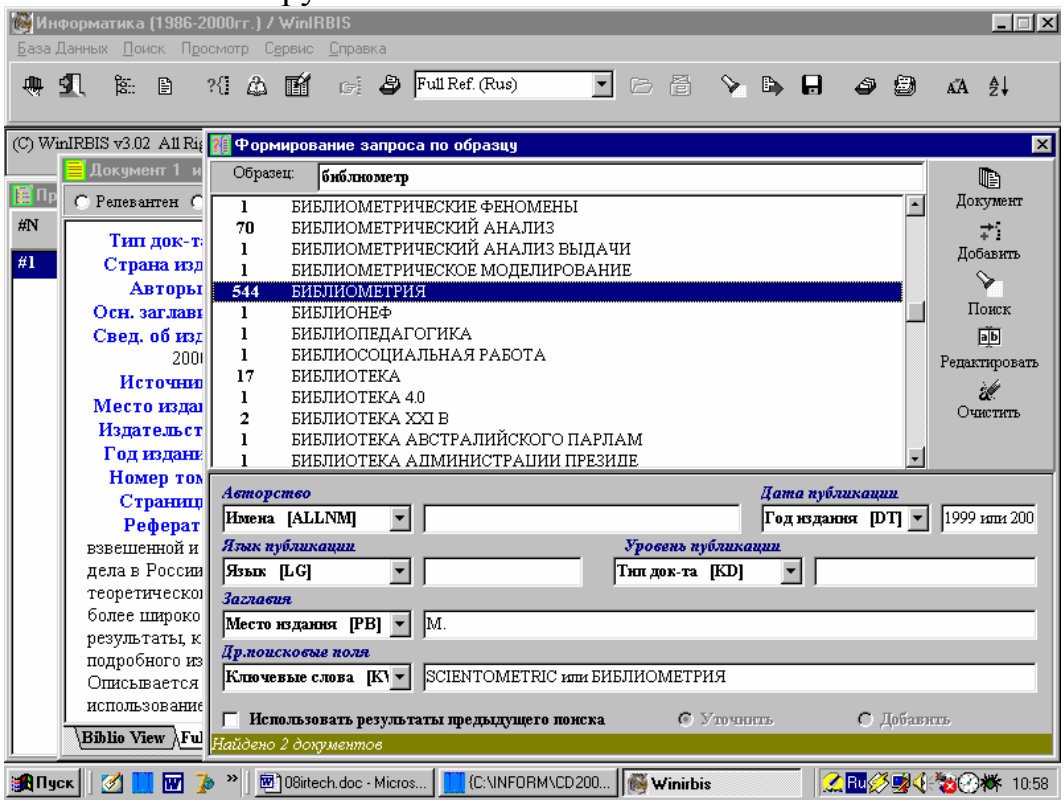


Рис. 8.1. Формирование запроса «по образцу» с использованием словаря

Формирование запроса начинается с выбора области поиска (имени поля) в одной из групп, представленных на панели шаблона. Список полей группы специфицируется текущей схемой БД.

Например, группа «Заглавия» может включать поля: «Осн. Заглавие [TI]», «Источник [SO]», «Место издания [PB]» и «Серия [SER]». Выбор поля вызывает отображение соответствующего словаря на панели просмотра словаря.

Любой термин или группу подряд выделенных курсором терминов в области отображения словаря средствами панели инструментов можно поместить в область поискового условия, создаваемого для этого поля. При этом, добавляемые термины автоматически связываются оператором ИЛИ, обеспечивая, таким образом, отбор документов из базы данных по условию присутствия в документе хотя бы одного из них.

Перед добавлением термина в запрос, его можно сначала отредактировать (включая добавление символов маскирования).

Для перехода к формированию логического выражения для другой области поиска необходимо активизировать комбинированный список другой группы имен полей и выбрать нужное поле (рис. 8.2).

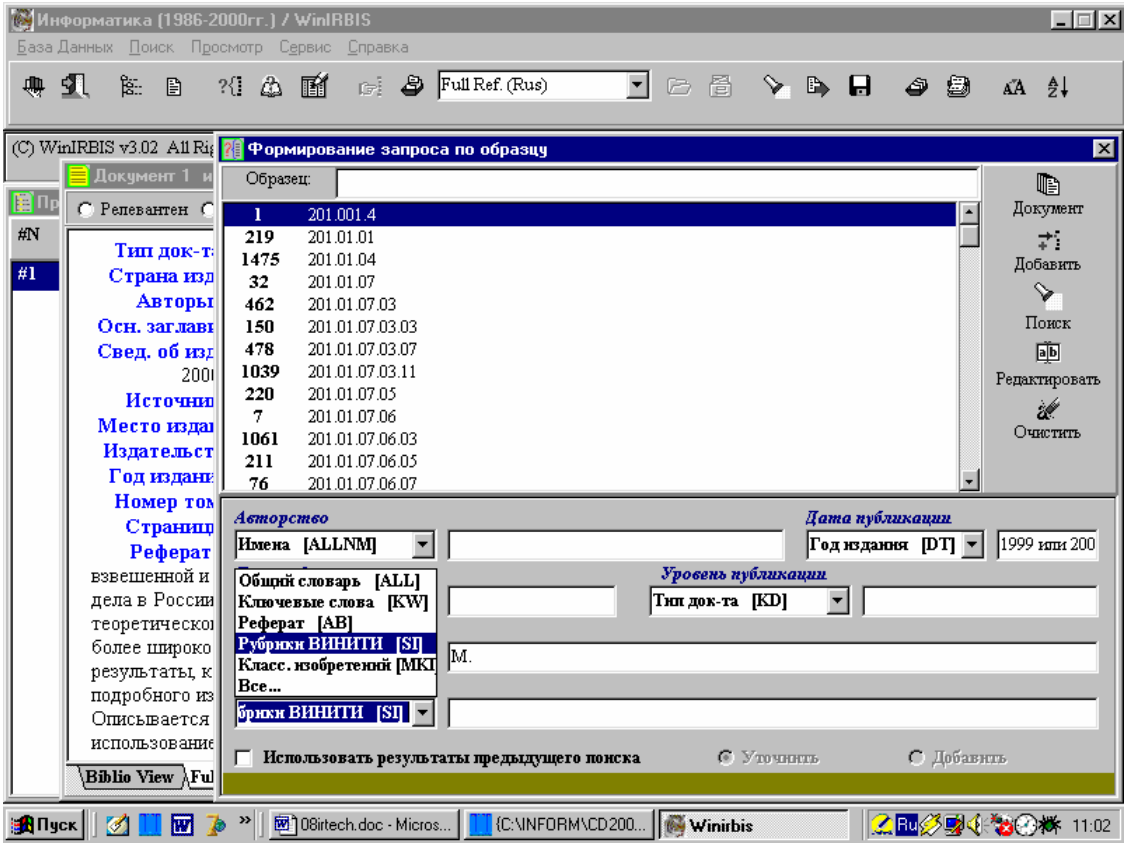


Рис. 8.2. Выбор имени поля в бланке запроса

Если в запросе заданы условия для нескольких поисковых областей, то (по правилам построения такого предложения запроса) соответствующие поисковые выражения всегда связываются оператором AND

(И), обеспечивая таким образом отбор документов из базы данных по принципу обязательного выполнения условий по всем специфицированным областям.

Поисковое условие для отдельного поля синтаксически представляет собой определенное ранее <*Выражение условия*> и может быть введено и отредактировано с помощью клавиатуры в выделенной области редактирования в соответствии со всеми возможностями ИПЯ, в том числе с использованием допустимых операций, символов маскирования и скобочных конструкций.

Для реализации технологии последовательного формирования поискового множества документов конструктор предусматривает возможность обязательного связывания результата поиска по сформированному предложению запроса с результатом поиска по предыдущему предложению. Соответствующий набор параметров позволяет либо *«Уточнить»* предыдущий результат (в этом случае в итоговый результат включаются документы, удовлетворяющие вновь построенному предложению запроса и присутствующие в предыдущей выдаче), либо *«Добавить»* новый результат к предыдущему (в итоговый результат включаются как ранее найденные документы, так и документы, удовлетворяющие новому условию).

8.1.2. Конструктор запроса «по шагам»

Конструктор запросов «по шагам» реализует технологию последовательного построения предложения запроса.

Предложение может включать логически связанные условия поиска, относящиеся к разным полям документа. В этом режиме запрос формируется либо последовательным добавлением в конец текущего условия терминов из поискового словаря, уточняя, расширяя или ограничивая значение предыдущего термина или ранее подготовленной части условия в целом, либо последовательным добавлением в предложение запроса нового условия поиска.

Интерфейсные средства конструктора представлены формой «*Запрос*» и двумя диалоговыми окнами: «*Конструктор запроса для области поиска*» и «*Область поиска*».

Форма «*Запрос*» (рис. *.5) предназначена для просмотра словаря и включает область словаря, область предложения запроса и панель инструментов.

Форма «*Запрос*» представляет собой интерактивное средство, позволяющее:

- просматривать частотные словари базы данных;
- просматривать документы базы данных, содержащие выделенный в окне просмотра словаря термин;
- отбирать термины словаря в предложение запроса;
- редактировать термины перед добавлением в предложение запроса.

При подготовке поискового предложения в форме «*Запрос*» инициированием соответствующей закладки в верхней части формы выбирается область поиска (в данном случае это может быть только отдельное поисковое поле), после чего из словаря этого поля в предложение запроса включаются выбранные дескрипторы (термины словаря, которые могут быть предварительно отредактированы).

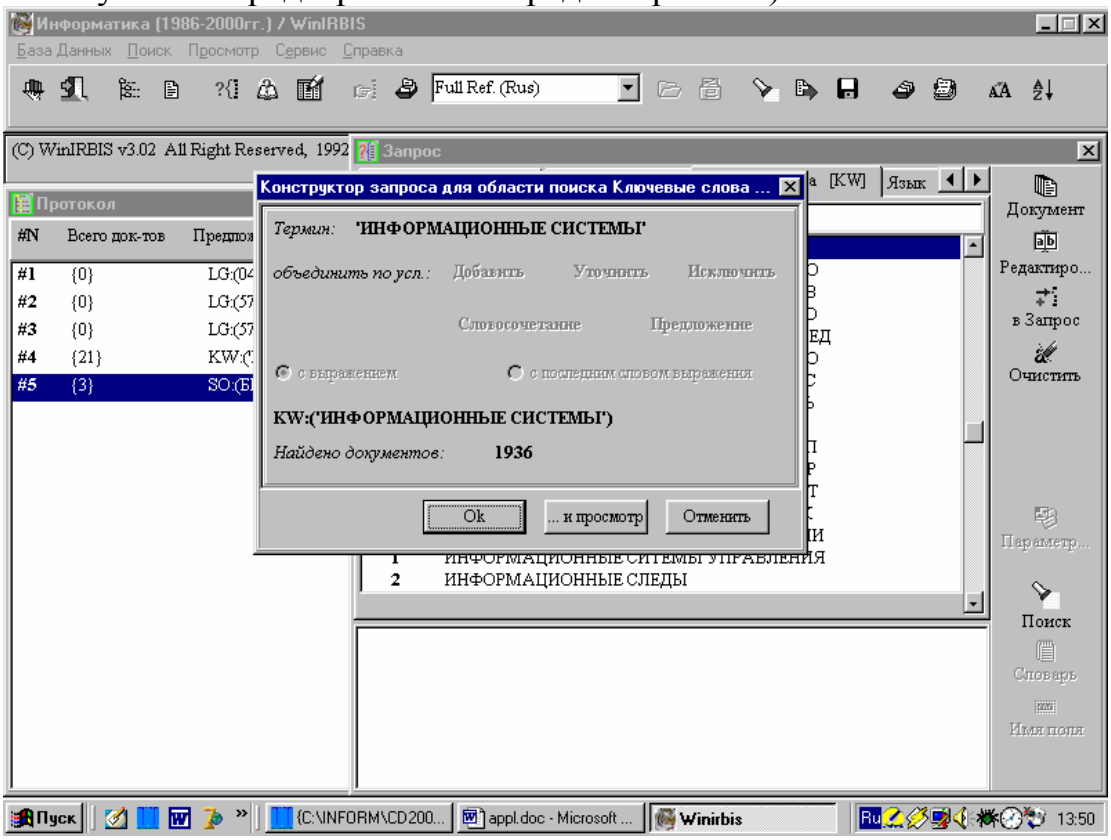


Рис. 8.3. Конструктор запроса «по шагам» - добавление термина

Каждый дескриптор включается в запрос путем связывания его с предыдущим термином логическим или контекстным оператором в соответствии с возможностями, определяемыми диалогом «*Конструктор запроса для области поиска*» (рис. 8.3).

Для включения в запрос дополнительного условия поиска инициируется диалог «*Область поиска*» для определения логической связи с ранее подготовленной частью предложения запроса (рис.8.4).

Например, в режиме Конструктора «по шагам» может быть сформировано следующее предложение запроса:

(ALL : ((ПРИВАТИЗАЦИЯ) и ТРАНСПОРТ)) и KD : (СТАТЬЯ)

Такой способ построения запроса характеризуют фиксированная расстановка скобок в логической формуле и, соответственно, зафиксированный порядок выполнения операций.

Семантика окна «*Конструктор запроса для области поиска*» следующая:

- *Добавить*: включение выбранного дескриптора в запрос при помощи оператора OR.

- *Уточнить*: включение выбранного дескриптора в запрос при помощи оператора AND.

- *Исключить*: включение выбранного дескриптора в запрос при помощи оператора NOT.

- *Словосочетание*: включение выбранного дескриптора в запрос при помощи оператора CTX.

- *Предложение*: включение выбранного дескриптора в запрос при помощи оператора SENT.

- *С выражением*: позволяет указать порядок связывания включаемого термина с ранее сформированным выражением, т.е. в скобки заключается все ранее сформированное выражение условия.

- *С последним словом выражения*: позволяет изменить приоритет выполнения операций, заключив в скобки последний и текущий дескрипторы выражения условия.

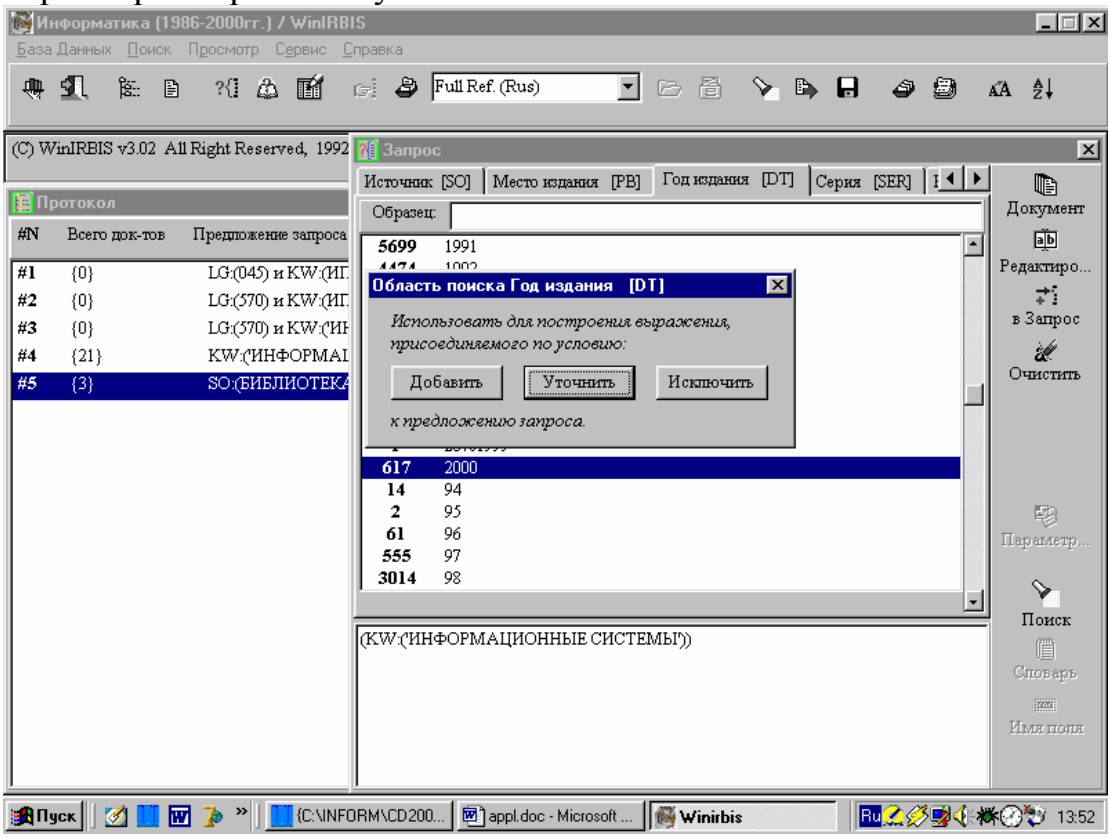


Рис. 8.4. Конструктор запроса «по шагам» - добавление условия поиска

8.1.3. Конструктор запроса «Логическое выражение»

Конструктор запроса «Логическое выражение» ориентирован на профессионального пользователя, знакомого с синтаксисом и семантикой булевоподобных выражений.

Конструктор представлен формами «*Запрос*» и «*Параметры поиска*».

Форма «*Запрос*» предназначена для просмотра словаря и редактирования поискового предложения запроса. Форма включает область словаря, область редактирования предложения запроса и панель инструментов и функционально отличается от формы «Запрос» конструктора «по шагам», во-первых, использованием механизма «выбора и вставки» имен полей и поисковых терминов по положению текстового курсора, и во-вторых – возможностью редактировать поисковое условие в области предложения запроса (рис. 8.5).

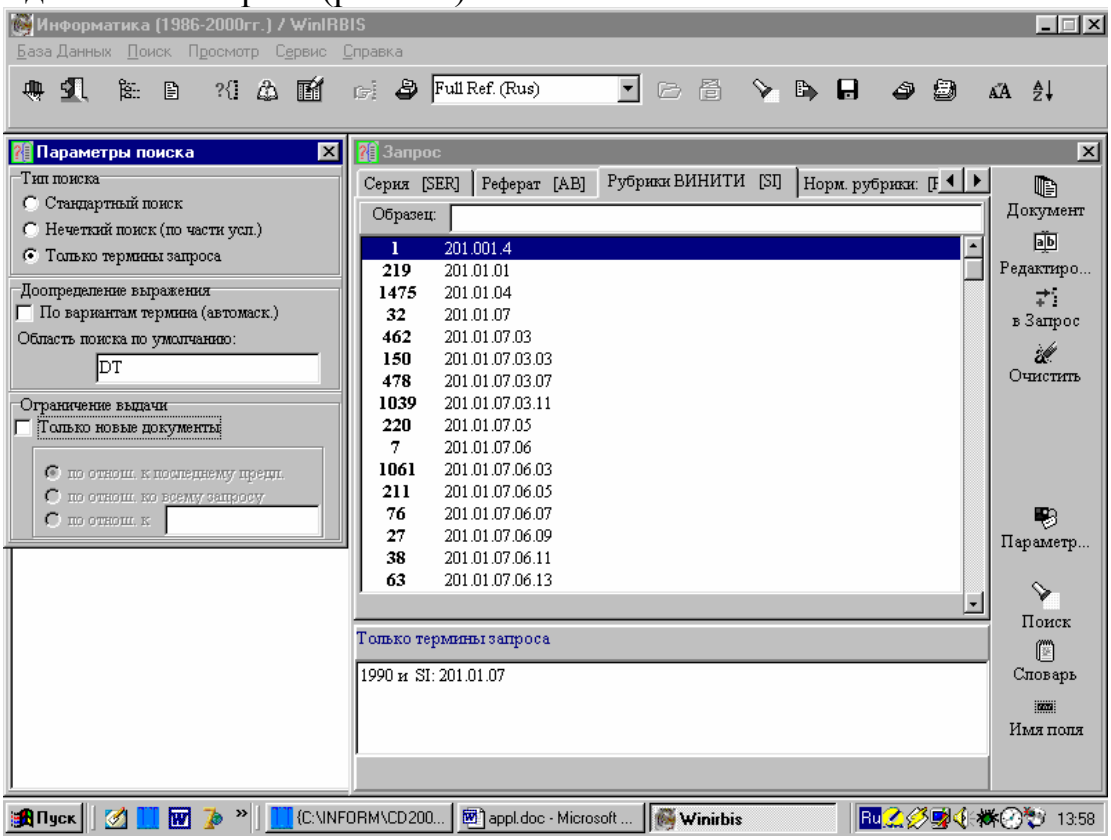


Рис. 8.5. Формирование логического выражения запроса

Кроме того, конструктор позволяет устанавливать дополнительные параметры поиска, используя форму «*Параметры поиска*».

Семантика формы «*Параметры поиска*» представлена списком параметров, управляющих процессом поиска в режиме логического выражения. Такими параметрами являются: «*Тип поиска*», «*Доопределение выражения*» и «*Ограничение выдачи*».

Параметр «*Тип поиска*» может принимать одно из трех значений:

- *Стандартный поиск* – означает, что проводится поиск по сформулированному в окне формирования запроса логическому выражению (реализуется модель механизма поиска по логическому выражению).
- *Нечеткий поиск* – означает, что скобки в предложении запроса игнорируются, а логические или контекстные операторы заменяются оператором OR (ИЛИ) (т.е. будут отобраны документы, содержащие хотя бы один из терминов предложения запроса в заданной области поиска). При этом множество документов результата делится на кластеры в зависимости от количества содержащихся в них терминов, соответствующих дескрипторам запроса, а результирующая выдача сортируется в порядке убывания номера кластера (реализуется модель механизма поиска по совпадению терминов).
- *Только термины запроса* – означает, что в результате поиска по булевоподобному выражению будут отобраны документы, содержащие в указанных полях только термины, заданные в предложении запроса.

Параметр «*Доопределение выражения*» позволяет:

- Установить (или отменить) режим поиска с автомаскированием, т.е. с применением при поиске встроенных процедур нормализации дескрипторов запроса. В главе 5 представлены правила применения нормализации при формулировке поискового запроса.
- Задать логическое выражение (или одно имя поля) для области поиска по умолчанию. Область поиска по умолчанию применяется в предложении запроса к терминам, для которых не задано выражение для области поиска.

Параметр «*Ограничение выдачи*» позволяет установить (или отменить) режим поиска, при котором по предложению запроса отбираются только новые документы, т.е. документы, которые не были включены либо ни в одну выдачу в рамках текущего запроса, либо в выдачу по последнему предложению запроса.

8.1.4. Использование формулировок ранее сохраненных запросов

Механизм сохранения/чтения запросов позволяет многократно использовать поисковые запросы. Запросы могут храниться либо в отдельных файлах (один из файлов с зафиксированным системным идентификатором выделен для общей Папки Запросов), либо в БД в структуре частотных словарей. В случае сохранения запроса в отдельном файле могут быть сохранены только тексты предложений запросов или предложения запросов вместе с поисковыми результатами.

Использовать ранее сохраненный запрос в дальнейшем можно как единое целое (весь набор поисковых предложений) или по отдельным предложениям.

Для работы с ранее сохраненными запросами служит интерфейсная форма «*Запрос для БД*» (рис. 8.6). Семантика формы следующая:

- *Выделить* - выделение одного и более предложений запроса для дальнейшего использования;
- *Редактировать* - редактирование выделенных предложений запроса;
- *Читать Запрос* – перенести выделенные предложения запроса вместе с их поисковыми результатами в текущий запрос;
- *Поиск по Запросу* – инициировать автоматическое выполнение поисковой процедуры для выделенных предложений запроса.

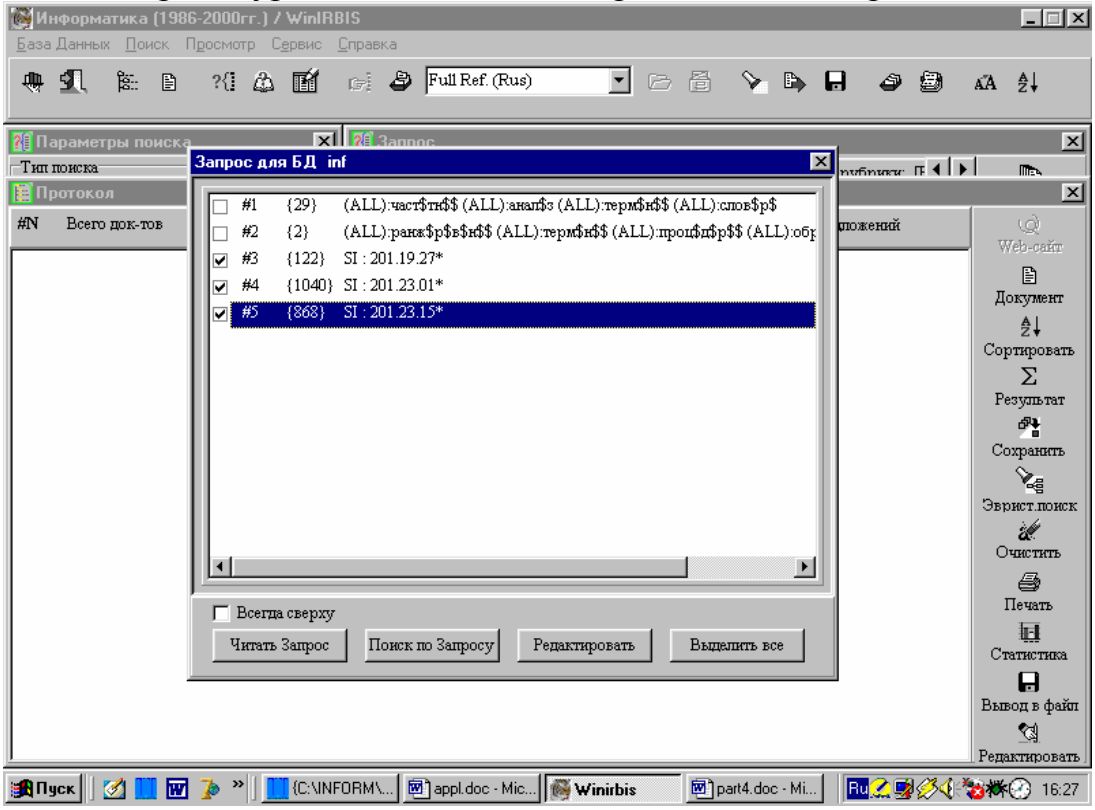


Рис. 8.7. Использование ранее сохраненных запросов

Общая «*Папка запросов*» – это средство, которое позволяет накапливать и в дальнейшем использовать поисковые (логические) выражения

запросов, независимо от того, для какой базы данных они были подготовлены. При этом в папку заносятся только тексты предложений запросов, а не результат - список идентификаторов документов, найденных по ним.

Использовать предложения запросов, помещенные в папку, можно по общей технологии работы с сохраненными запросами, за исключением возможности «Читать запрос» (рис. 8.8).

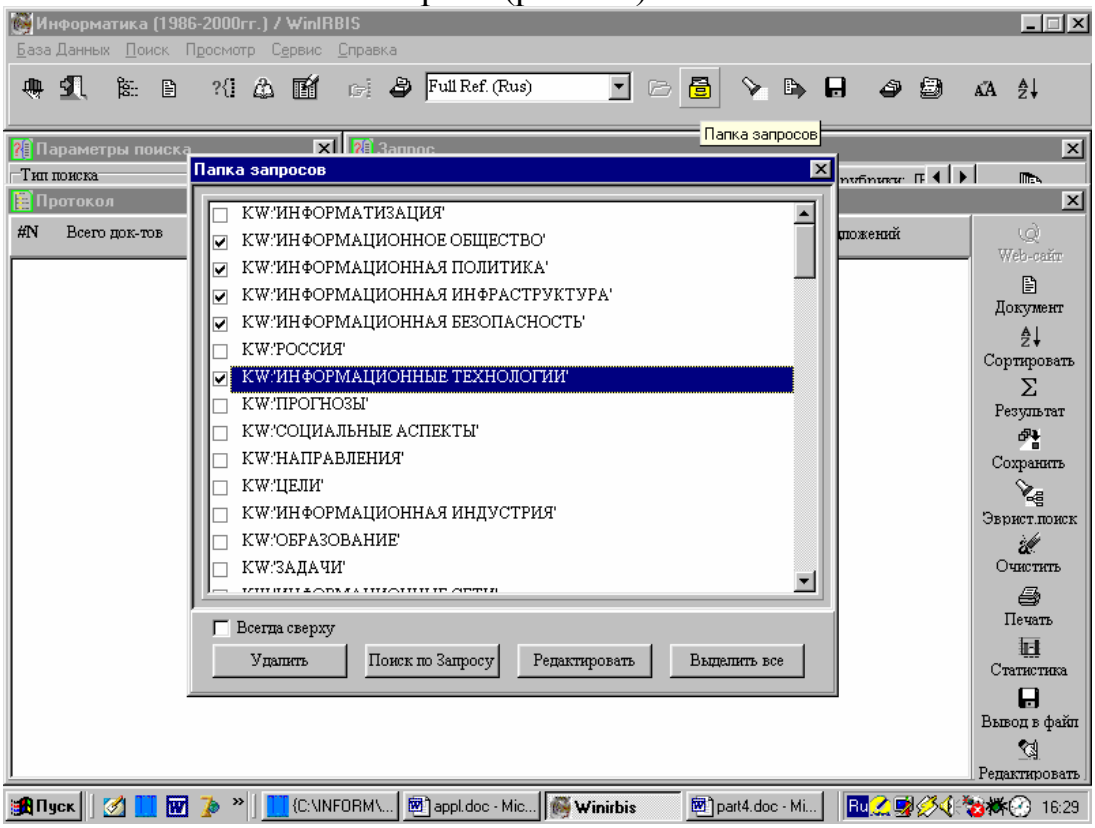


Рис. 8.8. Работа с папкой запросов

8.2. Средства и технологии поиска документов по сходству

Стратегия творческого поиска в больших массивах документов обычно нацелена на выявление нового знания или новых логических связей, порождаемых непосредственно в процессе просмотра, т. е. прямого взаимодействия пользователя с документами. Просмотровая функция рассматривается в данном контексте как эвристическая модель, ориентированная на изменение у пользователя существующих границ субъективного знания путем отыскания документов, не являющихся формально релевантными, а также путем динамического управления информационной средой.

Как было показано при анализе информационной модели воспроизводства знаний, требование «адекватного информирования» в автоматизированных информационно-поисковых системах может быть реализовано двойной реформулировкой запроса по технологии обратной связи по релевантности: построением на основе лексики истинно релевантных документов словников, термины которых в свою очередь размеча-

ются с точки зрения их способности смысловыражения темы и, далее, используются в качестве выражения запроса для следующей итерации поиска.

Существенно то, что здесь мы 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 выделить значимые термины (ранжированием или кластеризацией по статистическим показателям). Соответственно, для построения словников могут использоваться разные методы, что позволяет, в свою очередь, иметь разные «стратегии» реформулирования, реализуемые разными технологическими (интерфейсными) средствами. Возможность совместного использования нескольких стратегий поиска позволяет реализовать процесс итерационного повышения эффективности поиска путем генерации новых ПОЗов, учитывающих как «ситуационную» (проблемную) ориентацию запроса, так и «тематические» свойства массива документов.

Для класса документальных систем, где основными (базовыми) объектами являются документ и термин, может быть построено конечное множество механизмов поиска, основанных на технологии реформулирования запроса по обратной связи.

Технологическая схема поиска, иллюстрирующая использование различных механизмов поиска, приведена на рис. 8.14.

8.2.1. Поиск аналогов

Функция поиска аналогов позволяет осуществить поиск документов по сходству с содержимым заданных полей текущего документа, который в текущее время доступен в окне просмотра документов. Условие отбора задается в виде:

<Имя поля >:<число>

где <число > - количество терминов, совпадающих с терминами указанного поля. Допускается логическая комбинация условий (рис. 8.9).

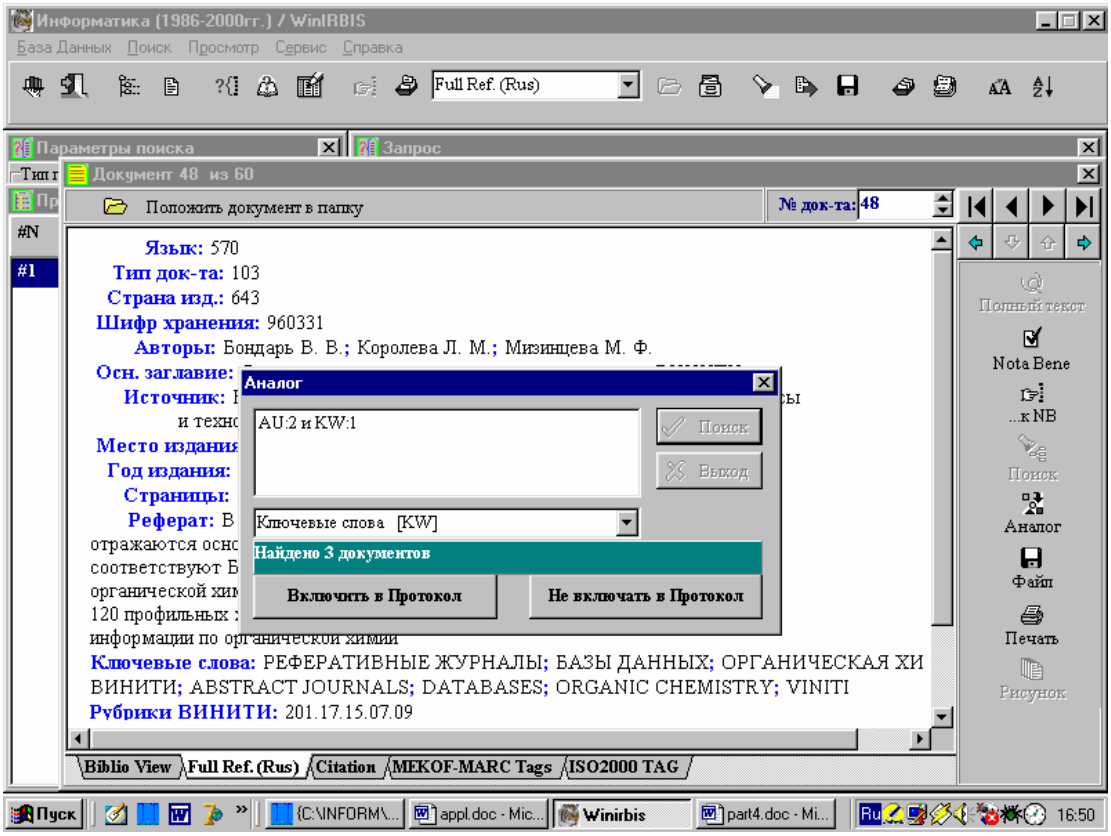


Рис. 8.9. Поиск документов-аналогов по логическому выражению

При поиске аналогов «по умолчанию» будут взяты поисковые поля, объявленные в схеме используемыми «по умолчанию», и пороговые значения, дающие не пустой результат.

8.2.2. Эвристический поиск

Эвристический поиск – это поиск документов по динамически формируемому критерию, использующему статистически значимые слова, выбираемые из текстов документов, отмеченных пользователем при просмотре как релевантные.

Найденные документы упорядочиваются в соответствии со значимостью, определяемой статистически в контексте релевантности документов - результатов поиска по текущему (выделенному в запросе) предложению. Количество документов в выдаче ограничивается соответствующим системным параметром.

Для проведения эвристического поиска необходимо при просмотре документов отметить как релевантные те документы, которые действительно соответствуют теме (а не только содержат термины поискового условия) – т.е. сформировать множество документов для эвристического поиска.

Эвристический поиск можно проводить по любому предложению запроса.

8.2.3. Поиск по обратной связи

Поиск по обратной связи реализуется, например, через последовательность взаимосвязанных этапов:

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

Особенностью реализации является возможность использования результатов, полученных на каждом этапе, для дальнейшего изменения стратегии поиска (так, например, термины, отобранные при просмотре словника релевантных документов, могут быть просто использованы для расширения запроса и проведения поиска по логическому выражению).

Рассмотрим порядок выполнения этапов на следующем примере.

Пусть исходное предложение запроса на поиск по логическому выражению в БД «Информатика» было следующим:

KW : ‘Поисковые стратегии’

При просмотре текстов документов-результатов поиска все из них были отмечены пользователем как релевантные. Тогда будет построен и выведен пользователю в ранжированном представлении следующий словник (для ранжирования использован интегральный коэффициент[62]):

2	**ВЕКТОРНЫЕ МОДЕЛИ**	**1,037520473**
5	ИНФОРМАЦИЯ О РАБОЧИХ МЕСТАХ	0,626009986
5	ПРЕДСТАВЛЕНИЕ ДОКУМЕНТОВ	0,602280990
7	РЕШЕНИЕ ПРОБЛЕМ	0,492237706
8	**АССОЦИАТИВНЫЙ ПОИСК**	**0,454205158**
23	РАЗРАБОТКА	0,440147604
12	**КЛАСТЕРИЗАЦИЯ**	**0,329185167**
15	**ПОИСК ИНФОРМАЦИИ**	**0,310164008**
18	ЗАВИСИМОСТЬ	0,256065252
24	**ПОИСКОВЫЕ СРЕДСТВА**	**0,232370300**
23	ПРОИЗВОДИТЕЛЬНОСТЬ	0,219695378
59	ОПРЕДЕЛЕНИЯ	0,120537948
59	ЭКСПЛУАТАЦИЯ	0,120537948
102	ОБЕСПЕЧЕНИЕ ДОСТУПА	0,092969592
88	СРЕДСТВА	0,092775247
90	ПОНЯТИЯ	0,091405883
255	АИПС	0,091326155
134	СВОДНЫЕ КАТАЛОГИ	0,077708893
169	ТИПЫ	0,059747985
241	ТЕНДЕНЦИИ РАЗВИТИЯ	0,052260958
261	ИНТЕРАКТИВНЫЙ ДОСТУП	0,049447866
231	РЕЗУЛЬТАТЫ	0,048036882
285	ИНФОРМАЦИОННЫЕ ЗАПРОСЫ	0,046496765

[62] Голицына О.Л. Моделирование и разработка средств и технологий поиска документальной информации - диссертация на соискание ученой степени кандидата технических наук. – М.: РГГУ, 2004

282	АИС	0,041655815
371	ИНТЕРАКТИВНЫЕ КАТАЛОГИ 0,	038532128
1973	ИНФОРМАЦИОННЫЙ ПОИСК	0,034875966
388	ИПС	0,032947558
423	ЭКСПЕРИМЕНТЫ	0,030870876
1309	ПОЛЬЗОВАТЕЛИ	0,026564388
591	МОДЕЛИ	0,023797976
805	ИНТЕРФЕЙСЫ	0,018443519
960	АНАЛИЗ	0,015820678
1252	ИНФОРМАЦИОННЫЕ СЕТИ	0,014472937
1241	ИНФОРМАТИКА	0,012474870
4886	ОБЗОРЫ	0,006685146
4153	БАЗЫ ДАННЫХ	0,002837892

В первой колонке словника – частота термина в БД, во второй колонке – сам термин, а в третьей колонке – значение весового коэффициента.

Далее пользователь отмечает некоторые термины словника (в списке они выделены курсивом) и либо добавляет их в исходный запрос самостоятельно, либо запускает процедуру построения информационного пространства и кластеризации. После этого пользователь получает доступ к просмотру через протокол ненулевых результатов следующих предложений запроса (идентифицируемых отдельными терминами или сочетаниями терминов, выделенных в словнике курсивом):

KW : (‘ВЕКТОРНЫЕ МОДЕЛИ’ and ‘КЛАСТЕРИЗАЦИЯ’)	{1}
KW : (‘АССОЦИАТИВНЫЙ ПОИСК’ and ‘ПОИСКОВЫЕ СРЕДСТВА’)	{1}
KW : (‘ВЕКТОРНЫЕ МОДЕЛИ’)	{2}
KW : (‘АССОЦИАТИВНЫЙ ПОИСК’)	{8}
KW : (‘КЛАСТЕРИЗАЦИЯ’)	{12}
KW : (‘ПОИСК ИНФОРМАЦИИ’)	{15}
KW : (‘ПОИСКОВЫЕ СРЕДСТВА’)	{24}

После просмотра документов отдельных предложений запроса и выделения нового множества релевантных документов процедуры построения словника и кластеризации могут быть заново инициированы, и так до тех пор, пока полученный совокупный результат (а все полученные пользователем результаты хранятся системой и в любой момент могут быть объединены в общее множество) не удовлетворит пользователя.

8.3. Технологические объекты построения предложения запроса

Для задания дескрипторов предложения запроса могут использоваться следующие технологические объекты:

- частотный словарь;
- тематический рубрикатор;
- тезаурус;
- иерархический словник.

Как показывают многочисленные исследования, наибольший эффект достигается при совместном использовании словарных и рубрикационно-классификационных систем.

Функциональные (интерфейсные) решения, обеспечивающие гибкое использование внешних по отношению к БД объектов, представлены в ИПС IRBIS в виде отдельной функционально-интерфейсной формы, позволяющей унифицировать отображение иерархических словарных структур и имеющей средства построения поисковых запросов (рис. 8.10).

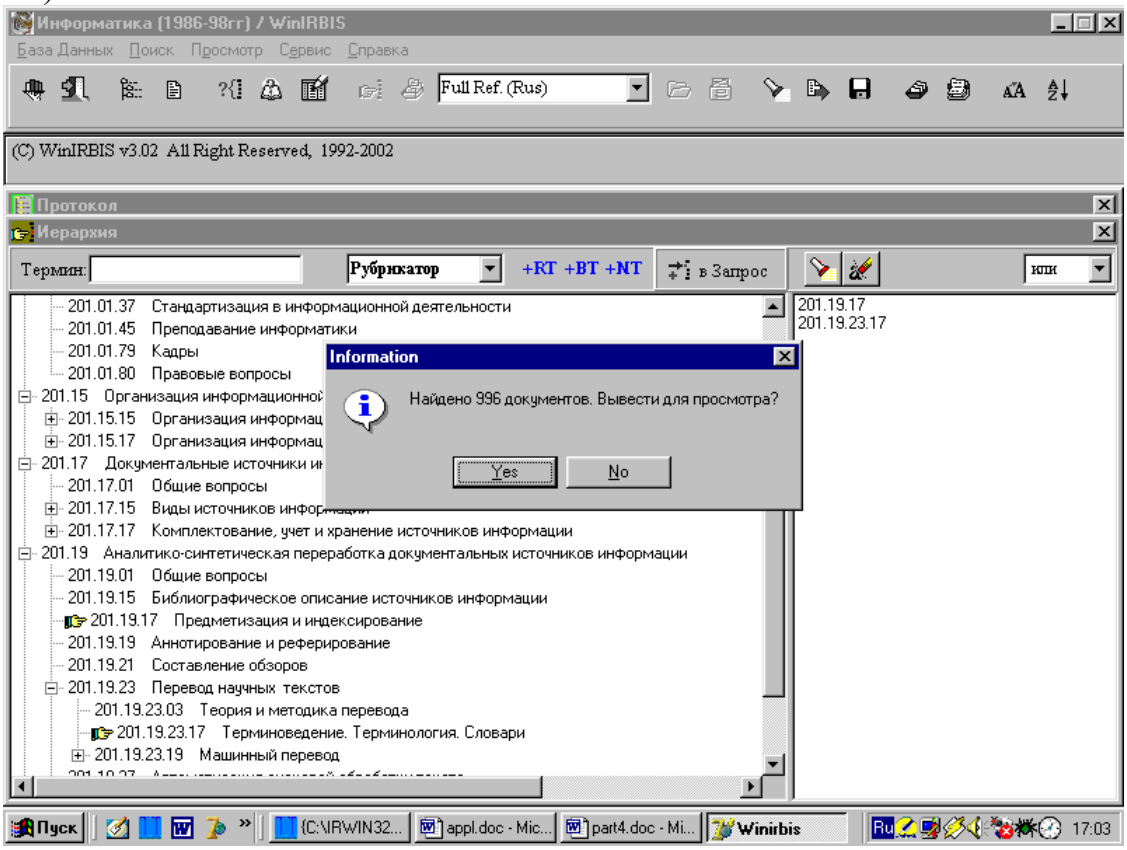


Рис. 8.10. Отображение и использование тематического рубрикатора

8.3.1. Частотный словарь

Словарь включает термины (слова и словосочетания) из документов базы данных с частотой их встречаемости.

Использование словаря как средства построения предложения запроса особенно эффективно для начинающих пользователей (для ознакомления с терминологией базы данных и контроля правильности написания поисковых терминов). Словарь доступен для просмотра (и редактирования термина перед включением его в предложение запроса) при переходе к одному из основных режимов формирования запроса (см. рис.8.1, 8.5).

Наличие при термине такого параметра, как частота встречаемости позволяет оценить размер предполагаемой выдачи, если данный термин будет включаться в предложение запроса.

8.3.2. Тематический рубрикатор

Тематический рубрикатор представляет собой иерархический тематический указатель, подготовленный в специальном формате для одной или нескольких баз данных и проведения тематического поиска. Поиск можно проводить как по отдельной рубрике, так и набору рубрик.

Физически тематический рубрикатор представляет собой текстовый файл операционной системы, подготовленный с использованием символов табуляции для идентификации уровней иерархии. Такой рубрикатор может быть создан пользователем самостоятельно (см. рис. 8.10).

8.3.3. Тезаурус

Напомним, что тезаурус представляет собой контролируемый иерархический словарь терминов, называемых дескрипторами. Значение тезауруса, как одного из главных компонентов документальной информационно-поисковой системы, весьма велико. Во-первых, тезаурусы используются для индексирования и поиска информации, улучшая качество поиска, во-вторых - для снятия неоднозначности и задания различных отношений между терминами в естественном языке. Тезаурус также может быть использован как одно из основных средств организации диалога между пользователем и информационным ресурсом в ИПС.

В ИПС IRBIS Тезаурус реализован в виде двухуровневого дерева, на верхнем уровне которого находятся дескрипторы, а на нижнем – их связи. Пример фрагмента Тезауруса по информатике представлен ниже:

ВЫДАЧА ЛИТЕРАТУРЫ
    SN выдача книг
    СМ выдача документов
    ВТ БИБЛИОТЕЧНОЕ ОБСЛУЖИВАНИЕ

RT АБОНЕМЕНТ
RT БИБЛИОТЕКИ
RT КНИГОВЫДАЧА
ВЫСОКАЯ ПЕЧАТЬ
BT ОПЕРАТИВНАЯ ПОЛИГРАФИЯ
BT ПЕЧАТЬ
ВЫСТАВКИ
SN экспозиции
NT ВДНХ СССР
NT МЕЖДУНАРОДНЫЕ ВЫСТАВКИ
NT ТЕМАТИЧЕСКИЕ ВЫСТАВКИ
RT НАУЧНО-ТЕХНИЧЕСКАЯ ПРОПАГАНДА
ВЫСШЕЕ ОБРАЗОВАНИЕ
BT НАРОДНОЕ ОБРАЗОВАНИЕ
RT ВУЗЫ
ВЫХОДНЫЕ ДАННЫЕ
BT БИБЛИОГРАФИЧЕСКИЕ ОПИСАНИЯ

Для повышения эффективности поиска информации используются связи дескрипторов Тезауруса, позволяющие автоматически включать, например, вышестоящие, нижестоящие, ассоциативные дескрипторы вместе с основным дескриптором в запрос рис. 8.11).

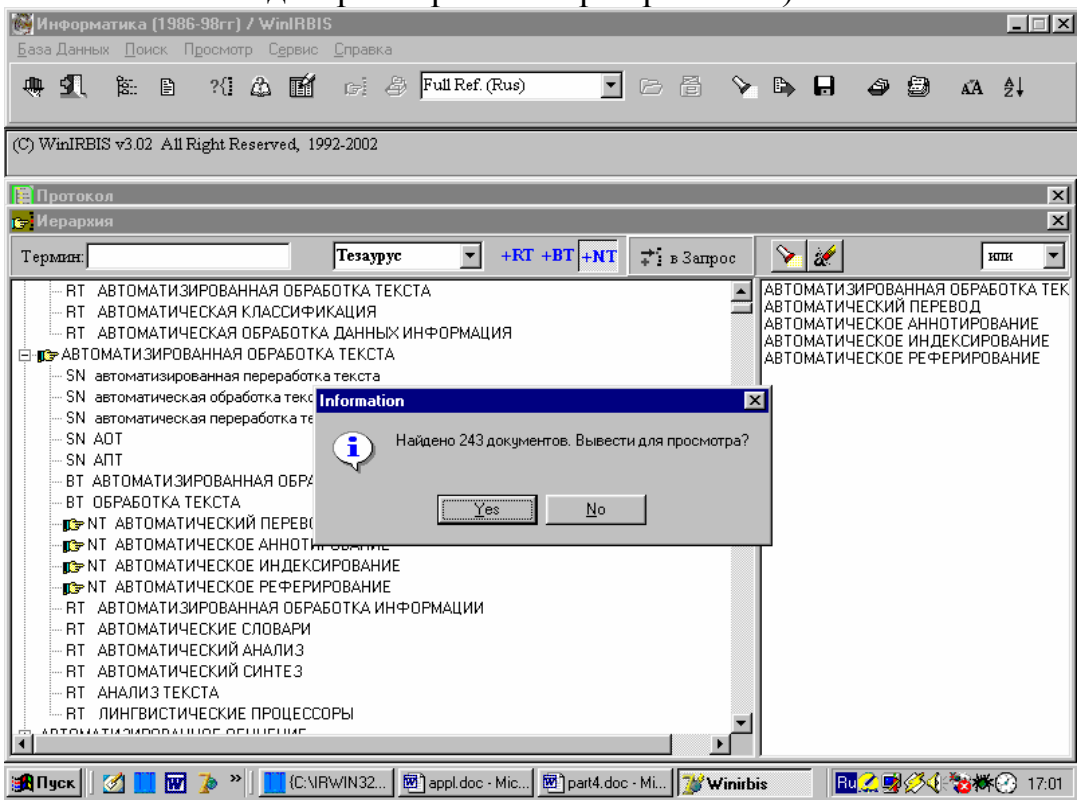


Рис. 8.11. Отображение и использование тезауруса

8.3.4. Иерархический словник

Иерархические словники служат дополнительными терминологическими структурами на этапе формирования предложения запроса и могут быть использованы в технологиях поиска по образцу и поиска по логическому выражению.

Построенная для некоторой тематической области словарная иерархия может быть отображена в интерфейсном окне, поддерживающем возможности отбора терминов и включения их в текущее предложение запроса.

Такого рода словарные структуры обеспечивают альтернативный частотному словарю вход в терминологию информационного массива, расширяя для пользователя среду формирования ПОЗа, но не выводя его при этом за пределы лексики предметной области.

Так, например, окружение термина «ИПТ» (информационно-поисковые тезаурусы) выглядит по-разному в частотном словаре и в иерархическом словнике. Для частотного словаря, политематического и упорядоченного по лексикографическим признакам, имеем:

262 ИПТ
1 ИПТ МАШИННЫЕ
138 ИПЯ
2 ИРАК
1 ИРАКСКАЯ АГРЕССИЯ
5 ИРАН
1 ИРАНСКИЙ УНИВЕРСИТЕТ
9 ИРБИС
1 ИРВИНА СЕМЕЙСТВО
96 ИРИ

Для иерархического словника, созданного по лексике тематической рубрики 201.23.15 «Информационно-поисковые языки» (включая подуровни), фрагмент, начинающийся термином «ИПТ», выглядит следующим образом:

ИПТ
    ДВУЯЗЫЧНЫЕ ИПТ
    ИПТ МАШИННЫЕ
    МНОГОЯЗЫЧНЫЕ ИПТ
    ОДНОЯЗЫЧНЫЕ ИПТ
    ОТРАСЛЕВЫЕ ИПТ
ИПЯ
    ДЕСКРИПТОРНЫЕ ИПЯ
    МЕЖДУНАРОДНЫЕ ИПЯ
    НАЦИОНАЛЬНЫЕ ИПЯ
    НЕКОНТРОЛИРУЕМЫЕ ИПЯ
    ОСОБЕННОСТИ ИПЯ

ПРЕДМЕТИЗАЦИОННЫЕ ИПЯ
ФАКТОГРАФИЧЕСКИЕ ИПЯ
ИСПАНСКИЙ ВАРИАНТ
ИСПАНСКИЙ ЯЗЫК

Если в частотном словаре пользователь может ориентироваться лишь по частоте употребления термина, то в иерархическом словнике рядом с термином представлены все его лексические расширения, что, с одной стороны, более удобно для пользователя, не вполне знакомого с лексикой темы, а с другой – позволяет сразу определиться в пространстве терминов для формулировки более точного поискового запроса (рис. 8.12).

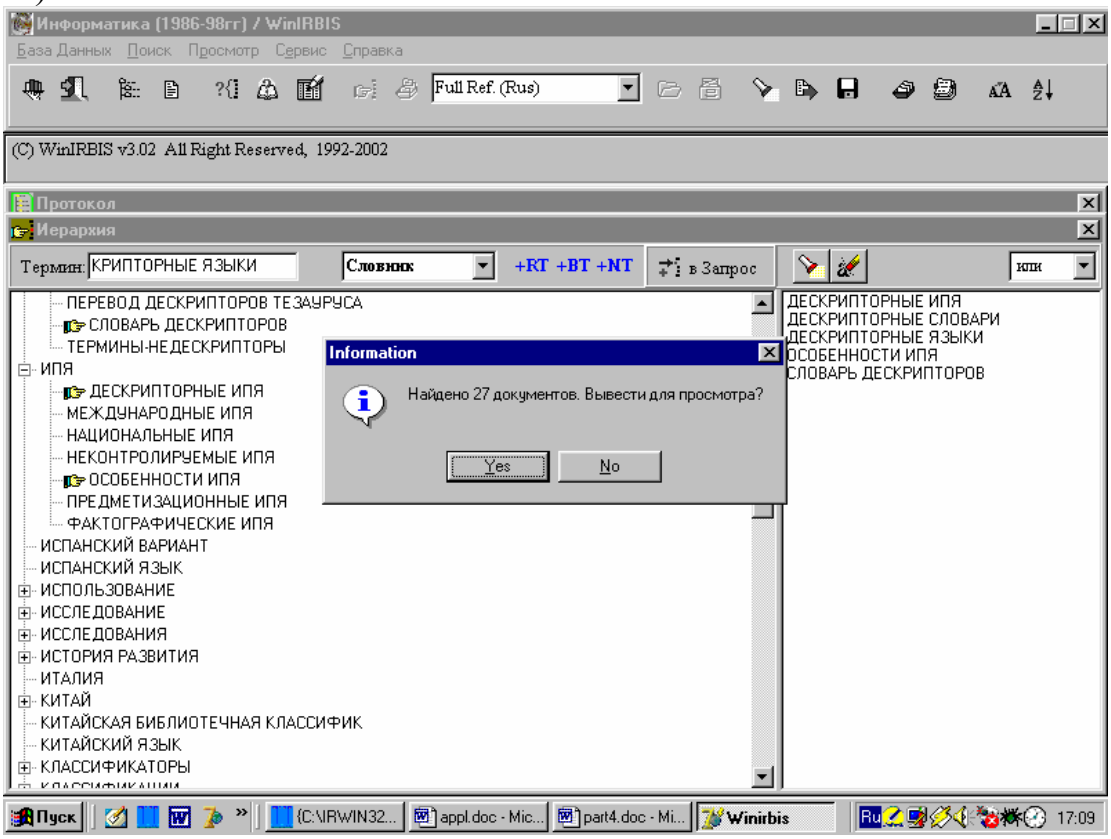


Рис. 8.12. Отображение и использование иерархического словника (мини-тезауруса)

Использование таких словников выглядит в ряде случаев предпочтительнее использования тезаурусов, т.к. словники предполагается строить «на лету» по лексике тематической подборки актуального информационного массива, в то время как тезаурусы не всегда с терминологической точки зрения успевают отразить текущее состояние лексики быстро развивающейся предметной области. Так, например, Тезаурус по информатике (как проиллюстрировано в главе 5) уже не может считаться актуальным лингвистическим средством ввиду того, что довольно низкий процент его терминов в настоящее время используется при индексировании документов.

Для построения словников тематической области предусмотрена специальная функция формирования статистики по результату произвольного предложения поискового запроса. Статистика представляет собой словник терминов тематической выборки с частотами употребления терминов в ней. Выборка определяется как множество документов, полученных при поиске по предложению запроса. Словник в свою очередь может быть построен по любому полю документа.

Такой словник может быть, во-первых, исходным средством для построения иерархических словарных структур, а во-вторых – результатом количественного анализа распределения документов тематической выборки.

При проведении статистического анализа выборки можно дополнительно установить требуемый минимальный порог для частоты встречаемости термина и порядок сортировки («по алфавиту» или «по частоте»).

Результаты статистического анализа отображаются в интерфейсной форме и могут быть сохранены в файле для дальнейшего использования (рис. 8.13).

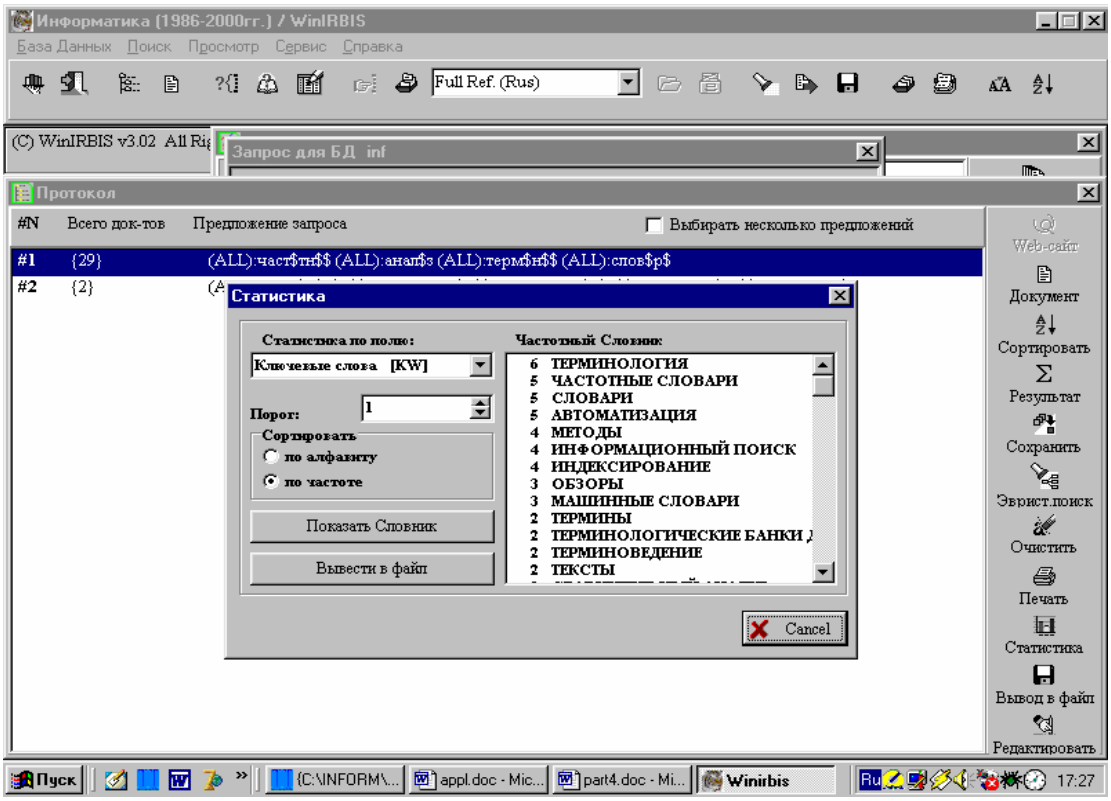


Рис. 8.13. Построения тематического словника по результату предложения запроса

8.4. Обобщенная характеристика развития поискового процесса

С точки зрения взаимодействия «пользователь - система» процесс информационного поиска, в общем случае включает следующие этапы (см. рис. 6.4):

1. определение темы запроса, ее локализация в предметной области и формализация на уровне понятий основной и смежных областей, а также идентификация ресурса.;
2. формирование, а также структурное и лексическое адаптирование выражения запроса;
3. отбор документов с помощью одного из механизмов поиска по критерию, адекватному степени неопределенности информационной потребности;
4. формирование и управление выдачей найденных документов;
5. оценку результата поиска на уровне отдельного документа, где система обеспечивает возможность фиксировать значение степени соответствия запросу пользователя и непосредственное использование лексики документов для непосредственной модификации выражения запроса;
6. итоговую оценку результатов поиска на уровне всего запроса или отдельных предложений с точки зрения принятия решения о завершении поискового процесса (например, исчерпывающее удовлетворение информационной потребности, или несоответствие цели поиска);
7. развитие запроса по технологии реформулирования по обратной связи по релевантности или использование каких-либо других ресурсов, например, ассоциированных баз данных вторичной или справочной информации.

Развитие запроса путем реформулирования запроса на основе лексики документов, релевантность которых подтверждена пользователем, реализуется двумя типами поисковых технологий.

1. Поиск по терминам отдельного документа, который реализуется либо на основе комбинативной схемы (поиск документов-аналогов, содержащих указанное количество терминов текущего – просматриваемого пользователем документа), либо «жестким» отбором - по терминам, указанным пользователем в текущем документе.
2. Поиск по терминам, статистически взвешенным на множестве документов, истинная релевантность которых была подтверждена пользователем. Статистически значимые термины, включенные системой в словник из текстов релевантных документов, ранжируются в соответствии с их весом. Далее - в случае эвристического поиска - автоматически кластеризуется и формируется документальная выдача, либо, в случае контекстного поиска, в сформированном словнике пользователь отмечает как релевантные наиболее информативные термины, которые система использует для генерации кластеров терминов, каждому из которых ста-

вится в соответствие кластер документов, представляемый в запросе как отдельный результат.

Таким образом, в процессе развития запроса используются два типа обратной связи: внешняя, отражающая оценку пользователя, и внутренняя, позволяющая выделять и ранжировать значимые поисковые термины.

Схема поиска, представленная на рис 7.6, отражает следующие требования к интерфейсным компонентам и организации процесса поиска в целом:

- подготовка следующего шага поиска выполняется непосредственно при обработке результата предыдущего: для развития поиска в качестве основного интерфейсного объекта в первую очередь используются документы;
- операционные объекты однородны на каждом шаге;
- на каждом шаге возможен возврат к ранее полученным результатам или оперативное переключение на другую тему и операцию;
- оценка степени завершенности (сходимости) процесса поиска возможна по критерию исчерпания как лексики, так и документального пространства темы.

Фактически классическая схема выдачи документов «по запросу-выражению» расширена до динамически управляемого процесса кластеризации пространства документов и терминов. Процесс поиска может развиваться по принципу «расходящихся кругов», обеспечивая выявление «центров активации» искомого образа в семантической сети базы данных, т.е. построение множеств или цепочек документов, которые в свою очередь могут служить мостом к понятиям (документам), возможно, уже не содержащим терминов исходного запроса. При этом, для случая «проблемного» поиска, когда доказательство полноты не может быть сведено к формально вычисляемым показателям, оно компенсируется подтверждаемостью – получением результата другим путем, например, вхождением в информационное пространство БД через информационные объекты разной природы и/или использованием поисковых механизмов разного типа.

Процесс поиска построен симметрично и реализует двойственную задачу: при подготовке запроса можно формировать коллекцию документов (непосредственным включением документов, к которым можно обращаться через «прямые» входы, такие как, словари, рубрикаторы, указатели и т.д.), а при формировании поисковой выдачи – реформулировать запрос и формировать компоненты лингвистического обеспечения (словники, рубрикаторы и т.д.).

Пользователь может продвигаться по пути (реализовать навигацию), предлагаемому системой, или же изменять его, в том числе и вы-

бирая из сформированных системой альтернатив, либо инициируя новый путь через процедуру поиска или прямого отбора.

При этом, доказательство полноты, которое не может быть сведено к формально вычисляемым показателям, компенсируется подтверждаемостью – получением результата другим путем, например, вхождением в информационное пространство БД через информационные объекты разной природы и/или использованием поисковых механизмов разного типа.

Такие подходы в сочетании со средствами идентификации и избирательной обработки объектов не заставляет пользователя постоянно заботиться об отложенных действиях при оперативной работе с запросом и результатами.

Таким образом, процесс поиска – это итеративная последовательность создания и использования технологических объектов, реализующая *целенаправленное* и *управляемое* перемещение (навигацию) в документальном и лексическом пространстве базы данных и обеспечивающая оцениваемый уровень удовлетворения информационной потребности или объективно подтверждающая отсутствие информации.

Целенаправленность здесь предполагает представление некоторой цели (обычно в сфере основной деятельности), как комплекса информационных целей, имеющих тематический (многоаспектный поиск) и/или технологический характер. Это означает, что для того, чтобы обеспечить целенаправленность избирательного обращения к поисковым объектам, интерфейсные средства системы должны позволять фиксировать и идентифицировать цель в виде технологических объектов, используемых в процессе ее достижения.

Управляемость - это, с одной стороны, возможность выбора средств и/или параметров работы, а с другой – обоснованное обращение к технологическим объектам, в том числе повторное обращение к результатам и их обработка, например, статистическая или структурно-форматная. Кроме того, поскольку выделенная последовательность результатов (физически соответствующая отдельным шагам, а логически – отдельной цели) образует ряд, то это позволяет использовать статистически вычисляемые показатели разностного типа, характеризующие сходимость процесса поиска и, в частности, обеспечивающие обоснованность решения об окончании процесса совершенствования запроса.

Контрольные вопросы

1. Перечислите основные технологические (интерфейсные) объекты при поиске информации.
2. Охарактеризуйте основные интерфейсные средства подготовки и модификации поисковых запросов
3. Охарактеризуйте основные интерфейсные средства развития поисковых запросов
4. Приведите типологию сценариев формирования выражения поискового запроса на ИПЯ.
5. Охарактеризуйте сценарии типа «укажи и выбери».
6. Охарактеризуйте сценарии типа «укажи и получи».
7. Охарактеризуйте интерфейсные средства использования тезаурусных связей при модификации поисковых запросов.
8. Охарактеризуйте интерфейсные средства использования терминологических структур при подготовке и модификации запросов.
9. Дайте обобщенную характеристику поискового процесса.

Список сокращений

АИПС – автоматизированная информационно-поисковая система.
АИС – автоматизированная информационная система.
БД – база данных.
ЕЯ – естественный язык.
ИД – информационная деятельность.
ИП – информационная потребность.
ИПП – информационная потребность пользователя.
ИПС – информационно-поисковая система.
ИПТ – информационно-поисковый тезаурус.
ИР – информационный ресурс.
ИС – информационная система.
ИПЯ – информационно-поисковый язык.
ЛО – лингвистическое обеспечение.
ОД – основная деятельность.
ПО – поисковый образ.
ПОЗ – поисковый образ запроса.
ПОД – поисковый образ документа.
ПОТ – поисковый образ темы.
ПрО – предметная область.
ТСС – тематико-статистический спектр множества документов.

Литература

1. Абдеев Р.Ф. Философия информационной цивилизации. - М.: Владос, 1994.
2. Аветисян Д.О., Аветисян Р.Д. Теоретические основы информатики. – М: РГГУ, 1999.
3. Аветисян Д.О. Проблемы информационного поиска. – М.: Финансы и статистика, 1981.
4. Афанасьев В.Г. Социальная информация и управление обществом. – М.: Политиздат, 1975.
5. Базмаджян Р.А. и др. Универсальная система «Кристалл». – Ереван: АрмНИИНТИ, 1969.
6. Белнап Н., Стил Т. Логика вопросов и ответов. – М.: Прогресс, 1981.
7. Белоозеров В.Н., Косарская Ю.П. Опыт разработки словаря с разветвленной системой тезаурусных связей // НТИ. Сер. 2, 2001. - N 8. - С. 28-31.
8. Большая Советская Энциклопедия. – М.: Советская энциклопедия, 1980.
9. Браславский П. И., Гольдштейн С. Л., Ткаченко Т. Я.. Тезаурус как средство описания систем знаний. //НТИ, Сер. 2, №11, 1997 г.
10. Брежнева В. В.; Минкина А. В. Современные информационные продукты и услуги: попытка классификации и анализа тенденций развития. Инф. ресурсы России Номер 6, 1995, С. 26-29.
11. Бриллюэн Л. Научная неопределенность и информация. - М., 1966.
12. Винер Н. Кибернетика и общество. – М.:ИЛ, 1958.
13. Вирт Н. Алгоритмы и структуры данных: пер. с англ. – М.: Мир, 1989.
14. Воробьев Г.Г. Проблема документальной информации / сб. Кибернетика и документалистика. Механизмы процесса накопления, хранения и поиска научной информации. - М.: Наука. –1966., с.5-34.
15. Голицына О.Л., Максимов Н.В. Человеко-машинный информационный поиск в документальных базах данных // Теория и практика общественно-научной информации. - Вып.12. - М.: ИНИОН РАН, 1996.
16. Горский Д. П.. Формальная логика и язык. В кн.: «Философские вопросы современной формальной логики». М., Изд-во АН СССР, 1962, стр. 61.
17. Григорьев В.А. Стереотипы и фантазия в интеллектуальных системах. // НТИ, Сер.2 -1999, №7, стр.10-27.
18. Громов Г.Р. Национальные информационные ресурсы: проблемы промышленной эксплуатации. – М.: Наука, 1985, 240с.
19. Гультяев А.К., Машин В.А. Проектирование и дизайн пользовательского интерфейса. – СПб.: Коронапринт, 2000. –352с.
20. Дернер Д. Логика неудач. М.: Смысл, 1997.

21. Димитрова К. Стратегии информационного поиска // Библиотека, 1995, 3, №1, С.16-22.

22. Дорожкин А.М. Научный поиск как постановка и решение проблем. Н.Новгород: Нижегородский гуманитарный центр, 1995.

23. Дружинин В.В., Конторов Д.С. Проблемы системологии. - М.: "Советское радио", 1976.

24. Захаров В.П. Языковые средства современных информационно-поисковых систем. Автореферат дисс. на соискание ученой степени кандидата филологических наук по спец. 10.02.21. – СПб.: СПбГУ, 1997.

25. Зинченко В.П. От классической к органической психологии // Вопр. психол. 1996, №6, стр. 6-25.

26. Информатика. Базовый курс. Учебник для ВУЗов / Симонович С.В. и др. – СПб.: «Издательство Питер», 1999.

27. Информатика. Энциклопедический словарь для начинающих. / Под ред. Д.А. Поспелова -М.: Педагогика-Пресс, 1994.

28. Информационно-библиотечная деятельность, библиография. Термины и определения. / Межгосударственный стандарт ГОСТ 7.0-99 / Система стандартов по информации, библиотечному и издательскому делу. – Минск, 2000.

29. Информационно-поисковый тезаурус по информатике. / Сост. Пащенко Н.А., Ксенофонтова Е.Б.. Скоробогатая В.Ф., научный редактор Черный А.И. - М.:ВИНИТИ, 1987.

30. Карначук В.И. Классификация информационно-поисковых стратегий. – Новосибирск, 1986.

31. Кацнельсон С. Д. Содержание слова, значение и обозначение. М.—Л., «Наука», 1965, стр. 6.

32. Козачук М.В. Концептуальный анализ текстов в системах автоматической обработки научно-технической информации. Дисс. на соискание ученой степени кандидата техн. наук по спец. 05.25.05. – М.: ВИНИТИ, 2002.

33. Колин К.К. Информационные проблемы социально-экономического развития общества // Проблемы социальной информатики. Вып.1. - М. 1995.

34. Котик М.А., Емельянов А.М. Природа ошибок человека-оператора. - М.: Транспорт, 1993.

35. Коутс Р., Влейминк И. Интерфейс «человек-компьютер». – М.:Мир, 1990. - 501с.

36. Криницкий Н.А., Миронов Г.Д., Фролов Г.Д. Автоматизированные информационные системы /Под ред. Дородницына А.А. - М.: Наука, 1982. –384 с.

37. Кулик А. Н. Информационные сети и языковая совместимость дескрипторных информационно-поисковых систем. М.: Сов. Радио, 1977 г.

38. Ломов Б.Ф. Когнитивные процессы как процессы психологического отражения. // Когнитивная психология. Материалы финско-советского симпозиума. – М.:Наука, 1986. с.7-21.

39. Мазур М. Качественная теория информация. - М.: Мир, 1974, 239с.

40. Максимов Н.В. Компоненты и технологии интерактивного поиска документальной информации. // МФД. – 2001, №3.

41. Мартин Дж. Организация баз данных в вычислительных системах. –М.: Мир, 1980.

42. Михайлов А.М., Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968, с. 755.

43. Моисеев Н.И. Предисловие к сб. «Число и мысль». – М.: Знание, 1977.

44. Муранивский Т.В. Методы обработки документов на основе использования свойств и закономерностей информации. – М.: МГИАИ, 1984.

45. Муранивский Т.В. Теоретические основы научно-технической информации. – М.: МГИАИ, 1982.

46. Найссер У. Познание и реальность. - М., 1981.

47. Озкарахан Э. Машины баз данных и управление базами данных. / Пер. с англ. - М.: Мир, 1989. – С. 539-593.

48. Полонский В. М. Методика построения классификации исследований в общественных науках М.:ВИНИТИ, НТИ. Сер. 1, Номер 5,1989, С. 11-13.

49. Попов И. И., Романенко А. Г. Некоторые вопросы оптимизации комплектования информационных фондов. // Вопросы моделирования и оптимизации информационных систем, Вып. 4 - М.: Информэлектро, 1973.

50. Попов И.И. Информационные ресурсы и системы: реализация, моделирование, управление. – М.: ТПК АЛЬЯНС, 1996, 408с.

51. Попов Э.В. Экспертные системы: решение неформализованных задач в диалоге с ЭВМ. – М.: Наука, 1987, 288с.

52. Поспелов Г.С., Ириков В. А. Программно-целевое планирование и управление. - М.: Советское радио, 1975.

53. Разумовский О.С. От конкурирования к альтернативам. Экстремальные принципы и проблема единства научного знания. – Новосибирск: Наука, 1983.

54. Рассолов М.М., Элькин В.Д., Рассолов И.М. Правовая информатика и управление в сфере предпринимательства. - М., - 1998.

55. Ратцева И.И. Проблема выбора значения слова и смысловые расстояния. НТИ, 1966, №5.

56. Саати Т. Принятие решений. Метод анализа иерархии / пер. с англ. – М.: Радио и связь, 1989. – 316с.

57. Скороходько Э.Ф. Лингвистические проблемы обработки текстов в автоматизированных информационно-поисковых системах. // Вопросы информационной теории и практики. Сб.№25, - М.: ВИНИТИ. 1974.

58. Словарь иностранных слов. / Под ред. Спиркина А.Г., -М.: Русский язык, 1987.

59. Смирнов В.А., Финн В.К. Предисловие к книге Белнап Н., Стил Т. Логика вопросов и ответов. – М.: Прогресс, 1981.

60. Смирнов С.Н. Элементы философского содержания понятия «система» как ступени развития познания и общественной практики. // Системный анализ и научное знание. – М.: Наука, 1978, с.60-83.

61. Солсо Р.Л. Когнитивная психология. - М.: Тривола, 1996.

62. Солтон Дж. Динамические библиотечно-информационные системы. / Пер. Хисамутдинов В.Р. -М.: Мир, 1979, 557с.

63. Стрелков Ю.К. Инженерная психология. / www.psy.msu.ru/science/public/strelkov, 2000г.

64. Сукиасян Э. Р. Новый этап модернизации Библиотечно-библиографической классификации // Библиотековедение, 1998. - 40-45; № 3.

65. Сухманева Е.Г. Уровни обобщения категорий универсальных классификаций и методы синтеза в библиотечно-библиографической классификации. - Л., 1987.

66. Сысойкина М.А. Автореферат диссертации на соискание ученой степени кандидата технических наук. – М.: РГГУ, 2003.

67. Сэлтон Г. Автоматическая обработка, хранение и поиск информации. - М.: Советское радио, 1973.

68. 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.

69. Терминологический словарь по основам информатики и вычислительной техники. / Под ред. А.П .Ершова, Н.М. Шанского. - М.: Просвещение , 1991.

70. Томсон Дж. Предвидимое будущее. – М.: ИЛ, 1958.

71. Урманцев Ю.А. Начала общей теории систем. / Сб. Системный анализ и научное знание. – М.: Наука, 1978, с.7-41.

72. Урсул А.Д. Отражение и информация. - М.: Мысль, 1973.

73. Урсул А.Д. Проблемы информации в современной науке. – М.:Наука, 1976.

74. Успенский В. А. К проблеме построения машинного языка для информационной машины. «Проблемы кибернетики», 1959, вып. 2, стр. 45.

75. Хорошилов А.В. и др. Информационные системы в экономике. - М.: МЭСИ, 1998.

76. Шамурин. Е. И. Словарь книговедческих терминов. М., «Сов. Россия», 1958, стр. 229.

77. Шемакин Ю. И. Тезаурус в автоматизированных системах управления и обработки информации. М.: Воениздат, 1974 г.

78. Шкаренкова Л. Оптимизация стратегии поиска при работе с зарубежными базами данных. - София, 1987.

79. Bates M. How to use information search tactics online. // Online, 1987, 11, №3.

80. Bates M. Search strategies for dialog's view fee. // Online, 1995, 1, pp.22-31.

81. Bates M. The design of browsing and berrypicking techniques for the online search interface. Online Rev. V.13, № 5, 1989, p. 407-424.

82. Borlund P., Ingwersen P. The development of a method for the evaluation of interactive information retrieval systems // J. Doc., 1997, 53, 3, pp.225-250.

83. Boughanem M., Chrisment C., Soule-Dupuy C. Query modification based on relevance back-propagation in an ad hoc environment // Inf. Process. and Manag., 1999, 35, pp.121-139.

84. Cory K. Discovering hidden analogies in an online humanities database // Libr. Trends, 1999, 48, pp.60-71.

85. Courtois M.P., Berry M.W. Results ranking in Web search engines // Online, 1999, 23, 3, pp.39-46.

86. Cranach M., Kalbermatten U. Indermuhle Zielgerichtetes Handeln. - Bern, 1980. - 77p.

87. David A. Krooks, F.W. Lancaster. The Evolution of Guidelines for Thesaurus Construction. Libri. 1993, Vol. 43, No. 4, pp. 326-342.

88. Doucet Manon; Filion Rolande; Garon Denise. Classification et analyse de collections d'objets de jeu selon le systeme ESAR: rapport de recherchй. Doc. et bibl. Том 35, Номер 4, 1989, С. 173-185.

89. Ellis D., Cox D., Hall K. A comparison of the information seeking patterns of researchers in the physical and social sciences // J. Doc., 1993, 49(3), pp.356-369.

90. Hacker W. Activity: A fruitful concept in industrial psychology. / M.Frese & J.Sabini (Eds.), Goal directed behavior: The concept of action in psychology. - Hillsdale, New Fersey: LEA., 1985. - 262p.

91. Harris Kevin. A faceted classification for special literature collections: The Dickens House Classification Int. Libr. Rev. Том: 19 Номер: 4, 1987.

92. Ingwersen P. Wormell I. Modern indexing and retrieval techniques matching different types of information needs. // 44th FID Conf. and Congr., Aug. 28 - Sept. 1, 1988. Pt 1. – 1988, pp.192-203.

93. Kerr S.T. Wayfinding in an electronic database: the relative importance of navigational cues vs. mental models // Inform. processing a. management, 1990, Vol.26, N 4. - P. 511-533.

94. Luhn H. P. A statistical approach to mechanized encoding and searching of literary information. «IBM Journal of Research and Development», 1957, v. 1, № 4, стр. 310.

95. Miyamoto S. Application of rough sets to information retrieval // J. of the Amer. soc. for inform. science Vol. 49, N 3. - P. 195-205.

96. Mohan K. C. Free-text retrieval systems: R&D in information retrieval // J. Sci. and Ind. Res., 1993, 52, №5, pp. 338-349.

97. Nitecki Andre An introduction to contemporary classificatory thought = Введение в современные принципы классификации // Int. Conf. Libr. Classif. and its Funct., Edmonton, June 20-21, 1989. - Edmonton, 1989. - C. 1-6.

98. Pejtersen A.M. Design of intelligent retrieval systems for libraries based on models of users search strategies. New York, 1986.

99. Rijsbergen K.V. Logics for information retrieval. // Note recens. e notiz, 1988, 37 (1-2), pp.121-124.

100. Robertson S.E., Thompson C.L., Macaskill M.J.; Bovey J.D. Weighting, ranking and relevance feedback in a front - end system. J. Inf. Sci. V.12, №1-2, 1986.

101. Saracevic T., Spink A., Wu M. Users and Intermediaries in Information Retrieval: What are They Talking About? / Jameson A. Paris C., Tasso C. User Modeling: Proceedings of the Sixth International Conference, UM97. CISM, -1997.

102. Shaw W. M. Retrieval expectations, clusterbased effectiveness, and performance standards in the CF database // Inf. Process. and Manag., 1994, 30, №5, pp.711-723.

103. Spink A. Study of interactive feedback during mediated information retrieval // J. of the Amer. soc. for inform. Science, 1997, Vol. 48, N 5. - P. 382-394.

104. Spink A., Saracevic T. Interaction in information retrieval: selection and effectiveness of search terms // J. of the Amer. soc. for inform. Science, 1997, Vol. 48, №8. - pp.741-761.

105. Svenonius Elaine Bibliographical Classification = Библиографическая классификация // Int. Conf. Libr. Classif. and its Funct., Edmonton, June 20-21, 1989. - Edmonton, 1989. - C. 21-53.

106. Swanson D. R. Undiscovered public knowledge // Libr. Quart. 1986, 56, №2.

107. Tailor R.S. Question-negotiation and information seeking in libraries. // College and Research Libraries, 1968, 29, pp.178-194.

108. Tenopir C., Shu Man Evena. Magazines in full text: uses and reach strategies, Online Rev., V.13, № 2,1989, p. 107-118.

109. Yoon K. Nilan M.S. Towards a reconceptualization of information seeking research: focus on the exchange of meaning // Inf. Process. and Manag., 1999, №35, pp.871-890.

Глоссарий

Алфавитно-предметная классификация – информационно-поисковый язык, основной словарный состав которого представлен упорядоченным по алфавиту множеством слов, словосочетаний и фраз естественного языка, обозначающих предметы какой-либо отрасли науки или практической деятельности.

Библиотечно-библиографическая классификация – классификация, применяемая в библиотечно-библиографической практике и служащая для расстановки как самих документов, так и их описаний в систематических каталогах.

Вторичный документ - документ, являющийся результатом аналитико-синтетической переработки одного или нескольких первичных документов.

Грамматика – система способов и средств построения слов и предложений в рассматриваемом языке.

Данные - информация, обработанная и представленная в формализованном виде для дальнейшей обработки.

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.

Документная информация - информация, содержащаяся в документах.

Естественный язык - язык, словарь и грамматические правила которого обусловлены практикой применения и не всегда формально зафиксированы.

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.

Иерархическая классификация – классификация, в которой каждое подразделение высшего порядка состоит из непересекающихся между собой подразделений низшего порядка.

Индексирование - формирование описания документа как совокупности дескрипторов, выбираемых из заранее созданных словарей понятий, либо из текстов документов.

Информатизация - Комплекс мер, направленных на обеспечение оперативного доступа к информационным ресурсам.

Информации переработка аналитико-синтетическая - преобразование документов в процессе их анализа и извлечения необходимой информации, а также оценка, сопоставление, обобщение и представление информации в виде, соответствующем запросу.

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, ранение, накопление, обработку, поиск, вывод, копирование, передачу и распространение информации.

Информационно-поисковый тезаурус – словарь-справочник, в котором перечислены все лексические единицы дескрипторного ИПЯ с синонимичными им словами, а также эксплицитно выражены важнейшие смысловые отношения между дескрипторами.

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).

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 анализ - выявление в документах и фиксация в виде данных информации, относящейся к определенной предметной области.

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 - информация, получаемая и (или) используемая в области науки и (или) техники.

Информационный запрос – записанный на естественном языке текст, выражающий некоторую информационную потребность.

Искусственный язык - язык, специально созданный и регулируемый на основе согласованных принципов.

Классификация - процесс соотнесения содержания документов с понятиями, зафиксированными в заранее составленных систематических схемах.

Ключевое слово – предметное слово, выбираемое из некоторого текста (документа) и используемое для координатного индексирования этого текста (документа).

Код - система представления информации в виде данных, состоящая из набора условных знаков и правил присвоения им значений.

Координатное индексирование – индексирование, при котором основное содержание документа представляется в виде сочетания ключевых слов или дескрипторов.

Лексика (или словарный состав) — вся совокупность слов, входящих в состав языка.

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 системе модели и алгоритмы процесса формирования выдачи документов в ответ на поисковый запрос.

Морфология — совокупность действующих в языке способов и средств построения слов.

Навигация - 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).

Парадигматические отношения (аналитические отношения, базисные отношения, ассоциативные отношения) – логические отношения, существующие между лексическими единицами ИПЯ, независимо от их контекста.

Перечислительные классификации - иерархические классификации, в которых имеются отдельные исчерпывающие классы для всех возможных объектов, т. е. перечислены в классификационных таблицах.

Поисковый образ документа (ПОД) – описание документа, выраженное средствами ИПЯ и характеризующее основное смысловое содержание или какие-либо другин 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.

Понятие - форма мышления, отражающая существенные свойства, связи и отношения предметов и явлений.

Посткоординируемый ИПЯ — ИПЯ, в котором лексические единицы (термины, слова) объединяются в «предложения» (поисковые образы) лишь во время индексирования документов или даже в процессе их поиска.

Предкоординированный ИПЯ – ИПЯ, в котором лексические единицы (термины, слова) поисковых образов связаны координатной (соподчинительной) или какой-либо иной связью до процесса индексирования.

Признак - свойство объекта, обусловливающее его различие или общность с другими объектами.

Символ - элемент согласованного набора материальных предметов, используемых для представления знаков.

Синтагматические отношения (текстуальные отношения, синтетические отношения, синтаксические отношения) – отношения между лексическими единицами ИПЯ, выражающие действительные логические связи между соответствующими понятиями в тексте сообщения.

Синтаксис — совокупность действующих в языке способов и средств построения предложений.

Сообщение - информация, переданная и (или) полученная в процессе коммуникации.

Средства поиска –взаимозависимый комплекс ИПЯ и языков определения/управления данными, обеспечивающий структурные и семантические преобразования объектов обработки (документов, словарей, совокупностей результатов поиска), а также объекты пользовательского интерфейса как технологические решения, обеспечивающие управление последовательностью выбора операционных объектов конкретной АИПС.

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.

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.

Терминосистема - организованная совокупность терминов в специальном языке определенной области знания.

Технологии поисковые – унифицированные (оптимизированные в рамках конкретной АИПС) последовательности эффективного использования в процессе взаимодействия пользователя с системой отдельных средств поиска для устойчивого получения конечного и, возможно, промежуточных результатов.

Фасет – совокупность подклассов, получаемая при делении основного класса по одному и тому же ряду характеристик.

Фасетная классификация – классификация, дающая возможность классифицировать объекты параллельно по нескольким различным признакам.

Фасетная формула – фиксированная схема последовательности расположения фасетов в индексах документов при их многоаспектном индексировании.

Формат - способ расположения и представления данных на носителе информации.

Язык - Система знаков, обеспечивающая коммуникацию и включающая набор знаков (словарь) и правила их употребления и интерпретации (грамматика).

Приложения

Приложение 1

Таблица 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. Методы, основанные на применении компьютеров

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. Онлайновые 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

=> 004.4'27 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

004.05 Качество систем и программ

004.051 Эффективность

004.052 Надёжность и точность

004.052.2 Устойчивость к ошибкам, живучесть

004.052.3 Отказоустойчивость

004.052.32 Контроль неисправностей

004.052.34 Порог неисправностей

004.052.4 Защита от ошибок. Обработка ошибок

004.052.42 Обнаружение ошибок.

Контроль допустимости. Проверка достоверности данных. Верификация данных

004.052.44 Исправление ошибок

004.053 Возможность многократного использования

004.054 Тестируемость, оцениваемость

004.055 Дружественность пользователю

004.056 Безопасность, защищенность данных

004.056.2 Целостность данных

004.056.3 Резервирование и восстановление данных

004.056.4 Избыточность

004.056.5 Защита данных

004.056.52 Ограничение доступа

004.056.53 Защита от несанкционированного доступа. Хакинг, кракинг

Хакинг - поиск уязвимости в вычислительных системах и осуществление их взлома (хакерами - в конструктивных целях, кракерами - в преступных)

004.056.55 Шифрование

=> (083.73) Коды

=> 003.26 Криптография. Тайнопись

=> 004.421.5 Алгоритмы генерации случайных (псевдослучайных) чисел

=> 651.928 Деловые коды

004.056.57 Защита компьютеров от заражения вирусами. Защита от компьютерных инфекций

=> 004.49 Компьютерные инфекции. Компьютерная вирусология

004.057 Совместимость

004.057.2 Стандартизация. Соответствие стандартам

=> (083.74) Стандарты

=> 006 Стандартизация

004.057.3 Форматы

004.057.4 Протоколы

004.057.5 Переносимость, мобильность

004.057.6 Конвертирование

004.057.7 Эмуляция

004.057.8 Открытые системы. Открытая архитектура

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

004.081 Перфоносители

004.081.1 Перфокарты

=> 676.815.4 Перфокарты для программного управления машинами, например, счетными, текстильными

004.081.2 Перфоленты

=> 676.816.5 Бумажные бобины, предназначенные для перфорации, для телетайпов, пишущих машинок, счетных машин и т. п. Перфорационные ленты (перфоленты)

004.083 Магнитные носители

004.083.1 Магнитные карты

004.083.2 Магнитные полоски

004.083.3 Магнитные чернила

004.083.4 Магнитные ленты

004.083.42 Катушечные ленты

004.083.44 Двухкатушечные компакт-кассеты

004.083.45 Ленточные картриджи

004.083.5 Магнитные сердечники

004.083.6 Магнитные барабаны

004.083.7 Магнитные диски

004.083.72 Жесткие диски

004.083.722 Пакеты дисков

004.083.723 Матрицы дисков

004.083.74 Гибкие диски. Дискеты

004.083.8 Память на магнитных пленках. Тонкопленочная память. Включая: Память на цилиндрических магнитных доменах. ЦМД-ЗУ

004.084 Магнитооптические носители. Стираемые оптические диски

004.085 Оптические носители (нестираемые)

004.085.2 Компакт-диски

004.085.22 Постоянное запоминающее устройство на компакт-диске КД-ПЗУ (CD-ROM)

004.085.23 Компакт-диски однократной оптической записи (CD-R)

004.085.3 Диски однократной записи и многократного считывания (WORM)

004.085.4 Видеодиски

004.085.7 Цифробумага

Полиэфирная пленка, покрытая слоем красителя. Запись производится с помощью лазера путем выжигания в нем углублений

004.085.72 Цифробумажная лента

004.085.74 Цифробумажные диски

004.087 Электронные носители

004.087.2 Полупроводниковая память. Твердотельные диски. Флэш-память на дисках

004.087.4 Платы памяти. ПК-карты

004.087.5 Интеллектуальные карточки. ИС-карты

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 (одиночный поток команд и множественный поток данных, SIMD)

004.272.33 МКОД-архитектура (множественный поток команд и одиночный 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 Потоковые 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 корпусы

004.3'142.25 Корпусы с двухрядным расположением выводов. Двухрядные корпусы

004.3'142.26 Корпусы с матричным расположением выводов. Матричные корпусы

004.3'142.27 Плоские корпусы

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

Основной ряд индексов

004.31 Блоки обработки данных. Процессоры

004.312 Логические схемы, блоки

004.312.2 Комбинационные схемы

004.312.22 Логические вентили

004.312.222 Схемы И, ИЛИ, НЕ

004.312.225 Схемы И-НЕ, ИЛИ-НЕ

004.312.24 Мультиплексоры. Коммутаторы

004.312.26 Кодеры. Декодеры

004.312.4 Последовательностные схемы

004.312.42 Триггеры

004.312.43 Пересчетные устройства. Счетчики

004.312.44 Арифметические и логические схемы для арифметических устройств

=> 004.315 Арифметические устройства

004.312.46 Цепи управления для математических операций

004.312.462 Монтажная логика

004.312.463 Логика наборной панели

004.312.466 Логика микропрограммного управления

004.314 Регистры

004.314.2 Регистры общего назначения
004.314.3 Аккумуляторы. Накапливающие регистры
004.314.4 Сдвиговые регистры
004.314.5 Регистры чисел с плавающей запятой
004.314.6 Стековые регистры
004.315 Арифметические устройства
=> 004.312.44 Арифметические и логические цепи арифметических устройств
004.315.2 Сумматоры. Полусумматоры
004.315.4 Схемы образования дополнения
004.315.5 Множительные устройства. Делительные устройства
004.315.7 Векторные арифметические устройства
004.315.8 Сдвигающие устройства. Сравнивающие устройства
004.318 Центральный процессор
004.32 Магистрали ЭВМ
004.322 Каналы
004.324 Шины
004.33 Блоки памяти. Накопители. Запоминающие устройства
Характеристики памяти обозначать специальными определителями .07..., носители информации - определителями .08...
004.332 Типы памяти в зависимости от возможности доступа
004.332.2 Память прямого доступа
004.332.3 Память произвольного доступа (RAM)
004.332.33 Динамическая память произвольного доступа (DRAM)
004.332.34 Статическая память произвольного доступа (SRAM)
004.332.5 Память последовательного доступа
004.333 Постоянная память
004.334 Магазинная память прямого и обратного типа
004.335 Ассоциативная память. Память, адресуемая содержимым
004.337 Внутренняя память
004.338 Внешняя память
004.35 Периферия. Устройства ввода-вывода
004.351 Читающие устройства
Например
004.351.081.1 Устройства для считывания перфокарт
004.352 Сканеры. Сканирующие устройства. Устройства развертки
004.352.2 Оптические сканеры
004.352.22 Графические сканеры
004.352.24 Распознающие оптические сканеры. Световые карандаши
004.352.242 Оптическое распознавание знаков (OCR)
004.352.243 Распознавание рукописного текста
004.352.246 Распознавание штриховых кодов
004.352.4 Сканеры для специфических носителей данных

Например

004.352.4.083.3 Сканеры для распознавания знаков, нанесенных магнитными чернилами (MICR)

004.352.4.087.5 Сканеры для ИС-карт

004.353 Пульты управления. Видеотерминалы. Устройства визуального отображения. Видеомониторы

004.353.2 Экраны

004.353.22 Разрешение экрана

004.353.24 Моды отображения

004.353.242 Побитовое отображение. Поэлементное отображение

004.353.244 Растровое отображение

004.353.245 Векторное отображение. Каллиграфическое отображение

004.353.25 Типы экранов

004.353.253 Индикаторы с электроннолучевой трубкой (ЭЛТ)

004.353.254 Плоские индикаторы

004.353.254.2 Плазменные индикаторы

004.353.254.3 Жидкокристаллические индикаторы. ЖК-табло

004.353.254.4 Электролюминесцентные индикаторы

004.353.254.5 Светодиодные индикаторы (LED)

004.353.4 Клавиатура компьютеров

=> 681.61.065 Клавиши и клавиатура пишущих и наборных машин

004.353.44 Специальные клавиши

004.353.45 Функциональные клавиши

004.354 Устройства ввода координат. Координатно-указательные устройства. Устройства ввода графики

004.354.2 Световое перо

004.354.3 Сенсорные экраны

004.354.4 Графические планшеты. Кодирующие планшеты. Дигитайзеры. Устройства для оцифровки чертежей и рисунков

004.354.5 Мышь

004.354.6 Шары трассировки

004.354.7 Рычажные указатели. Джойстики

004.355 Накопители. Дисководы. Лентопротяжки

Подразделять при помощи .08..., например

004.355.083.4 Накопители на магнитных лентах. Лентопротяжки

004.355.083.723 Накопители на матрицах дисков. RAID-конфигурации (избыточные матрицы недорогих дисков)

004.355.083.74 Накопители на гибких магнитных дисках

004.355.085.22 Накопители на оптических компакт-дисках (CD-ROM)

004.356 Периферия для получения твердых копий

004.356.2 Печатающие устройства. Принтеры

Виды принтеров => 681.625.9...

004.356.4 Вывод на фоточувствительные материалы

Включая: Компьютерное выходное микрофильмирование (COM)

004.357 Акустическая и мультимедийная периферия. Устройства ввода данных с голоса

004.358 Периферия для имитационного моделирования и виртуальной реальности

Включая: Ввод и вывод данных через перчатки, одежду, шлем

004.38 Виды компьютеров

Поколения компьютеров обозначать специальными определителями .032.8...

004.382 Компьютеры общего назначения. Универсальные вычислительные машины

004.382.2 Суперкомпьютеры. Супер-ЭВМ

004.382.3 Большие вычислительные машины. Универсальные ЭВМ. Мейнфреймы

004.382.4 Миникомпьютеры. Мини-ЭВМ

004.382.6 Рабочие станции

004.382.7 Персональные компьютеры. ПЭВМ. Микро-ЭВМ. Микрокомпьютеры

004.382.72 настольные

004.382.73 наколенные (lap-top). ПК-блокноты

004.382.75 ручные (ладонные) (palm-top)

004.382.76 карманные (pocketable)

004.383 Специализированные вычислительные машины

004.383.1 Текстовые процессоры

004.383.3 Процессоры цифровой обработки сигналов

004.383.4 Моделирующие устройства

004.383.5 Процессоры изображений

004.383.6 Процессоры баз данных

004.383.8 Машины с элементами искусственного интеллекта

Например

004.383.8.032.26 Нейрокомпьютеры

004.384 Вычислительные машины для управления технологическими процессами. Промышленные ЭВМ

004.386 Гибридные вычислительные машины

004.387 Аналоговые вычислительные машины

004.388 Компьютероподобные устройства

004.388.2 Калькуляторы. Программируемые калькуляторы

004.388.4 Игровые автоматы

004.4 Программные средства

Специальные определители

004.4'2 Инструментальные средства разработки программного обеспечения

004.4'22 Средства автоматизированной разработки программного обеспечения (CASE)

004.4'23 Средства поддержки программирования

004.4'232 Редакторы программ

004.4'233 Отладочные программы

004.4'234 Программы просмотра. Браузеры

004.4'236 Средства визуального программирования

004.4'24 Средства автоматической разработки программного обеспечения

004.4'242 Генераторы программ. Генераторы прикладных программ. Автоматическое программирование

004.4'244 Генераторы систем

004.4'27 Средства разработки мультимедиа

004.4'272 Авторские системы (Authoring systems)

004.4'273 Редакторы изображений (Image editors)

004.4'274 Видеоредакторы (Video editors)

004.4'275 Программы анимации. Синтез динамических изображений

004.4'277 Средства обработки звука

004.4'277.2 Редакторы звука

004.4'277.4 Средства создания музыки

Включая: Программы упорядочения для интерфейса электромузыкальных инструментов (MIDI sequencers)

004.4'4 Трансляция программ

004.4'41 Процессы трансляции программ

004.4'412 Лексический анализ

004.4'413 Синтаксический анализ. Грамматический разбор

004.4'414 Семантический анализ

004.4'415 Запись программы в машинных кодах

004.4'416 Оптимизация программы в машинных кодах

004.4'417 Использование таблицы идентификаторов

004.4'418 Использование промежуточного языка

004.4'42 Трансляторы программ

004.4'422 Компиляторы

004.4'423 Интерпретаторы

004.4'424 Ассемблеры

004.4'426 Макропроцессоры

004.4'427 Препроцессоры

004.4'6 Условия выполнения программ. Среда выполнения. Оперативные средства. Включая: Динамическое распределение памяти. Вычисление адресов. Управление выполнением команд

Основной ряд индексов

004.41 Программотехника. Разработка вычислительных систем

004.412 Метрики программного обеспечения

004.412.2 Метрики, основанные на объеме программы. Метрики строк (текста) программ (LOC)

004.412.3 Метрики, основанные на функциональных точках

004.413 Организация разработки программного обеспечения

004.413.2 Управление планированием

004.413.4 Анализ рисков

004.413.5 Методы оценки программных разработок

004.414 Этап формулировки задания для разработки вычислительных систем и программного обеспечения

004.414.2 Анализ и проектирование вычислительной системы

004.414.22 Анализ требований к системе

Например

004.414.22:004.2 Требования к архитектуре компьютера

004.414.23 Моделирование и имитация вычислительных систем

004.414.28 Техническое описание системы. Спецификация системы

004.414.3 Анализ требований к программному обеспечению

004.414.32 Прототипирование. Макетирование

004.414.38 Спецификация требований к программному обеспечению

004.415 Этап разработки программного обеспечения и вычислительных систем

004.415.2 Проектирование программного обеспечения и вычислительных систем

Например

004.415.2.031.43 Проектирование систем реального времени

004.415.2.041 Процедурно-ориентированные методы проектирования

004.415.2.043 Методы проектирования, ориентированные на структуру данных

004.415.2.052.03 Отказоустойчивое проектирование

004.415.22 Абстракция

004.415.23 Разбиение на модули. Модуляризация

004.415.24 Упрятывание информации

004.415.25 Программирование отдельных компонентов системы

004.415.26 Языки проектирования программ. Псевдокод. Символический код

004.415.28 Спецификация проекта программного обеспечения

004.415.3 Трансляция программ на язык машинных кодов

=> 004.42 Программирование ЭВМ. Компьютерные программы

004.415.5 Гарантии качества программного обеспечения

Включая: Верификация. Аттестация

004.415.52 Формальная техническая проверка

Включая: Метод сквозного контроля

004.415.53 Тестирование программного обеспечения

004.415.532 Блочное тестирование. Тестирование элементов программного обеспечения

004.415.532.2 Тестирование методом «черного ящика» (без вникания в подробности работы отдельных компонентов и взаимодействия между ними)

004.415.532.3 Тестирование методом «белого ящика» (проверка правильности работы отдельных частей и взаимодействия между ними)

004.415.533 Комплексное тестирование

004.415.535 Аттестационное тестирование

004.415.538 Системное тестирование. Тестирование системы в целом

004.416 Сопровождение программного обеспечения

004.416.2 Корректировка. Исправление ошибок. Отладка

004.416.3 Адаптация

004.416.6 Развитие. Доработка. Расширение возможностей. Совершенствование программного обеспечения

004.418 Внедрение программного обеспечения

004.42 Программирование ЭВМ. Компьютерные программы

=> 519.85 Математическое программирование

004.421 Алгоритмы составления программ

Алгоритмы отдельных процессов относить к классам, соответствующим этим процессам, с добавлением специального определителя .021, например

004.424.5.021 Алгоритмы сортировки

004.421.2 Базовые математические алгоритмы

=> 512.54.05 Алгоритмические задачи в теории групп. Проблема слов

=> 519.16 Алгоритмические задачи комбинаторного анализа

=> 519.178 Алгоритмические вопросы теории графов

=> 519.712 Кибернетические вопросы теории алгоритмов

Математическая теория алгоритмов в целом => 510.5 Теория алгоритмов и вычислимые функции

Математические процессы классифицировать с помощью :51... (присоединение со знаком отношения математических индексов), например

004.421.2:517.443 Алгоритмы быстрого преобразования Фурье

004.421.2:517.535 Алгоритмы для рациональных функций в комплексной области

004.421.2:519.17 Алгоритмы теории графов

004.421.4 Алгоритмы выражения оценок

004.421.5 Алгоритмы генерации случайных (псевдослучайных) чисел

=> 004.056.55 Шифрование

004.421.6 Алгоритмы символьной обработки

004.422 Компоненты компьютерных программ

004.422.1 Элементарные единицы. Атомы

004.422.2 Константы

004.422.3 Переменные
004.422.32 Типы переменных
004.422.322 Глобальные и локальные переменные
004.422.324 Статические и динамические переменные
004.422.326 Свободные и общие переменные
004.422.33 Правило умолчания для переменных
004.422.35 Описание переменных
004.422.352 явное (эксплицитное)
004.422.353 неявное (имплицитное)
004.422.4 Выражения
004.422.42 Части выражений
004.422.422 Оператор
004.422.423 Операнд
004.422.44 Типы выражений
004.422.442 М-выражение
004.422.444 Лямбда-выражение
004.422.5 Операторы. Предложения
004.422.52 Выполняемые предложения, операторы
004.422.53 Декларативные предложения, описатели
004.422.55 Хорновское предложение. Хорновское выражение. Дизъюнкт Хорна
004.422.56 Макрокоманды. Вызов макрокоманд
004.422.6 Типы данных
004.422.61 Основные типы данных
004.422.612 Логические данные. Булевы данные
004.422.613 Численные данные
004.422.614 Символьные данные
004.422.615 Перечислимые данные
004.422.616 Указатели
004.422.63 Структурированные данные. Структуры данных
004.422.632 Массивы. Матрицы
004.422.632.2 Символьные строки
004.422.632.4 Многомерные матрицы
004.422.633 Записи
004.422.634 Множества
004.422.635 Динамические структуры данных
004.422.635.2 Цепной список. Структура связанных списков
004.422.635.3 Древовидная структура
004.422.635.32 Сбалансированные деревья (Б-деревья)
Включая: B+-деревья
004.422.635.33 Бинарные деревья
004.422.635.5 Потоковые структуры
004.422.636 Абстрактные типы данных
004.422.636.2 Списки
004.422.636.3 Стеки

004.422.636.4 Очереди
004.422.636.5 Деки. Двухсторонние очереди
004.422.636.7 Таблицы
004.422.639 Другие типы данных
004.422.8 Программные модули
004.422.81 Главная программа
004.422.83 Подпрограммы
004.422.832 Сопрограммы
004.422.833 Модули. Процедуры
004.422.834 Функции
004.422.837 Программные макросы
004.422.86 Параметры. Аргументы
004.423 Синтаксис и семантика программ
004.423.2 Синтаксис программ
004.423.22 Конкретный синтаксис
004.423.23 Абстрактный синтаксис
004.423.24 Регулярное выражение
004.423.25 Бесконтекстная (контекстно-свободная) грамматика
004.423.26 Атрибутная грамматика
004.423.4 Семантика программ
004.423.42 Операционная семантика
004.423.43 Денотационная семантика
004.423.45 Алгебраическая семантика
004.423.46 Аксиоматическая семантика
004.424 Методы программирования
004.424.2 Основные управляющие структуры
004.424.22 Повторения. Петли. Итерации. Циклы
004.424.23 Селекция. Выборка
004.424.25 Таблица решений
004.424.27 Переход. GOTO. Передача управления. Ветвление
004.424.3 Техника подпрограмм
004.424.32 Вызов подпрограмм
004.424.33 Аргументы
004.424.35 Реентерабельные подпрограммы
004.424.36 Рекурсивные процедуры
004.424.4 Методы поиска
004.424.42 Поисковые ключи
004.424.43 Поиск по таблице. Табличный поиск
004.424.44 Операции над множествами
004.424.45 Линейный поиск
004.424.46 Дихотомический поиск
004.424.47 Хеширование. Рандомизация
004.424.5 Методы сортировки

Например

004.424.5:004.337 Внутренняя сортировка

004.424.5:004.338 Внешняя сортировка

004.424.5.032.24 Параллельная сортировка

004.424.52 Сортировка с минимальным числом сравнений

004.424.53 Сеть сортировки

004.424.56 Слияние. Объединение

004.424.57 Пермутация. Перестановка

004.424.6 Сопоставление образцов. Сравнение с образцом

004.424.62 Сравнение строк. Сопоставление строк с эталоном

004.424.64 Сравнение деревьев с образцом

004.424.7 Методы запоминания

004.424.72 Упаковка в блоки

004.424.74 Динамическое распределение памяти

004.424.75 Манипулирование неупорядоченными массивами

004.424.8 Команды ввода-вывода

004.428 Библиотеки подпрограмм

004.428.2 Стандартные подпрограммы

004.428.4 Пользовательские подпрограммы

004.43 Языки программирования

В этом классе применимы специальные определители 004.4'4

=> 811.93 Искусственные языки для использования машинами. Языки программирования. Компьютерные языки

004.431 Языки низкого уровня

004.431.2 Машинные языки

004.431.4 Ассемблерные языки

004.432 Языки высокого уровня

004.432.2 Процедурные языки. Императивные языки

004.432.4 Непроцедурные языки. Декларативные языки

004.432.42 Функциональные языки

004.432.45 Языки четвертого поколения. Диалоговые языки

=> 004.655 Языки баз данных

004.434 Проблемно-ориентированные языки

Например

004.434:004.422.635.2 Языки обработки списков

004.434:004.8 Языки программирования для систем искусственного интеллекта

004.434:004.82 Языки программирования для систем представления знаний

004.434:004.94 Языки моделирования

004.434:5 Языки программирования для научно-исследовательских прикладных систем

004.434:65 Языки программирования в области делопроизводства, полиграфии, связи, транспорта, торговли и организации производства. Языки для бизнес-применений

004.435 Метаязыки
004.436 Дескриптивные языки
004.436.2 Языки описания аппаратных средств
004.436.4 Языки описания систем
004.438 Отдельные языки программирования

Подразделять с помощью алфавитного расширения (Таблица Ih). Конкретные языки программирования могут быть также включены с соответствующим алфавитным расширением в подходящие подклассы типов языков класса 004.43.

=> =93 Искусственные языки для использования машинами. Языки программирования. Компьютерные языки

004.45 Системное программное обеспечение
004.451 Операционные системы
004.451.1 Управление вводом-выводом

Например

004.451.1:004.237 Ввод-вывод по прерываниям
004.451.2 Организация процессов
004.451.21 Взаимное исключение
004.451.22 Равнодоступность (ресурсов)
004.451.23 Синхронизация
=> 004.032.32 Синхронные процессы
004.451.24 Взаимодействие между процессами
004.451.25 Управление процессом
004.451.26 Планирование процесса
004.451.27 Предотвращение зависаний. Выход из зависания
004.451.3 Управление памятью
004.451.31 Распределение памяти
004.451.33 Сборка мусора
004.451.34 Сжатие памяти. Дефрагментация
004.451.35 Управление виртуальной памятью
004.451.352 Страничная организация
004.451.353 Сегментация
004.451.354 Страничная сегментация
004.451.36 Защита памяти
004.451.37 Организация вспомогательной памяти
004.451.4 Организация выполнения задач
004.451.42 Управление заданиями
004.451.43 Спулинг. Подкачка данных. Откачка данных
004.451.44 Планирование заданий
004.451.45 Многопоточная обработка
004.451.46 Мультипрограммирование. Многозадачный режим
004.451.47 Языки управления заданиями. Командные языки
004.451.48 Директивы оператора
004.451.5 Управление данными
004.451.51 Организация томов

004.451.52 Организация файлов

Включая: Создание файлов, уничтожение файлов и манипулирование файлами

Например

004.451.52.056.3 Дампирование файлов

004.451.52.056.5 Защита файлов

004.451.53 Структура файлов

Включая: Структуры директорий

004.451.54 Организация записей. Манипулирование записями

004.451.55 Гипермедиа. Гипертекст

004.451.56 Методы доступа

004.451.57 Буферизация

004.451.6 Организация связей

004.451.62 Взаимодействие между процессами

004.451.622 Передача сообщений

004.451.624 Совместно используемая память

004.451.64 Привилегированные команды. Аппаратные средства защиты

004.451.642 Рабочие режимы

Включая: Режим монитора (режим супервизора, режим операционной системы). Пользовательский режим

004.451.644 Системные вызовы (вызовы супервизора)

004.451.7 Обработка данных в режиме разделения времени

Например

004.451.7:004.7 Обработка данных, поступающих по линиям связи

004.451.7.031.43 Обработка данных в режиме реального времени. Обработка транзакций. Диалоговая обработка запросов

004.451.8 Конструктивные особенности операционных систем

004.451.82 Иерархическое представление. Разбиение на слои

004.451.83 Конструкция «клиент-сервер»

004.451.84 Системные интерфейсы

004.451.86 Подсистемы

004.451.87 Ядро

004.451.88 Микроядро

004.451.9 Конкретные операционные системы

Подразделять при помощи алфавитного расширения (Таблица Ih)

Например

004.451.9UNIX Операционная система ЮНИКС

004.453 Программы управления программами

004.453.2 Редакторы связей

004.453.3 Перемещающие программы

004.453.4 Программы загрузки

004.453.5 Первоначальный загрузчик программы

004.454 Драйверы устройств

004.457 Обслуживающие программы. Утилиты

004.49 Компьютерные инфекции. Компьютерная вирусология

=> 004.056.57 Защита компьютеров от заражения вирусами. Защита от компьютерных инфекций

004.491 Программы, повреждающие компьютерные системы

004.491.2 Размножающиеся носители инфекции. Самокопирующиеся носители инфекции

004.491.22 Компьютерные вирусы

004.491.23 Компьютерные черви

004.491.4 Несамокопирующиеся носители инфекции

004.491.42 Троянские кони

004.491.43 Логические бомбы

004.492 Программы для борьбы с компьютерными инфекциями

004.492.2 Предохраняющие программы

004.492.3 Детектирующие программы

004.492.4 Излечивающие программы. Обезвреживающие программы

004.5 Человеко-машинное взаимодействие. Человеко-машинный интерфейс. Пользовательский интерфейс. Операционная среда пользователя

Периферийное оборудование => 004.35

004.51 Дисплейный интерфейс

004.512 Текстоориентированные интерфейсы. Символьные пользовательские интерфейсы

004.512.2 Подсказка команд. Приглашение пользователя на ввод команды

004.512.3 Меню

004.512.4 Диалог

004.512.5 Табличный диалог

004.514 Графический интерфейс пользователя

004.514.2 Указатели

004.514.4 Всплывающее меню. Ниспадающее меню

004.514.6 Оконная среда WIMP-интерфейс (многооконная среда, управляемая мышью)

004.514.62 Менеджер виртуального рабочего стола

004.514.64 Механизмы «выбери и перенеси». Отбуксировать и оставить

004.52 Звуковой интерфейс

004.522 Речевое взаимодействие

004.523 Использование неречевых звуков

004.55 Гипермедиа. Гипертекст

004.58 Помощь пользователю

004.582 Помощь на экране

004.584 Мастер. Оперативный консультант. Модуль оперативной помощи

004.588 Обучающие программы

004.6 Данные

004.62 Манипулирование данными

004.622 Подготовка данных

004.623 Загрузка и подкачка данных

004.624 Экспорт и импорт данных

004.627 Сжатие данных

004.63 Файлы

004.632 Доступ к файлам

004.632.2 Последовательный доступ. Последовательная выборка (Serial access)

Данные считываются в оперативную память в порядке их физического размещения на носителе внешнего запоминающего устройства

004.632.3 Последовательный доступ. Последовательная выборка (Sequential access)

Данные файла обрабатываются в порядке их записи при создании файла

004.632.4 Произвольный доступ. Прямой доступ

004.632.5 Индексно-последовательный доступ

004.633 Манипулирование файлами

004.633.2 Сортировка

004.633.3 Слияние

004.633.4 Обновление

004.65 Системы управления базами данных (СУБД)

Например

004.65:004.451 Операционные системы для реализации баз данных

004.651 Файловая структура базы данных

004.651.2 Секционированная (библиотечная) структура. Структурированность по разделам

004.651.3 Мультисписковая структура

004.651.4 Древовидная структура

004.651.5 Динамическая файловая структура

004.651.52 B-дерево

004.651.53 Самоорганизующееся дерево

004.651.54 Динамическое хеширование

004.652 Модели баз данных

004.652.2 Иерархическая модель

004.652.3 Сетевая модель

004.652.4 Реляционная модель

004.652.42 Реляционная целостность

004.652.43 Реляционная алгебра

004.652.44 Реляционное исчисление

004.652.5 Объектно-ориентированные модели баз данных

004.652.6 Логическая модель

004.652.7 Модель инвертированных файлов

004.652.8 Модель отношений объектов. Модель «сущность-связь»

004.654 Зависимость от данных. Зависимость по данным

004.655 Языки баз данных

004.655.2 Языки определения данных

004.655.3 Языки манипулирования данными

Языки запросов. Стандартный язык запросов. SQL - структурированный язык запросов

004.656 Словари данных

004.657 Обработка запросов к базе данных

004.658 Управление базой данных. Ведение базы данных

004.658.2 Функционирование базы данных

004.658.3 Реструктуризация базы данных

004.658.4 Стандартизация баз данных

004.658.6 Взаимодействие баз данных

004.67 Системы обработки численных данных

Включая: Системы электронных таблиц

004.7 Связь компьютеров. Сети ЭВМ. Вычислительные сети

004.71 Сетевая аппаратура

=> 621.39 Телекоммуникационное оборудование

004.712 Сетевые адаптеры. Сетевые платы

004.713 Коммутаторы данных

004.714 Концентраторы. Хабы

Включая: Активные концентраторы, пассивные концентраторы

004.715 Маршрутизаторы

004.716 Устройства сетевой связи

Включая: Мосты, шлюзы, реле

004.717 Фронтальные процессоры. Связные процессоры. Фронтальные вычислительные машины. Связные компьютеры

004.72 Архитектура сетей

=> 004.057.4 Сетевые протоколы

Например

004.72:004.451 Сетевые операционные системы

004.722 Топологии сетей

004.722.2 Сети с прямыми соединениями. Сети с двухточечным соединением (абонентов)

004.722.22 Звездообразные (радиальные) сети

004.722.23 Кольцевые сети

004.722.25 Древовидные сети

004.722.4 Широковещательные сети

004.722.42 Сети с маркерным кольцом. Кольцевые сети с маркерным доступом

004.722.43 Сети с маркерной шиной. Магистральные сети с маркерным доступом

004.722.45 Спутниковые сети

004.724 Методы коммутации данных

004.724.2 Коммутация каналов

004.724.3 Коммутация сообщений

004.724.4 Пакетная коммутация

004.725 Элементы сетей

004.725.2 Хосты

004.725.4 Узлы

004.725.5 Сети локального доступа

004.725.7 Базовые сети

004.728 Модель стандарта взаимодействия открытых систем. Эталонная модель OSI

004.728.1 Физический уровень

004.728.3 Канальный уровень

Например

004.728.3.057.4 Протоколы управления каналом передачи данных. Канальные регламентирующие протоколы

004.728.4 Сетевой уровень

004.728.5 Транспортный уровень

004.728.6 Сеансовый уровень

004.728.7 Уровень представления. Представительный уровень

004.728.8 Прикладной уровень

004.73 Виды сетей в зависимости от охватываемой территории

004.732 Локальные сети (LAN)

004.733 Сети городского уровня (MAN)

004.735 Широкомасштабные сети (WAN)

004.738 Взаимодействие сетей. Межсетевой обмен

004.738.2 Передача сообщений

Включая: Упаковка сообщений. Трансляция сообщений

004.738.5 Интернет

Например

004.738.5.057.4 Протокол Интернета. TCP/IP

004.738.52 Средства поиска в Интернете

Глобальная гипертекстовая система Интернета World Wide Web (WWW). «Всемирная паутина»

004.75 Распределенные системы обработки данных

Например

004.75:004.451 Распределенные операционные системы

004.77 Применения компьютерных сетей

004.771 Удаленный доступ

004.772 Передача файлов

004.773 Обмен сообщениями

004.773.2 Доска объявлений

004.773.3 Электронная почта (e-mail)

004.773.5 Организация телеконференций

004.78 Диалоговые вычислительные системы для специальных целей

Например

004.78:025.4.036 Автоматизированные информационно-поисковые системы

004.78:336.717 Автоматизированные системы банковских операций

004.8 Искусственный интеллект

004.81 Модели когнитивных процессов

Подразделять с помощью :159.9..., например

004.81:159.942.52 Модели чувств

004.81:159.953 Модели памяти

004.81:159.953.32 Модели ассоциативной памяти

004.81:159.953.52 Модели обучения. Модели приобретения знаний

004.81:159.953.6 Модели забывания

004.81:159.955 Модели мышления

004.82 Представление знаний

004.822 Сети знаний

Включая: Семантические сети

004.823 Фреймы. Фреймовые системы

Включая: Схемы. Сценарии

004.824 Множественные миры

004.825 Порождающие системы. Системы правил вывода

004.826 Модель черной доски

004.827 Представление неоднозначности. Неопределенность. Пробелы знаний

004.83 Рассуждение

004.832 Решение задач

004.832.2 Поиск решения

004.832.22 Представление задачи

004.832.23 Поиск в пространстве состояний (задачи)

004.832.24 Поиск в пространстве игры

004.832.25 Поиск на основе ограничений

004.832.28 Стратегии управления

004.832.3 Логический вывод

004.832.32 Методы вывода

004.832.34 Вывод на основе недостоверного знания

004.832.38 Управление логическим выводом

004.838 Виды рассуждений

004.838.2 Качественные рассуждения

004.838.3 Рассуждения по аналогии

004.838.5 Абдукция. Силлогизмы с вероятностной малой посылкой

004.85 Обучение

004.852 Статистическое обучение и параметрическое обучение

004.853 Приобретение и использование знаний

004.855 Обучение и индуктивный вывод

004.855.2 Грамматический вывод

004.855.3 Концептуальное обучение

004.855.5 Обучение на примерах

004.855.6 Обучение по аналогии. Обучение как открытие

004.89 Прикладные системы искусственного интеллекта. Интеллектуальные системы, основанные на использовании знаний

004.891 Экспертные системы

004.891.2 Консультативные экспертные системы

004.891.3 Диагностические экспертные системы

004.896 Искусственный интеллект в промышленных системах. Интеллектуальные САПР и АСУП. Интеллектуальные средства робототехники

004.9 Прикладные информационные (компьютерные) технологии. Методы, основанные на применении компьютеров

004.91 Обработка и создание документов

004.912 Обработка текста. Подготовка текстов

004.915 Настольная издательская деятельность. Подготовка публикаций с помощью настольных издательских систем

004.92 Компьютерная графика

004.921 Элементы и объекты компьютерной графики

004.922 Координаты в компьютерной графике. Преобразование изображений. Преобразование для просмотра

004.923 Стереоскопическая визуализация

004.924 Методы ввода графики

=> 004.352.22 Графические сканеры

=> 004.354.4 Графические планшеты. Кодирующие планшеты. Дигитайзеры. Устройства для оцифровки чертежей и рисунков

004.925 Методы компьютерной графики

004.925.2 Удаление невидимых линий. Удаление невидимых поверхностей

Например

004.925.2.021 Алгоритм удаления невидимых линий

004.925.3 Затенение. Построение теней. Обработка полутонов. Включая: Трассировка лучей

004.925.4 Наложение текстуры. Текстурирование

004.925.5 Отображение в цвете

004.925.6 Раскрашивание. Закрашивание

004.925.8 Геометрическое моделирование

004.925.82 Каркасное моделирование

004.925.83 Моделирование поверхностей. Построение модели трехмерного тела с помощью поверхностей

004.925.84 Моделирование сплошных тел. Твердотельное моделирование. Моделирование трехмерных тел

004.925.86 Моделирование кривых

004.928 Анимация. Мультипликация

004.93 Распознавание и преобразование образов

004.93'1 Распознавание образов. Опознание образов

004.93'11 Пространство образов. Включая: Пространство признаков. Расстояние

004.93'12 Различение образов. Установление различий образов

004.93'14 Кластеризация образов

004.932 Обработка изображений

004.932.1 Дискретизация изображений. Сегментация изображений

004.932.2 Анализ изображений

004.932.4 Редактирование изображений. Фильтрация изображений. Восстановление изображений. Повышение качества изображений. Исправление искажений изображений. Сглаживание

004.932.7 Типы изображений

004.932.72 Объекты на изображениях

Например

004.932.72'1 Распознавание объектов

004.932.75 Символы как изображения

Например

004.932.75'1 Распознавание символов

004.934 Распознавание и преобразование речи

004.934.1 Слова в речи

Например

004.934.1'1 Распознавание слов

004.934.2 Анализ речи

004.934.5 Синтез речи

004.934.8 Проблема говорящего

Например

004.934.8'1 Распознавание говорящего

004.94 Имитационное компьютерное моделирование

=> 004.358 Периферия для имитационного моделирования и виртуальной реальности

=> 004.383.4 Моделирующие устройства

004.942 Исследование поведения объекта на основе его математической модели

Сюда относятся вычислительные проблемы имитационного моделирования

=> .001.57 Исследования на моделях. Моделирование

=> 519.876.5 Цифровое имитирование и моделирование систем

004.946 Виртуальная реальность

Приложение 2

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. Библиотековедение

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