РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
ИНСТИТУТ ВЫЧИСЛИТЕЛЬНЫХ ТЕХНОЛОГИЙ

Ю. И. Шокин, А. М. Федотов, В. Б. Барахнин

ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ

Ответственный редактор
доктор технических наук *О. Л. Жижимов*

НОВОСИБИРСК
«НАУКА»
2010

УДК 004
ББК 29.81
УДК Ш78

Шокин Ю.И. Проблемы поиска информации / Ю. И. Шокин, А. М. Федотов, В. Б. Барахнин. Новосибирск: Наука, 2010. — 220 с.
ISBN 918–5–02–018969–0

В монографии рассматриваются вопросы разработки и создания информационно-поисковых систем, способных в автоматизированном режиме извлекать данные из слабоструктурированных электронных документов с целью включения этих документов в научно-информационный процесс и получения новой информации и знаний. Приводится подробное изложение моделей, структур и алгоритмов, описывающих информационно-поисковые системы указанного типа, предназначенные для информационного обеспечения научной деятельности. Представлены результаты использования предложенных информационных моделей и структур при создании целого ряда разделов Информационно-справочной системы Сибирского отделения РАН.

Книга предназначена для специалистов в области информационных технологий, занимающихся вопросами создания информационно-поисковых систем для различных отраслей науки, а также аспирантов и студентов старших курсов.

Табл. 4. Ил. 15. Библиогр.: 267 назв.

Р е ц е н з е н т ы :
доктор физико-математических наук *К. Е. Афанасьев*,
доктор технических наук *О. Л. Жижимов*,
доктор технических наук *М. В. Ульянов*

Утверждено к печати Ученым советом
Института вычислительных технологий СО РАН

ISBN 918–5–02–018969–0

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

Делать что-либо — это нетрудно, трудно только начать.

Кретя Патачкувна. Цит. по книге: «Мысли людей великих, средних и пса Фафика»

Начавшееся в середине 1990-х годов бурное развитие высоких технологий в области передачи и обработки информации, в частности создание современных телекоммуникационных систем (прежде всего сети Интернет), привело к появлению принципиально новых возможностей организации практически всех этапов научно-информационного процесса, что в свою очередь обусловило качественный рост информационных потребностей научных работников.

К наиболее перспективным направлениям развития информационного обеспечения научной деятельности относятся электронные информационные технологии. В данной монографии речь пойдет только о тех способах удовлетворения информационных потребностей научного сообщества, которые базируются на электронных технологиях. В рамках указанного подхода основным инструментом информационного обеспечения научной деятельности являются информационные системы.

В настоящее время научные сообщества наиболее развитых стран и регионов мира обладают достаточно мощными информационными системами, которые в той или иной степени удовлетворяют потребностям исследователей в информации, однако основные недостатки большинства систем — не всегда своевременная актуализация информации (этот недостаток не относится к библиотечным системам) и ограниченность возможностей обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами (иными словами, низкая интероперабельность).

К тому же ограниченность возможностей классических информационно-поисковых систем во многом обусловлена тем обстоятельством, что наука об обработке данных, особенно в ее прикладном аспекте, несколько отстает от соответствующих аппаратно-программных средств. Сами по себе данные (как набор битов) не представляют никакой информационной ценности без соответствующих моделей (феноменологических, информационных, математических и др.). Для возможности эффективного восприятия человеком данных нужно, чтобы они были превращены в «информа-

цию», которая может быть отображена в виде «знаний»[1] — «адекватного отражения действительности в сознании человека в виде представлений, понятии, суждений, теорий» [159].

Преодоление указанных проблем возможно путем создания *интеллектуальных информационных систем*, в качестве составных компонентов которых выступают, наряду с традиционной информационной системой, еще и рассуждающая информационная система (формализующая правила логического вывода), а также интеллектуальный интерфейс (диалог, графика и т. д.), благодаря которому компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека.

Развитие сети Интернет предоставило создателям интеллектуальных информационных систем новые возможности, связанные с одновременным доступом ко множеству разнородных источников данных, что открывает широкие перспективы в развитии более совершенных технологий получения знаний. Однако многие современные исследования в области интеллектуального поиска опираются на неявное предположение о возможности широкого распространения более или менее подробной стандартизации представления информации. Разумеется, реализация подобных проектов, прежде всего концепции Semantic Web консорциума W3, позволила бы вывести работу с информацией на качественно новый уровень, но одна из основных особенностей сети Интернет как феномена цивилизации заключается в том, что развитие сети изначально носит децентрализованный характер, поэтому многие веб-сайты, содержащие важную информацию из той или иной предметной области, не соответствуют рекомендациям консорциума W3. В частности, на большинстве сайтов документы являются *слабоструктурированными*, т. е. хотя и снабженными метаданными, но при этом имеющими неструктурированные элементы.

Таким образом, весьма актуальна задача разработки теоретических основ создания информационно-поисковых систем информационного обеспечения научной деятельности, способных в автоматизированном режиме извлекать метаданные из электронных документов (прежде всего интернет-ресурсов) достаточно произвольной структуры, что позволит получать на основании этих данных новую информацию и знания. Решению сформулированной задачи и посвящена данная монография, содержащая подробное

[1] Подробно о соотношении понятий «данные», «информация» и «знания» речь пойдет в разд. 1.4.

изложение моделей, структур и алгоритмов, описывающих информационно-поисковые системы указанного типа, предназначенные для информационного обеспечения научной деятельности. В книге также представлены результаты использования предложенных информационных моделей и структур при создании целого ряда разделов Информационно-справочной системы Сибирского отделения Российской академии наук.

Особо оговоримся, что содержащийся в монографии библиографический список содержит лишь ссылки на публикации и электронные ресурсы, имеющие наиболее непосредственное отношение к тематике данной монографии. В свое оправдание приведем высказывание из вышедшей в свет в 1974 г. монографии Б. В. Бирюкова «Кибернетика и методология науки» [39]: «“Никто не обнимет необъятного” — в этом изречении Козьма Прутков, наверное, имел в виду сочинителей ученых книг второй половины XX в., периода начавшегося “информационного взрыва”...». За прошедшие с тех пор три с половиной десятилетия ситуация только усугубилась...

Кратко опишем методологическую базу намеченной программы исследований.

Методологические основы информатики как науки о структуре и свойствах научной информации заложены в монографиях сотрудников ВИНИТИ А. И. Михайлова, А. И. Черного, Р. С. Гиляревского, Ю. М. Арского и др. [3, 103–105], а также Б. В. Бирюкова [39].

Методология системного анализа, у истоков которой стояли А. А. Богданов [41] и Л. фон Берталанфи [216, 217], развита в работах М. Месаровича и Я. Такахары [102], В. Н. Садовского [143] и др. Применительно к кибернетическим системам методология системного анализа описана в статье А. А. Ляпунова и С. В. Яблонского [96], а к информационным системам — в работах Ю. А. Шрейдера и др. [200, 205].

Методология автоматизации процессов обработки текстовой информации представлена в работах Дж. Солтона (Г. Сэлтона) [252, 253], а также (с учетом особенностей русского языка) Г. Г. Белоногова и др. [38, 37].

Методология разработки программных систем информационного обеспечения различных аспектов научной деятельности на базе новых интернет-технологий предложена в публикациях авторов данной монографии [170, 191–195]. Близкие к этой методологии подходы рассмотрены В. А. Серебряковым, А. Н. Бездушным и др. [35, 36, 75]; С. В. Мальцевой [98] и др.

Глава 1
ИНФОРМАЦИОННЫЙ ПОИСК: ИСТОРИЯ И ТЕХНОЛОГИЧЕСКИЕ ПОДХОДЫ

Послушайте, ребята,
Что вам расскажет дед.
Земля наша богата,
Порядка в ней лишь нет.
А эту правду, детки,
За тысячу уж лет
Смекнули наши предки:
Порядка-де, вишь, нет.

А. К. Толстой. История Государства Российского от Гостомысла до Тимашева

1.1. Постановка проблемы

Начальник. Вы что, потеряли документ?
Подчиненный. Нет. Но найти пока не могу.

Бюрократический фольклор

Проблема поиска информации — одна из вечных проблем человеческого сообщества. На протяжении своего многотысячелетнего развития его представители неустанно находятся в поиске того, где находится что-либо: *пища, жилище, пастбища, дороги, сокровища* и т. п. Обобщая задачи поиска, можно сказать, что человечество постоянно находится в поиске *знаний*, а в частности «информации о том, где лежат сокровища». Великий аргентинский писатель Хорхе Луис Борхес в своем эссе «Четыре цикла» писал, что в мировой литературе вечными являются четыре темы:

— Падение города.
— Возвращение героя.
— Поиск.
— Самопожертвование бога.

Нетрудно заметить, что наиболее часто встречающейся как в литературе, так и в реальности является третья тема — *поиск*, ибо четвертая тема выходит за рамки обычного человеческого опыта, а две первые проявляются лишь в «минуты мира роковые».

Хотя информационные ресурсы в том или ином виде существовали с доисторических времен, но практически до эпохи Возрожде-

ния они из-за своей специфичности не рассматривались как отдельная экономическая категория, несмотря на то что информация всегда использовалась людьми для управления и решения насущных задач. Однако по мере осознания экономической ценности информации и, следовательно, информационных ресурсов проблема поиска перекочевала и в эту область. Чтобы решить проблему доступа к информации, человечество создало библиотеки как универсальную систему хранения «знаний», их систематизации и каталогизации[1].

Ситуация кардинально начала меняться по мере освоения (точнее, создания) человеческой цивилизацией «информационного» пространства. Первыми островами информационного пространства цивилизации стали общественные библиотеки, крупнейшие из которых (Библиотека Британского музея, Национальная библиотека в Париже, Библиотека конгресса США, Российская государственная библиотека и др.) уже к началу XX в. располагали собраниями в миллионы томов.

Долгое время одним из мощных инструментов поиска информации в книжных хранилищах был непосредственный доступ читателей к книгам, когда они, затрачивая большое личное время, могли свободно рыться в библиотеке. Это и понятно, поскольку человека, нуждающегося в научной информации (в знаниях), интересует прежде всего не сама книга как таковая, а только некоторый ее фрагмент, содержащий требуемые ему знания. Причем сам он часто не в состоянии объяснить, как эти знания могут быть связаны с названием книги или ее автором.

Накопление книг привело к парадоксальному результату, связанному с отделение книжных хранилищ от широкого круга читателей. Универсальный инструмент поиска знаний, основанный на прямом доступе к информации, стал доступен только избранным. Основная масса жаждущих знаний была вынуждена довольствоваться только поиском в каталоге, который не мог удовлетворить возникающие информационные потребности. Для решения проблемы доступа читателей к информации были предприняты попытки

---

[1] Наиболее древний из сохранившихся каталогов — шумерская глиняная плитка со списком более 60 произведений — был создан 4 тыс. лет назад, однако лишь к концу XVIII в. библиотечные каталоги окончательно перестали выполнять функцию только инвентарных описей и превратились в основное средство раскрытия библиотечных фондов [104, с. 160–161].

классификации и систематизации информации — стали создаваться специализированные книжные залы, куда источники информации отбирались исходя из каких-то (не всегда очень ясных) критериев.

С одной стороны, как отметил известный историк и социолог науки Д. де Солла Прайс [247], начиная с середины XVIII в. любой достаточно большой сегмент науки в нормальных условиях растет экспоненциально, т. е. любые параметры науки, включая объем накопленной информации, за определенный промежуток времени удваиваются (закон экспоненциального роста науки). С другой стороны, в указанный период времени происходит увеличения числа людей, нуждающихся в научной информации. Речь идет не только о научных работниках (численность которых тоже подчиняется закону экспоненциального роста), но и о представителях многих других профессий умственного труда: инженерах, агрономах, врачах, управленцах и т. п.

По мере накопления книг, а стало быть, и содержащейся в них информации, возможности традиционных методов поиска: с использованием алфавитного каталога (поиск книги по известному имени автора) и систематического каталога (поиск книги или класса книг по определенному предмету) — перестали удовлетворять читателей, прежде всего научных работников, информационные потребности которых в процессе научного поиска характеризуются невысокой четкостью осознания и выражения (см., например, [3]).

Современные информационные технологии предоставляют исследователю мощный аппарат для «манипулирования данными». Данные, переведенные в электронную форму, приобретают новое качество, обеспечивая более широкое распространение и эффективное использование. На первый взгляд, может сложиться впечатление, что развитие информационных технологий уже само по себе способно вывести работу с научной информацией на качественно новый уровень, но, к сожалению, это совсем не так, поскольку информационные технологии пока не могут предоставить адекватный аппарат для оперирования с «информацией» и информационными ресурсами [171]. Сами по себе данные (как набор битов) не несут никакой информационной ценности без соответствующих моделей: например, А. Н. Колмогоров неоднократно отмечал, что данные представляют информационную ценность лишь тогда, когда они являются составной часть некоторой модели реального мира и свя-

заны с другими данными [83, 84]. Таким образом, применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А. А. Ляпунов (см., например, [95]): «нет модели — нет информации».

Существующую проблему отбора информации уже дано пытаются решить путем создания универсальных или специализированных информационно-поисковых систем. В результате опережающего развития технологий поиска по сравнению с методиками работы с семантической информацией образовался заметный разрыв между техникой работы с данными (поиском) и способностью работать с содержанием, заложенным в этих данных.

1.2. Предыстория

> Век живи — век учись! и ты наконец достигнешь того, что, подобно мудрецу, будешь иметь право сказать, что ничего не знаешь.
>
> *Козьма Прутков. Плоды раздумья*

Проблема поиска и доступа к информации является одной из серьезных проблем, с которой столкнулось современное «информационное общество».

По всей видимости, впервые возникшую проблему наиболее четко осознал бельгийский социолог Поль Отле, который в конце XIX — начале XX века [246] предложил дополнить науку[1], ведавшую научно-технической информацией, и традиционное библиотековедение совершенно новым методом, названным им *«Документацией»*:

«Цели Документации состоят в том, чтобы суметь предложить документированные ответы на запросы по любому предмету в любой области знания: 1) универсальные по содержанию; 2) точные и истинные; 3) полные; 4) оперативные; 5) отражающие последние данные; 6) доступные; 7) заранее собранные и готовые к передаче; 8) предоставленные как можно большему числу людей» (см. [117, с. 190]).

---

[1] Термин «информатика» (фр. informatique), принадлежавший когда-то скромной науке, ведавшей именно информацией, в основном научно-технической, родился в 1960 г. Он происходит от французских слов information (информация) и automatique (автоматизация) и дословно означает «информационная автоматизация».

Суть метода *Документации* заключалась в том, что содержание книги (отчуждаемое от автора) заносится на карточку, причем совокупность карточек можно упорядочивать так, чтобы при этом отражались предметные связи. Поль Отле предвидел революционное развитие технологий работы с информацией вплоть до ее мультимедийного представления и удаленного доступа к банкам данных: «человеческое знание позволит создать оборудование, действующее на расстоянии, в котором соединятся радио, рентгеновские лучи, кинематограф и микроскопическая фотография. Все предметы Вселенной, все предметы, созданные Человеком, будут регистрироваться на расстоянии с момента их создания. Тем самым будет создан движущийся образ мира — его память, его подлинная копия. Любой человек сможет прочесть отрывок, спроецированный на его личный экран» (см. [117, с. 16]).

Идеи Поля Отле не были восприняты тогдашним информационным (библиотечным) сообществом, в частности потому, что они совершенно не были подкреплены техническим обеспечением: информационные работники и библиотекари той эпохи располагали лишь пишущими машинками, фотоаппаратами и карточными каталогами. Появление после Первой мировой войны устройств обработки перфокарт (точнее, их простейшей разновидности — перфокарт с краевой перфорацией) также не стало принципиальным технологическим прорывом, поскольку даже спустя 40 лет, в 1960-е годы, подобные устройства могли обрабатывать сравнительно небольшие (до 30 тыс.) массивы документов (см. [104, с. 549]).

Проблема нарастающих объемов информации, грозивших захлестнуть читателей, продолжала волновать исследователей. В 1941 г. упомянутый Х. Л. Борхес создает свою знаменитую притчу «Вавилонская библиотека». В этой притче Вселенная представляется в виде Библиотеки, беспредельной и всеобъемлющей, на полках которой «можно обнаружить все возможные комбинации двадцати с чем-то орфографических знаков (число их, хотя и огромно, не бесконечно) или все, что поддается выражению — на всех языках». Философский смысл притчи, конечно же, гораздо глубже проблемы информационного поиска, но исходный образ взят автором из повседневной реальности[1]. Трудно удержаться, чтобы не

[1] Борхес был профессиональным библиотекарем (библиографом) и даже одно время занимал пост директора Национальной библиотеки Аргентины.

привести хотя бы краткие выдержки из притчи, соответствующие тематике книги.

«Когда было провозглашено, что Библиотека объемлет все книги, первым ощущением была безудержная радость. Каждый чувствовал себя владельцем тайного и нетронутого сокровища. Не было проблемы — личной или мировой, для которой не нашлось бы убедительного решения... Вселенная обрела смысл, вселенная стала внезапно огромной, как надежда. В это время много говорилось об Оправданиях: книгах апологии и пророчеств, которые навсегда оправдывали деяния каждого человека во вселенной и хранили чудесные тайны его будущего. Тысячи жаждущих покинули родные шестигранники и устремились вверх по лестницам, гонимые напрасным желанием найти свое Оправдание..., но те, кто пустился на поиски, забыли, что для человека вероятность найти свое Оправдание или какой-то его искаженный вариант равна нулю...

На смену надеждам, естественно, пришло безысходное отчаяние. Мысль, что на какой-то полке в каком-то шестиграннике скрываются драгоценные книги и что эти книги недосягаемы, оказалась почти невыносимой. Одна богохульная секта призывала всех бросить поиски и заняться перетасовкой букв и знаков, пока не создадутся благодаря невероятной случайности канонические книги. Другие, напротив, полагали, что прежде всего следует уничтожить бесполезные книги...

Известно и другое суеверие того времени: Человек Книги. На некоей полке в некоем шестиграннике (полагали люди) стоит книга, содержащая суть и краткое изложение всех остальных: некий библиотекарь прочел ее и стал подобен Богу. В языке этих мест можно заметить следы культа этого работника отдаленных времен. Многие предпринимали паломничество с целью найти Его. В течение века шли безрезультатные поиски. Как определить таинственный священный шестигранник, в котором Он обитает? Кем-то был предложен регрессивный метод: чтобы обнаружить книгу А, следует предварительно обратиться к книге В, которая укажет место А; чтобы разыскать книгу В, следует предварительно справиться в книге С, и так до бесконечности...»

Движущей силой произошедшей в середине XX в. «информационной революции» стали не хранители информации — библиотечные работники, а ее потребители — ученые и инженеры. В 1931 г. в Германии была создана статистическая машина Эммануэля

Гольдберга [186], обеспечивавшая чтение специальным образом подготовленной микропленки, на которой хранился массив документов. Особенность организации хранения информации заключалась в том, что на пленку вместе с микрофильмированным документом заносилось описание этого документа, закодированное посредством перфорации. Поиск документа осуществлялся путем сравнения запроса (также закодированного) с перфорацией пленки. Машину Гольдберга отличало высокое качество механики и оптики: пользователь имел возможность просматривать за час более 100 тыс. кадров 35-миллиметровой пленки. Статистическая машина Гольдберга была, по-видимому, первым действующим инструментом, позволяющим автоматизировать поиск в больших массивах данных по их разметке. Кстати сказать, по мнению некоторых исследователей, на идеи Эммануэля Гольдберга опирался Вэннивер Буш, автор знаменитой статьи «Пока мы мыслим» («As we may think») [219], написанной в 1939 г., в которой сформулирована идея гипертекста и предсказано появление персонального устройства, хранящего информацию и автоматизирующего процесс ее поиска. Вот как выглядит одна из его идей:

«Обсудим устройство персонального назначения. Пусть оно называется Memex и представляет собой что-то вроде автоматизированного архива или библиотеки. Memex хранит для своего хозяина все нужные книги, записи, корреспонденцию. Прибор автоматизирован до такой степени, что дает ответы на вопросы, заданные в простой форме, т. е. очень гибок в общении.

Скорость ответов высока и не заставляет ждать. Имеется графический экран, клавиатура и кнопки управления. Когда пользователь ищет нужную книгу, он должен ввести ее мнемонический код и нажать нужную для поиска кнопку. Перед ним на экране появится первая страница. Должна быть возможность листать книгу в любом направлении. Можно будет остановиться на выбранной странице, а потом пойти по ссылке и найти следующий интересующий материал. При этом всегда можно вернуться к предыдущей странице или одновременно рассматривать несколько страниц.

Появятся энциклопедии с готовыми ссылками для связывания информации и быстрого поиска. Их можно будет загружать в Memex и искать все, что нужно».

Нередко в литературе можно встретить высказывания, что В. Буш предсказал идею персонального компьютера, но это не со-

всем правильно, ибо фактическое время написания статьи «As we may think» относится к тому периоду, когда под руководством В. Буша в Массачусетском технологическом институте был создан действующий макет микрофильмового селектора «Memex» [104].

Продолжая разговор о поисковых устройствах той эпохи, следует отметить реализованную на суперпозиционных перфокартах систему поиска патентов, которую в 1939 г. создал У. Баттен для британского концерна «Imperial chemical industries, Ltd». Ее алгоритм работы основан на координатном индексировании — представлении содержания документа при помощи списка содержащихся в нем ключевых слов. Эта идея получила дальнейшее развитие в работах американского математика Кельвина Муэрса, создавшего и запатентовавшего в 1947 г. систему механизированного поиска документов, работавшую на особых картах с вырезами вдоль краев (так называемых «Zato-картах»).

В основе системы также лежал метод координатного индексирования. Именно К. Муэрс стал основоположником научного подхода к информационному поиску, введя в 1950 г. термины «информационный поиск», «информационно-поисковая система», «информационно-поисковый язык», «поисковый образ», «дескриптор», «дескрипторный словарь» и др. С этого времени началось активное развитие информатики как науки о структуре и свойствах семантической информации (прежде всего научной). Важное место в этой науке занимали вопросы информационного поиска, в процессе выполнения которого, собственно говоря, и происходит непосредственное удовлетворение информационных потребностей пользователя. Обобщение накопленных результатов проведено в монографии сотрудников Всесоюзного института научной и технической информации (ВИНИТИ) [104], описавших методологические основы теоретической информатики.

Возможности практической реализации алгоритмов информационного поиска резко расширились, когда в середине 1960-х — начале 1970-х годов вместо механических устройств стали достаточно широко применять электронно-вычислительные машины, на базе которых создавались автоматизированные системы сбора, анализа, классификации, хранения, передачи на расстояние, поиска и выдачи информации. В частности, исследовательская группа под руководством профессора Гарвардского университета Дж. Солтона разработала систему анализа и извлечения текста SMART (Salton's

Magic Automatic Retriever of Text), в которой были впервые реализованы многие базовые принципы современных поисковых систем. Теоретическое описание и осмысление этих принципов проведено Дж. Солтоном в монографии [253], причем особый акцент в ней был сделан на изложении новых подходов к вопросам классификации документов и запросов, анализе содержания, интерактивного поиска и выдачи информации. Эта книга и до сих пор не потеряла своей актуальности.

Технологической основой создания подобных информационно-поисковых систем было использование мэйнфреймов — многопользовательских централизованных вычислительных систем, в которых массивы данных и программы их обработки располагались на мощной центральной ЭВМ, а пользовательский доступ осуществлялся посредством алфавитно-цифровых терминалов (дисплеев), работающих под управлением машин-сателлитов. Бытует мнение, что информационно-поисковые системы того времени не получили должного развития из-за недостаточной мощности и памяти тогдашних ЭВМ и отсутствия качественных каналов связи (особенно дальней). Здесь проблемы были несколько другие. Во-первых, отсутствие универсальных сетевых протоколов сильно ограничивало удаленный доступ к таким системам. Во-вторых, большая загрузка вычислительными задачами не позволяла организовать работу таких систем в режиме реального времени. Все это придавало информационно-поисковым системам преимущественно локальный характер.

Несмотря на это, в информационных системах того времени был собран и систематизирован колоссальный объем информации. Например, в Новосибирском ВЦ СО РАН на машинах типа БЭСМ-6 хранилась вся подписка реферативных журналов ВИНИТИ, библиографические описания изданий, поступавших в ГПНТБ, и большое количество научно-технической документации. Основные проблемы, связанные с ее использованием, — это отсутствие интерактивной работы, поскольку, как правило, запрос посылался с терминала, а ответ приходил в виде «километровой» распечатки на АЦПУ. И это была жизненная необходимость, поскольку анализировать ответ за дисплеем не представлялось никакой возможности. Вторая проблема была связана с «безбумажной» визуализацией материала — практически отсутствовало программное обеспече-

ние, позволявшее просматривать информацию в близком к печатному изданию виде.

В 1980-е годы мир начали завоевывать персональные компьютеры, которые позволяли обрабатывать информацию непосредственно на рабочем месте, без связи с центральным процессором, а кроме того, обладали достаточно мощными (по тем временам) средствами визуализации информации. Это обусловило существенное снижение интереса к созданию централизованных информационных систем и, как следствие, приостановку фундаментальных научных исследований в области информационного поиска, которые возобновились лишь с появлением сети Интернет, приведшим к распределенному хранению информации.

1.3. Современные проблемы создания и функционирования информационно-поисковых систем научной тематики

> Гений мыслит и создает. Человек обыкновенный приводит в исполнение. Дурак пользуется и не благодарит.
>
> *Козьма Прутков. Из «Сборника неоконченного (d'inachevé)»*

Проблема поиска информации имеет особенную значимость применительно к деятельности научного исследователя. Важность комплексного исследования проблем, связанных с информационным обеспечением научной деятельности, была осознана отечественным научным сообществом еще в начале 1950-х годов, когда по представлению Академии наук СССР был создан Институт научной информации, ныне Всероссийский институт научной и технической информации (ВИНИТИ). Книги сотрудников этого института «Основы научной информации» [105], «Основы информатики» [104], «Научные коммуникации и информатика» [103] заложили основы информатики как науки о структуре и свойствах научной информации, а также о закономерностях научно-информационной деятельности, а монография «Инфосфера: Информационные структуры, системы и процессы в науке и обществе» [3] отразила

достижения и проблемы информатики по состоянию на середину 1990-х годов.

Однако начавшееся в середине 1990-х годов бурное развитие высоких технологий в области передачи и обработки информации, в частности создание современных телекоммуникационных систем (прежде всего сети Интернет), привело к появлению принципиально новых возможностей организации практически всех этапов научно-информационного процесса, что в свою очередь обусловило качественный рост информационных потребностей научного сообщества, ибо «потребности социальных субъектов (личностей, социальных групп)... зависят от уровня развития данного общества, а также от специфических социальных условий их деятельности» [208].

Кроме того, за указанный период времени в России произошла смена общественно-экономической формации, приведшая к изменению принципов функционирования и финансирования науки, что также не могло не сказаться на характере информационных потребностей ученых.

Следовательно, возникает необходимость комплексного анализа информационных потребностей научного сообщества с учетом влияния как новых возможностей, открывшихся благодаря революции в области информационных технологий, так и изменившихся условий функционирования науки.

Разумеется, ни в коей мере нельзя полагать, что классические способы удовлетворения информационных потребностей посредством получения информации на бумажных носителях, общения на конференциях и т. п. ушли в прошлое, однако наиболее перспективным направлением развития информационного обеспечения научной деятельности являются все-таки электронные информационные технологии. В данном исследовании мы будем вести речь только о тех способах удовлетворения информационных потребностей научного сообщества, которые базируются на электронных технологиях. В рамках указанного подхода основным инструментом информационного обеспечения научной деятельности являются *информационные системы*, т. е. системы обработки данных о какой-либо предметной области [156].

В настоящее время научные сообщества наиболее развитых стран и регионов мира обладают достаточно мощными информационными системами. Например, в Европе функционирует интегрированная система ERGO [228], являющаяся частью проекта

CORDIS [223] (об используемых в проекте стандартах см. [220]), среди американских разработок своими масштабами выделяется информационная система Библиотеки конгресса США [239], а в числе наиболее известных отечественных информационных систем можно назвать Единое научное информационное пространство (ЕНИП) РАН [61], Информационную систему «База данных организаций и сотрудников СО РАН» [6], систему «Информика» [55], Университетскую информационную систему РОССИЯ [168], Научную электронную библиотеку eLIBRARY [109], Соционет [158]. Эти системы в той или иной степени удовлетворяют потребности исследователей в информации, однако каждая из них страдает определенными недостатками.

Во-первых, существенной проблемой большинства программных систем информационного обеспечения научной деятельности, предназначенных для функционирования в течение неопределенно долгого времени, является недостаточно своевременная актуализация информации (исключение составляют лишь библиотечные системы). Причина возникновения этой проблемы очевидна: недостаток средств, прежде всего, для оплаты труда лиц, которые должны отслеживать изменения информации, а также предъявляемые к этим лицам высокие квалификационные требования, возрастающие с усложнением структуры и возможностей поддерживаемой информационной системы. В частности, опыт выполнения интеграционных проектов СО РАН, в рамках которых создавались программные системы той или иной научной тематики, показал, что такие системы могут развиваться лишь в случае актуализации содержащейся в них информации самими пользователями этих систем. Наиболее эффективная реализация подобных проектов возможна в том случае, когда «черновая» информационная работа, неизбежная при каталогизации электронных документов научной тематики, составлении тезаурусов предметной области и т. п., в значительной степени автоматизирована посредством использования соответствующих программных средств, притом основную долю функций контроля качества полученной информации способен выполнить даже лаборант и лишь в редких случая требуется корректировка результатов с участием эксперта — научного работника.

К сожалению, задача автоматизации вовлечения электронных документов в научно-информационный процесс всё еще далека от

сколько-нибудь удовлетворительного решения. Одна из основных причин сложившейся ситуации заключается в том, что с появлением в конце 1970-х годов персональных компьютеров появились мощные средства визуализации информации, вследствие чего были почти остановлены научные изыскания в области теории создания информационно-поисковых систем, которые возобновились лишь в середине 1990-х в связи с развитием информационных технологий сети Интернет и перехода к распределенному хранению информации. В настоящее время в указанной области получены важные результаты (см. монографии [234, 258] и др.), однако эти разработки обычно опираются на неявное предположение о возможности широкого распространения более или менее подробной стандартизации представления информации, например на основе словарей (концепция Semantic Web консорциума W3 [256]). К тому же наработки консорциума W3 носят лишь рекомендательный характер, а объявить их стандартами могут только организации, имеющие соответствующий статус, такие как ISO, ГОСТ или ANSI, поэтому реальное развитие большинства ресурсов Интернета, в том числе научной направленности, идет без учета подобных необязательных рекомендаций. Более того, свободный характер размещения материалов в сети Интернет превращает требование соблюдения даже обязательных стандартов представления информации всего лишь в благое пожелание (особенно это касается российской части Интернета).

Одним из наиболее неприятных следствий описанной ситуации является сложность поиска информации, содержащейся в текстовых документах сети Интернет. Это относится даже к традиционным методам поиска, характерным для библиотек: поиск по имени автора документа, названия документа или тематический поиск, поскольку *слабоструктурированный* электронный документ (т. е. документ, снабженный *метаданными*[1], но при этом имеющий неструктурированные элементы) может не содержать явно заполненных соответствующих полей метаданных, причем классификационные признаки документа зачастую вообще отсутствуют. Разумеется, обработка слабоструктурированных документов не может быть полностью автоматизирована, и основная задача разработчи-

[1] Метаданные («данные о данных») — структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими [259]. Подробнее о метаданных см. разд. 1.9.

ков соответствующих программных средств состоит в уменьшении необходимого участия человека в процессе контроля за качеством обработки информации.

Так как пользователи, принимающие участие в актуализации информации, могут находиться в разных регионах России и даже мира, то становится очевидным экономическая нецелесообразность использования коммерческих программных пакетов, предназначенных для частичной автоматизации процесса каталогизации электронных документов, создания и расширения тезаурусов (онтологий) и т. п., поскольку необходимость установки таких пакетов на компьютерах всех специалистов, поддерживающих данную информационную систему и при этом работающих в разных организациях (или использование сетевых версий, рассчитанных на большое число пользователей), связано с немалыми финансовыми затратами. Поэтому становится актуальной задача разработки и реализации алгоритмов, автоматизирующих основные этапы научно-информационного процесса (включая создание тезаурусов и онтологий), посредством интернет-приложений, доступных с любого компьютера сети (разумеется, после аутентификации и авторизации пользователя-эксперта).

Во-вторых, построение масштабных информационных систем для поддержки научной деятельности требует распределенного хранения информации. В частности, относительно систем научно-организационной направленности, создаваемых в рамках одной большой научной корпорации (например Сибирского отделения РАН), можно сделать вывод, что «эффективная эксплуатация информационных ресурсов возможна только в том случае, когда они постоянно поддерживаются авторами» [66]. Таким образом, информационная система научной корпорации должна строиться как объединение информационных систем отдельных организаций. В свою очередь, информационная система каждой организации состоит из нескольких разнородных подсистем (кадровая, библиографическая и т. п.).

Отсюда неизбежно возникает проблема *интероперабельности*, т. е. обеспечения взаимодействия разнородных информационных источников (как с целью их непосредственной интеграции, так и для организации поиска по однотипным подсистемам различных информационных систем). Теоретические вопросы интероперабельности обсуждаются, например, в работах [36, 179]. Коротко

резюмируя их содержание, можно отметить, что организация в них поиска обеспечивается посредством согласования схем метаданных (*семантическая интероперабельность*). Для интеграции разнородных систем, а также разнородных ресурсов внутри каждой отдельно взятой системы (что необходимо для извлечения из содержащихся в информационной системе данных новой информации и знаний) требуется согласование как моделей данных и форматов их представления (*синтаксическая интероперабельность*), так и протоколов доступа к ресурсам (*техническая интероперабельность*).

Наконец, **в-третьих**, при создании информационных систем зачастую недостаточное внимание уделяется вопросам организации взаимодействия разрабатываемой системы с людьми — потребителями информации. Как утверждал А. А. Ляпунов [95], «информация всегда относительна, она зависит от того, какой информационной системой она воспринимается». Разработчикам программных средств обработки данных зачастую недостает понимания того обстоятельства, что конечная цель работы, связанной с применением информационных технологий, — понимание того или иного явления (т. е. возможность извлечения из информации знаний, определяемых [3] как структурированная (связанная причинно-следственными и иными отношениями) информация), а не получение каких-либо чисел, гистограмм, отдельных фактов и т. п.

Изложенное, в частности, означает, что предполагаемая возможность извлечения из содержащихся в информационной системе данных новой информации и знаний влечет за собой необходимость наличия связей между документами, *содержащими упоминание* тех или иных сущностей, с документами, *описывающими* эти сущности. Например, необходима связь имен собственных (как элементов библиографического описания и т. п.) с информацией о конкретных носителях этих имен, ибо в противном случае имя несет лишь назывную, но не информационную функцию [103].

Более того, информационные потребности научных работников на этапе научного поиска и изучения имеющихся в данной области результатов характеризуются невысокой четкостью осознания и выражения (см., например, [3]). Возникает необходимость оснащения информационных систем функцией поиска «по аналогии», т. е. нахождения по данному документу (или множеству документов) класса документов, схожих с ним по содержанию.

Если же говорить об атрибутивном поиске, то на практике большинство рядовых пользователей испытывает затруднения в самостоятельном построении запросов более сложных, нежели простой контекстный поиск, даже если им предоставлен удобный интерфейс, не требующий непосредственного использования языка запросов. Трудности возникают на уровне понимания схем данных и использования логических операторов, без которых немыслимы более или менее сложные запросы. Поэтому необходимо, чтобы рядовой пользователь информационной системы имел возможность получить интересующую его информацию посредством элементарных действий (навигации), при этом квалифицированным пользователям должны быть предоставлены дополнительные сервисы, отвечающие современным технологическим требованиям.

Комплексное решение указанных проблем возможно путем создания интеллектуальных информационных систем [3], в качестве составных компонент которых входят, наряду с традиционной информационной системой, еще и рассуждающая информационная система (формализующая правила логического вывода), а также интеллектуальный интерфейс (диалог, графика и т. д.), благодаря которому компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека. При этом крайне важно, чтобы создаваемые системы могли обрабатывать в автоматизированном режиме слабоструктурированные документы.

К тому же следует учитывать, что широта и многогранность информационных потребностей научного сообщества (см., например, [22]) вызывает необходимость массового создания информационных систем, разнообразных как по тематике, так и по целевому назначению, что приводит к необходимости систематического изучения всех стадий процесса разработки информационных систем, включающего стадии создания концептуальной модели, информационной модели и практической реализации системы.

Таким образом, в настоящее время возникла насущная необходимость осмысления процесса обработки компьютерной информации как технологии (см., например, [197]). Заметим, что аналогичное осмысление вычислительного моделирования было осуществлено в начале 1980-х годов в работах Н. Н. Яненко [209] и А. А. Самарского [153] и стало важной вехой в развитии прикладной математики.

1.4. Уточнение используемой терминологии на основе семиотического подхода

> Если бы комплименты были правдой, это были бы не комплименты, а информация.
>
> *Из книги Крети Патачкувны «Моя кибернетика». Цит. по книге: «Мысли людей великих, средних и пса Фафика»*

В соответствии с [162] будем понимать под технологией совокупность методов обработки, изготовления, изменения состояния, свойств и формы сырья, материалов или полуфабрикатов в процессе производства продукции. Разумеется, одним из важнейших свойств технологии является ее воспроизводимость (это вытекает, например, из определения технологии как научной дисциплины, согласно которому технология изучает различные закономерности, действующие в технологических процессах [162]). Иными словами, любая технология по своей сути — воспроизводимый инструмент, применяемый для превращения потребляемых факторов в продукцию или для достижения планируемых результатов [64].

Сошлемся еще на одно, пожалуй, наиболее краткое из определений технологии: «технология — способ преобразования данного в необходимое» (см., например, [163]), которое подтверждает, что применительно к поставленной задаче по-настоящему технологичным можно назвать лишь тот подход, который способен «перерабатывать» максимально широкие пласты интернет-ресурсов научной тематики.

Что же выступает исходным материалом для технологии переработки информации? Ответ, на первый взгляд, очевиден: сама информация. Однако и на вопрос о конечном продукте напрашивается тот же ответ! Конечно, человек, владеющий теоретическим основами информатики, после некоторого размышления ответит, что исходным материалом служат данные, а конечным продуктом — знания (или, по крайней мере, семантическая информация). Тем не менее описанная коллизия показывает, что проблемы возникают уже на терминологическом уровне.

Поскольку существует множество подходов к понятию «информация» с философских, социологических, биологических, физических, математических или кибернетических позиций (см. [3, с. 393]), включая так называемую «техническую» теорию информа-

ции, которая является, по сути, теорией передачи и хранения данных, постольку можно обнаружить десятки порой противоречащих друг другу определений того, что является информацией или знанием. Даже специалисты по информатике, работающие в разных ее областях, например научно-технической информации и экспертных систем, вкладывают в термин «знания» несколько разный смысл (ср., в частности, [3] и [50]). При этом в трактовании термина «данные» (понимаемые как факты и идеи, представленные в формализованном виде [156]) столь значительных расхождений обычно не наблюдается, что позволяет рассматривать информационные ресурсы (в широком смысле) как совокупность данных, организованных для эффективного получения достоверной информации.

Вряд ли существует некая «абсолютная» точка зрения, с которой возможно было бы судить о том, какое из многочисленных определений понятий «информация» или «знание», «тезаурус» или «онтология» и т п. является «более правильным». Поэтому цель проводимого ниже экскурса в историю терминологии (см. также статью авторов [31]) заключается, прежде всего, в том, чтобы уточнить соответствующие определения применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний, установив при этом основания выбора определений, принятых именно в этой области.

Еще в конце 1960-х годов в информатике был принят подход (см. [104]), восходящий к определению Л. Бриллюэна [218], согласно которому информация «...есть сырой материал и состоит из простого собрания данных, тогда как знание предполагает некоторое размышление и рассуждение, организующее данные путем их сравнения и классификации» (нетрудно заметить, что такой подход основан на работах К. Шеннона — создателя статистической теории информации, связывавшего это понятие с мерой неопрделенности). Однако уже к середине 1970-х годов созрело понимание (см., например, работу П. Чена [221] и приведенную в ней библиографию), что содержательная ценность данных, не сопровождаемых семантикой в виде модели предметной области, которую эти данные описывают, весьма невысока, т. е. такие данные фактически не являются информацией в традиционно-узком (не шенноновском) значении этого слова.

Решение этой проблемы возникло на пути применения в информатике методов *семиотики* — общей теории знаковых систем.

Об использовании семиотических методов указывалось в [104], но лишь применительно к построению формализованных информационных языков. Однако связь семиотики и информатики гораздо глубже. В классических работах [54, 218] уже присутствовало осознание того обстоятельства, что ценность информации (в отличие от ее количества) зависит от субъекта, ее воспринимающего, но поскольку семиотические аспекты этого вопроса явно ускользали от внимания авторов, он не мог найти решения и оставался на уровне постановки. Так, в работе [54] приводится следующий пример (который цитируется и в [104]): «...Сообщение о том, родила ли жена Джона Смита мальчика или девочку, содержит столько же двоичных единиц информации, сколько и сообщение о том, кого родила ваша жена. Вместе с тем последнее сообщение представляет для вас неизмеримо большую ценность, чем первое». Вполне очевидно, однако, что два указанных сообщения имеют куда большее сходство, чем простое равенство количества битов.

Реальное осознание сложностей в преодолении разрыва между шенноновским понятием информации и концепцией семантической информации как средства социальной коммуникации возникло в середине 1960-х годов: см., например, работы Ю. А. Шрейдера [201–203] и У. Шрамма [254], причем Ю. А. Шрейдер показал, что о количестве семантической информации в данном сообщении есть смысл говорить лишь применительно к конкретному приемнику сообщения.

В монографии Б. В. Бирюкова [39] показана ограниченность классической теории информации применительно к вопросам, связанным с человеческой психикой, познанием, явлениями смысла и ценности. Вместе с тем в ней отмечено, что в определенных условиях шенноновские понятия могут быть с успехом использованы в социогуманитарных исследованиях.

Попытку «телеологического» описания особенностей восприятия сообщения субъектом предприняли Р. Акофф и Ф. Эмери [210]. Они предложили классифицировать сообщения по видам изменений в получателе, которые делятся на несколько типов (при этом сообщение может принадлежать сразу к нескольким типам):

1) *информация* (изменения вероятности выбора);
2) *инструкция* (изменения в эффективности выбора);
3) *мотивация* (изменения в удельных ценностях).

Были приведены формулы для количественной оценки каждой из названных величин, однако практическая ценность этих формул оказалась не слишком высокой ввиду чрезвычайной сложности реальной оценки изменения состояния субъекта — носителя интеллекта (семиотико-информационные оценки количества информации были предложены также в диссертационной работе работе Е. В. Ляпунцовой [97]). Основная заслуга Р. Акоффа и Ф. Эмери состоит в том, что они, по-видимому, одними из первых специалистов в области информатики обратили внимание на многоуровневость восприятия сообщения получателем и сделали попытку описать эти уровни.

Наконец, в начале 1980-х годов немецкий исследователь В. Гитт разработал пятиуровневую модель [229], наиболее полно отражающую различные аспекты термина «информация». Структура модели представлена на рис. 1.1.

Анализируя эту модель, нетрудно видеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих — семиотической триаде (синтактика — семантика — прагматика), а верхний уровень носит, скорее, философский характер. При этом наличие в некотором сообщении ин-

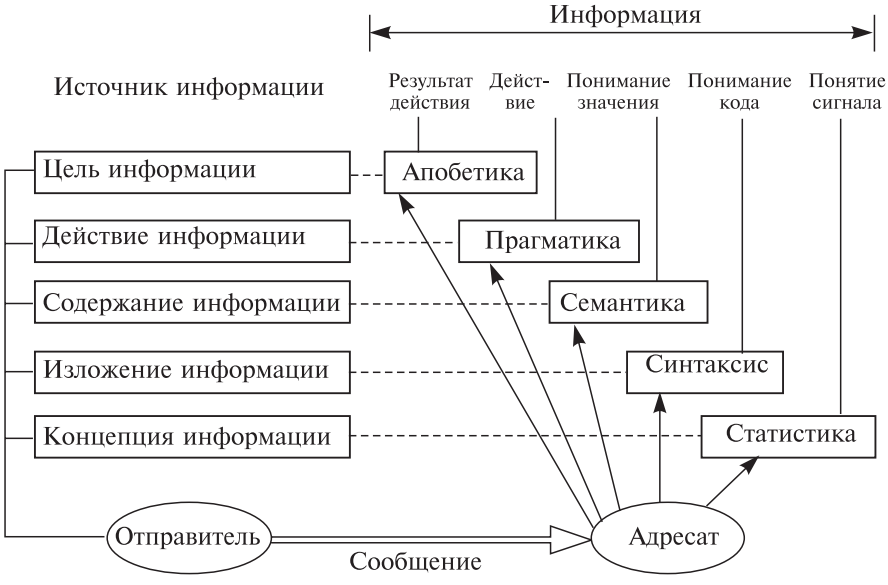

Рис. 1.1. Пятиуровневая модель информации.

формации высокого уровня влечет за собой наличие информации всех низших уровней, но, разумеется, не наоборот (еще раз напомним: объем информации зависит от характеристик адресата, причем это касается всех уровней информации).

Следует отметить, что модель В. Гитта не получила широкого распространения (во многом потому, что он пытался с ее помощью, делая акцент на пятый уровень, доказать невозможность самопроизвольного возникновения такой сложной информации, как генетический код, что явно противоречит общепринятым в современной науке представлениям).

Идеи, весьма близкие к тем, которые воплощены в модели В. Гитта, однако в несколько менее стройной форме, изложены в монографии Ю. А. Шрейдера и А. А. Шарова [205], изданной в 1982 г.

Таким образом, с начала 1980-х годов семиотическая триада заняла прочное место в кибернетике, о чем свидетельствуют соответствующие статьи в «Словаре по кибернетике» [156], хотя в первое время семиотическая терминология применялась, скорее, при описании языка (понимаемого как частный случай знаковой системы) в целом, нежели при анализе отдельных сообщений. К настоящему моменту описание непосредственно информации с помощью семиотической терминологии получило широкое распространение в отечественной литературе как научной, так и учебной (см., например, монографию К. Е. Афанасьева и Л. Е. Шмаковой [5]).

Важно подчеркнуть, что семиотический подход фактически использован при определении базисных понятий в фундаментальной монографии [3], изданной ВИНИТИ. *Данные* понимаются в ней (в соответствии с традиционным подходом) как факты и идеи, представленные в символьной форме, позволяющей проводить их передачу, обработку и интерпретацию, а *информация* — как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная (связанная причинно-следственными и иными отношениями) информация, образующая систему, составляет *знания*.

Исходя из этого понимания терминов «данные», «информация», «знания», которого мы будем придерживаться в дальнейшем, можно сказать, что *данные соответствуют синтаксическому уровню сообщения, информация (в узком смысле!) — семантическому, а знания — прагматическому*.

Среди новейших исследований семиотических оснований информатики можно выделить работы И. М. Зацмана [69, 70]. По мнению автора, сообщение может относиться к одной из трех основных сред, соответствующих семиотической триаде:

1) *цифровая среда* (двоично кодируемые объекты которой соотносятся со знаками, их формами и значениями);

2) *среда социальных коммуникаций* (объекты которой сенсорно воспринимаемы человеком);

3) *ментальная среда знаний человека*.

Названные среды достаточно четко разграничены (т. е. конкретные сообщения или сведения принадлежат только одной определенной среде). С учетом этого автор формулирует определения базисных терминов информатики, а также типологизирует основные классы элементарных технологий обработки информации, переводящие сообщение (сведения) из одной среды в другую.

Изложенная концепция позволила И. М. Зацману сделать интересные выводы гносеологического и общеметодологического характера, однако в ее рамках довольно сложно обосновать методы и алгоритмы, предназначенные для «массовой» автоматической переработки информации.

Следующей важной задачей является установление определенности в понимании и разграничении использования терминов «тезаурус» и «онтология». Изначально в информатике использовался лишь термин «тезаурус», пришедший из лингвистики, где он с начала XIX в. означал особый тип словаря, в котором понятию ставилось в соответствие слово (в том числе *различные* слова), его обозначающие. Наибольшую известность получил составленный в 1852 г. англичанином П. М. Роже́ «Тезаурус английских слов и фраз» [250], который в 1956 г. был использован в работах по автоматическому переводу Кембриджской группой по исследованию языка [122, с. 215]; в 1957 г. Ч. Берньер предложил применять словари тезаурусного типа для информационного поиска [215].

С этого времени термин «тезаурус» прочно вошел в профессиональную лексику специалистов в области информатики, причем определения тезауруса несколько варьировались в зависимости от класса задач, для решения которых предназначался тезаурус. В частности, применительно к задачам информационного поиска под тезаурусом обычно понимается так называемый *нормативный* тезаурус (см. [104, с. 432]) — словарь-справочник, содержащий все

лексические единицы информационно-поискового языка — дескрипторы (вместе с ключевыми словами, которые в пределах данной информационно-поисковой системы считаются синонимами этих дескрипторов), причем дескрипторы в словаре должны быть систематизированы по смыслу, а смысловые связи между ними эксплицитно выражены.

Однако в 1990-х годах в информатике, наряду с термином «тезаурус», стал употребляться близкий по смыслу термин «онтология». Первоначально этот термин, заимствованный из философии, появился в среде специалистов по искусственному интеллекту. Наиболее широко известно следующее определение Т. Грубера [231]: «Онтология — это явная спецификация концептуализации» (т. е. абстрактного представления предметной области). Довольно быстро термин «онтология» перешел и в другие области информатики, включая разработку информационных систем, и стал означать [114] «способ, который используется для описания некоторой области знаний..., в частности базовых понятий этой области, их свойств и связей между ними». В настоящее время, как отмечено в [58], под онтологией нередко понимается широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [265]:

1) словарь с определениями;
2) простая таксономия;
3) тезаурус (таксономия с терминами);
4) модель с произвольным набором отношений;
5) таксономия и произвольный набор отношений;
6) полностью аксиоматизированная теория.

В работах многих авторов термин «онтология» начал употребляться вместо термина «тезаурус» (что, в общем, неудивительно, ибо определения онтологии в той или иной степени сходны с определением тезауруса, а первоначальное значение термина «онтология» — «учение о бытии», звучит куда более многообещающе, чем заурядный «тезаурус» — «запас»).

Возникла ситуация, когда разными терминами стали называть один и тот же объект. Попытка разрешения коллизии сделана в работах А. С. Нариньяни [106, 107], причем в основе проделанного анализа лежит семиотическая методология. Подчеркивая, что «еще недавно сегодняшняя онтология именовалась тезаурусом», автор предлагает следующее разграничение этих понятий: «тезаурус

скорее более закреплен за лексикой в проекции на семантику, а онтология в ее новом, информационном употреблении — это семантика и прагматика, возможно до известной степени в проекции на язык», причем система сущностей, описываемая онтологией, должна быть связана универсальными зависимостями типа «общее — частное», «часть — целое», «причина — следствие» и т. п. [106].

Развивая и уточняя эту концепцию, А. С. Нариньяни показывает в [107], что соотношение между понятиями

Модель предметной области — Онтология — Тезаурус

симметрично известному треугольнику Г. Фреге:

Сущность — Понятние — Слово,

т. е. онтология рассматривается как общая часть модели предметной области и тезауруса, связывающая знания о мире со знаниями о языке, причем полноценный тезаурус невозможен без онтологии, поскольку она, пусть даже в простейшей форме, является скелетом всякой системы данных и (или) знаний.

Заметим, что ранее подобные функции приписывались непосредственно тезаурусу. Так, еще в середине 1970-х годов Н. С. Панова и Ю. А. Шрейдер отмечали [121], что тезаурус можно интерпретировать как запас семантической информации, содержащейся в документах на данную тему, т. е. как описание структуры знаний. Аналогичный подход используется и некоторыми современными исследователями. В частности, в диссертации С. В. Жмайло [67] показано, что информационно-поисковый тезаурус можно с достаточным основанием считать моделью предметной области знаний, или базой знаний, при этом подчеркивается, что в тезаурусе обязательно должны быть эксплицированы парагматические отношения (синонимы, квазисинонимы и т. п.), а также отношения и связи «род — вид», «часть — целое» и др.

Обобщая изложенное, можно сделать следующий практический вывод: *тезаурус становится онтологией тогда, когда связи между дескрипторами не просто эксплицированы* (как это предусмотрено в классическом определении [104]), *но и классифицированы*. В дальнейшем, если особо не оговорено иное, мы будем употреблять термин «онтология» именно в этом смысле. Следует отметить, что многие тезаурусы, например по науковедению и лексикографии, созданные С. Е. Никитиной в [111], или по безопасности инженерных

систем, созданный С. В. Жмайло [67], ввиду своей структурной сложности могут быть охарактеризованы как онтологии.

Наконец, уточним определение основного объекта исследования данной главы — *интеллектуальной информационной системы*.

Под *информационными системами* понимают комплексы аппаратно-программных средств для обработки данных, структурированных при помощи той или иной формальной модели [156]. В 1960–1980-е годы наибольшее распространение получили *информационно-поисковые системы* (ИПС), осуществляющие поиск, переработку и хранение информации. При этом имеются важные ограничения, связанные со сложностью информационно-поисковых систем как снизу, так и сверху.

С одной стороны, под общее определение информационно-поисковой системы формально подпадают и такие комплексы аппаратно-программных средств, в которых реализованы лишь простейшие поисковые запросы типа выдачи документа по его известному имени. Однако выведение подобных комплексов за рамки информационно-поисковых систем сделано еще в 1960-е годы в классической монографии А. И. Михайлова и др. «Основы информатики» [104], где подчеркнуто, что устройства и машины, предназначенные лишь для отыскания документов по известным адресам их хранения, информационно-поисковыми системами не являются. Такой подход соответствует фундаментальным положениям работы А. А. Ляпунова и С. В. Яблонского [96], согласно которой характерной чертой управляющих (в широком понимании этого термина) систем, требующих специальных кибернетических рассмотрений, является их сложность, проявляющаяся в большом количестве элементов, сложной системе связей, больших потоках информации и реализации сложных функций. Подчеркивание этой особенности информационных систем характерно и для англоязычной терминологии (см., например, словарь Вебстера [264]).

С другой стороны, автоматизированные системы, способные строить даже простые категорические силлогизмы, отнесены в монографии А. И. Михайлова и др. «Научные коммуникации и информатика» [103, с. 149–150] к особому классу *информационно-логических систем*), отличному, согласно «Словарю по кибернетике» [156], не только от класса информационно-поисковых систем, но и от автоматизированных информационных систем в целом. Таким образом, *ИПС выдавали в качестве продукта переработки дан-*

ных именно информацию, но не знания, полностью оправдывая свое название (следует отметить, что в [156] допускается возможность наличия и в обычной ИПС некоторых простейших видов логического и эвристического вывода, но степень «простоты» этих правил не конкретизирована).

Развитие алгоритмических, программных и аппаратных средств информатики привело в 1980-е годы к возможности создания *интеллектуальных информационных систем* (см., например, [52, 92, 141]), в которых компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека. Интеллектуальные системы (ИнтС) функционируют по следующей схеме [3]:

$$\boxed{\text{ИнтС}} = \boxed{\text{РИС}} + \boxed{\text{ИПС}} + \boxed{\text{ИнИн}}, \qquad (1.1)$$

где РИС — рассуждающая информационная система (формализующая правила логического вывода), а ИнИн — интеллектуальный интерфейс (диалог, графика и т. д.). При этом ИПС как подсистема ИнтС должна обладать механизмом поиска как фактов, так и документов.

Более развитые ИнтС должны обладать и механизмом пополнения базы данных, функционируя по схеме

$$\overline{\text{ИнтС}} = \overline{\text{РИС}} + \overline{\text{ИПС}} + \overline{\text{ИнИн}} + \overline{\text{АП}}, \qquad (1.2)$$

где АП — автоматическое извлечение фактов из текстов и соответствующее пополнение базы данных посредством этих фактов и выводов из них (подробнее об автоматизированном выводе из фактов см., например, [116]).

Таким образом, интеллектуальная система обладает по сравнению с обычной ИПС новыми возможностями, позволяя удовлетворить квалифицированного пользователя в соответствии со схемой «документ — факт — рассуждение[1]» [3, с. 343], т. е. согласно приведенным выше определениям, *интеллектуальные информационные системы позволяют не только извлекать из данных информацию, но и получать новые знания*.

При этом следует отметить, что мы придерживаемся принятого в информатике понимания факта как единичного значения данных, поскольку, как подчеркнуто в [103, с. 138], «объектом сбора, хране-

[1] Под рассуждением здесь понимается логический вывод.

ния, поиска и выдачи в так называемых фактографических информационно-поисковых системах... могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать... любой фрагмент такого текста».

На основании изложенного можно сделать вывод, что функционирование интеллектуальной информационной системы основано на двух противоположных процессах: *при пополнении ИнтС новыми сведениями происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс — извлечение из данных нужной пользователю информации и знаний.*

Далее вместо термина «интеллектуальная информационная система» там, где это не вызовет недоразумения, мы будем для краткости использовать термины «информационная система», употребляя термин «информационно-поисковая система» как обозначение соответствующего компонента интеллектуальной информационной системы.

1.5. Общие принципы организации информационно-поисковых систем

При виде исправной амуниции
Как презренны все конституции!
Ф. К. Прутков. Военные афоризмы

Как уже отмечалось, созданные в трудах К. Муэрса и Дж. Солтона фундаментальные основы поиска информации актуальны и по сей день. Однако здесь есть небольшой нюанс в использовании терминологии этих трудов. Следует подчеркнуть различие между поиском как автоматизированной процедурой и выделением требуемой информации в найденных документах. Суть различий состоит в следующем.

— Выделение информации — это деятельность человека, использующего поисковую машину. Она является интерактивной, итерационной и связана с другими видами интеллектуальной деятельности человека.

— Читатель ищет не документы как таковые, а содержащуюся в них информацию для каких-то собственных целей (обучения, принятия решений и др.).

— Читатель нуждается в доступе к разным источникам данных, чтобы получить всеобъемлющее представление об объекте поиска.

— Какими бы совершенными ни были аппаратное и программное обеспечение, используемые человеком, они остаются инструментами, а интеллект является атрибутом Читателя.

В трудах упомянутых выше «классиков» употребляется термин Information Retrieval System (IRS). В 1950–1970-е годы англоязычный термин Information Retrieval (IR) переводили на русский язык как «информационный поиск», и соответственно системы данного класса называли информационно-поисковыми. В этих системах использовались ручные процедуры индексирования документов и создания тезаурусов. Но, что чрезвычайно важно, такие системы предназначались для *выделения* информации (именно информации и именно выделения) из разных документов. «Выделение» — это более точное значение слова retrieval. Сейчас в энциклопедиях IR определяется как искусство и наука поиска информации в документах и поиска собственно документов и их описаний в базах данных (в том числе сетевых). Подмножеством IR является выделение информации в тексте (Text Retrieval, TR) и выделение информации в документах (Document Retrieval, DR).

Наиболее радикальный этап «информационный революции» начался в 1990-е годы. Он был связан с созданием WWW-сервиса сети Интернет, а также с по-настоящему массовым распространением мощных и недорогих персональных компьютеров, благодаря чему пользователи получили доступ к ресурсам этого сервера. Именно WWW-сервис, отличающийся от печатных изданий оперативностью размещения и доставки информации практически любого характера, а от классических электронных СМИ — возможностью передачи печатного текста, делает все более реальной перспективу создания единого информационного пространства человеческой цивилизации.

В настоящее время Интернет — главный источник электронных документов. Количество документов в сети поддается лишь косвенным, притом явно заниженным оценкам. Так, по состоянию на начало августа 2005 г. число документов, проиндексированных поисковой системой Yahoo, превысило 20 млрд, из них 19,2 млрд — текстовые документы, 1,6 млрд — изображения и около 50 млн —

аудио- и видеофайлы [244]. При этом, разумеется, нельзя утверждать, что Yahoo индексирует все интернет-документы[1].

Активно развивается и российский сектор сети Интернет — так называемый Рунет. По данным исследования, проводимого компанией Яндекс [78], на осень 2009 г. в Рунете зафиксировано около 15 млн сайтов, что составляет примерно 6,5 % от всего объема Интернета. Только в текстовом формате (без учета графики, видео и др.) в Рунете размещено более 140 тыс. гигабайт информации.

Однако такое обилие потенциально доступных документов сделало особенно актуальной задачу предоставления пользователям сети адекватных средств информационного поиска, без которых Интернет мог бы превратиться в реальное воплощение «Вавилонской библиотеки» из притчи Борхеса. Говоря о средствах информационного поиска в сети Интернет, обычно подразумевают *поисковые системы*, предоставляющие возможность поиска информации по всему Интернету (по крайней мере, по всем www-страницам). Такие системы известны всем пользователям Интернета: это Google, Yahoo, Yandex и др. Однако для поиска документов, относящихся к той или иной предметной области, пользователи Интернета нередко обращаются к тематическим каталогам интернет-ресурсов — структурированным наборам ссылок на документы соответствующей тематики.

Чтобы описать принципы работы средств информационного поиска, необходимо, прежде всего, уточнить соответствующую терминологию. Основные термины и определения в области поиска и распространения информации с помощью автоматизированных информационных систем, а также информационно-поисковых языков регламентированы официальными документами Российской Федерации (действующими и в большинстве других стран СНГ): государственными стандартами ГОСТ 7.73–96 «Поиск и распространение информации» и ГОСТ 7.74–96 «Информационно-поисковые языки».

Итак, согласно определению ГОСТ 7.73–96, *информационно-поисковая система (ИПС)* — это совокупность справочно-информационного фонда и технических средств информационного поиска в нем. В свою очередь, *справочно-информационный фонд (СИФ)* — это совокупность *информационных массивов* (т. е. упо-

[1] Интернет-документом будем называть любой документ, опубликованный в сети Интернет.

рядоченных совокупностей документов[1], фактов или сведений о них) и связанного с ними *справочно-поискового аппарата* (т. е. данных об адресах хранения документов с определенными поисковыми образами документа). Наконец, согласно определению ГОСТ 7.74–96, *поисковый образ документа* — это текст, состоящий из лексических единиц *информационно-поискового языка* (т. е. специального формализованного искусственного языка), выражающий основное смысловое содержание документа и предназначенный для реализации информационного поиска. Процесс выражения содержания документа на информационно-поисковом языке называется *индексированием*.

Заметим, что под содержанием документа в данном контексте обычно подразумевают не только более или менее краткое изложение того, о чем повествует документ, но и его библиографическое описание: название документа, фамилии его авторов, выходные данные и т. п. Еще раз напомним, что совокупность извлекаемых в процессе индексации характеристик документа вместе с формальным описанием структуры этих характеристик обычно называют метаданными.

Структурирование метаданных призвано облегчить поиск документов, ибо одно и то же слово (например, «Пушкин») может входить в список авторов документа, в его заглавие, в аннотацию или даже в выходные данные (город Пушкин в Ленинградской области как место издания документа). Эти случаи могут быть разграничены именно благодаря структурированию метаданных.

Нетрудно понять, что документ становится доступным для поиска с помощью той или иной информационно-поисковой системы, если его метаописание (т. е. совокупность метаданных) попадает в справочно-информационный фонд этой системы. Но каким образом осуществляются поиск и индексация интернет-документов, заносимых в СИФ? Так, поисковые системы сети Интернет используют поисковые роботы (их англи. название «crawler», т. е. «червяк», буквально «ползун»), которые последовательно просматривают интернет-документы, переходя от одного к другому посредством гиперссылок, и извлекают их метаданные. Разумеется, поисковые роботы периодически просматривают и документы, уже занесенные в СИФ информационной системы, чтобы установить, сущест-

[1] Подчеркнем, что в этой главе слово «документ» используется в общеупотребительном смысле; формализация этого понятия будет проведена в разд. 1.9.

вуют ли они в настоящее время и не претерпели ли каких-либо существенных изменений. При составлении тематических каталогов интернет-ресурсов также зачастую используются поисковые роботы, которые, однако, собирают данные о документах лишь с сайтов соответствующей тематики. Сетевые имена таких сайтов, как правило, указываются экспертами в данной предметной области, при этом допускается и непосредственное занесение экспертами сведений об отдельных интернет-документах. Наконец, некоторые специализированные информационно-поисковые системы создаются исключительно вручную, при этом размер их поисковых массивов может быть весьма внушителен. Например, одна из крупнейших баз данных научных публикаций Web of Science [263] содержит (по состоянию на июль 2009 г.) более 46 млн записей. Крупнейшая российская научная электронная библиотека eLIBRARY.RU [109] включает в себя (по состоянию на июль 2010 г.) рефераты и полные тексты более 12 млн научных статей и публикаций. Весьма популярная в среде математиков база данных журнала «Zentralblatt MATH» [266] содержит (по состоянию на июль 2010 г.) почти 3 млн записей — библиографических сведений (включая довольно подробные аннотации) о математических публикациях, вышедших в свет за последние полтора века.

Но все-таки справочно-информационные фонды большинства информационно-поисковых систем, работающих с электронными документами, пополняются не вручную, а с помощью тех или иных программ, автоматизирующих поиск и индексацию документов. И здесь-то, в процессе индексации документа, проявляется основная проблема использования таких программ: автоматическое структурирование метаданных оказывается весьма непростой задачей. Чтобы убедиться в этом, достаточно просмотреть небольшое число интернет-документов, например научной тематики. Можно легко увидеть, что в некоторых случаях фамилии авторов пишутся перед названием документа, а в некоторых, наоборот, после названия. Каким образом программа должна определять, что именно заносить в поле «авторы» данного документа, а что — в поле «название»? Заметим, что простейшие варианты решения этой проблемы (типа «дополнить индексирующую программу словарем фамилий») оказываются малоэффективными. И дело не только в необходимости огромного (и не существующего на практике) объединенного словаря фамилий разных наций с вариантами транскрипций на

других языках. Проблема состоит еще в том, что многие фамилии (особенно в языках со слабовыраженным изменением словоформ при помощи окончаний) совпадают с «обычными» словами языка. Кроме того, фамилия может являться названием документа, например книги или статьи биографического характера.

Наличие указанных проблем привело к тому, что обычной практикой универсальных поисковых систем является представление поискового образа документа в виде неструктурированного набора *ключевых слов* — информативных слов, приведенных к стандартной лексикографической форме. *Информативными словами*, согласно ГОСТ 7.74–96, называются слова, словосочетания или специальные обозначения в тексте документа (или запроса), выражающие понятия, существенные для передачи содержания документа. Конкретные критерии включения слова или словосочетания к множеству информативных слов зависят от вида ИПС. Так, в универсальных поисковых системах в качестве информативных рассматриваются практически все слова, включая служебные. Напротив, в специализированных информационно-поисковых системах, для которых набор ключевых слов — один из компонентов структуры метаданных документа, множество информативных слов обычно строится на основе предметного указателя соответствующей предметной области (содержащего наряду с одиночными словами и весьма сложные словосочетания), в то время как слова, относящиеся к «общеупотребительной» лексике, в число информативных не включаются.

Ввиду совершенно очевидных преимуществ структурированного описания документа перед неструктурированным (о чем уже упоминалось выше), организации, пытающиеся выступать в качестве «законодателя мод» в сети Интернет, прежде всего консорциум W3C, неоднократно предпринимали попытки предоставить создателям интернет-документов возможность *явно* указывать значения основных элементов метаданных документа, что позволило бы значительно повысить эффективность функционирования поисковых роботов. Так, еще в середине 1990-х годов в спецификации языка гипертекстовой разметки документов HTML было четко прописано, что каждый HTML-документ обязан иметь ровно один элемент TITLE («название») в поле HEAD («заголовок»). Более того, в описании языка HTML появился элемент META, предназначенный для записи парных элементов вида NAME:CONTENT («назва-

ние : значение»), описывающих свойства данного документа: фамилия автора, список ключевых слов и т. п.

Заметим, однако, что спецификация языка HTML не предусматривала каких-либо *конкретных* названий для обозначения элементов, содержащих информацию о фамилии автора, ключевых словах и пр. Поэтому даже при наличии в индексируемом документе элементов META задача автоматического определения его структуры оставалась трудноразрешимой. Наиболее известным подходом к ее решению стал предложенный в 1995 г. на семинаре, проводившемся Национальным центром суперкомпьютерных приложений (NSCA) в г. Дублин (штат Огайо, США), базовый набор из 15 полей метаданных, предназначенный для описания ресурсов, публикуемых в Интернете. В этот набор вошли такие общие свойства документов, как название, дата публикации, автор, издатель, владелец. Таким образом, в любом документе должно было существовать ядро метаданных, о которых заранее известно, как их следует интерпретировать. Эти предложения были опубликованы под рабочим названием Dublin Core metadata, которые впоследствии стали фундаментом проекта Dublin Core Metadata Initiative [227].

Названные идеи получили развитие и в проекте Semantic Web, суть которого заключается в создании сети документов, содержащих метаданные «исходных» документов сети Интернет и существующей параллельно с ними. Эта «параллельная» сеть предназначена специально для построения поисковыми роботами (и другими интеллектуальными агентами) однозначных логических заключений о свойствах «исходных» документов. Основные принципы создания Semantic Web (до практической реализации которой, впрочем, еще очень далеко) основаны на повсеместном использовании, во-первых, универсальных идентификаторов ресурсов (URI) посредством расширения этого понятия на объекты, недоступные для скачивания из Интернета (персоны, географические сущности и т. п.), а во-вторых, онтологий (т. е. формальных моделей описания тех или иных предметных областей) и языков описания метаданных.

К сожалению, ни один из перечисленных подходов не стал по-настоящему широко распространенным. В этом без труда можно убедиться, просмотрев произвольный набор интернет-документов. Почти наверняка в большинстве из них будут отсутствовать

элементы META, содержащие фамилии авторов, список ключевых слов и т. п. Причины сложившейся ситуации широко обсуждаются в интернет-сообществе, но, несомненно, к числу основных причин относится «человеческий фактор».

Во-первых, ввиду широкой распространенности интернет-технологий теоретическая подготовка многих создателей интернет-ресурсов оставляет желать лучшего, и они зачастую просто не задумываются о назначении элемента META в языке HTML. Во-вторых, явное указание значений метаданных — процесс весьма трудоемкий, поэтому даже те создатели ресурсов, которые знают о технологии метаданных, не всегда считают нужным тратить время и силы на работу с ними, тем более что разработчики универсальных поисковых систем, исходя из описанной ситуации, не слишком-то полагаются на возможность автоматического получения структурированного поискового образа индексируемого документа, ибо процент документов, подробно описанных создателями, весьма невелик. В итоге складывается своеобразный порочный круг, который в ближайшее время вряд ли будет разорван.

В несколько лучшем положении находятся создатели тематических каталогов интернет-ресурсов, поскольку количество организаций, работающих в той или иной области человеческой деятельности, а также веб-сайтов, публикующих действительно ценную и/или новую информацию соответствующей тематики, как правило, довольно невелико. Важно отметить, что реальные технологии создания подавляющего большинства сайтов таковы, что однородные документы с одного сайта имеют практически одинаковую HTML-разметку. При этом неважно, генерируются ли документы динамически (в этом случае однородность разметки — естественное следствие работы соответствующей программы) или же они создаются вручную посредством создания копии уже имеющегося документа с последующей заменой текста (что также сохраняет разметку). Данное обстоятельство позволяет автоматизировать процесс индексации метаданных электронного документа посредством указания шаблона документов того или иного сайта, т. е. явного указания команд (тегов) языка HTML, обрамляющих основные характеристики документа: авторы, название, ключевые слова, аннотация, коды того или иного классификатора и т. п. [30].

1.6. Составление поисковых предписаний

Жена посылает математика за продуктами.
— Сходи в магазин и купи батон колбасы. Да, если там будут яйца, возьми десяток.
Математик послушно приходит в магазин и спрашивает у продавщицы:
— Скажите, у вас яйца есть?
— Да, есть, — говорит она.
— Тогда дайте мне десяток батонов колбасы.
Математический фольклор. Цит. по книге:
С. Н. Федин. Математики тоже шутят

Из предыдущего раздела мы получили некоторые сведения о том, как устроен справочно-информационный фонд ИПС. Чтобы сделать запрос, мы должны составить поисковый образ запроса, т. е. его формальное представление в терминах информационно-поискового языка. После этого составляется *поисковое предписание*, включающее поисковый образ запроса и указания о логических операциях, подлежащих выполнению в процессе информационного поиска. ИПС сравнивает поисковое предписание с хранящимися в ее справочно-поисковом аппарате поисковыми образами документов (при этом в большинстве поисковых систем ключевые слова по умолчанию приводятся к стандартной лексикографической форме) и выдает сведения: адреса хранения и, как правило, краткие описания — о документах, поисковые образы которых соответствуют (т. е., фактически, не противоречат) поисковому предписанию.

Например, поисковое предписание для ИПС электронного магазина, торгующего мужскими костюмами, может выглядеть примерно так:

(рост = 176) и (размер = 104) и ((цвет = ‘черный’) или (цвет = ’темно-синий’)) и (страна-производитель = не ‘Китай’) и (цена < 7000 руб.)

При этом, коль скоро не указаны значения таких элементов метаданных, как материал и тип костюма (пара или тройка), то подразумевается, что пользователя устраивают любые значения этих элементов метаданных.

Простейшая формальная модель с использованием структурированных метаданных документов выглядит следующим образом.

Пусть в справочно-поисковом аппарате ИПС хранится информация о документах d_i. При этом любой документ d_i представляется как $d_i = < m_i^{j,k} >$, где $m_i^{j,k}$ — принадлежит множеству значений элементов метаданных M^j, k — количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов, используемых для составления поискового предписания (с учетом заданных логических операций). Для фиксированного элемента метаданных M^j, где $M^j \subset M_C$, множество документов разбивается на классы эквивалентности, соответствующие различным значениям этого элемента метаданных.

Будем считать два документа *толерантными*, если у них совпадает значение хотя бы одного из элементов метаданных, входящих в M_C (напомним, что толерантность — отношение, которое обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности). Каждое такое значение порождает класс толерантности [204].

Рассмотрим всевозможные сочетания значений элементов метаданных, входящих в M_C. Множества документов, обладающих одинаковым набором значений, суть ядра толерантности, которые служат классами эквивалентности на множестве документов.

Таким образом, поисковое предписание, содержащее подмножество метаданных, определяющее набор классификационных признаков, с указанием сочетания значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос.

К сожалению, в ИПС общего назначения поисковые образы документов, как уже отмечалось, структурированы весьма слабо. Обычно пользователь таких систем имеет возможность включить в поисковый образ запроса (точнее, в ту его часть, которую описывает *содержание* требуемого документа) лишь ключевые слова или словосочетания, указав при этом, где именно они должны содержаться: в заголовка веб-страницы или в ее тексте. Остальные поля в форме поискового запроса касаются языка документа, региона расположения сервера размещения документа, формата файла и т. п., т. е. не имеют непосредственного отношения к содержанию документа.

Впрочем, построение более или менее сложного поискового предписания способно вызвать затруднение у большинства рядовых пользователей, даже если им предоставлен удобный интерфейс, не требующий непосредственного использования языка запросов, включающего логические предикаты. Трудности возникают на уровне понимания схем данных и использования логических операторов. В частности, преподавательский опыт одного из авторов показывает, что даже студенты старших курсов, специализирующиеся в области информатики, при выполнении задания типа «*сделать запрос, выдающий данные за 3 и 5 октября*», нередко связывают даты логическим оператором «И».

Развитыми возможностями построения поисковых предписаний обладают, как правило, специализированные ИПС, справочно-информационный фонд которых содержит хорошо структурированные поисковые образы документов, причем возможности поискового интерфейса напрямую зависят от априорно оцениваемой возможности построения рядовыми пользователями сложных логических запросов. Так, в уже упоминавшейся базе данных журнала «Zentralblatt MATH», предназначенной для профессиональных математиков, функция «Расширенный поиск» позволяет соединять в поисковом предписании при помощи логических связок до 5 значений элементов медатанных (притом сами эти элементы, с возможными их повторениями, выбираются пользователем самостоятельно из общего списка), дополнительно указывая тип искомого документа и временной интервал его публикации.

И все же нельзя не отметить, что умение формально записать поисковый запрос, пусть и весьма сложный, — дело, собственно говоря, не слишком сложное, требующее лишь известного опыта и небольших технических навыков. Гораздо нетривиальнее задача правильно выразить свою информационную потребность, т. е. *неформально* задать «характеристики предметной области, значения которых необходимо установить для выполнения поставленной задачи в практической деятельности» (ГОСТ 7.73–96).

Наиболее простая ситуация возникает, когда пользователь хочет найти адрес хранения конкретного документа. В этом случае задание в поисковом предписании в качестве ключевых слов имени автора документа и его названия, как правило, позволяет довольно быстро добиться нужного результата, даже если ИПС не предоставляет возможности структурировать вхождение перечисленных

ключевых слов применительно к соответствующим полям метаданных. В последнем случае наибольшие проблемы могут возникнуть, если искомый документ относится к разряду хрестоматийных (как, например, «Гамлет» У. Шекспира, «Фауст» И.-В. Гёте или «Евгений Онегин» А. С. Пушкина) и существует масса документов, просто *упоминающих* о нем. Один из эффективных приемов решения подобной проблемы состоит в дополнении поискового предписания какой-либо достаточно длинной *цитатой* из текста (по возможности, не самой общеупотребительной)[1].

Однако на практике пользователю обычно требуется найти не какой-то конкретный, заранее известный документ, а некие *сведения* (*факты*), знание которых необходимо для решения поставленной задачи (или же для удовлетворения любопытства). Возникающая при этом ситуация напоминает сюжет известной русской сказки «Пойди туда — не знаю куда, принеси то — не знаю что» (впрочем, подобные сказки известны в фольклоре многих народов мира — от Ирландии до Китая [108]), причем акцент ставится на первой части фразы, поскольку о том, что́ именно ему нужно, пользователь все-таки имеет некоторое представление. Сказочного Федота-стрельца вел к цели волшебный мячик. А как же следует составить поисковый запрос, чтобы скорее достигнуть поставленной цели?

«Лобовая атака» в форме постановки прямого запроса типа «*Какова девичья фамилия жены М. Е. Салтыкова-Щедрина?*» обычно не приведет к желаемому результату, поскольку современный уровень развития поисковых систем общего назначения не предполагает диалога с пользователем на естественном языке. Отметим, что поставленный вопрос — не совсем тривиальный, ибо ответы на «совсем тривиальные» вопросы типа «*Где родился М. Е. Салтыков-Щедрин?*» поисковые системы обычно все-таки находят, поскольку подавляющее большинство биографий писателя начинаются примерно так: «М. Е. Салтыков-Щедрин родился в январе 1826 г. в селе Спас-Угол Тверской губернии» (слово «где» как служебное поисковой системой во внимание обычно не принимается). Кроме того, создатели некоторых веб-страниц, содержащих часто разыскиваемую в Сети информацию (обычного не научного, а «бытового» характера),

[1] Здесь уместно напомнить английский анекдот о пожилой американке, заявившей после прочтения «Гамлета»: «Не понимаю, почему все так восхищаются этой пьесой: ведь она вся состоит из затертых цитат!».

иногда включают предполагаемый вид пользовательского запроса (точнее, вопроса) в поисковый образ документа.

Более надежным способом составления поискового предписания является включение в поисковый образ запроса ключевых слов (или словосочетаний), которые, по мнению пользователя, непременно должны входить в текст документа, содержащего нужные сведения. Однако здесь возникает следующая дилемма: если включить в поисковый запрос небольшое количество «наиболее вероятных» слов, то его результатом будут сотни (а то и тысячи) документов, далеко не все из которых будут содержать ответ именно на поставленный вопрос. Если же включить в запрос много «предполагаемых» ключевых слов (или даже целую фразу), то мы рискуем получить на выходе пустое множество документов, поскольку авторы документов требуемой тематики могли описывать интересующий пользователя предмет фразами, несколько отличающимися от заданной в запросе.

Итак, в процессе поиска документов, содержащих некие интересующие нас факты, стоит задача сформулировать поисковое предписание таким образом, чтобы получить в результате его выполнения не пустое множество документов, в котором процент «нужных» документов как можно более велик. Это резко повышает шансы сократить количество документов, просмотренных «впустую», т. е. прежде чем мы наткнемся на «нужный» документ. Проблемы, связанные с получением количественных оценок эффективности поиска, будут рассмотрены далее.

1.7. Оценка эффективности поиска

> Семинар по алгебре у программистов. Преподаватель пишет на доске уравнение: $\sin X = 0$.
>
> — Кто из вас может найти X?
>
> Один из студентов выбегает к доске и радостно тычет пальцем в формулу:
>
> — Да вот же он, вот X!
>
> *Математический фольклор. Цит. по книге:*
> С. Н. Федин. Математики тоже шутят

Два основных понятия, в которых дается оценка эффективности поиска, определены в ГОСТ 7.73–96, причем эти определения остались практически неизменными с 1960-х годов (см. [104]): *релевантными* называются документы, содержание которых соответ-

ствует информационному запросу, а *пертинентными* — содержание которых соответствует информационной потребности. Разумеется, два этих понятия хотя и близки, но отнюдь не эквивалентны. Источник появления в выдаче нерелевантных документов — ошибки в описаниях и программном коде поисковых систем, а также прочие организационно-технические причины. При этом в тех случаях, когда поиск производится путем задания конкретного поискового запроса, возможно объективно судить о релевантности того или иного документа, вошедшего в выдачу, поскольку причиной выдачи нерелевантных документов (совокупность которого называется поисковым шумом) являются погрешности в индексировании документов (ручном или автоматическом), проявляющиеся, например, во внесении в поисковый образ документа «лишних» слов. Такая ситуация может возникнуть не только в результате явных ошибок, но и «языковых коллизий». Например, слова «вино» и «вина» имеют в некоторых падежах совпадающие словоформы, вследствие чего в поисковый образ документа, содержащего выражение «*в вине*», при автоматическом индексировании (которое, как правило, не сопровождается семантическим анализом текста) будут включены оба названных слова. Тем самым при включении в поисковый запрос слова «вино» будут выданы в том числе документы, содержащие слово с начальной формой «вина», которые являются, вообще говоря, нерелевантными. Обратите внимание, что при построении примера мы не могли ограничиться простыми омонимами, поскольку, например, при запросе «*лук*» релевантными будут документы как об оружии, так и о растении.

Что же касается пертинентности, то понятие это — сугубо субъективное, поскольку потребности (не обязательно информационные) разных людей, пусть даже и выраженные одними и теми же словами-запросами, могут быть весьма различны. Так, потребность в супе с точки зрения среднестатистического русского удовлетворяется посредством щей или борща, а с точки зрения среднестатистического француза — посредством супа-пюре.

Уже из этого примера видно, что пертинентность выдачи может быть повышена путем коррекции поискового предписания, формулируемого в соответствии с предполагаемым пониманием соответствующей потребности информационной системой (или, если угодно, разработчиками системы). Яркой иллюстрацией этого тезиса служит известный анекдот, в котором на вопрос пролетавших над не-

знакомой местностью воздухоплавателей: «*Где мы находимся?*» прохожий-математик дал абсолютно релевантный, но не пертинентный ответ: «*В корзине воздушного шара*». Конечно, объектом шутки здесь является буквализм математика, но ведь именно такое поведение характерно и для компьютерных алгоритмов. Поэтому правильно сформулированый запрос типа: «*Каковы наши географические координаты?*» или (если уж ориентироваться как на буквалиста, так и на обычного прохожего): «*Вблизи какого населенного пункта мы пролетаем?*» мог бы привести к пертинентному ответу.

В заключение перечислим основные количественные характеристики информационного поиска:

— коэффициент полноты: отношение числа найденных релевантных документов к общему числу релевантных документов, имеющихся в информационном массиве:

$$Recall = |D_{\text{rel}} \cap D_{\text{retr}}| / |D_{\text{rel}}|,$$

где D_{rel} — множество релевантных документов в информационном массиве, а D_{retr} — множество найденных документов;

— коэффициент точности: отношение числа найденных релевантных документов к общему числу документов в выдаче:

$$Precision = |D_{\text{rel}} \cap D_{\text{retr}}| / |D_{\text{retr}}|;$$

— коэффициент шума: отношение числа нерелевантных документов в выдаче к общему числу документов в выдаче:

$$Noise = |D_{\text{nrel}} \cap D_{\text{retr}}| / |D_{\text{retr}}|,$$

где D_{nrel} — множество нерелевантных документов в информационном массиве.

Заметим, что ни точность, ни полнота, взятые отдельно, не гарантируют высокого качества поиска. Так, выдача всех документов, имеющихся в информационном массиве, даст значение коэффициента полноты, равное 1, но точность при этом будет невысокой. Напротив, если выдан только один документ, и притом релевантный, то коэффициент точности равен 1, но при большом количестве ненайденных релевантных документов коэффициент полноты будет очень мал. Чтобы соблюсти баланс между полнотой и точностью, на практике используют так называемую *F*-меру (меру Ван Ризбергена), являющуюся средним гармоническим полноты и точности:

$$F = 2 \times Recall \times Precision \ / \ (Recall + Precision).$$

1.8. Поиск документов «по аналогии»

Если позавчера твой муж вернулся домой вчера утром, вчера вернулся сегодня, то можешь быть уверенной, что сегодня вернется завтра.

Из полезных советов Крети Патачкувны.
Цит. по книге: «Мысли людей великих, средних и пса Фафика»

1.8.1. Постановка проблемы

Кабы схемку иль чертеж,
Мы б затеяли вертеж,
Ну а так — ищи сколь хочешь,
Черта лысого найдешь!
Где искать и как добыть
То-Чаво-Не-Может-Быть?

Л. А. Филатов. Про Федота-стрельца,
удалого молодца

До сих пор мы рассматривали ситуацию, когда поисковый образ запроса задается пользователем как некое «идеальное представление» о поисковом образе искомого документа. Однако, как уже отмечалось, информационные потребности научных работников, когда они в процессе исследования находятся на этапах изучения уже имеющихся в данной области результатов и научного поиска, характеризуются невысокой четкостью осознания и выражения. Опять-таки имеет место ситуация «Пойди туда — не знаю куда, принеси то — не знаю что», но теперь акцент ставится уже *на второй части* фразы, поскольку известно, что описания документов, относящихся к той или иной научной тематике, заносятся в соответствующие реферативные базы данных, в частности:

— базы данных российских реферативных журналов;

— базы данных «Current Contents»;

— специализированные сетевые базы данных типа «Zentralblatt MATH».

С другой стороны, у каждого исследователя за годы его работы образуется картотека библиографических описаний статей, книг и т. д., представляющих для него интерес. Основной критерий их отбора — личные интересы ученого. В настоящее

время такие картотеки хранятся, как правило, на электронных носителях, что позволяет организовывать интегрированные картотеки путем объединения ресурсов совместно работающих исследователей. Так как документы в такие картотеки заносятся из различных источников, имеющих различные классификаторы, а иногда и не имеющих их вообще (что характерно, например, для многих научных журналов), то поиск и классификация документов по формальным классификационным признакам зачастую невозможны.

Таким образом, возникает задача нахождения по данному множеству документов класса схожих по содержанию документов (поиск «по аналогии»). Процесс разбиения множества документов электронной базы на классы, при котором элементы, объединяемые в один класс, имеют большее сходство, нежели элементы, принадлежащие разным классам, называется *кластеризацией*.

В качестве информационного запроса предполагается задание непустого множества документов, а в качестве результата выполнения запроса выдаются документы, каждый из которых в определенном смысле близок к одному из документов, входящих в заданное множество. В качестве информационного запроса предполагается задание непустого множества документов, а в качестве результата выполнения запроса выдаются документы, каждый из которых в определенном смысле близок к одному из документов, входящих в заданное множество. При этом надо иметь в виду, что при интеграции нескольких баз данных можно столкнуться с наличием в объединенном множестве дубликатов одного и того же документа, которые для удобства пользователя следует исключать из окончательных результатов поиска. Тем самым мы сталкиваемся с ситуацией, когда следует вводить ограничения на «излишнее» сходство (в данном случае тождество) находимых документов.

В классической монографии Дж. Солтона [253] для оценки отношения сходства между парами документов рекомендуется использовать множества их ключевых слов (понимаемых в данном случае как входящие в текст документа термины, относящиеся по смыслу к данной предметной области). Однако на сходство документов научной тематики могут также влиять и другие факторы, например наличие у документов общего автора (или, тем более, нескольких авторов).

Однако ситуация, когда требуется поиск документов «по аналогии», возникает не только применительно к научным публикациям. Например, в настоящее время в Рунете существует несколько десятков новостных информационных порталов, основной принцип работы которых заключается в аккумулировании новостной информации, публикуемой на сайтах информационных агентств, и объединении сообщений, освещающих ход развития того или иного события, в так называемые сюжеты (такое объединение обычно реализуется посредством публикации в конце сообщения гиперссылок на другие сообщения, относящиеся к этому же сюжету). Огромные объемы поступающей информации требуют автоматизации процесса выявления сообщений сходной тематики, причем, как и в задаче поиска научных публикаций, здесь также возникает проблема удаления избыточной информации. Однако в данном случае дело осложняется тем, что в отличие от предыдущей задачи, при формировании новостных сюжетов требуется удалять не только полные, но и нечеткие дубликаты, возникающие, например, вследствие того, что разные информационные агентства независимо друг от друга сообщили одну и ту же новость (естественно, несколько разными словами).

Более того, информационный поиск «по аналогии» все чаще используется и для художественных произведений. Это вполне естественно, поскольку одним из основных критериев, которым руководствуется большинство читателей при выборе произведений художественной литературы, является сходство выбираемых книг с книгами, ранее понравившимися данному читателю. Так как количество наименований художественных книг постоянно растет, то сделать выбор традиционным путем, просматривая книги в магазине или библиотеке, становится все более затруднительно. И здесь на помощь также могут прийти современные информационные технологии, поскольку Интернет содержит большое количество документов, непосредственно представляющих произведения художественной литературы (в виде их полных текстов) или достаточно подробно описывающих такие произведения (в виде выходных данных и аннотаций книг, продаваемых в интернет-магазинах).

Особенности алгоритмов для каждой из этих задач мы рассмотрим ниже, после того как будет описана формализация понятия сходства и приведены подходы к количественной оценке степени сходства объектов.

1.8.2. Формализация понятий *аналогии* и *сходства*

> Два математика пьют чай. Один из них роняет чашку и видит, что у нее отлетела ручка.
>
> — Ничего страшного! — радостно говорит он, поднимая чашку. — Хотя топологически она изменилась и негомеоморфна прежней, но пить из нее можно.
>
> Второй, приглядевшись, замечает, что у чашки откололось еще и дно.
>
> — Нет, ты не прав. Пить из нее как раз нельзя. Зато топологически она осталась прежней.
>
> *Математический фольклор. Цит. по книге:*
> С. Н. Федин. Математики тоже шутят

Напомним, что аналогия означает отношение сходства между объектами; рассуждение по аналогии — вывод о свойствах одного объекта по его сходству с другими объектами [130].

Методика умозаключений по аналогии, изложенная, например, в [112], состоит в следующем. Если более изученному объекту A присущи свойства $P_1, P_2, \ldots, P_n, P_{n+1}$, а изучение менее изученного объекта B показало, что ему присущи свойства $P_1, P_2, \ldots, P_n$, то можно сделать предположение о наличии у объекта B свойства P_{n+1}. Степень вероятности правильного умозаключения по аналогии будет тем выше, чем: 1) больше известно общих свойств у сравниваемых объектов, 2) существеннее обнаруженные у них общие свойства и 3) глубже познана взаимная закономерная связь этих сходных свойств. Для повышения надежности умозаключения по аналогии желательно, чтобы общие свойства $P_1, P_2, \ldots, P_n$ были возможно более специфичными для сравниваемых объектов, а свойство P_{n+1}, напротив, должно быть наименее специфичным.

Важно подчеркнуть, что в приведенном определении понятия *аналогия* речь идет лишь о *сходстве* сравниваемых объектов, поскольку устанавливать *тождество* возможно лишь при высокой степени абстрактности определения рассматриваемых свойств, что зачастую весьма затруднительно в конкретных исследованиях. Однако сразу же возникает вопрос: что же именно следует понимать под *сходством* некоторых объектов (в частности, документов)?

Общелитературное толкование слова *сходство*, приведенное, например, в словаре Ушакова: «одинаковость, подобие, соответствие в чем-нибудь с кем-чем-нибудь» [164], с формальной точки зрения

носит характер *circulus vitiosus*, поскольку в том же словаре *одинаковость* определяется как «...тождество, совпадение чего-нибудь», *подобие* — как «образ чего-нибудь, нечто похожее, сходное», *соответствие* — как «соотношение между чем-нибудь, выражающее согласованность, равенство», *равенство* — как «одинаковость, полное сходство» и т. д. Изложенное, разумеется, не является упреком в адрес составителей словаря, которые не ставили перед собой цель построения «формально-аксиоматической» модели семантики русского языка, но служит свидетельством того, что *сходство* относится к числу «базисных» слов, понимание которых априори предполагается для любого носителя языка. Таким образом, рассматривая далее основные постановки задач информационного поиска по аналогии, мы вправе подразумевать общелитературное понимание слова сходство, памятуя, однако, о том, что в основе алгоритмов автоматизации нахождения «похожих» документов должно лежать более формализованное толкование термина, которое будет подробно обсуждаться ниже.

Начнем с определения *сходства*, данного в Философском энциклопедическом словаре (ФЭС) [113]: «Сходство, отношение, родственное равенству. При наличии у пары объектов хотя бы одного общего признака можно говорить о сходстве объектов этой пары. Ввиду многообразия признаков на одной паре могут индуцироваться разные отношения сходства, а ввиду повторяемости признаков — одно отношение сходства на разных парах. Отношения сходства на разных парах интенсионально[1] совпадают, если они определены одним и тем же признаком. Если этот признак — четко определенное свойство, присущее каждому элементу рассматриваемых пар, то отношение сходства всегда рефлексивно, симметрично и транзитивно, т. е. совпадает с отношением равенства (эквивалентности) на множестве объектов, входящих во все такие пары. Если же этот признак — отношение, то сходство по отношению, вообще говоря,... рефлексивно, но не обязательно транзитивно или даже симметрично...».

Какие же важные характеристики понятия сходства содержатся в этом определении?

---

[1] Интенсиональность — зависимость истинности высказываний не только от истинности составляющих их более простых высказываний, но и от психологических, прагматических и модальных оттенков смысла этих высказываний.

Во-первых, отмечается интенсиональность отношения сходства, т. е. его контекстная зависимость, что фактически в той или иной мере обусловливает субъективность или, по крайней мере, интерсубъективность рассматриваемого отношения. Особо подчеркнем, что в период публикации ФЭС в отечественной гносеологии безраздельно господствовала ленинская теория отражения [93], согласно которой процесс восприятия объектов внешнего мира и их свойств носит объективный характер: «...логично предположить, что вся материя обладает свойством, по существу родственным с ощущением, свойством отражения» [там же, с. 91]. Тем не менее применительно к восприятию отношения сходства даже тогда нельзя было не отметить элемент субъективности (который, впрочем, в той или иной степени присущ всему процессу познания: «информация всегда относительна, она зависит... от того, какой информационной системой она воспринимается» [95]. Житейской иллюстрацией интенсиональности сходства может служить не вышедший на экраны эпизод из сценария известного советского фильма «Мимино» (1977, режиссер Г. Данелия), когда главные герои фильма Валико (исполнитель роли В. Кикабидзе) и Рубик (исполнитель роли Ф. Мктрчян) входят в лифт гостиницы, где стоят два японца, похожие друг на друга, с точки зрения советского (притом отнюдь не только русского!) зрителя, как близнецы. Увидев входящих, японцы говорят друг другу по-японски (с русскими субтитрами): «Как все эти русские похожи друг на друга!».

Во-вторых, в ФЭС подчеркивается, что отношение сходства, определяемое всего лишь одним признаком, далеко не всегда обладает свойствами транзитивности и даже симметричности (проводимое в известной книге Ю. А. Шрейдера «Равенство, сходство, порядок» [204] подробное обсуждение нетранзитивности отношения сходства не столь показательно, поскольку основано на использовании при сравнении разных пар объектов *разных* признаков). Пример нетранзитивного сходства довольно очевиден: вполне естественно (в соответствующих контекстах) звучит как выражение «четыре часа ночи» (ср. «час ночи»), так и выражение «четыре часа утра» (ср. «семь часов утра»), т. е. временной признак события «произойти в 4.00» сходен (в зависимости от контекста) в смысле принадлежности к одному и тому же времени суток как с признаком «произойти в 1.00», так и с признаком «произойти в 7.00». Вместе с тем невозможно назвать время суток, к которому одновременно можно было бы отнести события, происшедшие в 1.00 и в 7.00.

Случаи несимметричного сходства не столь тривиальны, и их обсуждению мы посвятим отдельный пункт.

Из определения ФЭС, однако, неясно, как конкретно можно установить сходство между объектами. Некоторую ясность вносит определение сходства из Большой советской энциклопедии (БСЭ) [166]: «…соответствие отображения, образа своему оригиналу… Оно включает три основные отношения: соответствие качественных характеристик отображения особенностям оригинала (например, ощущение зеленого цвета листьев растения соответствует определенной длине электромагнитных волн, излучаемых поверхностью листьев); соответствие структур отображения структурам оригинала (например, структура географической карты соответствует геометрическим структурам местности), причем разные виды соответствия структур могут описываться с помощью различных математических отображений — изоморфизма, гомоморфизма и др.; соответствие количественных характеристик отображения и оригинала (например, количественные значения состояний термостата соответствуют измеряемой температуре тела)».

Важный момент, отмеченный в определении БСЭ, — представление процедуры установления сходства как сравнения объекта с неким «оригиналом». Очевидно, что такой «оригинал» далеко не всегда имеет отношение к реальному происхождению сравниваемых объектов (так, внешнее сходство двух людей, заведомо не являющихся близкими родственниками, обусловлено отнюдь не наличием в обозримом прошлом их гипотетического общего предка). Тем самым в качестве «оригиналов (прежде всего, при установлении сходства по качественным характеристикам) выступают так называемые общие понятия (говоря философским языком, *универсалии*). Мы не будем здесь обсуждать вопрос о природе универсалий, являющийся, начиная с античности, предметом острейших философских споров (см., например, [59] и имеющуюся в статье библиографию). Отметим лишь, что наличие универсалий, рассматриваемых, по крайней мере, как некие общепринятые обозначения, позволяет определять множество универсалий, с помощью которых описывается то или иное свойство объекта, в качестве *измерительной шкалы*. Известно (см., например, [49]), что шкалы, по которым устанавливается сходство, бывают *номинальные* (задающие только наименования значений отслеживаемых свойств), *порядковые* (на множестве значений свойств задано отношение по-

рядка, но отсутствует возможность количественного сравнения значений) и *арифметические* (множество значений которых допускает ту или иную форму количественного сравнения). При этом набор возможных значений неарифметических шкал не обязательно полностью задан априори: например, при изучении документов множество значений шкалы «авторы» может пополняться, если автор нового документа впервые фигурирует в таком качестве для данной коллекции документов. Кстати, на примере этой же шкалы можно видеть, что некоторые свойства того или иного объекта могут иметь сразу несколько различных значений (в данном случае — когда у документа несколько авторов).

Определение БСЭ интересно еще и тем, что в нем четко выделяются два типа сходства: по качественным и по количественным характеристикам (правда, пример, иллюстрирующий соответствие качественных характеристик, подобран, на наш взгляд, не совсем удачно: в нем соответствие устанавливается фактически по количественной характеристике — длине волны). Что же касается сходства по структурным характеристикам, то оно сводится к качественному (а иногда, при необходимости уточнения совпадающих качественных характеристик, — и к количественному) соответствию структурных элементов сравниваемых объектов. Но именно структурное сходство наиболее важно для возможности делать умозаключения по аналогии. Вот что об этом писал известный американский математик венгерского происхождения Д. Пойа в книге «Математика и правдоподобные рассуждения» [128, с. 35–36]: «Рассматривая в музее естественной истории скелеты различных млекопитающих, вы можете обнаружить, что все они страшны. Если в этом все сходство, которое вы между ними обнаружили, то вы видите не такую уж сильную аналогию. Однако вы можете подметить удивительно много говорящую аналогию, если рассмотрите руку человека, лапу кошки, переднюю ногу лошади, плавник кита и крыло летучей мыши — эти столь различно используемые органы, как состоящие из сходных частей, имеющих сходное отношение друг к другу. Последний пример иллюстрирует наиболее типичный случай выясненной аналогии; *две системы аналогичны, если они согласуются в ясно определенных отношениях соответствующих частей*».

Таким образом, начальный этап формализации процедуры поиска объектов «по аналогии» состоит в выделении основных составных частей или свойств, присущих данному типу объектов, и зада-

нии соответствующих шкал путем установления множества возможных значений каждого из этих свойств.

Проведенный анализ значений этого термина (и родственных ему понятий) может, на первый взгляд, показаться излишне подробным. Однако без такого анализа практически невозможно обоснование адекватности выбора той или иной формальной модели сходства, так как при поиске документов «по аналогии» оценка релевантности документа носит интенсиональный, а зачастую и весьма субъективный характер, поскольку такой поиск допускает произвол в выборе элементов структуры, по которым устанавливается сходство, а также, как мы увидим в дальнейшем, в способе задания количественной оценки степени (меры) сходства, в установлении ее порогового значения, отделяющего «похожие» документы от «непохожих» и т. п.

1.8.3. О несимметричном сходстве

— Изоморфны ли группы *A* и *B*?
— Группа *A* изоморфна, а *B* — нет.

Математический фольклор. Цит. по книге: С. Н. Федин. Математики тоже шутят

В определении сходства из БСЭ отмечено, что если признак, по которому определяется сходство, является отношением, то такое сходство может быть несимметричным. В качестве иллюстрации этого утверждения приведем следующий пример. На множестве городов — административных центров субъектов Российской Федерации установим в качестве признака сходства близость расстояний между городами. С помощью географической карты нетрудно увидеть, что для Калининграда наиболее близкими (а следовательно, и сходными в заданном понимании) городами будут Псков и Смоленск. Однако для Пскова в качестве «наиболее сходных» городов будут названы отнюдь не Калининград, а Новгород и Санкт-Петербург, для Смоленска — Брянск и Калуга.

Приведенный пример несимметричного сходства построен на том, что отношение сходства определялось на достаточно широком множестве объектов. Однако гораздо более интересно рассмотреть вопрос о том, может ли сложиться такая ситуация, когда при наличии пары объектов *A* и *B* мы сделаем вывод о том, что *A* похож на *B*, но при этом *B* не похож на *A*?

На первый взгляд, возможность того, что сходство двух отдельно взятых объектов бывает несимметричным, кажется оксюморо-

ном или, по меньшей мере, парадоксом. Именно так воспринималось известное еще, по-видимому, с античных времен высказывание «Помпей и Цезарь очень похожи, особенно Цезарь» (для дальнейших рассуждений важно отметить, что современники событий отнюдь не признавали кажущееся нам почти бесспорным превосходство Цезаря. Как писал Плутарх («Сравнительные жизнеописания. Помпей», гл. I), «никто из римлян, кроме Помпея, не пользовался такой любовью народа, — любовью, которая возникла бы столь рано, столь стремительно возрастала в счастье и оказалась бы столь надежною в несчастьях»).

Особую остроту тема «несимметричного сходства» обрела в христианском богословии в связи с библейскими стихами «сотворил Бог человека по образу Своему» (Быт. 1, 27) и «когда Бог сотворил человека, по подобию Божию создал его» (Быт. 5, 1), а также с многочисленными высказываниями отцов Церкви о том, что антихрист, явившись в мир, будет стремиться во всем походить на Христа. Мысль о том, что имеет место и «обратное» сходство (которая следует из обыденного понимания сходства), казалась многим в первом случае, как минимум, вольнодумством, а во втором — откровенным кощунством.

Однозначного разрешения эта коллизия не получила даже в отдельно взятой католической традиции. Например, французский философ и публицист Жозеф де Местр, долгое время проживший в России в качестве посланника сардинского короля, в книге «Санкт-Петербургские вечера» (изд. 1821 г.) так обосновывал возможность несимметричного сходства: «Сходство между человеком и его Создателем есть сходство изображения и образца… Если же кто-то считает, будто мы говорим, что человек похож на свой портрет, то нелепостью этой он обязан самому себе, ибо мы утверждаем прямо противоположное». Тем не менее уже в XX в. находились мыслители, продолжавшие отстаивать (применительно к рассматриваемой коллизии) приоритет логической связи сходства и тождества. В частности, известный английский писатель и христианский мыслитель Г. К. Честертон в эссе «Франциск Ассизский» заметил: «В одной из своих блестящих полемических работ кардинал Ньюмен обронил фразу, которая может служить примером смелости и логической ясности католичества. Рассуждая о том, как легко принять истину за нечто противное ей, он говорит: “Если Антихрист похож на Христа, то и Христос, наверное, похож на Антихриста”.

Религиозному чувству неприятен конец этой фразы, но опровергнуть ее может лишь тот, кто сказал, что Помпей и Цезарь очень похожи, особенно Цезарь».

Почему же так трудно признать возможность несимметричного сходства двух объектов? Ответ на этот вопрос, по-видимому, вытекает из приведенных рассуждений Г. К. Честертона: отношение сходства нередко воспринимается как некое обобщение (а то и полный аналог, как в процитированных выше определениях из словаря Ушакова) отношения тождества, которое, безусловно, симметрично с точки зрения классической логики. При этом, однако, не учитывается, что общее понятие может и не обладать некими свойствами частного.

Интересно отметить, что в русской художественной литературе XX в. возможность несимметричного сходства признавалась вполне допустимой (хотя и воспринималась как любопытный феномен). Так, известный русский писатель В. А. Солоухин в эссе «Третья охота» отмечал: «…я, много раз принимавший издали валуи за белые грибы, хочу сказать, что ни разу еще, увидев настоящий белый гриб, я не принял его за валуй. У Глазкова[1] есть четверостишье о необратимости сравнения. Там говорится о том, что свистящий на плите чайник напоминает сирену, но настоящая сирена не напоминает свистящий чайник. Так и здесь».

Приведенные примеры «несимметричного сходства» позволяют провести формальное описание тех ситуаций, в каких может наблюдаться этот вид сходства. Именно свойства объектов (зачастую носящие весьма сложный, «комплексный» характер) устанавливаются с помощью порядковой шкалы. При этом признается, что объект, обладающий свойством с меньшим значением, сходен с объектом, обладающим свойством с бóльшим значением, но обратного сходства может, вообще говоря, и не быть.

Применительно к поиску документов указанная ситуация может наблюдаться, например, при установлении сходства между кратким изложением документа (будь то реферат научной статьи, детское издание «Путешествий Гулливера» Дж. Свифта или «краткий пересказ» «Войны и мира» Л. Н. Толстого) и его полной версией. В большинстве случаев, имея краткое изложение заинтересовавшего его документа, пользователь считает целесообразным найти его полную версию. Напротив, пользователь, имеющий полный текст научной публикации, вряд ли будет рассматривать реферат этой же

[1] Н. И. Глазков — русский поэт, основоположник «самиздата» (1939 г.).

публикации в качестве пертинентного (т. е. соответствующего информационной потребности) результата поиска сходных с ней документов. Аналогично читатель, которому понравилось классическое произведение художественной литературы, вряд ли захочет найти его детское издание (если, конечно, такая цель не ставится специально; но применительно и к этой ситуации следует отметить, что знаменитый английский писатель, профессор филологии Оксфорда Дж. Р. Р. Толкин в эссе «О волшебных сказках» достаточно негативно отзывался о тенденции создания «смягченных» обработок классических сказок). Наконец, использование «кратких пересказов» классики является, на наш взгляд, делом малоэтичным и, следовательно, ответственный разработчик информационно-поисковой системы вправе поставить ограничение на возможность удовлетворения такого рода «информационных потребностей».

1.8.4. Определение меры близости между объектами

> Три измерения: длина, высота и ширина?
> Верно, но есть и четвертое — вес.
>
> *Дама, желающая похудеть. Цит. по книге: «Мысли людей великих, средних и пса Фафика»*

Вернемся к изложению процедуры поиска объектов «по аналогии». Итак, пусть мы уже выделили основные свойства, присущие данному типу объектов, и задали подходящие шкалы, описывающие множества возможных значений каждого из свойств (если рассматриваемые объекты — документы, то в качестве шкал для определения меры сходства обычно используются атрибуты библиографического описания документов). Далее необходимо провести нормализацию шкал, введя на каждой из них «частную» меру сходства (иногда называемую нормативной операцией сопоставления двух значений свойства Ψ_i), т. е. функцию, заданную на множестве значений i-й шкалы Ψ_i (а фактически — на множестве сравниваемых объектов D) следующим образом:

$$\mu_i : \Psi_i \times \Psi_i \to [0, 1],$$

причем функция μ_i в случае полного сходства принимает значение 1, в случае полного различия — значение 0.

Процедура нормализации зависит от типа шкалы (подробнее см. [49]). Так, для арифметических и порядковых свойств Ψ_i на

множестве их значений $\{\psi_i\}$ (здесь $\psi_i^n = \psi_i(d_n)$, где n — номер объекта) всегда существуют минимальное ψ_i^* и максимальное ψ_i^{**} значения. Тогда для арифметических свойств можно положить

$$\mu_i(\psi_i^k, \psi_i^l) = \frac{|\psi_i^k - \psi_i^l|}{\psi_i^{**} - \psi_i^*},$$

а для порядковых

$$\mu_i(\psi_i^k, \psi_i^l) = \frac{\Delta n_i^{k,l} + 1}{\Delta n_i^{*,**} + 1},$$

где $\Delta n_i^{k,l}$ — число различных значений ψ_i^n, лежащих между ψ_i^k и ψ_i^l (если $\psi_i^k = \psi_i^l$, то полагается $\Delta n_i^{k,l} = -1$), а $\Delta n_i^{*,**}$ — число различных значений ψ_i^n, лежащих между ψ_i^* и ψ_i^{**}.

Наконец, для номинальных шкал мера сходства определяется следующим образом: если значения свойств объектов совпадают, то мера близости по этой шкале равна 1, иначе 0. При этом необходимо учитывать, что значения свойств объектов для номинальной шкалы могут быть составными (например, документ может иметь сразу нескольких авторов). В таком случае $\mu_i = n_{i1}/n_{i0}$, где $n_{i0} = \max\{n_{i0}(d_1), n_{i0}(d_2)\}$, $n_{i0}(d_j)$ — общее количество элементов, составляющих значение i-го атрибута документа d_j, n_{i1} — количество совпадающих элементов.

После того как подсчитана мера сходства по каждой из шкал, можно приступить к вычислению меры сходства $\mu(d_1, d_2)$ между объектами, входящими в заданное множество, и объектами, среди которых мы ищем аналогичные заданным. Для этого обычно используется одна из стандартных формул вычисления расстояний с весовыми коэффициентами, которые обеспечивают, чтобы вычисленное значение меры не превосходило 1. Весовые коэффициенты (они, разумеется, неотрицательны) в простейшем случае равны между собой, однако путем задания весовых коэффициентов, отличных друг от друга, мы можем указать априорную относительную важность шкал. Более того, значения весовых коэффициентов могут определяться и предполагаемой апостериорной достоверностью данных соответствующей шкалы, т. е. в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документа d_1 и до-

кумента d_2 более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одного автора). Использование для вычисления меры сходства между объектами d_1 и d_2 стандартной евклидовой метрики

$$\mu_E(d_1, d_2) = \sqrt{\sum (a_i \mu_i(d_1, d_2))^2}, \quad \text{где} \quad \sum a_i^2 = 1 \tag{1.3}$$

оказывается не всегда удобным из-за заметного влияния отдельных больших значений μ_i . Этот недостаток менее заметен при использовании расстояния Хемминга

$$\mu_H(d_1, d_2) = \sum a_i \mu_i(d_1, d_2), \quad \text{где} \quad \sum a_i = 1. \tag{1.4}$$

Напротив, если понимание сходства для конкретной задачи подразумевает отсутствие больших различий по любой отдельно взятой шкале, то целесообразно использовать расстояние Чебышева

$$\mu_\infty(d_1, d_2) = \max|a_i \mu_i(d_1, d_2)|, \quad \text{где} \quad \max|a_i| \le 1. \tag{1.5}$$

В формулах (1.3)–(1.5) выражение $\mu_i(d_1, d_2)$ означает $\mu_i(\psi_i(d_1), \psi_i(d_2))$.

1.8.5. Установление аналогии и оценка эффективности поиска

Если мать иль дочь какая,
У начальника умрет,
Расскажи ему, вздыхая,
Подходящий анекдот;

Но смотри, чтоб ловко было,
Не рассказывай, грубя:
Например, что вот кобыла
Также пала у тебя.

Потому что, если пылок
Твой начальник и сердит,
Проводить тебя в затылок
Он курьеру повелит.

А. К. Толстой. Мудрость жизни

Для непосредственной процедуры нахождения объектов, аналогичных объектам из заданного множества, необходимо задать пороговое значение меры сходства $r \in (0,1)$. Если заданное множество

D_* состоит из одного объекта d_*, то при $\mu(d_*, d_j) \le r$ делается вывод, что объект d_j аналогичен заданному, в противном случае считается, что аналогия отсутствует. Ситуация осложняется, если множество D_* содержит более одного объекта. Тогда критерием аналогичности объекта d_j элементам множества D_* служит неравенство $\mu(D_*, d_j) \le r$, в котором $\mu(D_*, d_j)$ — расстояние от объекта d_j до множества D_* (обычно под этим подразумевается минимум расстояний от объекта d_j до элементов множества D_*, хотя иногда в качестве $\mu(D_*, d_j)$ целесообразно рассматривать расстояние от объекта d_j до определенного тем или иным способом «центра» множества D_*). Независимо от количества элементов в множестве D_* возможно задание «градаций аналогичности», определяемых посредством набора чисел $\{r_i\}$, $i = 1, \ldots, n$, где $r_k < r_l$ при $k < l$. Если $r_k < \mu(D_*, d_1) \le r_{k+1}$, а $r_l < \mu(D^*, d2) \le r_{l+1}$ при $k < l$, то считается, что объект d_1 более схож с элементами множества D_*, чем объект d_2. Введение градаций аналогичности используется, например, для установления приоритета просмотра документов, найденных в процессе информационного поиска.

Указанные процедуры поиска аналогичных документов могут быть снабжены дополнительными условиями, связанными, например, с исключением из поисковой выдачи соответствующих документов при реализации ситуации несимметричного сходства.

Несколько иной подход к нахождению аналогичных объектов связан с кластеризацией объектов объединенного множества, включающего в себя как элементы множества D_*, так и объекты, относительно которых необходимо установить наличие или отсутствие аналогии с элементами множества D_* (напомним, что кластеризацией называется разбиение множества объектов на классы, при котором элементы, объединяемые в один класс, имеют большее (в определенном смысле) сходство, нежели элементы, принадлежащие разным классам). При этом объектами, аналогичными элементам множества D_*, признаются объекты, принадлежащие классам, содержащим определенное количество элементов D_* (это количество может быть задано как абсолютная величина или как доля элементов D_* в данном классе).

Подробный обзор алгоритмов кластеризации содержится, например, в монографии [33]. Сравнение нескольких алгоритмов кластеризации применительно к задаче установления сходства документов сделано в работе [18] и будет подробно изложено в разд. 4.5.

Для задания меры сходства на множестве документов научной тематики использовано расстояние Хемминга, подсчитываемое по формуле (1.4), где в качестве шкал применялись следующие атрибуты библиографического описания:

— авторы;
— ключевые слова;
— аннотация.

Так как сравнение аннотаций в явном виде (т. е. как текстовых строк) бессмысленно, то они сравнивались как составные атрибуты на основании вхождения в их текст терминов из тезауруса соответствующей предметной области. При задании меры сходства принимался во внимание тот факт, что значения весовых коэффициентов в формуле (1.4) определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы и в определенных случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных (подробнее см. раздел 4.5).

Задача объединения новостных сообщений в сюжеты отличается от предыдущей, в частности, тем, что новостные сообщения обычно структурированы намного слабее, чем научные публикации ввиду отсутствия у них основных атрибутов библиографического описания (авторы, аннотация, ключевые слова и т. п.).

Методы, применяемые для решения задачи объединения новостных сообщений в сюжеты, рассматриваются, например, в работах [72, 86]. В [72] выделены два основных подхода: «синтаксический» и «лексический».

Суть «синтаксического» подхода состоит в представлении документа в виде множества всевозможных последовательностей фиксированной длины k, состоящих из соседних слов (такие последовательности называются «шинглами»). Нетрудно видеть, что шинглы суть значения соответствующей номинальной шкалы. Два документа считаются похожими, если множества их шинглов существенно пересекаются. При этом, если число совпадений слишком велико, документы считаются нечеткими дубликатами.

В рамках «лексического» подхода строится словарь (список различных слов) L коллекции документов, из которого исключены слова, встречающиеся в коллекции слишком редко и слишком часто (как правило, содержание документа наиболее адекватно отражают слова со средним значением частоты встречаемости). Далее для каждого документа формируется множество входящих в него

различных слов U и определяется пересечение P этого списка с построенным словарем L. На основании близости таких списков можно судить о сходстве документов. В частности, для выявления нечетких дубликатов списки слов P_i, соответствующие различным документам, упорядочиваются и для каждого из них вычисляется хеш-функция, т. е. преобразование входного массива данных произвольной длины в выходную битовую строку фиксированной длины. В случае совпадения хеш-функций документы объявляются нечеткими дубликатами.

Задача поиска художественных произведений по сравнению с предыдущими гораздо менее формализуема, поскольку свойствами художественных документов (книг, фильмов и т. п.), определяющими их сходство или различие, являются не только содержание, фамилии авторов и т. п., но и эстетическое впечатление, а в некоторых случаях — их идейная направленность. Поэтому основным приемом определения сходства таких документа является экспертная оценка пользователей. Простейший прием такого рода встречается на некоторых сайтах, содержащих коллекции с описаниями художественных фильмов: пользователю предлагается указать известные ему фильмы, похожие на данный. Разумеется, такие оценки носят крайне субъективный характер.

Несколько более объективный подход используется в некоторых интернет-магазинах: в качестве одного из элементов описания книги или компакт-диска указывается список книг или дисков, которые пользователи обычно покупают вместе с данной книгой (диском).

Наконец, некоторые пользователи «Живого Журнала» и подобных ему ресурсов блогосферы не просто приводят списки своих любимых книг и фильмов, но и сопровождают их комментариями вида: «если Ваш список совпадает с моим более чем на треть, пожалуйста, сообщите свой список мне».

Важно подчеркнуть, что при поиске «по аналогии» оценка релевантности, а тем более пертинентности документа носит интенсиональный, а зачастую и весьма субъективный характер, поскольку процедура поиска допускает произвол в выборе элементов структуры, по которым устанавливается сходство, в способе задания количественной меры сходства, в установлении ее порогового значения, отделяющего «похожие» документы от «непохожих» и т. п. Но даже если мы сочтем все эти параметры неотъемлемой ча-

стью поискового предписания, т. е. декларируем их «объективный» (для данного конкретного предписания) характер, то все равно останется практически неустранимая зависимость результата поиска «по аналогии» от всей совокупности документов, входящих в информационный массив. Попросту говоря, вывод о схожести объекта «*кошка*» с объектом «*корова*» различается в случае, когда «информационный массив» есть множество: *лев, корова*, и в случае, когда «информационный массив» — *корова, кобра* (или даже *лев, корова, кобра*).

Для сравнения заметим, что в случае обычного поиска посредством задания конкретного поискового образа можно достаточно объективно судить о релевантности того или иного документа, вошедшего в выдачу, поскольку причиной выдачи нерелевантных документов являются погрешности в индексировании документов, проявляющиеся, например, во внесении в поисковый образ документа «лишних» слов (в результате явных ошибок, многозначности естественного языка и т. п.).

В каждом конкретном случае оценка пертинентности документа может быть более или менее объективно дана пользователем, ознакомившимся с этим документом, однако использование таких оценок для улучшения работы алгоритмов поиска «по аналогии» является иногда весьма непростой задачей, так как связь между параметрами алгоритма и результатами выдачи далеко не всегда носит очевидный характер (особенно для алгоритмов, основанных на кластеризации).

1.9. Метаданные и обработка электронных ресурсов

> Существуют три вида автомобилистов: те, которые сами моют свою машину, те, которые отдают ее мыть, и те, которые ждут дождя.
>
> *Из книги «Будь Карузо баранки». Цит. по книге: «Мысли людей великих, средних и пса Фафика»*

Но каким же образом справочно-информационный фонд ИПС пополняется поисковыми образами новых интернет-документов? В разд. 1.5 мы изложили общую схему организации этого процесса, теперь перейдем к его более подробному рассмотрению.

Целесообразно рассматривать два определения интернет-документа: на синтаксическом и на семантическом уровнях информации. С синтаксической точки зрения интернет-документ — информационный ресурс, имеющий уникальный идентификатор и обладающий некоторой структурой и содержанием [214]. С семантической точки зрения интернет-документ — целостный информационный объект, помещенный в информационное пространство сети Интернет, который описывает, представляет, отображает или моделирует некоторую сущность реального мира [198]. Таким образом, интернет-документ — частный случай обычного документа (при этом в определении первого указываются некоторые свойства, характерные для этого класса документов).

Научно-информационный процесс включает в себя следующие этапы работы с документами [104]:

1. Сбор документов.
2. Аналитико-синтетическая переработка документальной информации.
3. Хранение и поиск информации.
4. Репродуцирование и распространение информационных материалов.

Следует учесть, что за рамки этого определения выведена первоначальная стадия информационной деятельности — подготовка научных документов к их размещению в сети Интернет. Хотя, как уже неоднократно отмечалось в этой книге, развитие сети Интернет изначально носит децентрализованный характер и выработка общих стандартов представления информации не более чем благое пожелание, но при создании интернет-документов следует стремиться к тому, чтобы работа с ними была максимально удобной для пользователей, что достигается, в частности, включением документов в информационные системы, основные принципы создания которых будут изложены в гл. 3.

Здесь мы лишь коротко отметим, что для наиболее эффективного функционировании ИнтС целесообразно рассматривать в качестве логической единицы хранения *документ*, понимаемый как информационный ресурс, имеющий (по определению [214]) уникальный идентификатор и обладающий некоторой структурой и содержанием. Разумеется, документ — информационный ресурс — представляет собой поисковый образ исходного документа, причем в некоторых случаях содержание последнего может вхо-

дить в поисковый образ в качестве одного из элементов (это противоречит ограничению из классической монографии [104], но из контекста следует, что подобное ограничение было вызвано необходимостью уменьшения объема поисковых образов с целью уменьшения трудоемкости процесса их обработки). С другой стороны, поисковый образ документа тоже является документом (описывающим исходный документ), поэтому далее, где это не вызовет недоразумения, мы будем использовать термин «документ» в значении «поисковый образ исходного документа». Отметим, что в фундаментальных работах по информатике и кибернетике [104, 156], вышедших в конце 1980-х годов, поисковый образ документа не рассматривается даже в качестве вторичного документа.

Особо подчеркнем, что в этап сбора интернет-документов мы будем включать и первоначальную стадию их аналитико-синтетической переработки: каталогизацию, предусматривающую занесение в каталожную карточку сетевого имени (url—адреса) документа. Согласно стандартам построения открытых систем (OSI) [88], структура и содержание документа должны описываться в соответствии с международными схемами данных. Совокупность извлекаемых в процессе индексации характеристик документа вместе с формальным описанием структуры этих характеристик обычно называют метаданными. Более формально, метаданные — это структурированные данные, представляющие собой характеристики описываемых сущностей для целей их идентификации, поиска, оценки, управления ими [259]. Как отмечено в обзоре Ю. Е. Хохлова и С. А. Арнаутова [183], метаданные нельзя рассматривать как обычную разновидность каталожного описания документов ввиду специфики области их применения, используемых подходов и т. п. Таким образом, сбор интернет-документов сводится к сбору их метаданных, поскольку, как будет показано в разд. 3.2, информационная система работает не с данными, а исключительно с метаданными; к тому же непосредственное копирование документов может вызвать серьезные вопросы относительно соблюдения авторских прав.

Метаданные определяют структуру и смысловое содержание документа, а также правила работы с ним и в соответствии с этим иногда подразделяются по своему функциональному назначению на *структурные*, *описательные* и *административные* [183].

Структура метаданных иерархична: наиболее общий характер имеют метаданные, задающие структуру документа, т. е. описывающие метаданные более низкого уровня (атрибуты документа), которые определяют содержание документа; наконец, значения этих атрибутов являются фактически метаданными по отношению к исходному документу (рис. 1.2). При этом описательные метаданные, характеризующие документ, могут быть частью документа и в то же время содержать в соответствии с выбранной схемой данных сведения о документе (основные и дополнительные, такие как, например, авторы, название, дата создания и т. д.). Это свойство метаданных послужило для некоторых исследователей [139] основанием для заключения об относительности понятия «метаданные»: оно значимо только в контексте и дает понять, чем, собственно, являются данные.

Следует отметить, что понятие «метаданные» возникло в процессе развития методологии обработки именно электронных документов, когда появилась необходимость в знании способа кодирования информации в тексте. В отечественной литературе это понятие (названное, правда, «метаинформацией») было, по-видимому, впервые сформулировано в статье Ю. А. Шрейдера [199] (термины типа «метасообщение», «метаязыковый текст» и т. п. из более ранней работы С. И. Гиндина [53] имели несколько иное значение, ибо касались пояснения семантики сообщения через информацию более

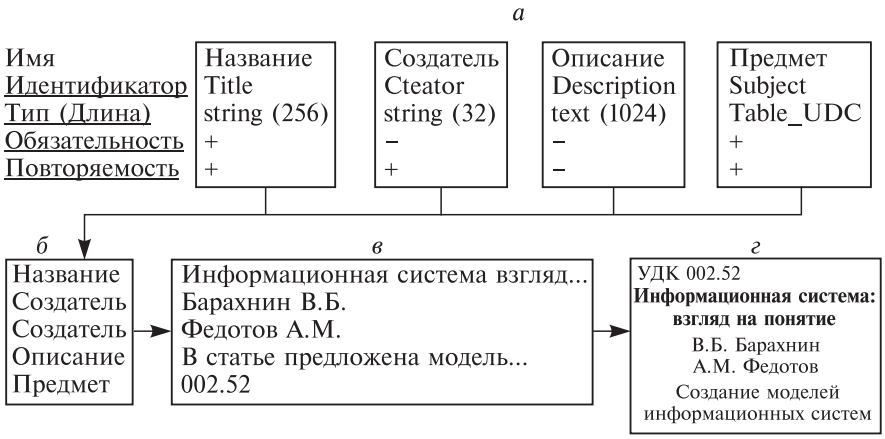

Рис. 1.2. Иерархия метаданных документа.
а — структура; *б* — атрибуты; *в* — содержание; *г* — документ.

высокого уровня, например общекультурный контекст или цели отправителя). В 1980-е годы термин «метаданные» прочно вошел в профессиональную лексику специалистов в области информатики (см., например, «Словарь по кибернетике» (1989) [156]).

Перейдем к рассмотрению понятия метаданных с точки зрения семантического уровня информации. Достаточно подробный обзор существующих форматов метаданных приведен в работе Ю. Е. Хохлова и С. А. Арнаутова [183]. Вопросы формирования наборов метаданных для научных информационных ресурсов рассмотрены в статье авторов данной монографии [191] и в работах возглавляемого А. Б. Жижченко и В. А. Серебряковым коллектива сотрудников ВЦ им. А. А. Дородницына РАН [35, 36, 75].

В [35] сформулированы основные требования к стандартам на метаданные для описания научной информации, которые должны обеспечивать:

— полноту описания основных типов научной информации;

— открытость для доступа;

— расширяемость описаний;

— возможность интеграции информации;

— уникальную идентификацию информации;

— возможность размещения и поиска информации в распределенной среде;

— ориентированность на семантические технологии описания и использования информации (Semantic Web [256]);

— интероперабельность с внешней средой.

В наибольшей степени перечисленным требованиям удовлетворяет набор элементов метаданных так называемого Дублинского ядра (Dublin Core) [227], который может быть расширен схемами конкретной предметной области. Среди других форматов метаданных, которые могут быть использованы при создании специализированных ресурсов научной тематики, можно отметить Electronic Business Card (vCard) — формат описания информации о деловых контактах [262], используемый почтовыми клиентами MS Outlook Express и Netscape Communicator для управления контактной информацией пользователей, и Global Information Locator Service (GILS)[1] [230] — схема общего описания информационных ресурсов. Перечисленные форматы являются не более чем *рекомендациями*,

[1] Профиль протокола Z39.50 (ISO 23950) — Government Information Locator Service.

хотя при их создании могут быть использованы *международные стандарты*: например, при определении элементов метаданных Дублинского ядра использован набор атрибутов из стандарта спецификации элементов данных ISO/IEC 11179 [232].

В [35] предложено также несколько уровней поддержки схем данных:

— *минимальная* — минимально достаточная для обмена метаданными, поддержки взаимосвязей ресурсов;

— *базовая* — достаточная для эффективной работы неспециалистов в конкретной предметной области;

— *расширенная* — достаточная для основной работы специалистов предметной подобласти;

— *специализированная* — ориентированная на специалистов предметной области, используется для создания специализированных систем.

Конкретное содержание набора метаданных, которое должно являться расширением схемы данных Дублинского ядра, определяется предметной областью, для которой создается та или иная информационная система; схемы метаданных для информационных систем различной научной и научно-организационной тематики, создаваемых в Сибирском отделении РАН, обсуждаются в статьях [9, 28, 91, 196] и др. и будут изложены в разд. 3.4.

Однако, как было подробно описано в разд. 1.3, существующие подходы к построению программных систем информационного обеспечения научной деятельности не способны в полной мере удовлетворить потребности научного сообщества, поскольку имеют следующие недостатки: ограниченность возможностей обеспечения интеграции ресурсов как внутри каждой из систем, так и с внешними системами (иными словами, низкая интероперабельность) и (или) отсутствие механизмов своевременной актуализации информации. Нетрудно заметить, что перечисленные недостатки так или иначе связаны с процессом сбора метаданных документов.

Как же организован сбор метаданных документов в информационных системах различных типов? Электронные библиографические базы (Current Contents, Zentralblatt MATH, Реферативные журналы) содержат составленные экспертами краткие аннотации «бумажных» документов без ссылок на электронные (обычно более подробные, чем аннотация) версии документов и уж тем более без метаданных, задающих ссылки на документы, описывающие пер-

соны авторов. ИПС научной тематики (каталоги ресурсов) работают с документами после непосредственного согласования форматов метаданных, при этом активно развиваемые в последнее время системы, использующие концепцию Semantic Web, могут работать только с документами, у которых значения метаданных суть элементы заданных словарей. Наконец, поисковые системы общего назначения работают с любыми документами, но слабо используют анализ метаданных, что приводит к низкой пертинентности найденных документов.

Продолжим рассмотрение особенностей процесса обработки интернет-документов. В тех случаях, когда документ сети Интернет представляет конкретную сущность (книгу, статью и т. п.) или же отображает ее (т. е. является точной копией или электронным образом другого документа), подходы к изучению его информационного содержания аналогичны тем, которые применяются в библиотечном деле при изучении информационного содержания полиграфического издания. В частности, к интернет-документу могут быть применены методики каталогизации и классификации (индексирования).

Однако полиграфические издания не обязательно представляются (отображаются) одним интернет-документом. Например, каждую публикацию в составе электронного журнала, сборника и т. п. целесообразно представлять как отдельный документ. Напротив, многотомное издание одного произведения, имеющего большой объем, представляется как один документ (возможно, состоящий из нескольких файлов, связанных гиперссылками).

Если интернет-документ описывает реальную сущность, например персону, организацию, артефакт, природный объект и т. д., его стандартного библиографического описания может оказаться явно недостаточно для создания адекватного поискового образа документа. Во избежание этого библиографическое описание дополняется необходимой информацией, относящейся к описываемой сущности, для чего используются стандарты или принятые правила соответствующей предметной области.

Важной особенностью интернет-документа является наличие у него сетевого имени, без занесения которого в каталожную карточку документа каталогизация становится бессмысленной. Заметим, что речь идет не только о статических именах: имя, образуемое при

запросе динамически формируемого интернет-документа, также может быть использовано для каталогизации.

Другой особенностью интернет-документа, резко отличающей его от полиграфического издания, является возможность внесения в него изменений. Здесь речь идет, прежде всего, о документах справочного характера. Разумеется, незначительные изменения, не влияющие на каталожное описание документа, не вызовут проблем, однако возможны случаи, когда существенное изменение информации приводит к необходимости составления нового каталожного описания. Наконец, не исключены ситуации, когда документ удаляется его владельцем из сети Интернет. Это возможно, в частности, если сущность, описываемая документом, исчезает (например, ликвидируется подразделение организации) или перестает представлять интерес для владельца документа (например, сотрудник увольняется из организации). Необходимость оперативного отслеживания подобных изменений — важная составляющая процесса каталогизации

Нетрудно видеть, однако, что задача более или менее полной каталогизации интернет-документов научной и научно-организационной тематики в соответствии с библиографическими стандартами крайне сложна ввиду следующих причин:

1. Огромное количество документов, причем в качестве новых интернет-документов могут выступать давно опубликованные полиграфические документы (так, многие научные журналы постепенно выкладывают на свои сайты статьи, вышедшие в старых номерах).

2. Отсутствие специальных структур, отслеживающих появление новых документов в сети; например, каталогизацией научных интернет-документов обычно занимаются заинтересованные специалисты, работающие в соответствующей предметной области.

3. Необязательность авторской классификации интернет-документов (в отличие от печатных изданий) посредством их аннотирования, приписывания кодов классификатора и т. п., что значительно осложняет процесс каталогизации.

4. Проблема отслеживания изменений документов.

Трудоемкость процесса каталогизации отдельных документов приводит к необходимости его частичной автоматизации. Основные сложности при решении этой задачи состоят в разработке алгоритма, позволяющего автоматически извлекать из документа основ-

ные элементы его библиографического описания, а также в классификации документов при отсутствии у них явно указанных классификационных признаков.

Таким образом, наиболее перспективным подходом к сбору метаданных является автоматизированное их извлечение из документов достаточно произвольной структуры, что значительно упрощает процесс работы с информацией, представленной во внешних системах (в том числе сети Интернет), включая механизм актуализации информации.

1.10. Методология изучения интернет-сайтов

> ...Из окна чудесный вид на будущий стадион, вот только бульдозеры окончат работу и будет осушено болото.
>
> *Дж. Б. Пристли. Сэр Майкл и сэр Джордж*

Задачу изучения отдельных интернет-документов практически невозможно решить без изучения структур, объединяющих эти документы, т. е. интернет-сайтов, понимаемых как группы документов, имеющих общего владельца и объединенных точкой входа. Более того, ввиду огромного количества научных интернет-ресурсов, а также слабой структурированности некоторых из них, препятствующей выделению основных элементов метаданных, приходится заносить в каталоги интеллектуальных информационных систем сведения не только об отдельных документах, но и о сайтах.

Количество сайтов в сети Интернет весьма сложно оценить хотя бы потому, что один сайт может быть составной частью другого (например, сайт журнала «Вычислительные технологии» является частью сайта Института вычислительных технологий СО РАН). Косвенную оценку снизу числа сайтов можно получить исходя из количества доменных имен, которое, по данным компании VeriSign на июль 2005 г., достигло 82,9 млн [85]. Согласно более современной оценке, на осень 2009 г. количество сайтов в Интернете составляет порядка 230 млн, в том числе в Рунете — около 15 млн [87].

При рассмотрении сайта исключительно с точки зрения его содержания как совокупности документов, составляющих сайт, в процессе его изучения могут быть использованы практически все

приведенные ранее в разд. 1.9 соображения относительно научно-информационных процессов с участием интернет-документов.

Реальные технологии создания подавляющего большинства сайтов таковы, что однородные документы с одного сайта имеют практически одинаковую HTML-разметку. При этом неважно, генерируются ли документы динамически (в этом случае однородность разметки — естественное следствие работы соответствующей программы) или же они создаются вручную посредством копии уже имеющегося документа с последующей заменой текста (что также сохраняет разметку). Данное обстоятельство позволяет автоматизировать процесс извлечения метаданных интернет-документа с целью занесения их в каталог. Соответствующая технология будет изложена в гл. 4.

Однако интернет-сайт представляет собой не просто набор документов, а достаточно сложную систему, при изучении которой целесообразно анализировать следующие вопросы (см. работу авторов [30]):

1) информационное наполнение сайта;

2) методы хранения и обработки данных (рассматриваемые вместе с программными средствами, с использованием которых они реализованы);

3) значение сайта для информационного обеспечения соответствующего вида деятельности.

Эти вопросы тесно взаимосвязаны.

Необходимо выяснить: является ли комплексное изучение сайтов, охватывающее все перечисленные выше аспекты, научной проблемой?

Можно выделить три основных аспекта научного анализа артефактов:

1) технология производства;

2) сравнительный анализ артефактов на основе их функциональных свойств (систематизация, классификация и т. п.);

3) влияние (в широком смысле) на человека и общество.

Отметим, что далеко не все типы артефактов становятся предметом комплексного научного изучения. В большинстве случаев научный интерес представляют лишь некоторые из перечисленных аспектов, притом, как правило, они рассматриваются по отдельности.

Например, технология производства того или иного типа артефактов является предметом изучения соответствующей отрасли прикладной науки. В то же время особенности функционального применения артефактов далеко не всегда представляют предмет научного исследования. Если же использование артефактов — предмет научного анализа (применение оружия и боевой техники, изучаемое военными науками, или применение компьютеров, изучаемое информатикой), то соответствующие разделы наук весьма слабо связаны с технологией производства данного артефакта. Наконец, изучение воздействия артефактов на общество обычно весьма опосредованно связано как с технологическими, так и функциональными аспектами. В тех же случаях, когда такая связь присутствует, она обычно рассматривается как отдельный феномен.

На наш взгляд, наиболее комплексно все три указанных аспекта изучаются в архитектуре, ибо «функциональные, конструктивные и эстетические качества архитектуры (прочность, польза, красота) взаимосвязаны» [74].

Остается выяснить, является ли сходство в комплексном изучении сайтов и объектов архитектуры случайным или за ним стоят весьма глубокие закономерности?

Во-первых, если архитектура представляет собой организацию среды физического пребывания человека, то сеть Интернет, отличающаяся от печатных изданий оперативностью размещения и доставки информации практически любого характера, а от классических электронных средств массовой информации — возможностью передачи печатного текста, впервые создает единое информационное пространство человеческой цивилизации.

Во-вторых, разнообразие информационных потребностей (не меньшее, чем потребностей физических) вызывает необходимость массового производства интернет-сайтов самого разнообразного назначения, сопоставимого с массовым строительством. Это резко отличает две рассматриваемые сферы от производства подавляющего большинства сколько-нибудь сложных технических средств (включая компьютерные программы), создаваемых на весьма ограниченном количестве предприятий и затем распространяемых по всему миру. С другой стороны, среди технологий, применяемых в мелкосерийном или штучном («кустарном») производстве, технологии строительства и создания интернет-сайтов относятся к числу

наиболее сложных, что приводит к необходимости систематического их изучения.

В-третьих, архитектурные сооружения и интернет-сайты имеют определенное структурное сходство, так как представляют совокупность более или менее однородных объектов (соответственно помещений и документов), связанных между собой определенным образом.

В-четвертых, терминология описания интернет-сайтов во многом заимствована из архитектуры (термины «архитектура сайтов», «строительство сайтов», «портал» и проч.). Но, как подчеркивают лингвисты, занимающиеся исследованиями в области семантики языка (см., например, монографию Ю. Д. Апресяна [2]), естественный язык типологизирует ситуации внешнего мира. Указанная типологизация имеет особое значение в языке науки. Как отмечено С. Е. Никитиной [111], перенос значений (метафора) «занимает почетное место среди способов объективации и развития научного знания». Именно метафоры выступают в языке науки как средство переосмысления старых терминов применительно к новым референтам и могут быть охарактеризованы как мостики, соединяющие старые и новые теории [123, 126].

Исходя из сказанного выше, можно прийти к выводу, что сходство между архитектурными сооружениями и веб-сайтами имеет весьма глубокий характер. Особенно ярко данное обстоятельство проявилось в истории развития идеи шаблонов проектирования, первоначальной предложенной К. Александером для решения задач архитектуры [211], но получившей наиболее широкое развитие в задачах программирования, причем первый шаг в этом направлении был сделан в работе К. Бека и У. Каннингема [213] применительно к технологии создания пользовательских интерфейсов.

Отмеченное сходство имеет, в частности, методологическое значение, так как позволяет переносить некоторые приемы научных исследований из области архитектуры в область изучения веб-сайтов. Например, введенная в УДК классификация раздела «Архитектура» (Классификация зданий по назначению; Теория, философия, эстетика архитектуры; Методы работы; Строительные материалы; Повреждения; История архитектуры; Архитектурные детали; Части зданий; Этические и социологические аспекты; Занятия (профессии), имеющие отношение к архитектуре) может быть с минимальными изменениями применена для вычленения

различных аспектов изучения интернет-сайтов, рассматриваемых с различных точек зрения: как источник данных, как техническое средство обработки и распространения информации и как социокультурный феномен.

1.11. Проблемы разработки теоретических основ создания интеллектуальных систем

> Составлять бюджет — отравлять себе жизнь, еще не приступив к расходам.
>
> *Ян Камычек. Из рубрики «Savoir vivre» («Хорошие манеры») журнала «Пшекруй»*

Подведем итог изложенному в этой главе. Нами было показано, что осмысление процесса обработки компьютерной информации как технологии невозможно без разработки теоретических основ создания информационно-поисковых систем информационного обеспечения научной деятельности, способных в автоматизированном режиме извлекать метаданные из электронных документов (прежде всего, интернет-ресурсов) достаточно произвольной структуры, что, в свою очередь, позволит получать на основании этих данных новую информацию и знания (в последнее время в качестве обозначения таких систем пытаются закрепить термин «портал» [43]). Следует отметить, что для некоторых других классов информационных систем (в которых семантика информации сравнительно проста) аналогичные задачи уже получили достаточно полное решение. В частности, вопросы анализа информационных систем, работающих с результатами измерений (так называемых *информационно-измерительных систем* [184]), рассмотрены в монографии В. П. Бакалова [7].

Поставленная задача носит комплексный характер и включает в себя целый ряд подзадач.

Во-первых, следует решить вопросы построения моделей основных компонентов интеллектуальной системы: как информационно-поисковой системы (рассматриваемой *в абстрактном виде* (см. [104, с. 253]), т. е. без учета средств технической реализации), так и логических компонент, отвечающих за поиск информации, вывод новых знаний и диалог с пользователем.

Имеющиеся наработки в области теории моделирования баз данных, как классические, которые изложены, например, в монографии Д. Цикритзиса и Ф. Лоховски [185] (некоторые аспекты дальнейшего развития положений этой монографии, а также важные терминологические уточнения содержатся в статье М. Р. Когаловского [81]), так и современные, представленные, в частности, в диссертационной работе С. В. Зыкина [73], не могут в полной мере отвечать требованиям поставленной задачи. Дело в том, что эти работы рассматривают в качестве логической единицы хранения, т. е. основного элемента системы, *запись в базе данных*, в то время как развитие интернет-технологий требует рассматривать в этом качестве *документ*, т. е. в данном контексте информационный ресурс, имеющий (по определению [214]) уникальный идентификатор и обладающий некоторой структурой и содержанием.

Однако и рассмотрение ресурса в качестве основного элемента системы не решает всех проблем. Например, в модели RDF консорциума W3 [248] элементы суть ресурсы, которые могут представлять и сущности, и их характеристики. Неудобства такого подхода очевидны: приходится иметь дело с большим количеством равноправных мелких элементов, между которыми нужно устанавливать чрезвычайно много связей, вследствие чего структура модели далека от естественной. В модели ИСИР РАН [75] элементы — «ресурсы, аналогичные документоподобным объектам», а связи задаются с помощью отношений между типами ресурсов, т. е. также имеют внешний характер по отношению к ресурсу.

Подход к построению моделей логических компонентов также принципиально отличается от подхода, применяемого специалистами в области искусственного интеллекта для разработки экспертных систем (см., например, монографию [50]): последние предназначены для решения узкоспециализированных задач, содержат относительно небольшой объем документов, и основной упор при их создании делается на развитие сложных продукционных правил, в то время как интеллектуальные системы, работающие с документальной информацией, могут обладать достаточно простыми продукционными правилами, а получение новых знаний становится возможным благодаря большому объему документов, способных выступать в качестве аргументов проверяемых утверждений.

Следовательно, необходима разработка оригинальных моделей основных компонентов интеллектуальной информационно-поисковой системы, учитывающих перечисленные особенности.

Во-вторых, поскольку документы информационной системы связаны между собой, неизбежно встает проблема возможного рассогласования информации. Так, включение в документы информации о разнородных сущностях может привести к появлению множественной информации об одном и том же объекте. Кроме того, для представления сложных документов, когда один документ является частью другого (полностью или частично, в том числе и в виде гиперссылки), необходимо выработать подходы к установлению связей между ними.

Таким образом, становится актуальной разработка технологии идентификации, спецификации и визуализации горизонтальных отношений между сущностями, информация о которых содержится во множестве документов, а также между документами, которые являются составной частью сложных документов. Одним из основных элементов этой технологии является разработка информационной модели отношений и тематических связей между документами системы.

Отметим, что в библиотечных системах, построенных на основе протокола Z39.50 и его версий [65], выполняется полное дублирование служебной информации. Аналогичная ситуация возникает в информационных системах, построенных на основе LDAP-каталогов [46], в которых имеется мощная система перекрестных ссылок, но используемая иерархическая модель не допускает отношений «многие-ко-многим». Если такие отношения все же возникают, то появляется необходимость дублирования информации, что может привести к рассогласованию информации.

Ввиду этого целесообразно хранить информацию в единственном экземпляре, устанавливая в нужных случаях отношения «многие-ко-многим». Традиционный подход, применяемый при проектировании реляционных баз данных (см., например, [4, 101, 167]), заключается в рассмотрении многоместных отношений с их последующей декомпозицией в процессе нормализации. Его недостаток состоит в излишней привязке к структуре данных, поэтому актуальна задача разработки модели связей, обладающей более высоким уровнем абстрагируемости от структуры данных.

В-третьих, поскольку основной особенностью научно-информационного процесса с участием интернет-документов является необходимость и возможность частичной автоматизации процессов извлечения их метаданных и классификации, возникает задача

создания соответствующих программных средств. При этом важно подчеркнуть, что в качестве эксперта, координирующего функционирование таких средств, может выступать любой пользователь системы, обладающий необходимым уровнем квалификации.

Разумеется, существуют программные инструменты, решающие те или иные частные вопросы автоматизации указанных процессов. Так, для извлечения из текстов информации на основании гипертекстовой разметки обрабатываемых документов созданы пакеты RoadRuuner, Lixto и др. (см., например, [225, 251]). Однако коммерческий характер таких программ и необходимость их специальной установки на компьютере каждого пользователя-эксперта (а количество таких пользователей может исчислять десятками и даже сотнями, притом они могут находиться в разных регионах России и даже мира) делает актуальной задачу реализации алгоритмов, автоматизирующих основные этапы научно-информационного процесса, посредством интернет-приложений, доступных с любого компьютера сети (разумеется, после аутентификации и авторизации пользователя-эксперта).

К этому же кругу задач относится разработка автоматизированной технологии создания тезауруса и онтологии той или иной предметной области, которая обеспечивала бы высококвалифицированное описание предметной области с использованием надежно выверенных терминов, позволяла бы минимизировать трудозатраты специалистов—экспертов.

В-четвертых, важной проблемой остается разработка структур представления научной и научно-организационной информации. Поскольку на практике большинство рядовых пользователей испытывают затруднения в самостоятельном построении запросов более сложных, нежели простой контекстный или атрибутивный поиск, постольку необходимо, чтобы базовая структура представления информации отвечала такой совокупности заранее сформулированных информационных запросов, которая была бы в состоянии удовлетворить основные информационные потребности пользователей системы.

Разумеется, конкретное решение всех перечисленных задач невозможно без подробного анализа современного уровня информационных потребностей научного сообщества, чему и будет посвящена следующая глава.

Глава 2

АНАЛИЗ ИНФОРМАЦИОННЫХ ПОТРЕБНОСТЕЙ НАУЧНОГО СООБЩЕСТВА

> Большинство людей не получают того, чего хочет, а всё потому, что сами не знают чего хотят. Как исполнять желания, если их нет?
>
> *Дж. Б. Пристли. Тридцать первое июня*

2.1. Основные характеристики информационных потребностей в сфере науки

> Не вкусивши сладкого, горькое еще можно есть, а раз вкусивши сладкое, кислое уже неприятно.
>
> *Козьма Прутков. Из «Сборника неоконченного (d'inachevé)»*

Одним из наиболее распространенных способов исследования информационных потребностей является пришедшая из англоязычной литературы формула «5 “W” + 1 “H”», восходящая к хрестоматийному стихотворению Дж. Р. Киплинга «Шестерка слуг» («Six serving men»):

I have six honest serving men,
They taught me all I knew.
Their names are What and Why and When,
And How and Where and Who.

Эта формула заключается в постановке шести вопросов, которые, как отмечено в [3], предъявляются лицу, запросившему информацию, но, очевидно, могут быть поставлены и априорно при анализе информационных потребностей некоторой социальной группы. Итак, вот эта формула, адаптированная к сфере науки:

— WHERE (Где?) Где работает потенциальный потребитель информации: в научно-исследовательском учреждении, на предприятии, в правительственном ведомстве и т.д.?

— WHAT (Какая?) Какая информация может представлять интерес: об окружающей среде, о конкурентах, внутренняя? Какого

типа: в исходном виде, ретроспективная, текущая, прогностическая?

— WHO (Кто?) Кто может выступать в качестве потенциального потребителя информации: исследователь, специалист (инженер, агроном и т. п.), управленец?

— WHY (Зачем?) Зачем требуется информация: для научного исследования, разработки, изучения, планирования или управления?

— WHEN (Когда?) Когда требуется информация: немедленно, регулярно, по мере необходимости?

— HOW (Как?) В каких видах и форме требуется информация: в виде оригиналов, в машиночитаемом виде, подвергнутая анализу или другой обработке?

Для описания принципов организации информационного обеспечения научной (включая научно-организационную, инновационную и т. п.) деятельности необходимо дать ответы на перечисленные вопросы. Следует отметить, что в постановке некоторых вопросов, например «Где?», «Кто?» и «Зачем?», наблюдается определенное сходство, поэтому ответы на них в значительной мере коррелируют (конкретные особенности корреляции будут описаны ниже.)

Разумеется, наиболее важными и сложными являются вопросы «Какая?» (особенно в части, касающейся типа предоставляемой информации) и «Как?». Ответы на них, зависящие от вариантов ответов на вопросы «Где?», «Кто?» и «Зачем?», подразумевают подробное описание методов реализации предложенных решений.

При ответе на вопрос «Где?» необходимо учитывать, что информационное обеспечение научной деятельности предполагает не только удовлетворение специальных информационных потребностей научного сообщества, но и предоставление информации работникам других отраслей экономики и управления, могущих сотрудничать с научными учреждениями, ибо такое сотрудничество способствует в конечном счете повышению эффективности научной деятельности. Здесь речь идет, прежде всего, о промышленных предприятиях и различных финансовых институтах, заинтересованных в получении информации об инновационных разработках с целью их внедрения или коммерциализации, а также об органах государственной власти и местного самоуправления, которые заинтересованы в получении разнообразной информации научного и на-

учно-организационного характера для принятия соответствующих управленческих решений.

Способы удовлетворения соответствующих информационных потребностей будут рассмотрены при ответах на вопросы «Кто?», «Зачем?» и «Какая?».

В постановке вопроса «Кто?» наблюдается заметная корреляция с вопросом «Где?» (хотя и неполная, поскольку в организациях каждого из типов, перечисленных в вопросе «Где?», могут работать разные категории потребителей информации, причем информационные потребности, даже применительно к научно-технической информации, например, управленцев, представляющих науку, промышленность и органы государственной власти, несколько различны). Эта корреляция позволила разделить информационные потребности в сфере науки и техники на три основных вида:

— информационные потребности ученых-исследователей;

— информационные потребности специалистов;

— информационные потребности управляющих (руководителей).

Основные различия в выражении и удовлетворении информационных потребностей указанных групп потребителей информации сведены в [3] в табл. 2.1.

Т а б л и ц а 2.1. **Основные различия информационных потребностей ученых, специалистов и управляющих**

Характеристика выражения и удовлетворения информационных потребностей	Ученые	Специалисты	Управляющие
Четкость осознания и выражения	Небольшая	Большая	Очень большая
Требующаяся полнота информации	Большая	Не больше, чем нужно	Самая нужная
Срочность удовлетворения	Не важна	Важна	Очень важна
Форма получения информации	Любая	Удобная для использования	Максимально удобная для восприятия
Степень переработки первичной информации	Минимальная	Большая	Очень большая
Виды предпочтительной информации	Первоисточники, библиография, численные данные	Фактографическая информация	Обзорно-аналитическая информация

Представленные в таблице характеристики «Четкость осознания» и «Полнота информации» призваны ответить на вопрос «Какая?», характеристика «Срочность» — на вопрос «Когда?», а характеристики «Форма получения», «Степень переработки» и «Виды предпочтительной информации» — на вопрос «Как?».

Нельзя не отметить, что столь резкое разграничение информационных потребностей ученых с аналогичными потребностями специалистов и управленцев (которое еще сильней выражено в монографии [103], изданной в 1976 г.), более характерно для общественно-экономических условий 1960–1980-х годов, когда наука и техника (технология) рассматривались как два существенно разных вида общественной деятельности (см. [103, с. 119]). Однако со середины 1990-х годов ситуация значительно изменилась. С одной стороны, благодаря господству высоких технологий размывается грань между наукой и производством [68]. С другой стороны, в России произошло изменение принципов функционирования и финансирования науки, вследствие чего ученые, используя результаты своих фундаментальных исследований, стали более активно заниматься опытно-конструкторскими работами и даже непосредственным производством уникальных наукоемких изделий (подробнее эти обстоятельства будут проанализированы далее в разд. 2.3). Кроме того, выполнение проектов Федерального агентства по науке и инновациям, грантов РФФИ и т. п. требует управленческих навыков, включая использование управленческой информации, не только от руководства научных учреждений, но и непосредственно от ученых-исследователей.

В частности, в работе авторов [28] на основании анализа результатов социологических опросов молодых ученых СО РАН [51] показано, что молодые исследователи, являющиеся наиболее активными пользователями сети Интернет, испытывают насущную потребность в разнообразной научно-организационной и управленческой информации.

Следовательно, ответ на вопрос «Зачем?» для специалистов и управляющих более или менее очевиден: информация им нужна соответственно для проведения разработок и для осуществления планирования и управления. Что же касается ученых, то им информация требуется как для научных исследований, так и для выполнения функций разработчиков и управленцев.

Итак, можно констатировать, что табл. 2.1, приведенная в монографии [3], в настоящее время применительно к ученым отвечает на вопрос «Зачем?» в соответствии с тем аспектом деятельности научных работников, для осуществления которого требуется данная информация. К тому же некоторые положения таблицы нуждаются в уточнении. Так, утверждение, что «четкость осознания и выражения информационных потребностей ученых небольшая», конечно же, верно применительно к процессу научного поиска, но не совсем применимо к повседневному труду ученого. Например, известный математик Д. Пойа указывал [127, 129], что процесс решения задачи начинается с распознавания ее элементов с использованием определений, т. е. в данном случае имеется четко выраженная информационная потребность. Аналогичная ситуация возникает, когда ученый ищет конкретную публикацию и т. п.

Таким образом, как отмечено в [103], существует два типа информационных потребностей:

1) потребности в сведениях об источниках необходимой научной информации;

2) потребности в само́й необходимой научной информации.

При этом ученым-исследователям свойственны оба вида информационных потребностей.

Однако наибольшее влияние современные высокие технологии оказали на решение вопроса «Когда?». Развитие телекоммуникационных систем, прежде всего сети Интернет, привело к появлению принципиальной возможности практически немедленного удовлетворения возникающих информационных потребностей, а также значительно упростило регулярное предоставление пользователю периодически обновляемой информации, например посредством рассылки электронных почтовых сообщений.

Как же быть с приведенным в табл. 2.1 утверждением, что для ученого срочность удовлетворения информации не важна (по крайней мере, когда речь идет об информации, необходимой непосредственно для поведения научных исследований)? Очевидно, что при прочих равных условиях предпочтительно скорейшее удовлетворение возникшей информационной (как, впрочем, и любой другой) потребности, поэтому вывод о «несрочности» потребности в научной информации был продиктован прежде всего экономическими соображениями. Это было вполне оправданно в тех условиях, когда отсутствовала возможность немедленного удовлетворения инфор-

мационных потребностей, и различие между «срочным» и «несрочным» предоставлением информации носило сугубо количественный характер. В настоящее время немедленное удовлетворение информационных потребностей по-прежнему несколько более затратно, чем предоставление информации по мере необходимости (например, создание веб-сайта научного журнала с полными текстами статей организационно сложнее, чем размещение на сайте одних только аннотаций, предполагающее получение пользователем твердых копий заинтересовавших его публикаций по обычным библиотечным каналам). Однако здесь вступают в дело психологические соображения.

Психологическое влияние своевременного удовлетворения информационных потребностей на производительность труда было отмечено еще в начале 1920-х годов А. А. Богдановым (предвосхитившим в своей «Тектологии» не только общую теорию систем, но и некоторые основные принципы кибернетики): «…если грамотный, культурный рабочий лишается привычной уже для него газеты, чтения брошюр, книг, то падает его “настроение” и опять понижается рабочая сила», вследствие чего обстоятельства, «вынуждающие к сокращению таких потребностей, могут иметь серьезное значение для работоспособности» [41, т. 1, с. 265]. Подчеркнем, что в процитированном фрагменте речь идет о влиянии на работоспособность *текущих* информационных потребностей (т. е. непосредственно не связанных с производственным процессом данного работника). Разумеется, неудовлетворенные *конкретные (специальные)* потребности (потребности в профессиональной информации), появившиеся с развитием сети Интернет, в частности невозможность оперативного получения нужной информации, тем более могут оказать негативное влияние на работоспособность. Таким образом, необходимо максимально срочное удовлетворение информационных потребностей научных работников.

Изложенное ни в коей мере не означает, что классические способы удовлетворения информационных потребностей посредством получения информации на бумажных носителях, общения на конференциях и т. п. ушли в прошлое, однако наиболее перспективным направлением развития информационного обеспечения научной деятельности являются все-таки электронные информационные технологии. В соответствии с целями данного исследования далее мы будем вести речь только о тех способах удовлетворения

информационных потребностей научного сообщества, которые базируются на электронных технологиях. В рамках указанного подхода основным инструментом информационного обеспечения научной деятельности являются *информационные системы*, включая их наиболее совершенный класс — *интеллектуальные системы* (подробнее см. разд. 1.4). В настоящее время в подавляющем большинстве случаев подразумевается, что информационная система обладает удаленным доступом через сеть Интернет.

Таким образом, возможность срочного удовлетворения информационных потребностей научного сообщества зависит от оперативности появления научной информации в сети Интернет, а также регистрации ее в каталоге той или иной информационной системы, ибо в противном случае соответствующий информационный ресурс практически не имеет шансов стать достоянием широкой научной общественности.

Исходя из сформулированных выше положений, приступим к ответу на основные вопросы, характеризующие информационные потребности: «Какая?» и «Как?».

2.2. Исследование информационных потребностей коллективных пользователей — научных учреждений СО РАН

> Пиши письма такие, какие тебе хотелось бы получать.
>
> *Кретя Патачкувна. Цит. по книге: «Мысли людей великих, средних и пса Фафика»*

Перейдем к более конкретному анализу потребностей научного сообщества в информации, распространяемой с помощью электронных технологий. Поскольку, как сказано в БСЭ [208], «потребности социальных субъектов (личностей, социальных групп)... зависят от уровня развития данного общества, а также от специфических социальных условий их деятельности», постольку развитие сети Интернет, а также резкий рост производительности персональных компьютеров и веб-серверов обусловили качественный рост информационных потребностей субъектов научной деятельности.

Изучение информационных потребностей ученых осложняется тем, что эти потребности зависят от множества разных факторов и носят в значительной мере персонифицированный характер (см. [3, с. 226]). При этом одним из более или менее объективных методов определения информационных потребностей является построение картины фактического использования учеными разных видов источников информации, а также собственная оценка учеными относительной важности этих источников.

В качестве объекта исследования рассмотрим «внутренние» информационные потребности Сибирского отделения РАН (под «внутренними» мы подразумеваем потребности ученых СО РАН в научной информации из источников самого́ СО РАН). Репрезентативность подобной выборки доказывается следующими фактами.

Сибирское отделение РАН — это расположенные на территории трех федеральных округов почти 90 научно-исследовательских и конструкторско-технологических учреждений, в которых работает более 20 тыс. человек, в том числе (по состоянию на 1 января 2009 г.) 8718 научных сотрудников, из них 130 членов РАН, 1853 доктора и 4725 кандидатов наук [47].

Для удовлетворения информационных потребностей Сибирского отделения создана Сеть передачи данных СО РАН [154], в которой зарегистрировано около 150 организаций-абонентов. Только в Новосибирске Сеть обслуживает более 40 тыс. пользователей и насчитывает более 12 тыс. подключенных компьютеров. Кроме того, в региональных научных центрах Отделения находится еще около 30 тыс. пользователей. Суммарный объем информации, получаемой и отправляемой по каналам Сети, составляет более 700 Гбайт в сутки, при этом 58 % общего объема составляет информация, получаемая абонентами из Сети, а 42 % — передаваемая ими во внешний мир.

Высокий уровень информатизации СО РАН (и, следовательно, развитые информационные потребности работающих в нем ученых) подтверждает рейтинг Webometrics Кибернетической лаборатории Национального исследовательского совета Испании [260]. В этот рейтинг входят сайты ведущих научно-исследовательских центров всего мира, при его подсчете основное значение имеет число размещенных на сайте научных работ и количество ссылок на них. По состоянию на январь 2010 г. сайт Сибирского отделения РАН занимал 1-е место среди российских сайтов (17-е — в Европе, 49-е — в мире).

К сожалению, построение картины фактического использования учеными СО РАН различных источников «внутренней» информации пока не осуществлено, поскольку система мониторинга и сбора статистики Сети передачи данных СО РАН [189] создана совсем недавно и еще не накопила достаточного количества данных для проведения соответствующего анализа.

С другой стороны, имеются данные, позволяющие исследовать оценку научным сообществом СО РАН сравнительной важности тех или иных форм удовлетворения информационных потребностей коллективных пользователей. В качестве таких данных мы рассматриваем итоги четырех конкурсов интеграционных проектов СО РАН, проводимых раз в три года. Проекты-победители определялись Постановлениями Президиума СО РАН [131–137]. По итогам конкурса 2000 г. победителями признаны 88 проектов, 2003 г. — 180, 2006 г. — 247, 2009 г. — 267 проектов. В каждом проекте принимали участие несколько институтов СО РАН (а иногда и других научных организаций), притом институты — участники проекта — представляли, как правило, разные направления наук. Описанная процедура проведения конкурсов позволяет сделать вывод, что проекты, так или иначе связанные с информатикой, адекватно отражают информационные потребности коллективных пользователей — научных учреждений СО РАН.

Подавляющее большинство проектов—победителей, которые предусматривали получение новых результатов в области теоретической информатики или (и) использование методов теоретической информатики для создания программных систем информационного обеспечения научной деятельности на основе новых интернет-технологий, имело в качестве организаций—исполнителей один или несколько академических институтов, занимающихся исследованиями в области информатики: Институт математики СО РАН (ИМ), Институт вычислительных технологий СО РАН (ИВТ), Институт систем информатики СО РАН (ИСИ), Институт вычислительной математики и математической геофизики СО РАН (ИВМиМГ), Институт вычислительного моделирования СО РАН (ИВМ), Институт динамики систем и теории управления СО РАН (ИДСТУ), Институт математики и механики УрО РАН (ИММ), Институт автоматики и процессов управления ДВО РАН (ИАПУ), а также научно-исследовательские организации или вузы аналогичного профиля: Научно-исследовательский вычислительный

центр Московского государственного университета (НИВЦ МГУ), Новосибирский государственный университет (НГУ), Сибирский федеральный университет (СФУ), Томский университет систем управления и радиоэлектроники (ТУСУР).

Анализ списков проектов показал (см. [22]), что можно выделить пять основных типов задач из области информатики, решаемых в рамках этих проектов, причем некоторые проекты могут соответствовать сразу нескольким типам задач (далее в обозначениях проектов первое число означает номер таблицы, второе — порядковый номер проекта в таблице):

1) исследование и моделирование (в том числе когнитивное) интеллекта; 2) разработка средств анализа моделей информационных структур; 3) проведение компьютерного анализа большого массива данных в той или иной области с целью получения новых знаний ; 4) разработка и создание с использованием интернет-технологий специализированных информационных систем на основе современных алгоритмов обработки данных; 5) исследование общих принципов организации телекоммуникационных систем. Указанные данные сведены в табл. 2.2.

Отсюда следует вывод, что коллективные пользователи — научные учреждения СО РАН — особенно нуждаются в разработке специализированных информационных систем и в технологиях получения новых знаний из данных, причем первая из названных задач тесно увязана со второй. Таким образом, научное сообщество испытывает все более растущую потребность не просто в информационных системах, но в системах, извлекающих из имеющихся данных новые знания, то есть в интеллектуальных системах.

Т а б л и ц а 2.2. **Анализ тематики интеграционных проектов СО РАН, связанных с информационными технологиями, %**

Год	Моделирование интеллекта	Анализ моделей информационных структур	Телекоммуникационные системы	Анализ данных, извлечение знаний	Создание информационных систем	
						из них с анализом данных
2000	0	0	20	40	80	**25**
2003	8	0	17	50	67	**50**
2006	9	9	27	55	45	**80**
2009	20	0	30	80	50	**100**

2.3. Информационная модель описания деятельности научного сообщества

> Мистер Кенфорд принадлежал к методистам старого закала и подозрительно, даже враждебно относился ко всякой человеческой деятельности, кроме круглогодичной работы на ферме, купли-продажи, прибыли-экономии, морального осуждения и принятия пищи.
>
> *Дж. Б. Пристли. Трое в новых костюмах*

Прежде всего, сформулируем, информация о каких сущностях (точнее, классах сущностей) требуется при описании той или иной отрасли человеческой деятельности (сразу подчеркнем, что здесь речь идет исключительно о деятельности, связанной с *информационными объектами*). .

Любая деятельность человека предполагает определенное противопоставление субъекта и объекта деятельности [115], причем в качестве субъекта деятельности могут выступать как отдельные люди, так и группы (коллективы) людей. В условиях современного общества производственно-технические отношения между людьми возникают, как правило, посредством вхождения этих людей в одну группу, а характер отношений определяется функциями конкретного человека в группе. В свою очередь группы также могут вступать между собой в те или иные общественные отношения (подчиненности, учредительства и т. п.).

Таким образом, процесс деятельности организации может быть охарактеризован описаниями следующих сущностей (см. работу авторов [15]):

1) субъекты деятельности:
   а) группы,
   б) отдельные лица;
2) объекты деятельности:
   а) предметы деятельности,
   б) продукты деятельности,
   в) акты деятельности.

Между этими сущностями устанавливаются связи:

1) отношения между субъектами и объектами деятельности:
   а) группа — объект деятельности,
   б) лицо — объект деятельности;

2) отношения между субъектами деятельности:
   а) группа — группа,
   б) группа — лицо,
   в) лицо — лицо.

Что касается связей между объектами деятельности, то, ввиду сложности соответствующих моделей, а также их большой специфичности для каждой конкретной сферы деятельности, эти вопросы будут рассмотрены в разд. 4.1, где речь пойдет о создании тезаурусов и онтологий.

Выбор конкретного множества описаний из приведенного списка определяется родом деятельности организации. Далее мы выделим отличительные особенности систем информационного обеспечения научной деятельности (см. работы авторов [19, 22]). Сразу отметим, что весь последующий анализ, касающийся представления информации о деятельности коллективов (или, если использовать более привычный термин, организаций), основан на изучении российского сектора сети Интернет, поскольку отечественная корпоративная культура, влияющая, в частности, на особенности представления организациями информации о своей деятельности, во многом обусловлена особенностями исторического развития России и весьма отличается от корпоративной культуры западных стран.

Техническое разделение труда приводит к тому, что информационные системы *производственных организаций*, как правило, не содержат связей между субъектами и объектами (продуктами) деятельности, поскольку последние являются результатом коллективного труда, к тому же продукты деятельности в таких информационных системах обычно выступают в качестве товаров. Более того, на интернет-сайтах производственных организаций подробные списки подразделений и персон обычно отсутствуют, как не представляющие интереса для внешних пользователей (хотя такие списки могут быть доступны из соответствующей локальной сети). К примеру, на сайтах крупнейших российских корпораций [160], таких как Газпром, ЛУКОЙЛ, Роснефть и др., в подавляющем большинстве случаев содержится информация только о руководстве и совете директоров, в то время как основное наполнение таких сайтов направлено на отражение корпоративной политики и деятельности компании.

Напротив, информационные системы *политических, общественных, религиозных и т. п. организаций* предполагают, как правило, подробное представление информации о структурных подразделениях организации и о персональном составе руководства, но не содержат развернутой информации об объектах деятельности, так как деятельность организаций такого рода обычно описывается в виде сообщений об актах деятельности (т. е. событиях), однако такие сообщения не привязаны к конкретному лицу, а отображаются в виде новостной ленты.

Именно так устроен официальный сайт Русской Православной Церкви [119], содержащий развернутую персональную информацию о высших церковных иерархах, однако не предоставляющий возможности связать конкретную персону (за исключением Патриарха) с тем или иным событием. Сайты буддистских сообществ России, как правило, предоставляют информацию об основных учителях той или иной школы, в то время как порталы, посвященные иудаизму и исламу, содержат только общую информацию о религии, последние новости и т. д.

Для большинства сайтов зарегистрированных в России *политических партий* [124] характерно наличие информации о руководящем составе партии, однако во всех случаях данная информация организована по-разному. На сайте «Единой России» доступ к персональной информации конкретного лица может быть получен через дерево должностей организации, а в случае ЛДПР хоть сколько-нибудь развернутая информация имеется только о лидере партии. В редких случаях присутствует связь персоналий с актами их деятельности (выступлениями, заявлениями и т. д.), в частности, такая связь наблюдается на сайтах КПРФ (применительно к членам фракции КПРФ в Государственной Думе) и Российской объединенной демократической партии «Яблоко». Для последней все подобные события классифицированы как публикации, а также отсортированы по году; имеется достаточно подробная информация о биографии каждой персоны, включающая, помимо уже отмеченного раздела «Публикации», также информацию о работе над законопроектами, депутатских запросах и обращениях, контактных адресах и ссылках на другие источники, однако все эти данные представлены в виде статических HTML-страниц, что, очевидно, затрудняет процесс их актуализации.

В информационных системах *органов законодательной власти* наряду с подробной информацией о персональном составе содержатся сведения и о «продуктах деятельности» этих организаций (законодательных актах), но с конкретными лицами эта информация, как правило, не связана (исключение составляет сайт Государственной Думы [118]). Что же касается *органов исполнительной власти*, то сложившаяся в современной России практика такова, что содержащиеся на их информационных интернет-сайтах сведения о структуре этих организаций носят самый общий характер, сведения о персональном составе (за исключением узкого круга руководства) вообще отсутствуют, а «продукты деятельности» (постановления, распоряжения и пр.) отображаются на официальных сайтах весьма выборочно, в то время как информация о событиях — подробно.

Итак, отличительной особенностью построения информационных систем перечисленных типов организаций является отсутствие (в подавляющем большинстве случаев) связи между субъектами и объектами деятельности, и как следствие, фактический распад информационных систем на ряд отдельных подсистем, слабо связанных между собой (что характерно и для сайта Государственной Думы, где каждому созыву соответствует «новая» биография лица, даже если это лицо было депутатом и других созывов, при этом отсутствует привязка законодательных актов непосредственно к описанию лица — автора соответствующей законодательной инициативы).

Напротив, информационные системы организаций, которые исходя из специфики их деятельности должны содержать как подробную информацию о персональном составе (включая возможность вхождения в целый ряд структур, а также отслеживание служебных перемещений), так и связи между субъектами и объектами деятельности, должны иметь более сложную структуру.

К числу таких организаций можно отнести творческие и научные учреждения, объединения и т. п.

К сожалению, российские *творческие организации* представлены в Интернете довольно слабо. Так, на сайте Союза писателей России [120] присутствует только список членов руководящих органов, хотя, если судить по печатным версиям его справочников, имеется вполне адекватная информационная модель описания деятельности Союза писателей, учитывающая вхождение его членов в

те или иные структуры, созданные ими произведения, полученные награды и пр.

У большинства других творческих союзов (художников, композиторов) официальные сайты вообще отсутствуют, а в тех случаях, когда такие сайты удавалось найти, они, за редким исключением, не содержали никакой персональной информации даже о руководстве.

Следует отметить, что некоторые наиболее известные учреждения культуры, например Большой [145], Мариинский [147], Екатеринбургский оперный [146], Новосибирский оперный [149] театры, имеют весьма содержательные информационные системы, включающие творческие биографии ведущих членах художественного коллектива, подробное описание постановок, содержащее связи с задействованными в них персонами, и т. п. Однако историческая ретроспектива, а также средства поиска информации в них обычно представлены недостаточно.

В чем же состоят отличия информационной модели описания деятельности научных организаций от аналогичных моделей творческих организаций и союзов?

Предметы деятельности членов творческих союзов (тексты, музыка, изображения) по своей природе сходны с результатами научной деятельности (новыми знаниями в форме информационных продуктов) тем, что после создания они становятся независимыми от создателя и доступны для ознакомления, т. е. информация о них сохраняет актуальность неопределенно долгий срок. Однако создатели информационных продуктов художественного характера нацелены на их коммерческое распространение. В то же время, как отмечено в монографиях ВИНИТИ [3, 103], объектом купли-продажи является не сама научная информация, а ее носители (книги, журналы и т. п.), а также право на использование отдельных видов научной информации с целью получения прибыли. Сам же научный работник заинтересован в максимально широком распространении созданной им информации, причем обычно (за исключением научно-инновационных разработок) неважно, какую форму носит это распространение — коммерческую или бесплатную, тем более что в современных российских условиях ученый, как правило, не получает дохода от реализации носителей научной информации.

Поэтому оптимальной формой представления публикации в информационной системе научного сообщества является размещение описания публикации с указанием адреса полной электронной версии, в то время как художественный продукт обычно представляется посредством описания, а полная версия, если и присутствует в системе, то предоставляется на коммерческой основе (разумеется, речь идет об информационных системах, создатели которых соблюдают законодательство об авторских правах).

Отметим, что специфика функционирования отдельных подразделений научного сообщества, занимающихся коммерческим распространением научной информации, например библиотек, в настоящей работе не затрагивается. Эти вопросы подробно рассмотрены, в частности, в диссертационной работе Л. К. Боброва [40].

Продолжая сравнительный анализ информационных моделей описания деятельности научных и творческих объединений, следует подчеркнуть, что ввиду специфики художественного творчества (отсутствия или слабой выраженности коллективного начала) информация о принадлежности членов творческих союзов к той или иной организационной структуре гораздо менее важна, чем аналогичная информация о научных работниках (на сайтах научных организаций, как правило, присутствуют сведения не только об их административной структуре, но и сведения о вхождении персон в состав ученых советов, советов по защите диссертаций и т. п.). Тем не менее и для научной деятельности применительно к исторической ретроспективе характерна аналогичная особенность: информация, сконцентрированная вокруг организаций, сообществ и т. п., зачастую утрачивает с течением времени свою актуальность. Так, для нас может представлять интерес метод Бубнова — Галёркина решения операторных уравнений или сама биография И. Г. Бубнова, но вряд ли мы будем искать эту информацию посредством поиска сведений о Морской академии или Опытовом судостроительном бассейне, где служил Бубнов. В значительной мере теряется интерес и к актам деятельности. Поэтому при создании информационных систем по истории науки в первую очередь рассматриваются субъекты — отдельные лица, объекты — предметы и продукты деятельности.

С другой стороны, актуальность информации о продуктах деятельности организаций культуры (спектаклях, концертах и т. п.) сильно зависит от того, сохраняется ли данный продукт в репер-

туаре. Отсюда следует, что персональная информация в таких системах также утрачивает с течением времени свою актуальность (за исключением информации о наиболее выдающихся деятелях). Структурирование творческих коллективов (особенно в группы вне административной структуры) выражено весьма слабо, в отличие от научных организаций (и творческих союзов).

Результаты проведенного анализа обобщены в таблице 2.3 (знак «+» означает актуальность представления соответствующего аспекта, «±» актуальность при отсутствии в большинстве случаев практической реализации, «–» неактуальность).

Таким образом, информационная модель описания научной деятельности отличается следующей совокупностью признаков, указанных в работе авторов [15]:

1) наличие подробной информации о персонах, причем сведения об отношении персоны к структуре не утрачивают актуальность даже после прекращения данного отношения;

2) необходимость включения подробной информации о структуре групп, с учетом групп, не входящих в основную административную структуру (это требование не является обязательным для случаев описания ретроспективного аспекта деятельности);

3) возможность вхождения персоны сразу в несколько групп;

4) максимально подробное представление информации о предмете деятельности;

Таблица 2.3. **Сравнительный анализ информационных моделей описания различных сфер деятельности**

Сфера деятельности	Подробная информация о персонах	Отображение включения персоны в несколько групп	Сохранение неактуальных связей между персонами и группами	Связи между субъектами и объектами деятельности	Отображение подробной информации о предмете деятельности
Производственные организации	–	–	–	–	+
Общественные организации	+	–	–	–	–
Законодательные органы	+	±	+	±	+
Творческие коллективы	+	–	±	+	–
Научные организации	+	±	+	+	+

5) наличие сохраняющих актуальность связей между персонами и предметом деятельности.

Заметим, что к информационной модели описания деятельности научного сообщества наиболее близка модель описания деятельности органов законодательной власти, однако она, как уже отмечалось, реализована далеко не в полной мере даже на сайте Государственной Думы.

Отличительной особенностью предложенной информационной модели описания деятельности научного сообщества является, во-первых, четкое выделение субъекта и объекта деятельности. Предложенная модель эффективна при описании как научной деятельности в той или иной предметной области (когда основные субъекты деятельности — персоны), так и деятельности крупных научных корпораций (когда в качестве основных субъектов деятельности, наряду с персонами, выступают организации). Для сравнения заметим, что модель, использованная при создании Единого научного информационного пространства РАН [36] излишне персоноцентрична, например «организации» рассматриваются в одном ряду с «проектами».

Другой отличительной особенностью предложенной модели является неиерархичность структуры субъектов деятельности, возникающая из-за возможности вхождения персоны сразу в несколько групп. Ввиду этого требует решения проблема работы с персональными данными, которые могут одновременно принадлежать к разным ветвям иерархического дерева и вместе с тем должны однозначно определять персону, поскольку предполагаемая возможность извлечения из содержащихся в информационной системе данных новой информации и знаний влечет за собой необходимость наличия связи имен собственных (как элементов библиографического описания и т. п.) с информацией о конкретных носителях этих имен, ибо в противном случае имя несет лишь назывную, но не информационную функцию [103].

Разумеется, возможно создание и «облегченных» вариантов моделей описания научной деятельности, которые призваны отражать какие-либо отдельные ее аспекты. Дальнейшая конкретизация информационных моделей, описывающих научную деятельность, будет проведена в разд. 3.4.

Особо следует рассмотреть вопросы информационного обеспечения инновационной деятельности, под которой обычно понимают

процесс преобразования научной идеи в конкретный продукт, услугу или технологию с последующим обеспечением их практического использования в народном хозяйстве (см, например, [71]). Отметим, что «расширение связей между наукой и производством, участие в инновационной деятельности, в реализации достижений науки и техники» является одной из основных уставных задач Российской академии наук [169].

Подобный подход к инноватике, рассматривающий внедрение достижений науки в неразрывной связи с получением фундаментальных знаний, восходит к родоначальнику методологии эмпирической науки Фрэнсису Бэкону. В утопической повести «Новая Атлантида» Ф. Бэкон, излагающий первый в истории проект государственный организации науки, приписывает «Дому Соломона» (научно-техническому центру утопического общества) не только функции организации научно-исследовательской и изобретательской деятельности, но и внедрение полученных достижений в экономику и быт [45].

В последнее время получает распространение термин «научно-инновационная деятельность», призванный однозначно разграничить этапы исследований и разработок от этапа серийного производства. Его появление обусловлено особенностями изменения цикла воспроизводства инноваций, происшедшими в России за время перехода к рыночной экономике.

В советской экономике имел место следующий цикл воспроизводства инноваций [60]:

деньги (ресурсы) → фундаментальные исследования, возложенные на научное сообщество → опытные образцы и технологии, создаваемые в отраслевых НИИ → внедрение или создание новых производств → товары, услуги, продукция, дающие экономический эффект в условиях плановой экономики → деньги.

С точки зрения содержания начальных этапов этот цикл практически не отличался от аналогичного цикла, присущего либеральной экономике, например американской:

деньги → фундаментальные исследования и изобретения → создание опытных образцов, обработка технологий → внедрение, вывод на рынок, вытеснение предшествующих образцов → производство товаров → продажа на рынке → деньги.

Однако выполнение начальных этапов в условиях либеральной экономики обычно возложено не на государственные учреждения, а на научно-исследовательские и конструкторские организации, принадлежащие корпорациям.

При переходе от плановой экономики к рыночной в России сохранилось финансируемое государством научное сообщество, но резко сократилось количество отраслевых НИИ, выполнявших основную часть опытно-конструкторских работ. В то же время процесс создания аналогичных структур частными компаниями явно затягивается. В результате этого научные организации нередко вынуждены брать непосредственно на себя функции отраслевых НИИ и КБ по выполнению ОКР, также самостоятельно (или через дочерние структуры) проводить коммерциализацию своих разработок.

Таким образом, как показано в [1], сложилась научно-инновационная отрасль народного хозяйства, включающая в себя три вида деятельности: научные исследования, разработки, инфраструктурные услуги.

Перечислим основные отличительные особенности информационной модели описания инновационной деятельности, отмеченные авторами в работе [196]:

1) деперсонифицированность субъекта деятельности: в силу особенностей российского законодательства объект деятельности связан, как правило, с организацией, а не с персоной;

2) регулярное обновление информации, включая утрату отдельными источниками актуальности, что для собственно научных источников бывает крайне редко;

3) наличие большого количества «внешних» источников, например нормативно-правовых актов, регулирующих процесс инновационной деятельности, которые, тем не менее, могут быть непосредственно связаны с объектами деятельности (так, некоторые нормативно-правовые акты могут регулировать отношения лишь в сфере определенных технологий), в то время как при описании процессов собственно научной деятельности достаточно четко разграничены научные и научно-организационные аспекты.

Конкретизация информационной модели описания научно-ин новационной деятельности будет проведена в п. 3.4.3.

Глава 3

СТРУКТУРА ОСНОВНЫХ КОМПОНЕНТОВ ПРОГРАММНОЙ СИСТЕМЫ

> Как... обсуждать правительственные мероприятия, не владея ключом их взаимной связи? — «Не по частям водочерпательницы, но по совокупности ее частей суди об ее достоинствах». Это я сказал еще в 1842 г. и доселе верю в справедливость этого замечания.
>
> *Козьма Прутков. Проект: о введении единомыслия в России*

3.1. Формулировка требований к программной системе

> Если я призываю вас, дорогие братья и сестры, ограничить свои требования до минимума, я не имею в виду мини-юбки.
>
> *Из одной проповеди. Цит. по книге: «Мысли людей великих, средних и пса Фафика»*

Используя методику общей теории систем, на основании положений, изложенных в гл. 1 и 2, сформулируем требования к программной системе информационного обеспечения научной деятельности. Методология системного анализа применительно к кибернетическим системам была описана в [96]. Выделены два основных подхода к изучению кибернетических систем: *макроподход*, при котором система рассматривается как «черный ящик» для исследования ее взаимодействия с окружающей средой, и *микроподход*, при котором изучается внутреннее строение системы. В рамках этих подходов сформулированы двенадцать основных направлений исследования систем (в рамках макроподхода — информационные потоки, коды, функции, функционирование систем; в рамках микроподхода — элементы, связи между элементами, алгоритмизация, анализ, синтез, преобразования, эволюция, надежность систем). При этом подчеркнуто, что некоторые из перечисленных направлений (прежде всего преобразования и эволюция управляющих систем) актуальны далеко не для всех типов систем, включая лингвистические (к которым относятся и информационные системы).

Отметим, что одна из первых попыток применения системного анализа для описания информационно-поисковых систем была предпринята Ю. А. Шрейдером в работе [200], однако там затрагивались вопросы макроподхода, в то время как системный анализ в данной ситуации особенно эффективен в рамках микроподхода, поскольку классическое определение системы «множество объектов вместе с отношениями между объектами и между их атрибутами» [182] основано на тех же понятиях, что и, например, реляционная модель данных [222], нередко применяемая для описания элементов информационной системы и связей между ними.

Основные системные принципы включают в себя [44, 90, 144]:

— *целостность* (зависимость каждого элемента, свойства и отношения от его места и функций внутри целого);

— *структурность* (возможность описания системы через установление ее структуры, т. е. сетей связей и отношений системы);

— *иерархичность* (каждый компонент системы в свою очередь может рассматриваться как система, а исследуемая система — как компонент более широкой системы);

— *множественность описания* (посредством использования множества различных моделей);

— *взаимозависимость системы и среды* (система формирует и проявляет свои свойства в процессе взаимодействия со средой, являясь при этом активным компонентом взаимодействия).

Нетрудно видеть, что в рамках макроподхода изучаются вопросы взаимозависимости системы и среды, в то время как комплекс проблем, изучаемых в рамках микроподхода, фактически приводит к исследованию целостности системы. Рассматривая отдельные аспекты микроподхода, можно отметить, что изучение элементов системы, в том числе выявление масштаба основных элементов — «кирпичей»(в терминологии [96]), требует исследования ее иерархичности, изучение связей — исследования структурности, анализ системы требует использования множества различных моделей ее описания.

Далее рассмотрим подробнее, как при создании информационных систем реализуются основные системные принципы.

Вопросы макроподхода (взаимозависимости информационной системы и среды) могут представлять интерес в плане как взаимодействия с пользователем — разработки моделей информационных запросов и моделей представления информации, так и отражения

системой изменений во внешней среде (т. е. актуализации информации), а также интероперабельности — интеграции системы с внешними информационными системами. Эти вопросы подробно обсуждались в гл. 1 и 2. Еще раз подчеркнем, что в настоящее время наиболее эффективным способом организации взаимодействия информационной системы с пользователями является ее включение во всемирную сеть Интернет, поэтому при дальнейших рассуждениях указанное обстоятельство будет подразумеваться явно или неявно. Интересно отметить, что алгоритмы извлечения из данных информации и знаний могут играть важную роль и в рамках макроподхода, например при обнаружении закономерностей и распознавании аномальных событий в потоке данных сетевого трафика, что обеспечивает безопасность системы от негативных внешних воздействий (см. работу авторов [48]).

Актуализация информации является слабым местом практически всех информационных систем некоммерческой направленности (за исключением, разумеется, систем, поддерживаемых государственными органами), предназначенных для функционирования в течение неопределенно долгого времени. Причина этого очевидна — недостаток средств для оплаты труда лиц, которые должны отслеживать изменения информации, а также предъявляемые к этим лицам высокие квалификационные требования, возрастающие с усложнением структуры и возможностей поддерживаемой информационной системы.

На заре развития информатики в монографии А. И. Михайлова и др. [104] были введены термины *информатор* — «специалист в какой-либо отрасли науки или практической деятельности, занимающийся исключительно научно-информационной деятельностью и использующий в своей работе достижения информатики» и *информатик* — «специалист в области информатики. Кадры информатиков пополняются, главным образом, за счет наиболее квалифицированных информаторов». Там же приводятся и англоязычные аналоги этих терминов: соответственно *information officer* и *information scientist*.

Однако реальное развитие информатики показало, что ученые не слишком часто начинают заниматься *исключительно* научно-информационной деятельностью и тем более полностью оставляют свою научную работу ради занятий информатикой. Именно поэтому в следующей монографии А. И. Михайлова и др. [103] об упомянутых специальностях речь уже не идет.

Опыт выполнения интеграционных проектов СО РАН, в рамках которых использовались методы теоретической информатики для создания программных систем информационного обеспечения научной деятельности на основе новых интернет-технологий (см. разд. 2.2), показал, что специалисты в какой-либо отрасли науки готовы участвовать в подобных проектах при условии, когда «черновая» информационная работа, неизбежная при каталогизации электронных документов научной тематики, составлении тезаурусов предметной области и т. п., в значительной степени автоматизирована посредством использования соответствующих программных средств, притом основную долю функций контроля качества полученной информации способен выполнить даже лаборант, и лишь в редких случаях требуется корректировка результатов с участием эксперта — научного работника.

Изложенное приводит к выводу о необходимости максимальной автоматизации процесса актуализации информации. Некоторые аспекты этой проблемы рассмотрены в работе авторов [30]. Одним из основных требований к информационной системе, процесс пополнения которой автоматизирован, является наличие в системе одного или нескольких *каталогов*, т. е. множеств унифицированных структурированных документов-описаний (фактически объединяющих поисковые образы исходных документов). В противном случае автоматическое добавление в систему новых документов становится, очевидно, крайне проблематичным.

Еще один важный вывод по итогам анализа интеграционных проектов, в рамках которых созданы программные системы информационного обеспечения какой-либо отрасли науки, например «Электронный атлас биоразнообразия животного и растительного мира Сибири» [207] или «Электронная библиотека MathTree» [206], состоит в том (см., в частности, [63, 172]), что подобные системы могут развиваться лишь в случае актуализации содержащейся в них информации самими пользователями этих систем.

Более того, даже относительно систем научно-организационной направленности, создаваемых в рамках одной большой научной корпорации (Сибирского отделения РАН), сделан вывод, что «эффективная эксплуатация информационных ресурсов возможна только в том случае, когда они *постоянно поддерживаются авторами*» [66].

Кроме того, как уже отмечалось, в интеллектуальных информационных системах компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека, при

этом происходит автоматизированное пополнение базы данных. В силу указанных обстоятельств при работе с интеллектуальными информационными системами многих пользователей возможности систем резко возрастают.

Поскольку пользователи, принимающие участие в актуализации информации, могут находиться в разных регионах России и даже мира, постольку практическое взаимодействие таких программных систем с внешним миром в плане занесения в них новых данных целесообразно организовывать преимущественно или даже почти исключительно через веб-интерфейс (как наиболее распространенную реализацию «тонкого» клиента).

Распределенное хранение информации, а также разнородность ресурсов, интегрируемых в рамка программных систем, порождают проблему *интероперабельности*, т. е. обеспечения взаимодействия разнородных информационных источников (как с целью их непосредственной интеграции, так и для организации поиска по однотипным подсистемам различных информационных систем). Теоретические вопросы интероперабельности обсуждаются, например, в работах [36, 179]. Коротко резюмируя их содержание, можно отметить, что построение распределенных информационных систем, а также организация в них поиска обеспечиваются посредством согласования схем метаданных (*семантическая интероперабельность*). Наиболее естественной формой унификации представления метаданных является *каталог*. Для интеграции разнородных систем, а также разнородных ресурсов внутри каждой отдельно взятой системы (такая интеграция необходима для извлечения из содержащихся в информационной системе данных новой информации и знаний) требуется согласование как моделей данных и форматов их представления (*синтаксическая интероперабельность*), так и протоколов доступа к ресурсам (*техническая интероперабельность*).

Переходя к микроподходу, отметим, что подсистемы интеллектуальной системы, определяемые формулами (1.1) или (1.2), имеют различную структуру. Наибольший интерес представляет структура информационно-поисковой системы, поскольку именно в ней содержатся данные, из которых извлекаются информация и знания. Далее в этом разделе под информационной системой мы будем подразумевать ИПС, входящую в состав интеллектуальной системы (анализ структуры логических компонентов интеллектуальной системы будет проведен в разделе 3.3).

В монографии А. И. Михайлова и др. [104] указывалось, что информационная система должна оперировать не непосредственно с документами, а с их *поисковыми образами*. При отсутствии поставленных в соответствие документам поисковых образов поиск документа возможен лишь по его адресу, что противоречит приведенному в разд. 1.4 нижнему ограничению сложности информационно-поисковой системы.

Заметим, что в некоторых случаях содержание исходного документа может входить в поисковый образ в качестве одного из элементов (это противоречит ограничению из [104], но из контекста данной работы следует, что подобное ограничение вызвано необходимостью уменьшения объема поисковых образов с целью уменьшения трудоемкости процесса их обработки). С другой стороны, поисковый образ документа тоже является документом (описывающим исходный документ), поэтому далее, говоря об информационных системах, мы будем использовать термин «документ» в значении «поисковый образ исходного документа». Следует подчеркнуть, что в классических работах по информатике и кибернетике [104, 156], вышедшим в конце 1980-х годов, поисковый образ документа не рассматривается даже в качестве вторичного документа.

Структурность, а также иерархичность интеллектуальной информационной системы тесно связаны с возможностью ее расширения. Составить информационную модель, которая очень подробно описывает соответствующую предметную область, — задача, как правило, не слишком сложная. Всё упирается в ограниченность ресурсов создателей системы. Поэтому необходим компромисс между качеством решения поставленной задачи и разумными сроками ее выполнения. Данный принцип давно является основополагающим в деятельности ученых, работающих в области вычислительной математики (см., например, монографию Н. С. Бахвалова [34]), при этом улучшение решения возможно с течением времени и достигается, применительно к информационной системе, посредством ее дополнения новыми подсистемами, которые, разумеется, связаны с уже существующими.

Таким образом, можно сформулировать основные требования к программной системе информационного обеспечения научной деятельности:

— лежащая в основе системы информационная модель описания деятельности научного сообщества (концептуальная модель предметной области) должна отражать различные аспекты дея-

тельности научного сообщества, включая научно-организационную и научно-инновационную;

— отвечающая основным системным принципам модель информационной системы (выступающей в качестве основного компонента создаваемой программной системы) должна позволять работать с основными элементами системы — документами (т. е. ресурсами, снабженными метаданными) как с целостными информационными объектами;

— структура связей в модели должна обеспечивать возможность принадлежности персоны одновременно к нескольким ветвям иерархического дерева групп — субъектов деятельности и вместе с тем однозначно определять персону, позволяя связывать имена собственные (как элементы библиографического описания и т. п.) с информацией о конкретных носителях этих имен;

— структуры представления информации и логических компонентов интеллектуальной системы должны обеспечивать удовлетворение потребностей пользователей (независимо от их квалификации в области информатики) в информации и знаниях, получаемых на основе данных системы;

— алгоритмы, обеспечивающие включение в научно-информационный процесс слабоструктурированных документов, должны обеспечивать максимальную автоматизацию всех его этапов (включая извлечение метаданных, определение ключевых слов, классификацию, а также предварительный этап создания тезауруса и онтологии предметной области), причем программные средства, реализующие эти алгоритмы, должны создаваться и функционировать как интернет-приложения.

3.2. Модель информационной системы

> Не нам, господа, подражать Плинию,
> Наше дело выравнивать линию.
>
> *Ф. К. Прутков. Военные афоризмы*

Прежде всего определим, что же является основным элементом информационной системы. Для этого предпримем небольшой исторический экскурс. Первоначально основной логической единицей хранения в информационно-поисковых системах являлась запись в базе данных, представлявшая собой поисковый образ доку-

мента [156]. При этом важно отметить, что записи не имели непосредственной связи друг с другом, что, как уже отмечалось, резко сужало возможности информационно-поисковых систем в плане получения новых знаний. Таким образом, стало ясно, что основные элементы системы должны быть связаны между собой, причем крайне желательно, чтобы эта связь носила естественный характер.

К началу 1980-х годов были созданы разнообразные модели данных, позволявшие весьма адекватно отражать свойства объектов внешнего мира, однако наиболее сложные из них не могли быть полноценно реализованы на существовавших тогда компьютерах [185]. Поэтому наибольшее распространение получила модель «сущность — связь» [221], в которой связь между объектами — элементами системы рассматривается как нечто внешнее по отношению к ним, вследствие чего затрудняется описание процессов взаимодействия. В то же время использование инфологической модели данных (впервые предложенной Б. Лангефорсом в [236, 237]), которая наиболее естественным образом отражает свойства объектов реального мира в терминах структур, ограничений и операций, сдерживалось ввиду больших затрат вычислительных мощностей при ее практической реализации.

В 1980-е — начале 1990-х годов наблюдался некий спад в развитии моделей данных высокого уровня, поскольку первые персональные компьютеры, пришедшие в те годы на смену большим ЭВМ (мэйнфреймам), явно уступали последним по вычислительной мощности и требовали совершенствования технологий работы с моделями данных более низкого уровня, в частности автоматизации процессов проектирования баз данных [82].

В настоящее время достаточно широкое распространение получила модель RDF [248] консорциума W3, предлагающая рассматривать в качестве элементов системы ресурсы, которые могут представлять и сущности, и их характеристики. Неудобство такого подхода очевидно: появляется множество равноправных мелких элементов, между которыми устанавливается чрезвычайно много связей, структура модели далека от естественной.

Особо следует подчеркнуть, что эта модель (как и другие модели, основанные на концепции Semantic Web) ориентирована на работу с хорошо структурированными документами, значения атрибутов метаданных которых суть элементы заданных словарей, что практически делает труднодоступным для обработки множество

размещенных в сети Интернет слабоструктурированных документов.

Как мы неоднократно отмечали, к числу основных свойств программной системы относится требование автоматизированного получения из данных новой информации и знаний. Это требование влечет за собой необходимость того, чтобы информационные ресурсы, образующие систему, были снабжены метаданными, причем значения атрибутов этих метаданных, вообще говоря, не являются элементами заданных словарей (в отличие от подхода, принятого при разработке концепции Semantic Web). Отсюда вытекает, что основными структурными элементами информационной системы должны быть *документы*, понимаемые применительно к данной ситуации как информационные ресурсы, имеющие (по определению [214]) уникальный идентификатор и обладающие метаданными. Для сравнения отметим, что в модели RDF ресурсы, могут вообще говоря, и не сопровождаться метаданными.

Как уже упоминалось в разд. 1.9, структура метаданных иерархична. Наиболее общий характер имеют метаданные, задающие структуру документа, т. е. описывающие метаданные более низкого уровня (атрибуты документа), которые определяют содержание документа. Наконец, значения этих атрибутов является фактически метаданными по отношению к исходному документу. Отсюда следует *важнейшая отличительная черта информационной системы: она работает не с данными, а исключительно с метаданными*. Отметим, что в [139] использование метаданных в качестве «строительного материала» названо определяющей характеристикой цифровых библиотек (фактически информационно-поисковых систем).

Выбор документа в качестве основного структурного элемента информационной системы дает возможность задавать связи между сущностями, описываемыми системой, посредством установления связей между соответствующими документами, при этом один документ может являться частью другого полностью или частично, в том числе и в виде гиперссылки.

Теперь сформулируем основные особенности реализации системных принципов для информационных систем, изложенные авторами в статье [21].

— *Целостность* системы проявляется в зависимости каждого объекта, свойства и отношения от его места и функций внутри це-

лого и реализуется посредством использования единого набора метаданных

$$M = \cup M^i.$$

Тем самым любой документ d_i системы представляется как

$$d_i = < m_i^{j,k} >,$$

где $m_i^{j,k}$ — значения элементов метаданных M^j, k — количество значений (с учетом повторений) соответствующего элемента метаданных в описании документа.

— *Иерархичность* системы проявляется в том, что она состоит из разнородных подсистем, отвечающих тем или иным частным задачам. Документы, описываемые при помощи одних и тех же элементов метаданных, образующих множество $M_i \subseteq M$, образуют *класс* K_i. Если $M_1 \subset M$, $M_2 \subseteq M$ и $M_1 \subset M_2$, то класс K_2 является *подклассом* класса K_1. Множество унифицированных структурированных документов-описаний одного класса называют *каталогом*. Фактически каталог объединяет поисковые образы исходных документов.

— *Структурность* системы обеспечивается выбором модели связей между документами, позволяющей адекватно описывать различные аспекты соответствующих межсущностных отношений. Достаточно универсальный характер имеет, например, модель направленных связей [14], которая будет подробно изложена ниже. Краткая ее суть состоит в следующем: если документ $d_{i'}$ входит в качестве значения элемента M^j метаданных документа d_i, то можно говорить о связи межу этими документами вида $M^j < d_i, d_{i'}, m_{i,i'}^{l,k} >$, где $m_{i,i'}^{l,k}$ — атрибуты связи, являющиеся значениями соответствующих элементов метаданных. Таким образом, выстраиваемые отношения фактически переносятся на уровень элементов, определяющих структуру документов.

— *Множественность описания* системы подразумевает наличие множества различных аспектов построения системы (модель данных системы, информационная модель системы, ее содержательное наполнение и пр.). Наиболее общий характер имеет описание *модели информационной системы*, которая строится посредством задания классов K_i, определяемых соответствующими множествами элементов метаданных M_i, и типов возможных связей

между классами $M^j < K_i, K_{i'} >$ с указанием элементов метаданных $M^j_{i,i'}$, описывающих атрибуты соответствующих связей, т. е. модель данных информационной системы может быть отнесена к моделям инфологического типа [235].

Нетрудно видеть, что принципы построения модели вобрали в себя черты, свойственные как для традиционного объектно-ориентированного подхода, так и для используемого в Semantic Web языка RDFS [249]. В частности, мы описываем классы в терминах их структуры, как это принято в объектно-ориентированном прогаммировании, а не определяем свойства в терминах классов, что характерно для RDFS. Такой выбор связан с тем, что задание базовых структур создаваемой системы, опирающееся на разработанную модель предметной области, носит централизованный характер. С другой стороны, ограничения, накладываемые моделью на свойства классов, носят менее жесткий характер, чем при объектно-ориентированном подходе (например, может быть объявлено произвольное, в том числе нулевое, количество значений некоторого элемента метаданных), что сближает нашу модель с подходом RDFS.

Подчеркнем, что предложенная модель информационной системы обладает столь высоким уровнем абстракции, что с точки зрения этой модели несущественно разделение метаданных на функциональные типы (структурные, описательные, административные), поскольку все они рассматриваются как структурированные сведения о документе.

Конкретное наполнение информационной системы определяется содержанием ее каталогов, причем, поскольку иерархическая модель данных может быть представлена в реляционном виде, есть смысл говорить и о *каталогах связей*.

Кратко рассмотрим особенности конкретных реализаций моделей систем информационного обеспечения научной деятельности (подробно этот вопрос будет обсужден в разд. 3.4). Действующие системы, как правило, имеют целью подробно описать один из трех названных в разд. 3.3 типов сущностей (группы, персоны, объекты деятельности), характеризующих деятельность научного сообщества. Например, сайты «Члены Российской академии наук» [152] или «Научные сотрудники — математики СО РАН» [148] содержат персональную информацию. В базе данных «Организации СО РАН» [150] представлена организационная структура отделе-

ния, а в различных библиотечных системах или в базах данных инновационных разработок [151] описываются объекты деятельности.

При построении информационной модели, лежащей в основе систем узкой тематики, сущности соответствующего типа становятся независимыми, а сущности других типов — зависимыми. Это обстоятельство способно оказать существенное влияние на выбор конкретной модели данных. Так, для разработки информационных систем, отражающих различные аспекты деятельности персон, целесообразно использовать подход, развитый в сетевых операционных системах и приложениях, которые используют справочники для хранения информации о пользователях. Независимыми сущностями таких справочников являются персоны, объединяемые в группы. Поскольку между персонами нет прямых отношений подчиненности, справочники имеют плоскую структуру. Пример приложений — почтовые клиенты MS Outlook Express и Netscape Communicator, ориентированные на схему данных X.500 [224].

В подходе, реализуемом в справочных системах организаций, а также систем, подобных «желтым страницам», независимой сущностью является организация. Система упорядочения организаций связана с административным и территориальным делением, что подразумевает жесткую иерархическую структуру справочника. В общей схеме служб справочника для организаций играет большую роль протокол LDAP [240]. Поддержка стандарта LIPS (Lightweight Internet Person Schema) [226] обеспечивает относительную стандартизацию общей схемы представления персональных данных, таких как имя, организация, сертификат и контактная информация.

Наконец, хранение больших объемов информации об объектах научной деятельности организуется, как правило, с использованием протокола Z39.50 [65]. Независимой сущностью таких систем является объект деятельности. Персона (или организация) приобретает уровень словаря, помогающего идентифицировать персону как субъект данной деятельности, а не как отдельную сущность. В качестве стандарта описания данных используется схема GILS [230] — общие описания информационных ресурсов. Важное достоинство протокола Z39.50 — возможность организации атрибутивно-

го поиска, что позволяет, в частности, искать документы из разных коллекций, имеющих один или несколько общих атрибутов.

Таким образом, можно выделить два основных подхода к организации информации в системах информационного обеспечения научной деятельности: иерархический (характерный для протокола X.500 и его «облегченной» версии LDAP) и горизонтальный (характерный для протокола Z39.50). К сожалению, каждый из этих подходов страдает определенными недостатками. Отсутствие в Z39.50 возможности построения иерархической структуры приводит к дублированию информации, относящейся к объектам с общими свойствами или к одному и тому же объекту, входящему в разные группы. С другой стороны, отсутствие в X.500 горизонтальных связей влечет необходимость повторения записей, описывающих объекты, связанные с тем или иным объектом.

Возникает проблема установления связей между документами, относящимися к разным составным частям системы (особенно актуальная при связывании имен с информацией об их носителях в случае, когда соответствующие денотаты (персоны) входят одновременно в разные структурные группы), а также, в отдельных случаях, между документами, относящимися к одной и той же составной части системы (например, между документами, описывающими организацию и ее неструктурные подразделения). Тем самым становится актуальной разработка технологии идентификации, спецификации и визуализации горизонтальных отношений между документами. С этой целью авторами предложена уже упоминавшаяся модель направленных связей [14, 238], в которой выстраиваемые отношения фактически переносятся на уровень элементов, определяющих структуру документов.

Как уже отмечалось, структурность информационной системы обеспечивается оптимальным выбором модели связей между документами, позволяющей адекватно описывать различные аспекты соответствующих межсущностных отношений. Однако при этом неизбежно встает проблема возможного рассогласования информации. Во-первых, включение в документы информации о разнородных сущностях может привести к появлению множественной информации об одном и том же объекте. Такая ситуация возможна, например, когда человек работает в разных организациях, участвует в разных проектах, является автором множества публикаций. Это может вызвать серьезные проблемы в случае необходимости

появления различных версий информации, возникающих вследствие ее модификации.

Кроме того, для представления сложных документов, когда один документ — часть другого (полностью или частично, в том числе и в виде гиперссылки), необходимо выработать подходы к установлению связей между документами. Такая ситуация возникает, если о сущностях, описываемых документами, может быть построено истинное высказывание (представляющее интерес с точки зрения содержания системы) типа: «*Сущность А есть (*или *была) нечто (по отношению к) сущности Б*» или «*Сущность А имеет (*или *имела) в некотором качестве сущность Б*». Например: «Евклид — автор «Начал»» или «Институт математики СО РАН имел директором С. Л. Соболева». Нетрудно видеть, что типы таких связей могут быть различными, и это обстоятельство нужно учитывать в процессе разработки модели отношений между документами.

Решение данной проблемы заключается в том, чтобы хранить информацию о каждом факте, относящемся к той или иной сущности или к некоторому свойству сущности, в единственном документе, устанавливая в нужных случаях отношения между документами типа «многие-ко-многим».

Указанный подход является традиционным при проектировании реляционных баз данных (см., например, [4, 101, 167]), однако основной прием его реализации заключается в рассмотрении многоместных отношений с их последующей декомпозицией в процессе нормализации. Мы же строим информационную модель с использованием только бинарных отношений, приписывая им дополнительные атрибуты, не укладывающиеся в общую схему. Таким образом, декомпозиция проводится на более высоком уровне абстрагируемости от структуры данных, что делает нашу модель более универсальной.

В основу представленной модели отношений между документами в информационной системе легла модель RDF [248], которая описывает ресурсы и отношения между ними. Описание ресурса в RDF — это совокупность утверждений о свойствах ресурса. Каждое утверждение представляет собой тройку: ресурс, именованное свойство и его значение. Отношения между ресурсами — именованные свойства.

Основное отличие представленной модели от модели RDF состоит в том, что выстраиваемые нами отношения переносятся на

уровень элементов, определяющих структуру документов: связи между документами устанавливаются путем задания на множестве документов бинарных отношений. В соответствии с правилами RDF эти отношения могут быть записаны в виде $A(R, V)$: объект R имеет атрибут A со значением V. Например, тот факт, что Барахнин В.Б. занимает некоторую должность (post) в ИВТ СО РАН, записывается как Post ('ÈÂÒ ÑÎ ÐÀÍ', 'Áàðàõíèí Â.Á.'), где Post — то или иное значение из списка (тезауруса) должностей.

Мы выделяем два вида отношений:

1. *Отношение порядка* между документами, выстраивающее иерархию подчинения в коллекции, например отношение подчиненности между документами в коллекции «Организации»: Head ('Êàôåäðà ìàòåìàòè÷åñêîãî ìîäåëèðîâàíèÿ', 'ÌÌÔ ÍÃÓ'). Данный тип отношения предполагает установление только односторонней связи между документами.

2. *Отношение связи* между документами, например отношение типа принадлежности между документами коллекции «Организации» и документами коллекции «Персоны»: Post ('ÈÂÒ ÑÎ ÐÀÍ','Áàðàõíèí Â.Á.'). Данный тип отношения допускает установление двухсторонней связи между документами в том смысле, что одновременно может существовать и обратная связь, например Position ('Áàðàõíèí Â.Á.', 'ÈÂÒ ÑÎ ÐÀÍ'). Направленность связи определяется порядком записи аргументов отношения $A(R, V)$. Таким образом, любой объект также может играть роль значения.

Различие отношений первого и второго типа заключается в том, что отношениям первого типа изначально приписано свойство — иерархия, а отношениям второго типа никаких свойств изначально не приписано. Свойства отношений второго типа определяются для каждого конкретного отношения.

Отношение первого типа, как правило, имеет не более одного атрибута, например тип подчинения (территориальное, научно-методическое и т. д.). Отношение второго типа имеет несколько дополнительных атрибутов. Например, отношение типа Post не просто описывает принадлежность персоны к организации, но и обладает следующими атрибутами: название должности, ключевые слова, дата назначения, дата освобождения от должности, видимость и др.

Для отношения $A(R, V)$ будем называть R *головным* документом, а V — *подчиненным* документом.

Документ в системе может быть связан с любым количеством документов. Между двумя документами могут быть заданы прямые и обратные отношения.

Прямое отношение — отношение головного документа к подчиненному ему документу, например отношение документа «визитная карточка организации» к документу, содержащему множество подразделений, множество сотрудников или список дополнительной информации. Документ из коллекции «Персона» или «Организации» может быть связан отношением с документами из коллекции дополнительной информации, например списком дополнительных сведений.

Обратное отношение — отношение подчиненного документа к головному документу. Для односторонних отношений родительский документ всегда знает свои дочерние документы, а дочерний документ ничего не знает о своем родителе. Учет обратных отношений о документе необходим для обеспечения навигации по коллекциям.

Таким образом, выделяя два вида отношений между документами, мы решаем две задачи:

1) установление связей между документами (гиперссылки, вставки);

2) навигация по коллекциям (навигационное дерево).

Исходя из свойств отношений второго типа, в документе можно выделить два типа элементов:

1) элементы, содержание которых не зависит от значений атрибутов отношения;

2) элементы, содержание которых может зависеть от значений атрибутов отношения. Например, должностные лица организации могут иметь адрес электронной почты, связанный с должностью (т. е. не зависящий от конкретной персоны, занимающей должность).

При использовании предложенной модели направленных связей между документами для создания систем информационного обеспечения научной деятельности необходимо учитывать следующее обстоятельство. В связи с тем что в основе информационной модели таких систем находится персона (см. разд. 2.3), при поиске возникает необходимость сопоставить персоне все ее позиции (в том числе и относительно публикаций), т. е. пользоваться подходом, обратным описанному выше. Решение этой задачи с помощью контекстных запросов (даже к конкретному полю) не всегда удобно, так как может привести к выдаче непертинентных документов. Тем самым возни-

кает потребность в построении обратной модели отношений, которая носила бы достаточно универсальный характер.

Таким образом, документы информационной системы сгруппированы по следующему принципу: имеется специально выделенная коллекция «Персоны» и множество других коллекций: «Публикации», «Организации», «Сообщества» (т. е. советы, общества, журналы и др.), причем все отношения строятся вокруг персон.

Персона может занимать различные позиции: быть автором или редактором публикации, занимать некоторую должность в организации, быть председателем или членом совета и т. д. Все эти случаи представляются одним типом отношения Position, который может принимать различные наименования (*директор*, *аспирант*, *председатель совета*, *автор* и т. д.)

Важной особенностью рассматриваемой модели является возможность связи имен с информацией об их носителях в случае, когда соответствующие денотаты (персоны) входят одновременно в разные структурные группы. Модель данных позволяет не вводить дублирующие записи, а разделять информацию о персоне на две части: личную — связанную с самой персоной, и ролевую — связанную с позициями, занимаемыми персоной, причем каждой позиции соответствует новая ролевая запись.

На рис. 3.1 приведена схема, иллюстрирующая основные особенности рассматриваемой модели: примеры прямых и обратных

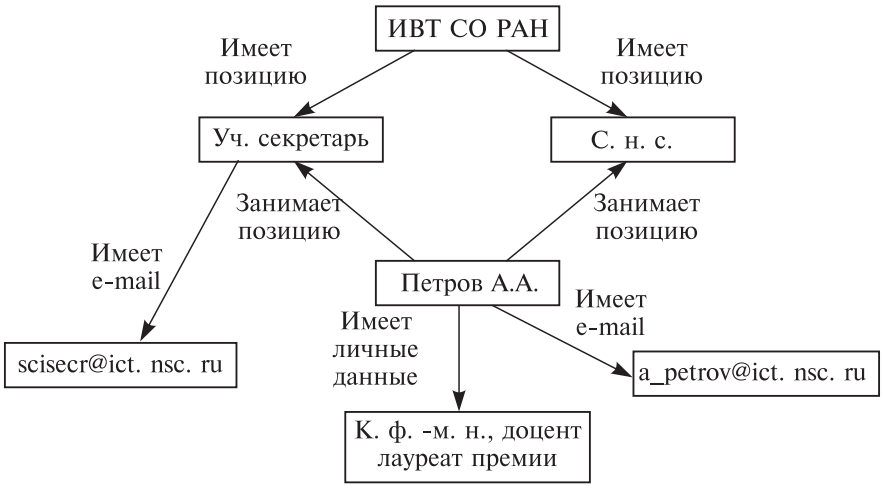

Рис. 3.1. Схема модели направленных связей.

отношений, отношений первого и второго типа, разделение информации о персоне на личную и ролевую.

Подводя итог, повторим краткое формальное описание модели направленных связей: если документ $d_{i'}$ входит в качестве значения элемента M^j метаданных документа d_i, то можно говорить о связи межу этими документами вида $M^j < d_i, d_{i'}, m_{i,i'}^{l,k} >$, где $m_{i,i'}^{l,k}$ — атрибуты этой связи, являющиеся значениями соответствующих элементов метаданных. Тем самым выстраиваемые отношения фактически переносятся на уровень элементов, определяющих структуру документов.

3.3. Структура логических компонентов программной системы

> П р о ф е с с о р. У какой бесконечно дифференцируемой функции ряд Тейлора имеет лишь конечное число ненулевых членов?
>
> С т у д е н т *(уверенно)*. Только у многочлена.
>
> П р о ф е с с о р. Сколько ненулевых членов в разложении синуса в ряд Тейлора?
>
> С т у д е н т *(так же твердо)*. Бесконечно много.
>
> П р о ф е с с о р. Ну и последний вопрос: а почему в разложении синуса бесконечно много ненулевых членов?
>
> С т у д е н т *(с торжествующим видом)*. Да потому, что синус — это одночлен!
>
> *Математический фольклор. Цит. по книге: С. Н. Федин. Математики тоже шутят.*

Интеллектуальная система, как уже отмечалось в разд. 1.4, включает в себя, наряду с информационно-поисковой системой, еще и рассуждающую информационную систему, формализующую правила логического вывода, а также интеллектуальный пользовательский интерфейс. Каковы же структуры этих подсистем?

С точки зрения пользователя, важнейшей чертой информационных систем (и, более того, их определяющей характеристикой) является наличие возможности осуществлять сложные информа-

ционные запросы. С точки зрения структурной сложности выделяются следующие типы запросов [138]:

1) дающие информацию об одном объекте;

2) указывающие объекты, которые удовлетворяют определенным признакам;

3) для объектов с заданными требованиями на значения признаков указывающие значение других признаков.

Сразу отметим, что формализация правил логического вывода обычно проводится непосредственно в высокоуровневом языке запросов, поэтому нам важно задать такую структуру документов, которая позволила бы осуществлять логический вывод из данных, содержащихся в документах. Иными словами, необходимо ответить на вопрос: какой набор поисковых признаков, образующих поисковый образ документа в каталоге, требуется выделить для того, чтобы пользователь имел возможность реализации поисковых запросов более сложных, чем выдача документа по известному имени?

Можно выделить три основных модели поиска в информационных системах:

1) контекстный поиск;

2) атрибутивный поиск;

3) поиск «по аналогии».

С точки зрения организации поиска атрибуты (поисковые признаки) удобно подразделять на три типа:

а) текстовые;

б) числовые;

в) табличные (т. е. выбираемые из заданного списка).

Контекстный поиск отбирает документы, у которых значения атрибутов текстового типа (любых или выбранных пользователем) содержат заданный в качестве поискового запроса текстовый фрагмент. В некоторых случаях по желанию пользователя учитываются возможные морфологические изменения текстового фрагмента. Запрос к каждому атрибуту может также включать несколько текстовых фрагментов, связанных логическими операциями конъюнкции, дизъюнкции и отрицания. Этими же операциями могут быть связаны запросы к нескольким атрибутам.

Контекстный поиск — наиболее простая модель поиска, которая может быть реализована даже при минимальной структуаризации документов, однако ее эффективность зачастую невелика.

Она максимальна в том случае, когда в качестве поискового запроса пользователь задает заранее известную точную цитату достаточной длины. В противном случае, если поисковый запрос слишком короток, информационная система выдаст большое количество документов, многие из которых наверняка будут не интересны пользователю. Напротив, для запросов большой длины часто вообще не обнаруживается соответствующих документов, поскольку авторы документов могли описывать интересующий пользователя предмет фразами, несколько отличающимися от заданной в запросе.

При *атрибутивном поиске* информационный запрос представляет собой набор значений одного или нескольких табличных атрибутов, выбираемых из списка, и (или) значений числовых атрибутов из заданного диапазона, связанных логическими операциями (обычно — конъюнкцией).

Атрибутивный поиск целесообразен в тех случаях, когда пользователя интересует не столько какой-либо конкретный документ, сколько класс документов, удовлетворяющий некоторому набору свойств.

Поиск «по аналогии», как уже упоминалось в разд. 1.8, предполагает задание в качестве информационного запроса непустого множества документов. В качестве результата запроса выдаются документы, каждый из которых в определенном смысле близок к одному из документов, входящих в заданное множество. Подробнее вопросы поиска «по аналогии» будут освещены в разд. 4.5.

Наконец, чтобы информационно-поисковая система могла быть использована для получения новых знаний, ее пользователю должны быть предоставлены возможности, во-первых, формулировать такие запросы, которые для объектов с заданными требованиями на значения признаков указывают значения других признаков, и, во-вторых (обобщая предыдущее требование), проверять, истинно или нет утверждение $R_s(d_{i_1}, \ldots, d_{i_n})$ относительно сущностей, описываемых документами $d_{i_1}, \ldots, d_{i_n}$. Высказыванию $R_s(d_{i_1}, \ldots, d_{i_n})$ формально соответствует n –местный предикат P_s, определенный на множестве документов, причем при его построении могут использоваться документы системы (точнее, значения атрибутов этих документов), информация из онтологии предметной области и т.п.

Если количество документов в системе, способных выступать в качестве аргументов предиката P_s, велико, то весьма перспективным методом извлечения новых знаний является проверка истинности предиката на различных наборах документов, автоматически перебираемых системой. Тем самым реализуется механизм автоматического извлечения данных из документов с целью пополнения базы данных посредством этих фактов, который характеризует интеллектуальные информационные системе высокого уровня, описываемые формулой (1.2) (подробнее об этом см., например, [116]).

Подчеркнем отличие описанного подхода от подхода, применяемого специалистами в области искусственного интеллекта для разработки экспертных систем (см. монографию [50]): последние предназначены для решения узкоспециализированных задач, содержат относительно небольшой объем документов, и основной упор при их создании делается на развитие сложных продукционных правил. В то же время, как показано в [58], «в настоящее (и ближайшее) время ни одна из существующих систем автоматической обработки текстов, извлечения знаний из текстов не может обеспечить такой уровень точности и полноты получения информации из текстов, на которых надежно можно было обосновывать работу таких правил вывода». Поэтому интеллектуальные системы, работающие с документальной информацией, могут обладать достаточно простыми продукционными правилами, а получение новых знаний становится возможным благодаря большому объему документов, способных выступать в качестве аргументов проверяемых утверждений.

Таким образом, для эффективной организации информационного поиска и получения новых знаний важна хорошая структуризация документов, предусматривающая, в частности, достаточно большое количество поисковых признаков, образующих поисковый образ документа. Это накладывает особые требования на каталог информационно-поисковой системы, которые мы сформулируем в следующем пункте.

Покажем, что реальные возможности контекстного поиска фактически сводятся к поиску документов с известным именем (частью имени) или известной фразой из текста документа. Очевидно, эффективность контекстного поиска максимальна, если в качестве поискового запроса пользователь задает заранее извест-

ную точную цитату достаточной длины. В противном случае, если поисковый запрос слишком короток, информационная система выдаст множество документов, большинство из которых наверняка будут не пертинентными (то есть не соответствующими информационной потребности пользователя), хотя и формально релевантными (то есть соответствующими «букве» информационного запроса). Для «произвольных» запросов большой длины зачастую вообще не обнаруживается даже релевантных документов, поскольку авторы документов могли описывать интересующий пользователя предмет фразами, несколько отличающимися от заданной в запросе.

Описываемая ситуация, когда система позволяет осуществлять только контекстный поиск, возникает, если информационно-поисковый язык системы практически совпадает с естественным языком.

Более того, аналогичные проблемы встают и тогда, когда в информационно-поисковом языке отсутствуют средства выражения имманентных отношений между предметами, т. е. язык не имеет парадигматических отношений. Примером такого языка может служить система унитермов — набора одиночных ключевых слов (в редких случаях словосочетаний).

Таким образом, возможность получения в результате поискового запроса пертинентных документов появляется лишь в том случае, когда информационно-поисковый язык имеет средства выражения имманентных отношений, т. е. обладает *онтологией*. Так как, согласно разд. 1.4, онтология включает в себя словарь-тезаурус, определение которого [104] предусматривает наличие смысловой классификации терминов, то создание тезауруса информационно-поискового языка дает возможность осуществлять атрибутивный поиск (а также поиска «по аналогии»). В качестве поисковых атрибутов могут выступать разделы соответствующего классификатора. Итак, наличие онтологии (и, следовательно, классификатора) в качестве составной части информационно-поискового языка, используемого при создании каталога, является обязательным условием выполнения основного требования к информационно-поисковым системам — возможности реализации сложных информационных запросов.

3.4. Структуры представления научной и научно-организационной информации

Алиса спросила:

— Скажите, пожалуйста, куда мне отсюда идти?

— А куда ты хочешь попасть? — ответил Кот.

— Мне все равно... — сказала Алиса.

— Тогда все равно, куда и идти, — заметил Кот.

— ...только бы попасть куда-нибудь, — пояснила Алиса.

— Куда-нибудь ты обязательно попадешь, — сказал Кот. — Нужно только достаточно долго идти.

Л. Кэролл. Алиса в стране чудес

3.4.1. Структура информационно-справочной системы по истории науки (на примере математики)

А История такое большое дело, что и Топтыгин, при упоминовении об ней, задумывался. Сам по себе, он знал об ней очень смутно, но от Осла слыхал, что даже Лев ее боится: «Не хорошо, говорит, в зверином образе на скрижали попасть!»

М. Е. Салтыков-Щедрин. Медведь на воеводстве

Как было отмечено в разд. 2.3, при создании информационных систем по истории науки в первую очередь рассматриваются субъекты — отдельные лица, объекты — предметы и продукты деятельности. Поскольку творцами науки являются отдельные выдающиеся личности, информация в справочных системах по истории науки должна группироваться вокруг персон, при этом требуется подробное структурирование биографических данных в плане хронологии, географии и т. п. Разумеется, биографии ученых немыслимы без библиографического списка, который включает в себя, наряду с публикациями данного ученого, и публикации о нем самом. Наконец, необходимо четко отразить связь научной деятельности исследователя с формализованным описанием предметной области, в которой этот исследователь работал.

В настоящее время в Интернете имеется целый ряд справочников, содержащих биографии деятелей различных областей наук.

Наиболее полно и качественно, на наш взгляд, представлены биографии математиков. Это объясняется, по меньшей мере, двумя причинами. Во-первых, основная масса ученых, занимающихся информационными технологиями, имеет математическое образование, и вполне естественно, что специалисты, создающие такие справочники (а подобная работа в большинстве случаев основана на энтузиазме небольших групп ученых), предпочитают работать с историей той области фундаментальной науки, которая им хорошо знакома. Во-вторых, формализованное описание предметных областей с учетом исторической ретроспективы, необходимое для адекватного отражения деятельности исследователей прошедших эпох, для многих разделов естественных наук затруднено тем, что естественно-научные теории постоянно развиваются, причем нередко ранее принятые концепции впоследствии полностью отвергаются, вследствие чего построение формальных моделей эволюции естественно-научных теорий является сложной проблемой, решение которой, по-видимому, не допускает сколько-нибудь общих подходов. В этом смысле математика является счастливым исключением, поскольку она оперирует, согласно терминологии И. Канта [80], априорными синтетическими суждениями, что обеспечивает последовательное (практически без отвержения полученных ранее результатов) развитие математическое науки.

Среди представленных в Интернете коллекций биографий математиков можно выделить портал истории математики, разработанный в шотландском Университете Святого Андрея [241]. Здесь представлена наиболее полная коллекция биографий. Ресурс обладает развитыми поисковыми возможностями, имеет сортировку по годам жизни математика, по стране его рождения (смерти), а также содержит хронологию основных математических открытий. Однако, к сожалению, данный портал не имеет ссылок на электронные публикации трудов, а также классификатора предметной области, что значительно снижает его справочную ценность.

Также стоит упомянуть проект The Mathematics Genealogy Project [242], разрабатываемый Университетом Северной Дакоты. Ресурс призван собрать информацию обо всех математиках, когда-либо получивших степень доктора математических наук. Обширная база включает информацию о диссертационной работе математика, название университета, присвоившего степень, а также указывается предметная область с использованием классификато-

ра MSC2000, список учеников и список работ. Однако более полные биографические материалы берутся с портала Университета Святого Андрея. К недостаткам ресурса относится скудность предоставляемой информации (лишь некоторые статьи имеют указания на предметную область деятельности данного математика, часть статей вообще не снабжена информацией даже о годе получения степени), а также то, что российские математики представлены в небольшом количестве: из общего числа 89 300 российских математиков упомянуто 1296 (данные по состоянию на начало 2008 г.).

Из российских ресурсов наиболее близкий аналог данного проекта — Математический портал [100], разрабатываемый при поддержке Отделения математических наук Российской академии наук и Московского центра непрерывного математического образования. Посвященный истории математики раздел содержит, помимо подробного текста биографий как исторических, так и некоторых современных математиков, их библиографию. Однако поисковые возможности портала не поддерживают атрибутивный поиск и ограничены лишь поиском по фамилии.

Таким образом, ни один из имеющихся биографических порталов в области математики не способен достаточно адекватно отразить персонифицированную ретроспективу развития математической науки. Порталы с подобными свойствами отсутствуют и для других разделов естественных наук.

Ниже представлено описание структуры информационно-справочных систем по истории науки (прежде всего математики), удовлетворяющих сформулированным выше требованиям, которое опубликовано в работах авторов [12, 23, 24]. Это описание включает в себя информационную модель справочника, особенности реализации его каталогов, а также основные виды информационных запросов, удовлетворяющих потребностям пользователя справочника.

Отметим, что при создании справочника нужно учитывать следующее соображение: стремление к максимальной полноте представления информации о каждом ученом может привести к неоправданной задержке работы над справочником. С другой стороны, ценность справочника для пользователя состоит в широте охвата биографий, позволяющей получить более или менее целостное представление о состоянии описываемой области науки в определенную эпоху, в том числе в конкретной стране, или же проследить

за развитием какого-либо раздела науки. Ввиду этого, наряду с полной версией модели мы будем приводить и варианты ее упрощения, предназначенные для ускоренного первоначального наполнения справочника.

Справочник содержит следующие основные каталоги:

а) персоны (анкетные данные);

б) публикации, так или иначе связанные с персонами, представленными в справочнике;

в) структурированное описание предметной области.

Кроме того, при создании подробной версии справочника целесообразно создание вспомогательных каталогов:

— учебные заведения;
— научные организации;
— научные сообщества (академии, редколлегии и т. п.);
— награды и премии;
— и др.

Документы из каталога «Персоны» связываются с документами из прочих каталогов с использованием модели направленных связей, в которой каждый тип связей обладает определенным набором атрибутов, характеризующих данное отношение. Отметим, что в роли значений атрибутов связи могут выступать документы, входящие в тот или иной каталог. Так, при установлении связи между персоной и полученным этой персоной результатом предметной области в качестве атрибутов связи могут выступать атрибуты публикации, содержащей данный результат.

Необходимый набор информации о представленной в справочнике персоне должен включать в себя ее основные анкетные данные: ФИО, дата и место рождения, для умерших — дата и место смерти, а также названия страны (стран), с которой принято связывать профессиональную деятельность данной персоны.

Расширенный набор анкетной информации включает сведения о полученном образовании, местах работы, членстве в научных сообществах, наградах. Первоначально такие сведения могут быть частью текстового описания деятельности персоны, однако по мере развития системы информация подобного рода будет представлена посредством установления связи персоны с соответствующими вспомогательными коллекциями. В качестве атрибутов связей выступают даты событий, а также занимаемые должности (в широком

смысле, включающем иерархический статус в академиях, редколлегиях и т. п.).

Элемент метаданных «портрет персоны» может иметь значения: файл или ссылка.

Библиографический каталог системы включает в себя описание публикаций, так или иначе связанных с персонами, представленными в справочнике. Можно выделить следующие виды публикаций:

— научные публикации, относящиеся к представленной в справочнике отрасли науки;
— прочие научные публикации;
— научно-популярные публикации;
— прочие публикации;
— интервью;
— биографии персоны;
— прочие публикации о персоне.

При этом одна публикация может относиться к разным видам: например, публикация об одной персоне может одновременно быть научно-популярной публикацией другой персоны, представленной в справочнике.

По мере развития системы возможно разбиение некоторых из перечисленных разделов на подразделы. Так, научные публикации могут подразделяться на монографии, статьи, учебники и т. п. (подробная классификация жанровых типов научных ресурсов будет приведена в п. 4.4.2).

Описание публикации представляет собой, как минимум, библиографическую ссылку, оформленную в соответствии со стандартом. Для классификации публикаций крайне желательно применять тот классификатор предметной области, который используется при создании ее онтологии.

В исключительных случаях, касающихся важнейших публикаций минувших эпох, описание может быть ограничено автором, названием и годом выпуска, при этом, по возможности, параллельно указывается современное общедоступное издание данной работы.

Структурированное описание предметной области целесообразно представлять в виде онтологии, причем, создавая онтологии большого объема, охватывающие историческую ретроспективу развития той или иной науки, в них можно включать устаревшие

понятия (в том числе отвергнутые современной наукой), снабжая их соответствующей пометкой.

Наконец, термины онтологии могут быть снабжены краткими словарными статьями, посвященными этим терминам.

Важнейшим этапом создания справочника является установление связей между основными каталогами, а также разработка многомерного классификатора документов каталога «Персоны».

Связь между персонами и публикациями вряд ли способна вызвать особые проблемы; в качестве атрибута этих связей предполагается указывать вид публикации из списка, приведенного выше.

Менее тривиальным является установление связей между персонами и терминами из тезауруса. Практически любой термин, входящий в тезаурус предметной области, был включен в научный оборот благодаря тому, что соответствующему объекту или явлению была посвящена научная публикация, имеющая одного или нескольких авторов. Таким образом, в качестве атрибутов связи между персоной и термином из тезауруса выступают атрибуты соответствующей публикации (по крайней мере, год ее выхода в свет). Разумеется, чем меньше объем имеющегося тезауруса, тем более общий характер имеют его термины, и каждый их них в принципе может быть связан с большим числом персон.

Установление связей между персонами и терминами тезауруса (которые обязательно имеют классификационный признак классификатора предметной области) способствует, в частности, наделению этими классификационными признаками персон. Решение этой задачи только на основании публикаций персоны не всегда возможно, ибо запись о публикации в каталоге системы не всегда может иметь классификационные признаки, тем более относящиеся к основному классификатору справочника (каковым является классификатор MSC2000, использованный при создании онтологии).

В процессе развития системы она дополняется вспомогательными каталогами: учебные заведения, научные организации и научные сообщества (академии, редколлегии и т. п.) соответствующего профиля, награды и премии и др. Установление связей между персонами и элементами соответствующих каталогов призвано повысить справочную ценность системы. Непременным (но, разумеется, не всегда единственным) атрибутом этих связей являются даты соответствующих событий.

Для удовлетворения основных информационных потребностей пользователя справочника следует ввести следующие классификаторы документов каталога «Персоны»:

1) тематический;
2) хронологический;
3) географический.

Таким образом, пользователь получает возможность изучить хронологию развития выбранного раздела предметной области (или представлений об ее отдельном понятии), в том числе в сужении на отдельной взятую страну, или же составить представление о состоянии соответствующей науки в тот или иной период времени.

3.4.2. Структуры представления информации о деятельности научного сообщества (на примере СО РАН)

Что рота на взводы разделяется,
В этом никто не сомневается.
Ф. К. Прутков. Военные афоризмы

Научные центры СО РАН, расположенные на территории трех федеральных округов, включают более 100 организаций, в том числе почти 90 научно-исследовательских и конструкторско-технологических учреждений. При этом каждая организация, являясь самостоятельным субъектом научной деятельности, традиционно обладает широкой самостоятельностью в выборе форм научно-организационной работы, включая информационное обеспечение. Это делает невозможным жесткую стандартизацию частных аспектов корпоративных решений в области информационных технологий. Поэтому для построения информационной системы СО РАН выбран путь интеграции информационных систем институтов в распределенную систему с единым каталогом ресурсов (точнее, набором каталогов, каждый из которых включает однородные ресурсы той или иной тематической направленности).

В соответствии с результатами работ [28, 176, 177] опишем структуры представления информации о деятельности СО РАН, основными компонентами которой, как показано в разд. 2.3, являются каталоги «Организации», «Персоны» и «Публикации».

Каталог «Организации» содержит в качестве элементов метаданных основные сведения об организациях: полное и сокращенное

названия, почтовые и электронные адреса, телефоны, список головных организаций и т. п., а также дополнительную информацию: описание деятельности, историческую справку, ключевые слова, список информационных ресурсов, описание инновационной деятельности.

Между организациями присутствуют иерархические связи подчинения следующих типов:

— территориальное подчинение (институты, входящие в региональный научный центр, подчинены его Президиуму);

— научно-методическое подчинение (организации научно-методически подчинены Объединенному ученому совету по направлениям наук либо Научно-координационному совету, которые, в свою очередь, научно-методически подчинены Президиуму СО РАН);

— административное (юридическое) подчинение (региональные филиалы или подразделения институтов подчинены головным институтам СО РАН).

Поскольку сотрудник СО РАН может участвовать в деятельности нескольких организаций (например своего Института, Объединенного ученого совета и Президиума СО РАН), постольку информация о научных работниках в каталоге «Персоны» разделяется на две части: личную (связанную с самой персоной) и служебную (связанную с должностью персоны).

Личная информация включает следующие элементы метаданных:

— анкетные данные персоны: ФИО, пол, место рождения, дата рождения (а также, в соответствующей ситуации, дата смерти), ученое звание и ученая степень, фотография;

— личные контактные данные (не доступные через веб-интерфейс рядовым пользовтелям);

— дополнительные сведения: область интересов и деятельности, образование, специальность, награды.

Служебная информация включает:

— связь персоны с организацией (должность);

— рабочий телефон, факс, адрес электронной почты, адрес веб-страницы;

— время работы персоны в организации на данной должности (дата принятия на должность и дата освобождения должности).

Если персона занимает несколько должностей, то для каждой должности создается свой экземпляр служебной информации.

Каталог «Публикации» содержит библиографическое описание публикаций сотрудников организаций СО РАН. Для их описания используется минимально допустимый набор полей, определенных библиографическими стандартами. Было выделено 17 жанровых типов публикаций:

1. Монографии.
2. Учебные пособия с грифом УМО.
3. Учебно-методическая литература.
4. Центральная печать.
5. Зарубежная печать.
6. Труды международных конференций.
7. Труды всероссийских и региональных конференций.
8. Авторефераты диссертаций.
9. Препринты.
10. Публикации в российских изданиях.
11. Тезисы конференций.
12. Электронные публикации.
13. Патенты или Свидетельства о регистрации программ для ЭВМ.
14. Отчеты НИР.
15. Депонированные издания.
16. Прочие научные издания.
17. Прочие издания.

Для публикации устанавливаются связи «публикация — автор публикации» и «публикация — организация».

К основным каталогам СО РАН присоединяются дополнительные каталоги, например «Проекты и программы», «Инновационные предложения», «Документы» (в узко-юридическом смысле) и др. Рассмотрим классификаторы, обеспечивающие представление документов из этих каталогов.

В разделе «Проекты и программы» представлены тематические информационные системы, связанные с научной деятельностью СО РАН, такие как «Электронный атлас биоразнообразия животного и растительного мира Сибири» [207], «Химия в СО РАН» [76], «Web-ресурсы математического содержания» [77] и др.

В качестве примера рассмотрим «Web-ресурсы математического содержания». Документы в каталоге классифицируются по их типу (персона, общество, институт, отдел, лаборатория, группа,

факультет, кафедра, научная школа, конференция, семинар, издательство, журнал, книга, статья, проект, пакет программ, библиотека, коллекция, база данных, форум). Кроме того, применяется тематическая классификация в соответствии с классификатором MSC2000. Предусмотрено установление внутренних связей между документами.

Каталог «Инновационные предложения» ведется в соответствии с международными стандартами, принятыми, в частности, в Российской сети трансфера технологий (подробнее об этом см. в п. 3.4.3). Отличительной особенностью ведения этого каталога является необходимость его регулярной актуализации, в том числе в плане удаления (архивации) тех инновационных предложений, которые по каким-либо причинам больше не предлагаются для коммерциализации. Для элементов этой коллекции устанавливается связь «инновационное предложение — организация».

Каталог «Документы» состоит из законодательных актов, указов, постановлений, рекомендаций и т. п., принимаемых органами государственной власти Российской Федерации, Президиумами РАН, СО РАН, а также органами государственной власти субъектов федерации и Президиумами научных центров СО РАН. Документ включается в коллекцию, если его содержание непосредственно затрагивает деятельность СО РАН или представляет интерес для широкого круга его сотрудников.

Прежде всего документы классифицируются по их региональной-ведомственной принадлежности:

— федеральный уровень;

— Российская академия наук;

— Сибирское отделение РАН;

— регионы и соответствующие научные центры РАН.

Внутри каждого раздела классификация ведется с учетом вида источника права:

— законы и кодексы (с выделением в отдельные подразделы Конституции Российской Федерации, а также Уставов организаций);

— указы, постановления и распоряжения;

— международные договоры.

Однородные документы упорядочены по дате принятия (в порядке убывания).

3.4.3. Структуры представления информации о научно-инновационной деятельности

> — Теперь нужно делом заниматься, а не языком трепать... Вот придумаю какое-нибудь изобретение, возьму патент и продам, к стыду России, куда-нибудь за границу... Ну, например, скажем, — изобрету такую какую-нибудь машинку, чтобы каждое утро, в положенный час, аккуратно меня будила. Покрутил с вечера ручку, а уж она сама и разбудит. А?
>
> — Папочка, — сказала дочь, — да ведь это просто будильник!
>
> *Тэффи. Взамен политики*

Наконец, рассмотрим структуры представления информации о научно-инновационной деятельности, предложенные авторами в работах [16, 187, 196].

В качестве основных компонентов систем информационного обеспечения научно-инновационной деятельности можно выделить следующие каталоги:

— организации (и их подразделения);
— документы (в узкоюридическом смысле);
— авторские публикации;
— инновационные разработки;
— события.

Документы классифицируются в соответствии с тематической классификацией ресурсов, в разделы которой объединяются разнообразные материалы, относящиеся к той или иной сфере инновационной деятельности:

— государственная политика в области инновационной деятельности;
— разработка высоких технологий;
— трансфер технологий;
— инновационный менеджмент;
— вопросы интеллектуальной собственности;
— подготовка кадров для инновационной деятельности;
— семинары и конференции;
— конкурсы, программы, гранты;
— межрегиональное сотрудничество;
— международное сотрудничество.

Дальнейшая классификация определяется разделом тематического классификатора. Так, организации классифицируются по признаку их ведомственной и (или) региональной принадлежности. Классификация нормативно-правовых документов проводится по правилам, описанным в п. 3.4.2. Классификация авторских публикаций двумерная:

1) жанр (публицистический, научный);

2) вид (аннотация, тезисы, статья, учебник, монография).

Следует особо остановиться на классификаторе предметной области. Очевидно, классификатор такого рода должен использоваться при классификации инновационных разработок. Однако, как уже отмечалось в разд. 2.3, организационные вопросы инноватики также достаточно тесно связаны непосредственно с предметной областью инновационной деятельности, например некоторые нормативно-правовые акты могут регулировать отношения лишь в сфере определенных технологий. Поэтому в набор метаданных документа из каталогов «Организации» или «Нормативно-правовые документы» также может входить элемент, значением которого является код раздела определенной предметной области, в соответствии, например, с классификатором Российской сети трансфера технологий [140]. Этот классификатор подразделяет сферу инновационной деятельности на следующие предметные области:

— промышленные технологии;
— информационные технологии;
— экология, охрана окружающей среды;
— медицина;
— биотехнологии;
— новые материалы;

которые, в свою очередь, включают в себя соответствующие разделы. Например, область «Информационные технологии» содержит разделы:

— электроника;
— микроэлектроника;
— обработка информации;
— информационные системы;
— телекоммуникации.

* * *

В данной главе мы подробно описали структуру программной системы информационного обеспечения научной деятельности. Однако подобная любая структура нуждается в наполнении ее информацией (в данном случае — документами). При этом, как уже неоднократно подчеркивалось, указанный процесс должен быть максимально автоматизирован. Описанию алгоритмов обработки слабоструктурированных документов и будет посвящена следующая глава.

Глава 4

МЕТОДОЛОГИЯ ОБРАБОТКИ СЛАБОСТРУКТУРИРОВАННЫХ ДОКУМЕНТОВ

Не спрашивай, какой там редут,
А иди куда ведут.

Ф. К. Прутков. Военные афоризмы

4.1. Автоматизированная технология построения тезаурусов и онтологий

Употребляй в разговоре киловатты, векторы и т. п. Мир уважает специалистов.

Али бен Марабут. Цит. по книге: «Мыслилюдей великих, средних и пса Фафика»

Полноценное включение в информационный процесс документов, в том числе электронных, практически невозможно без использования тезаурусов и онтологий соответствующих предметных областей (о том, какой смысл в рамках данной работы вкладывается в понятие «онтология», было сказано в разд. 1.4). При этом, как отмечено в [57], еще в начале 2000-х гг. очень незначительное количество информационных систем сопровождалось тезаурусом, поскольку традиционные информационно-поисковые тезаурусы разрабатывались для ручного индексирования, а объем потоков информации в настоящее время значительно превосходит возможности ручной обработки.

Возникшая необходимость создания средств автоматизированной интеллектуальной обработки данных обусловила бурное развитие исследований (см., например, монографии [234, 258]), связанных, в частности, с построением онтологий весьма сложной структуры. Как правило, эти исследования основаны на концепции Semantic Web [256], реализация которой позволила бы вывести работу с информацией на качественно новый уровень. Однако наработки консорциума W3 носят лишь *рекомендательный* характер, а объявить их *стандартами* могут только организации, имеющие соответствующий статус, например ISO, ГОСТ или ANSI, поэтому

реальное развитие большинства ресурсов Интернет, в том числе научной направленности, идет без учета подобных необязательных рекомендаций. Такие ресурсы зачастую не могут быть обработаны с использованием онтологий (тезаурусов) сложной структуры, включающих правила вывода (аксиомы), поскольку «в настоящее (и ближайшее) время ни одна из существующих систем автоматической обработки текстов, извлечения знаний из текстов не может обеспечить такой уровень точности и полноты получения информации из текстов, на которых надежно можно было обосновывать работу таких правил вывода», вследствие чего «значительные трудозатраты на такого рода формализацию информационно-поисковых тезаурусов не приведут к улучшению качества автоматической обработки текстов и созданию ресурсов, лучше приспособленных к автоматическим режимам работы, чем существующие информационно-поисковые тезаурусы» [58].

Именно поэтому по-прежнему представляет интерес развитие технологий создания «традиционных» тезаурусов и онтологий. В этой связи следует отметить, прежде всего, цикл работ коллектива сотрудников НИВЦ МГУ [56–58], создавших крупнейший тезаурус русского языка РуТез («общезначимую онтологию» [56]), предназначенный для автоматической обработки больших текстовых коллекций, который представляет собой иерархическую сеть 50 тыс. понятий с более чем 125 тыс. слов и выражений (статистика относится к 2007 г. [94]), а также диссертационную работу С В. Жмайло [67], посвященную исследованию и разработке теории и методики построения тезаурусов для информационного поиска в полнотекстовых базах данных.

Анализ перечисленных работ показал, что составление тезаурусов и онтологий «с чистого листа» может потребовать весьма значительных трудозатрат специалистов-экспертов, которые должны собрать все термины, достаточно полно охватывающие предметную область, согласовать их значения, установить связи и провести классификацию. С целью автоматизации названных процессов нами была разработана и реализована технология создания тезаурусов и онтологий на основе предметного указателя специализированных энциклопедий [8, 17]. Эта технология обеспечивает высококвалифицированное описание предметной области с использованием надежно выверенных терминов, позволяя провести начальный

этап построения онтологии с минимальным привлечением специалистов — экспертов в данной предметной области.

Алгоритм, лежащий в основе технологии, имеет оригинальный характер [8]. В работе создателей тезауруса РуТез [58] также говорится об использовании предметных указателей энциклопедий, но без механизма автоматизации установления связей.

Сразу отметим, что онтология, являясь в том или ином приближении моделью предметной области, никогда, как и любая модель, не может считаться совершенной. Поэтому, говоря об автоматизации процесса создания тезауруса и онтологии, мы подразумеваем лишь начальный этап этого процесса, который, однако, требует наибольших трудозатрат. Дальнейшее развитие онтологии осуществляется экспертами с учетом, например, рекомендаций работы [58].

Приведем описание разработанного нами алгоритма, а также реализующего его веб-приложения.

В качестве списка ключевых слов и словосочетаний для тезауруса предлагается использовать предметный указатель специализированной энциклопедии (или нескольких энциклопедий). Выбор конкретной энциклопедии осуществляет специалист по предметной области, и этот выбор зависит от целей, преследуемых при создании онтологии. Так, для решения комплексных экологических задач целесообразно использовать энциклопедии (или, при их отсутствии, — энциклопедические словари) по физике, химии, геологии, биологии, медицине, математике и т. п. При должном выборе предметный указатель вполне пригоден в качестве базового списка ключевых слов, который при необходимости будет пополняться.

Предметные указатели большинства энциклопедий устроены сходным образом — в них содержатся термины, являющиеся названиями статей энциклопедии (они обычно выделяются шрифтом), термины, определения которых даны в статьях, а также упомянутые в статьях наиболее важные результаты. Каждый термин сопровождает указание соответствующего тома и страницы (иногда таких указаний может быть несколько). На рис. 4.1 представлена структура предметных указателей математической [99], физической [180] и химической [181] энциклопедий.

В качестве дескрипторов (т. е. терминов, являющихся именами классов близких по смыслу понятий) полагаются названия статей энциклопедии, а связанными с ними по смыслу считаются слова из предметного указателя, встречающиеся в соответствующих стать-

а

Ядровая конъюнкция 1, 560
Язык автоматный 1, 1087
– алгоритмический 1, 222
– бесконтекстный 1, 1087
– в алфавите 5, 643
– линейный 1, 1091
– логико-математический 4, 578
– машинно-ориентированный 1, 223
– машинно-ориентируемый 3, 629
– машинный 1, 223
– многообразие 5, 644
– многосортный 5, 637
– модель 3, 769
– анализирующая 1, 247
– аналитическая 1, 247
– над данными символами и константами 1, 125

б

Автоионизационные состояния атомов (и ионов), I, 12
Автоионизация, II, 195(2)
Автоионизация колебательная, IV, 395(1)
Автоионный микроскоп, II, 209 (1)
Автоколебания, I, 12; IV, 695(2)
Автоколебания стохастические, IV, 695(2)
Автоколлимация, I, 15
Автокорреляционная функция, II, 466(2)
Автолокализация квазичастиц в твердых телах, I, 15; IV, 81(1)
Автоматизация эксперимента, I, 16

в

Абсорбция 1/4,5–14; 2/1300; 5/170
адиабатическая 4/755
аппаратура, см. *Аборберы*
и газов осушка 1/896, 897
– – очистка 1/931, 932
– – разделение 1/904, 905
и однонаправленная диффузия 2/1306
– охлаждение 5/597–599, 605
изотермическая 4/755–757
как вид сорбции 4/770
как метод концентрирования 2/916
как неэквимолярный массообмен 2/1298
масляная 1/902, 928, 932
с химической реакцией, см. *Хемосорбция*

Рис. 4.1. Структура предметных указателей энциклопедий.
555*а* — математической, *б* — физической, *в* — химической.

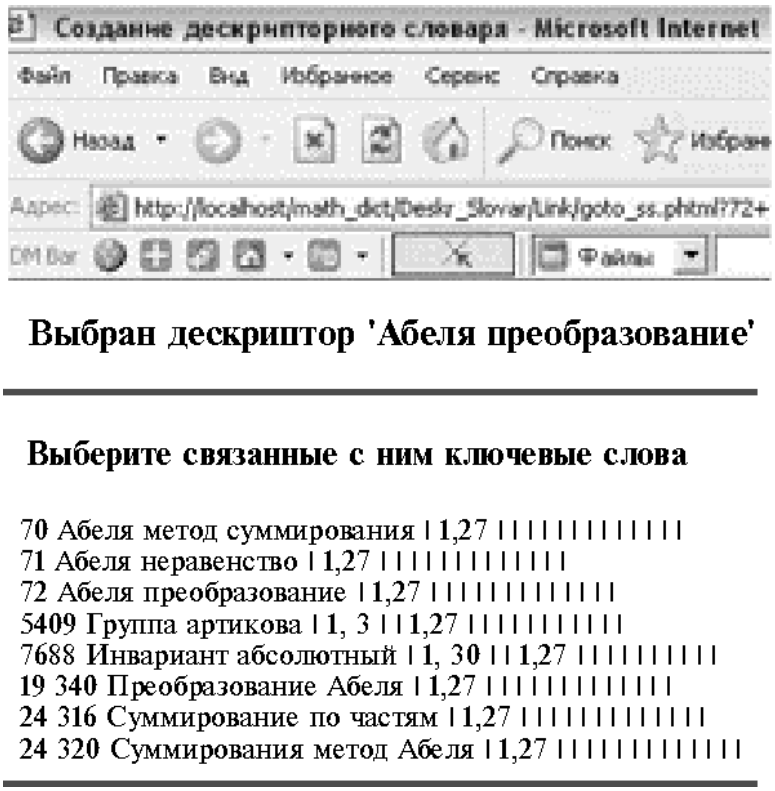

Рис. 4.2. Выбор ключевых слов, связанных с дескриптором.

ях. Основным преимуществом такого метода является то, что для установления связей между терминами не требуется быть экспертом (и даже специалистом) в данной предметной области: наличие в предметных указателях страниц, на которых встречаются термины, позволяет автоматически определять термины, относящиеся к статье, соответствующей дескриптору (с незначительным «избытком», поскольку на одной странице может быть несколько статей; избыточные термины отсеиваются вручную; см. рис. 4.2). Если же статья занимает несколько страниц, то термины, бесспорно к ней относящиеся (находящиеся не на первой или на последней страницах), связываются с дескриптором автоматически, что значительно сокращает объем ручной работы.

Следующий этап построения онтологии — классификация связей (отношений) между дескрипторами и ключевыми словами.

Обсуждение необходимого набора отношений началось еще в 1950–1960-е годы (соответствующий обзор сделан в [104, с. 432–454]) и продолжается по сей день (см., например, [67, 56]), однако выводы большинства современных работ (еще раз напомним: мы рассматриваем только «классический» подход к построению онтологий и тезаурусов!) весьма сходны друг с другом: минимально необходимый набор отношений должен включать в себя связи «род — вид», «часть — целое», синонимию (симметричную ассоциацию) и несимметричную ассоциацию (типа *дерево — лес*, которая, очевидно, не сводится к отношению «часть — целое» [56]).

Конкретный набор связей может определяться в соответствии с теми или иными стандартами, принятыми разработчиками информационной системы, для которой создается онтология. В частности, для работы с использованием протокола Z39.50 (принятого в качестве одного из корпоративных стандартов СО РАН [176]) связи устанавливаются в соответствии с рекомендациями схемы Zthes [267], которая выделяет следующие типы:

— BT — связь с родительским термином, т. е. с термином более широкого смысла;

— NT — связь с дочерним термином, т. е. с термином более узкого смысла. Связь BT — NT является взаимно-обратной;

— USE — связь с термином, который используется вместо этого;

— UF — взаимно-обратная связь USE;

— RT — связь, определяющая связанный по смыслу термин;

— LE — связь между лингвистически эквивалентными терминами.

Разумеется, могут (и должны!) быть установлены связи не только дескрипторов с ключевыми словами, но и дескрипторов между собой, однако это уже задача экспертов, выполняемая на следующем этапе, связанном с классификацией дескрипторов.

Наконец, при создании тезауруса и онтологии необходимо провести классификацию дескрипторов в соответствии с разделами данной предметной области. Выбор конкретного классификатора определяется целями создания онтологии, при этом возможно одновременное использование нескольких классификаторов (что, естественно, требует больши́х трудозатрат).

Сделаем краткий обзор некоторых классификаторов, наиболее употребительных в информационном обеспечении научной деятельности [212], рассматривая в качестве конкретных примеров

представление в этих классификаторах вопросов математики и информатики.

Документы научно-организационной направленности классифицируются в соответствии с правилами, установленными соответствующей научно-административной структурой. Такого рода классификаторы состоят из небольшого числа разделов, соответствующих наиболее общим разделам конкретной науки. Например, документы, связанные с защитой диссертаций (тексты диссертаций и авторефератов, страницы диссертационных советов, паспорта специальностей и т.п.), классифицируются с использованием Номенклатуры специальностей научных работников, утверждаемой Министерством образования и науки, которая содержит, в частности, 8 наименований математических специальностей и более 10 наименований специальностей, связанных с информатикой.

Исследования, проводимые в Российской академии наук, классифицируются в соответствии с утвержденными Президиумом РАН основными направлениями фундаментальных исследований. Они содержат 13 разделов математики и 6 разделов информатики, касающихся кибернетических вопросов.

Для классификации проектов, выполняемых при поддержке Российского фонда фундаментальных исследований, используется классификатор, содержащий 14 разделов математики и 9 — информатики. Однако отсутствие на сайте РФФИ подробной информации о содержании проектов приводит к выводу о нецелесообразности подключения в настоящее время данного классификатора.

Промежуточное положение по объему между краткими научно-организационными классификаторами и подробными классификаторами, предназначенными для систематизации документов собственно научного характера (электронных публикаций статей, книг и т.п.) занимает Государственный рубрикатор научно-технической информации (ГРНТИ), принятый для систематизации научно-технической информации в России и государствах СНГ. Он разработан и поддерживается Межведомственной комиссией по классификации и Научно-техническим центром РЕКТОР. ГРНТИ содержит, в частности, более 140 разделов в области математики и более 70 — в области кибернетики (иерархия в каждой из областей двухуровневая). ГРНТИ связан в своей кодовой части с Номенклатурой специальностей научных работников. Этот классификатор широко используется в распределенных информационных системах, рабо-

тающих с библиографическими базами данных (например, сервером ZooPARK, созданным на основе протокола Z39.50 [65]), но непосредственная классификация веб-публикаций с его помощью, как правило, не проводится.

Из классификаторов, используемых библиографами, наиболее популярна Универсальная десятичная классификация (УДК) [261], поддерживаемая Международной федерацией по информации и документации и Консорциумом УДК (русская версия УДК поддерживается ВИНИТИ). В частности, многие российские научные журналы указывают коды УДК публикуемых статей, что может быть использовано в процессе автоматического извлечения метаданных документа. Тем не менее для классификации электронных научных документов УДК не получил достаточно широкого распространения. Это связано, во-первых, с недостаточно оперативным его обновлением, во-вторых, с жестким искусственным ограничением структуры (количество классов очередного уровня ограничено 10), и, в-третьих, с неравномерностью глубины уровней.

Вследствие указанных недостатков в некоторых отраслях науки разрабатываются специализированные классификаторы, используемые для индексации как полиграфических, так и электронных документов. Мы не ставим целью сделать обзор всех специализированных классификаторов, поэтому подробно рассмотрим ситуацию на примере предметной области «Математика». Среди математиков весьма популярна Mathematics Subject Classification (MSC) [243], используемая ведущими мировыми реферативными изданиями: «Mathematics Review» и «Zentralblatt MATH», которая более приспособлена для функционирования в рамках интеллектуальных систем обработки документов из области «Математика» по следующим причинам [8].

1. В настоящее время мировое сообщество математиков гораздо чаще использует MSC, чем УДК. Например, с помощью MSC ведется база данных «Zentralblatt MATH» [266], содержащая рефераты почти 3 млн статей. Наличие такой базы чрезвычайно важно для решения задачи автоматической классификации веб-ресурсов с использованием удаленных библиографических описаний (см. разд. 4.3).

2. MSC значительно более точно отражает современное состояние математической науки, чем УДК — последняя переработка MSC состоялась в 2000 г. (причем уже ведутся работы по подготовке

в 2010 г. новой версии), а раздел «Математика» УДК; в 1975 г. (в 1988 г. он претерпел довольно незначительные изменения). Достаточно сказать, что в УДК отсутствует термин «псевдодифференциальные операторы» (в MSC им посвящен класс второго уровня 35Sxx); вся *K*-теория уместилась в один пункт 512.736 (в MSC ей выделен класс первого уровня 19-XX, включающий в себя 11 классов второго уровня, разделенных более чем на 50 классов третьего уровня); в разделе 519.6 — «Вычислительная математика, численный анализ» — приведены только области применения вычислительных методов, а указания на тип этих методов, в отличие от раздела MSC 65-XX, отсутствуют. Такие примеры можно продолжить.

3. В MSC имеются специальные классы, охватывающие историю и философию математики, смежные с математикой науки (информатику и теоретическую физику, включая различные отрасли механики), а также посвященные применению математических методов в естественных и общественных науках. Для того чтобы добиться при помощи УДК такого охвата понятий, пришлось бы использовать не только разделы 51 — «Математика», 52 — «Астрономия. Астрофизика. Исследование космического пространства. Геодезия», 53 — «Физика», 004 — «Информационные технологии. Вычислительная техника», но и многие другие разделы, касающиеся естественных и общественных наук, что резко увеличило бы объем избыточной информации. Альтернативное решение — включить в создаваемый тезаурус лишь отдельные, наиболее связанные с математикой, подразделы этих разделов УДК — нарушило бы целостность классификатора.

4. Как уже отмечалось, с точки зрения числа допустимых для отображения в структуре членов классификационного деления на каждой его ступени УДК относится к классификациям с искусственными ограничениями, так как в ней количество классов очередного уровня ограничено 10. В то же время MSC, в которой количество классов первого и третьего уровня ограничивается 100, а второго — 26, что вполне удовлетворяет практическим потребностям, может быть отнесена к классификациям с естественными ограничениями.

5. Трехуровневая структура MSC значительно более приспособлена для автоматизации поиска и классификации документов, чем структура УДК, которая в отдельных случаях насчитывает до

7 уровней (число уровней подсчитывается внутри раздела УДК «Математика»).

Однако классификатор MSC имеет и свои недостатки. Некоторые разделы математики классифицируются в нескольких несвязанных местах. Например, обратным задачам соответствуют классы 35R30, 49N45, 65M32, 86A22. Это затрудняет понимание развития теории обратных задач и поиск информации по этой тематике. Впрочем, подобный недостаток присущ и УДК.

Иногда в MSC классы одного уровня соответствуют как большим, так и частным разделам математики. Например, для неассоциативных колец и алгебр и для численного анализа, очень обширного и состоящего из множества подразделов, выбраны классы одного уровня: 16-XX и 65-XX соответственно.

Наконец, часто выдающиеся математические результаты получаются на стыке разных областей математики. Это свидетельствует о том, что взаимопроникновение разных ветвей математики дает толчок для развития всей математики. Существует необходимость проследить тенденции в развитии математики и увидеть наиболее важные точки роста, предсказать появление новых научных дисциплин как внутри самой математики, так и в связи с другими областями знаний. Сделать это, используя только классификатор MSC, крайне затруднительно.

Несмотря на указанные недостатки, при создании тезауруса и онтологии предметной области «Математика» целесообразно предпочесть именно Mathematics Subject Classification, ныне действующая версия которой нередко обозначается как MSC2000.

Что касается автоматического (т. е. основанного только на кодах классификатора и не опирающегося на другие метаданные документа) переиндексирования документов (включая отдельно взятые термины) из классификации *A* в классификацию *B*, то оно возможно лишь для тех разделов верхнего уровня классификатора *A*, которые связаны с соответствующими разделами классификатора *B* связями вида «один ко многим» (подобно тому, как проблематичен внеконтекстный перевод многозначных слов, например «*лук*», на иностранный язык). На практике полное отсутствие связей «один ко многим» может встретиться лишь в том случае, когда *A* — подробный научный классификатор, а *B* — очень краткий научно-организационный. Поэтому использование при создании онтологии (или каталога) двух подробных классификаторов, например

MSC и УДК, влечет за собой большой объем дополнительного труда экспертов.

Вернемся к алгоритму построения онтологии. После того как выбран классификатор, названия его разделов также включаются в тезаурус в качестве дескрипторов. Между дескрипторами, являющимися названиями разделов классификатора, и прочими дескрипторами устанавливаются связи вида NT, RT, LE.

Далее проводится классификация ключевых слов. С целью экономии трудозатрат экспертов на первом этапе работы возможно ограничиться классификацией дескрипторов, при этом для классификации следует использовать, по возможности, разделы классификатора максимально низкого уровня. После того как дескриптор будет классифицирован, ключевым словам, связанным с ним отношениями BT, USE, RT и LE, приписывается тот же классификационный индекс, что и дескриптору. Впрочем, это не исключает такой ситуации, что если дескриптор отнесен к классу не самого низкого уровня, то при последующей работе эксперта термины, связанные с дескриптором отношениями BT и USE, могут быть отнесены к классу более низкого уровня. В этом случае указанные термины сами станут дескрипторами.

Наконец, проводится определение типа термина в соответствии с рекомендациями Zthes, получившими развитие в [65]. Выделяются следующие типы терминов:

— TT — термин верхнего уровня, т. е. термин, не имеющий связанных терминов более широкого класса (терминов с типом связей BT);

— NT — не термин верхнего уровня, т. е. дескриптор, имеющий связи типа ÂÒ;

— ND — не основной термин;

— NL — фиктивный термин, т. е. термин, не используемый для индексации документов, но включенный в иерархию, чтобы указать логический базис раздела классов.

Определение типа терминов позволяет существенно упростить работу с онтологией.

В результате все термины предметного указателя энциклопедии оказываются связанными с другими терминами и классифицироваными в соответствии с разделами выбранного классификатора. Это означает, что базисная структура онтологии создана. Далее в случае необходимости эксперты могут пополнять ее новыми терми-

нами, классифицируя их и устанавливая связи посредством веб-интерфейса, а также редактировать версии онтологии на иностранных языках (частичная автоматизации процесса перевода терминов будет описана в разд. 4.3).

Общая схема изложенного алгоритма представлена на рис. 4.3.

Работоспособность данного алгоритма и веб-приложения была проверена путем создания тезауруса ряда разделов предметной области «Математика» («Дифференциальные уравнения», «Уравнения в частных производных», «Численный анализ», «Механика жидкости» и др.) на основе предметного указателя «Математической энциклопедии». Установлено, что для классификации терми-

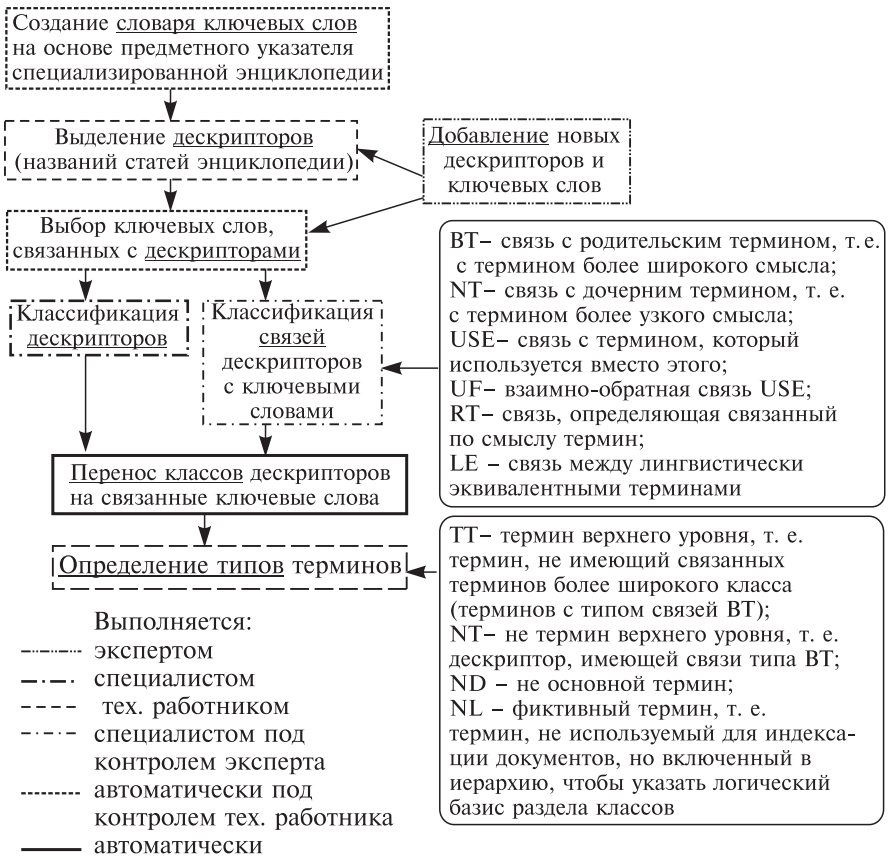

Рис. 4.3. Схема алгоритма автоматизированного создания тезаурусов и онтологий.

нов и определения связей между ними достаточно квалификации бакалавра математики (при условии привлечения в редких случаях для консультаций эксперта с ученой степенью). Это доказывает высокую эффективность разработанного алгоритма.

4.2. Автоматизация процесса извлечения метаданных из слабоструктурированных документов

Фуражировка и ремонтерство
Требуют сноровки и прозорства.
Ф. К. Прутков. Военные афоризмы

Как уже отмечалось выше, первоначальный этап научно-информационного процесса с участием электронных документов включает в себя их сбор и каталогизацию, сводящуюся к извлечению из документов их метаданных.

Подчеркнем, что каждую публикацию в составе электронного журнала, сборника и т. п. целесообразно представлять как отдельный документ. Это существенно облегчает процесс поиска пользователем нужной информации, позволяя вести атрибутивный поиск отдельных статей по авторам, названию, классификационным признакам, ключевым словам (понимаемым в этой главе в узкобиблиографическом смысле, в отличие от «информационного» употребления этого термина в определении тезауруса из разд. 1.4) и т. п. Разумеется, аналогичный подход весьма желателен и при работе с полиграфическими изданиями, однако в этом случае данное требование трудноосуществимо из-за огромных трудозатрат: как показано в [104], один человек за рабочий день способен описать не более 50–70 документов на родном языке и не более 20–30 — на иностранном. При обработке электронных документов, в том числе интернет-документов, возможна частичная автоматизация процесса каталогизации отдельных публикаций.

Поскольку количество организаций, работающих в той или иной конкретной области науки, а также журналов, публикующих статьи соответствующей тематики, как правило, сравнительно невелико, постольку задача первичного поиска и каталогизации научных ресурсов (прежде всего, сайтов научно-исследовательских институтов и электронных версий журналов) не представляет большой сложности для

специалиста, активно работающего в данной области науки. Однако полноценное удовлетворение информационных потребностей пользователя возможно лишь при каталогизации отдельных документов, в частности статей. К сожалению, далеко не все журналы размещают на своих сайтах полные тексты статей; многие ограничиваются лишь аннотациями, тем не менее практика размещения в сети Интернет полнотекстовых версий статей получает все более широкое распространение. В большинстве случаев такая публикация представляет собой HTML-страницу с аннотацией документа, на которой имеется ссылка на полный текст, например в формате pdf или ps.

Трудоемкость процесса извлечения метаданных из документов приводит к необходимости его частичной автоматизации. Основные сложности при решении этой задачи состоят в разработке алгоритма, позволяющего в автоматизированном режиме извлекать из слабоструктурированного документа основные элементы его библиографического описания.

Так как однородные документы, размещенные на одной сайте, имеют однородную структуру, то наиболее целесообразно использовать алгоритмы, учитывающие информацию о гипертекстовой разметке обрабатываемых документов (см., например [225, 251]), при этом надо иметь в виду, что метаданные в мета-тегах документа могут отсутствовать, поэтому следует ориентироваться только на HTML-разметку документа.

Один из возможных алгоритмов решения задачи частичной автоматизации процесса извлечения метаданных разработан и изложен нами в работах [10, 11, 30]. Его преимущество перед подходами, применяемыми в коммерческих пакетах, предназначенных для решения аналогичной задачи, состоит, прежде всего, в возможности получения метаданных обрабатываемого документа из удаленных библиографических баз данных.

Алгоритм, основанный на типичном для интеллектуальных информационных систем человеко-машинном взаимодействии, сводится к выполнению последовательных операций:

— создание шаблона для обрабатываемого сайта;

— создание списка адресов, где расположены документы;

— обработка документов, включая возможное извлечение метаданных из удаленных библиографических источников (подробнеесм. разд. 4.3);

— поддержание актуальности информации.

Ниже приводится описание этапов этого алгоритма, общая схема которого представлена на рис. 4.4.

Шаблон документа необходим для автоматического выделения его основных метаданных. Например, для документа — журнальной статьи метаданные включают в себя следующие элементы биб-

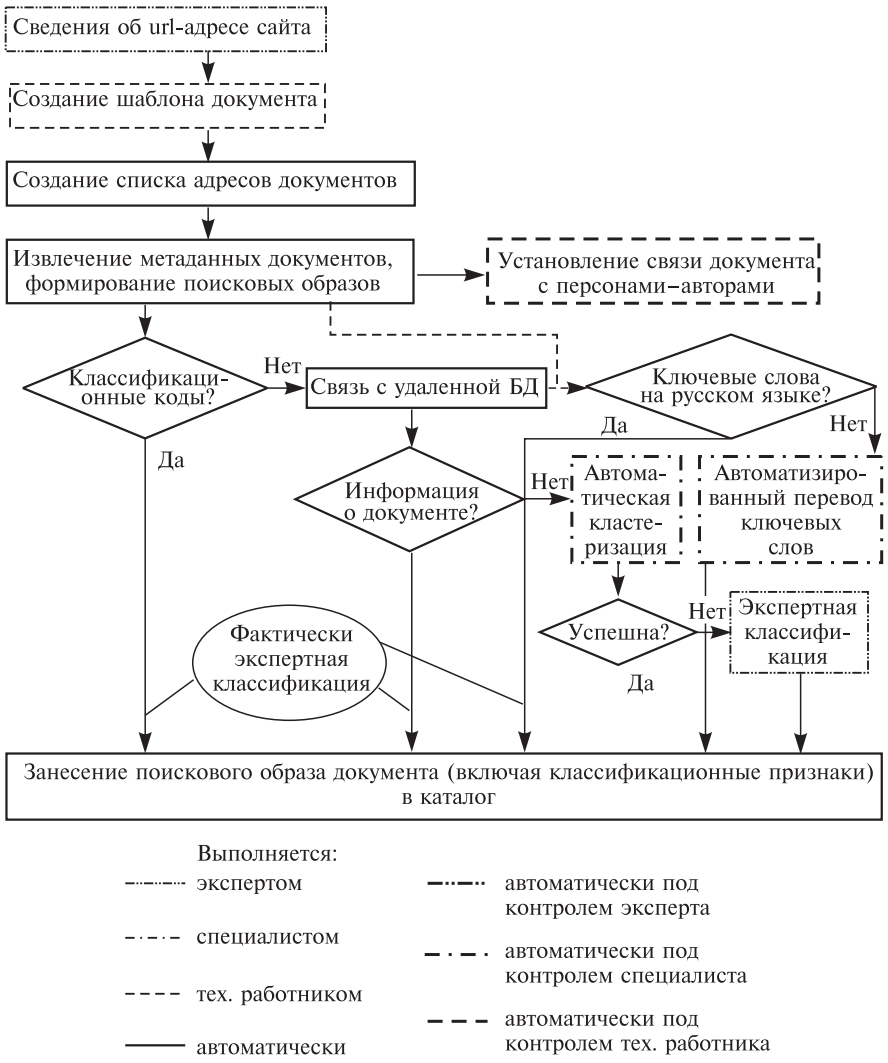

Рис. 4.4. Схема алгоритма автоматизированного извлечения метаданных из слабоструктурированных документов.

лиографического описания: авторы, название, ключевые слова, аннотация, коды классификатора, выходные данные полиграфической версии статьи и т.п., причем текстовые поля могут заполняться и на русском, и на английском языках. В каждом конкретном случае шаблон создается сравнительно легко с использованием языка регулярных выражений в том или ином формате (RegEx, Posix и т. п.), однако проблема заключается в том, что разные сайты, пусть даже и однотипные (например, электронные версии разных журналов), имеют сильно различающуюся структуру описания и представления документов. Поэтому желательно создание алгоритма, реализуемого в виде веб-приложения, позволяющего пользователю, даже не владеющему языками обработки регулярных выражений, генерировать шаблоны для различных сайтов. Разумеется, если пользователь знает основы использования регулярных выражений, то он может указать выражения, описывающие формат данных, поскольку такое указание способно в отдельных ситуациях повысить эффективность работы алгоритма.

В общем случае пользователю-каталогизатору достаточно занести в поля формы (см. пример на рис. 4.5) теги, окружающие в HTML-коде статьи значения каждого из элементов метаданных одного документа обрабатываемого сайта, а также указать разделитель данных в случае множественности некоторого элемента метаданных, после чего создается и сохраняется шаблон веб-страницы документов сайта. Надежность такого алгоритма особенно высока в случае автоматической верстки веб-страниц каталогизируемого сайта, когда их структура практически однородна.

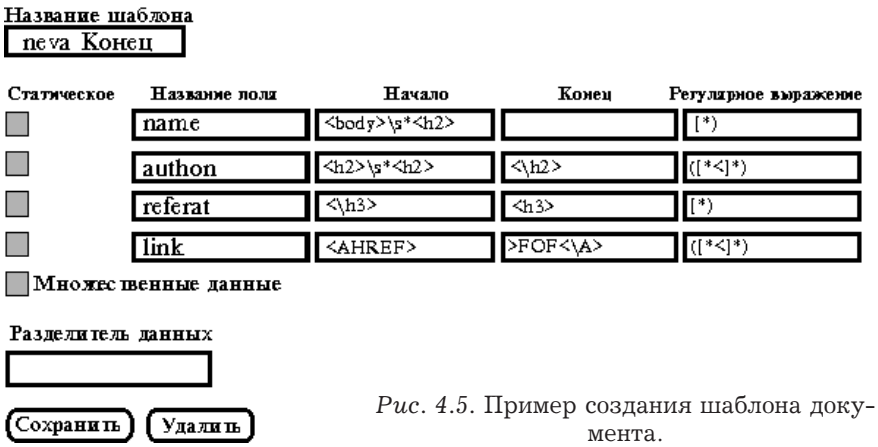

Рис. 4.5. Пример создания шаблона документа.

Нельзя исключить ситуацию, когда на одном обрабатываемом сайте хранится несколько коллекций документов (например, статьи и краткие биографические данные авторов). В этом случае соответствующий шаблон создается для каждой коллекции.

Сайты, отображающие коллекции документов, имеют, как правило, достаточно четкую иерархическую структуру вложенных гиперссылок (например, для журнала эта структура представляется как том — номер — статья), причем количество уровней обычно невелико. Поэтому создание списка адресов, где расположены документы, сводится к автоматическому нахождению гиперссылок, расположенных на данной странице, с занесением в базу данных адресов нижнего уровня, непосредственно содержащих каталогизируемые документы. На каждом этапе каталогизатор отмечает те из найденных адресов, с которых следует вести дальнейший сбор информации (рис. 4.6).

Полная автоматизация процесса создания списка адресов оказалась нецелесообразной, так как веб-страницы могут содержать «посторонние» ссылки (переход к другим разделам сайта и т. п.), что способно привести к зацикливанию процесса. Попытка автоматического отсеивания таких адресов чревата ошибками первого рода (пропуском нужных веб-страниц и документов), что недопустимо.

Этап обработки документов носит чисто технический характер. Он заключается в загрузке ранее необработанных веб-страниц из базы адресов и их разборе с использованием регулярных выражений по соответствующему шаблону. Извлеченные таким образом метаданные заносятся в каталог.

Регулярная проверка актуальности занесенных в каталог документов также осуществляется посредством стандартных приемов.

Отметим, что автоматизированная каталогизация электронных документов может привести к появлению в каталоге дублирующих записей (например, ссылка на одну и ту же статью содержится на сайте журнала, на персональной странице автора статьи и в списке публикаций организации, где работает автор). Ситуация осложняется тем, что форматы записи в разных источниках могут несколько отличаться друг от друга, в некоторых источниках текст записи может содержать грамматические ошибки и т. п. Возникает задача выявления нечетких дубликатов библиографи-

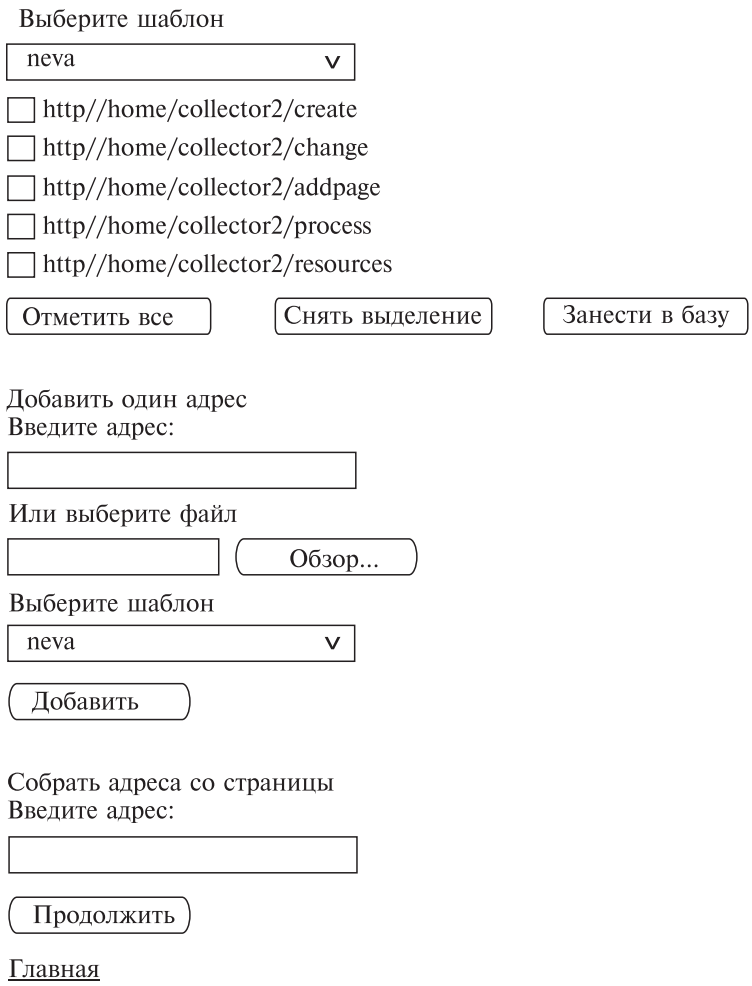

Рис. 4.6. Занесение адресов документов.

ческих записей с целью исключения из каталога (или из результатов запроса) тех документов, которые содержат менее полную информацию (например, только аннотацию при наличии полнотекстового документа). Один из методов решения этой задачи, основанный на использовании в качестве функции похожести наибольшей общей подпоследовательности двух строк, изложен в работе авторов [142].

4.3. Автоматизация процесса получения метаданных документа с использованием удаленных библиографических описаний

Сноб — это пес, блохи которого привезены исключительно из Лондона.

Фафик, терьер. Цит. по книге: «Мысли людей великих, средних и пса Фафика»

Следует обратить особое внимание на извлечение таких метаданных, как классификационные признаки (т. е. коды того или иного классификатора) документа и ключевые слова. Без этих элементов метаданных ценность каталожного описания документа минимальна, поскольку процесс поиска документа человеком или его обработка рассуждающей информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа. К сожалению, даже наиболее структурированные документы — журнальные статьи — далеко не всегда содержат ключевые слова и классификационные признаки. И даже в тех случаях, когда эти признаки указаны, классификатор, используемый журналом, может не соответствовать классификатору каталога. Например, как было сказано в разд. 4.1, в некоторых математических журналах используется классификатор УДК [261], в то время как в международном математическом сообществе более распространен классификатор MSC2000 [243].

Разумеется, наиболее качественно решить задачу классификации может эксперт-человек, поэтому прежде всего следует проверить, не внесена ли информация о полиграфической версии документа в ту или иную электронную библиографическую базу данных удаленного доступа, в которой документы классифицированы в соответствии с нужным классификатором, например базы данных российских реферативных журналов, «Current Contents» и т. п. Исходя из особенностей описания документа в той или иной конкретной библиографической базе, к ней формируется запрос, содержащий соответствующие сведения о классифицируемом документе. Так, в среде математиков достаточно широко известна база данных журнала «Zentralblatt MATH» [266], содержащая около 3 млн записей. Статью в этой базе можно однозначно идентифицировать по ISSN журнала, его номеру и страницам, на которых расположена статья.

К сожалению, не все электронные версии журналов содержат номера страниц полиграфических версий статей, поэтому при отсутствии сведений о страницах в процессе идентификации следует опираться на фамилию автора (авторов) в латинской транскрипции.

Особо отметим, что полная репликация метаданных документа из библиографической базы не может служить эффективной заменой процессу непосредственного извлечения метаданных из интернет-документа, поскольку в большинстве случаев библиографические базы не содержат сведений об url—адресе электронной версии документа.

Процесс определения метаданных документа с использованием удаленной библиографической базы также может быть частично автоматизирован.

Для обращения к этой базе данных с целью получения классификационных признаков документа автоматически формируется строка запроса к серверу библиографической базы, использующая в качестве параметров запроса уже извлеченные с веб-страницы журнала библиографические данные. При наличии сведений о запрошенном документе в базе данных сервер выдает страницу с его описанием, на которой присутствуют, среди прочих библиографических данных, и классификационные признаки документа (так, например, база данных журнала «Zentralblatt MATH» содержит классификационные коды по классификатору MSC2000 и ключевые слова на английском языке). Обработка полученной страницы, т. е. извлечение недостающих метаданных документа, производится по стандартному шаблону с помощью регулярных выражений.

Если же поиск в библиографической базе данных успехом не увенчался, то классификация документа возможна с помощью методов, излагаемых в разд. 4.5.

После получения ключевых слов документа из англоязычной библиографической базы данных может возникнуть проблема их перевода на русский язык. Разумеется, прежде всего следует проверять наличие переводимого термина в англоязычной части тезауруса (онтологии) предметной области (см. разд. 4.1). Процесс перевода отсутствующих в тезаурусе терминов может быть частично автоматизирован с использованием словарей, доступных через Интернет, например словаря «Лингво» компании «Яндекс» [155]. Аналогично описанной ситуации автоматически формируется строка запроса к удаленному словарю с последующей обработкой результатов запроса.

Соответствующий алгоритм включает в себя следующие этапы.

1. После получения списка ключевых слов для перевода программа разбивает этот список на отдельные словосочетания.

2. Каждое словосочетание из списка программа сначала пытается найти в англоязычной части тезауруса.

3. Если словосочетание в тезаурусе не нашлось, то делается запрос к удаленному словарю.

4. Если в удаленном словаре не удалось найти перевод словосочетания целиком, то оно разбивается на отдельные слова и для каждого их них выполняются вышеперечисленные действия, т. е. слово сначала ищется в англоязычной части тезауруса, и если его найти не удается, то делается запрос к удаленному словарю. Потом из переводов отдельных слов вновь составляется словосочетание.

5. После того как все словосочетания переведены, пользователю предлагается скорректировать переводы (рис. 4.7).

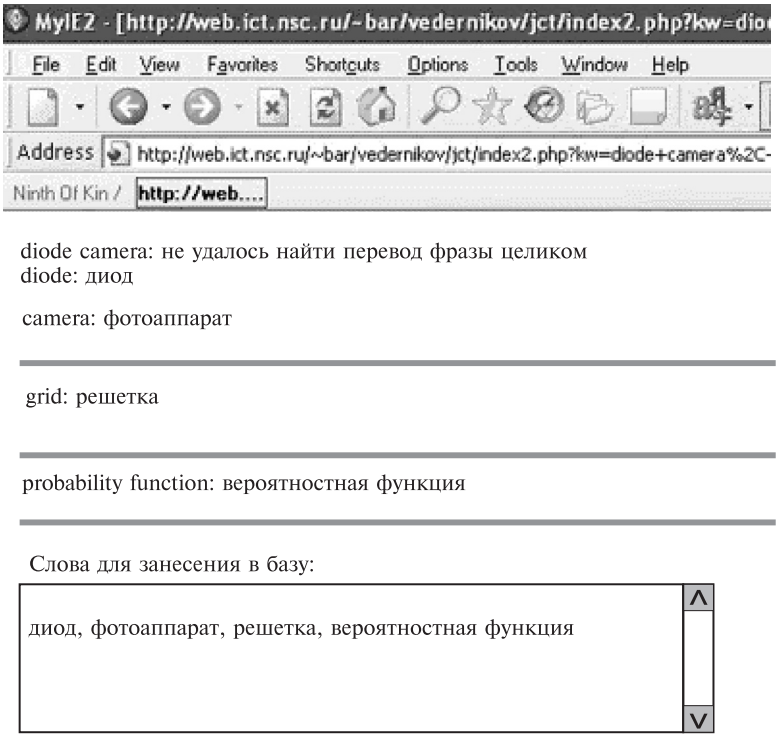

Рис. 4.7. Корректировка перевода ключевых слов.

6. Ключевые слова заносятся в метаописание документа и при наличии русских соответствий в тезаурусе — в его англоязычную часть.

Таким образом, происходит процесс обучения системы: чем больше слов и словосочетаний переведено, тем меньше программа обращается к удаленному словарю через Интернет, так как уже переведенные слова и словосочетания заносятся в тезаурус.

4.4. Автоматическое извлечение из текстов ключевых слов

> Усердный в службе не должен бояться своего незнанья; ибо каждое новое дело он прочтет.
>
> *Козьма Прутков. Плоды раздумья*

Важной задачей обработки текстовых документов, без решения которой практически невозможна автоматизация процесса извлечения из них информации и знаний, является координатное индексирование, т. е. извлечение из текстов ключевых слов (всех содержащихся в индексируемом тексте терминов, входящих в словарь онтологии данной предметной области). Как отмечено в [3, с. 280], координатное индексирование документов может производиться автоматически, поскольку оно дает почти такие же результаты, как и ручное, но имеет перед ним ряд преимуществ:

— обеспечивает единообразие индексирования, почти невозможное для человеческого интеллекта;

— обходится, по меньшей мере, в 3 раза дешевле.

Ввиду того что в русском языке имена существительные и прилагательные при склонении изменяют свою форму, разработка эффективного алгоритма автоматизации извлечения ключевых слов представляет нетривиальную задачу, ибо необходимо учитывать и те случаи, когда слова, образующие термин, находятся не только в именительном, но и в косвенных падежах.

Среди некоммерческих программных продуктов, решающих указанную задачу, можно назвать стимер компании «Яндекс» [161], алгоритм работы которого изложен в работе [255]. Однако он анализирует текст только на синтаксическом уровне, позволяя извлекть словосочетания заданной структуры (например, *(прилагательное) + (существительное)* или *(существительное) + (существи-*

тельное в родительном падеже), но не проверяя принадлежность словосочетаний к тому или иному лексическому словарю.

В основу алгоритма [13] положено использование двух индексов, содержащих триады «номер текста» — «позиция в тексте» — «номер слова из лексического словаря» и «номер термина» — «позиция слова в термине» — «номер слова из лексического словаря». При этом если первый индекс встречается практически во всех информационно-поисковых системах, то введение второго индекса, позволяющее резко повысить эффективность алгоритма, имеет оригинальный характер. Индекс терминов наряду с их списком размещается в хранилище данных программной библиотеки, реализующей алгоритм, и пополняется по мере изменения этого списка.

Алгоритм построения индекса терминов состоит из следующих этапов:

1. Разбиение термина на отдельные слова.

2. Создание предварительного индекса, содержащего триады «номер термина» — «позиция слова в термине» — «слово в символьном представлении».

3. Добавление встретившихся неизвестных слов в лексический словарь библиотеки, где им присваиваются идентификационные номера.

4. Переработка индекса в формат «номер термина» — «позиция в тексте» — «номер слова из лексического словаря».

5. Сбор статистики о длинах терминов для реализации поиска и идентификации составных терминов (т. е. терминов, состоящих более чем из одного слова).

6. Сбор статистики о количестве вхождений отдельных слов в термины для оптимизации поиска путем исключения из рассмотрения терминов, заведомо отсутствующих в тексте.

Алгоритм построения индекса текстов аналогичен, но в нем отсутствует этап 3.

Заключительная стадия работы программной библиотеки — подсчет количества вхождений терминов в текст (тексты). Ее этапы:

1. Подсчет возможных комбинаций «текст» — «термин», основанный на статистике вхождения отдельных слов (см. этап 6 алгоритма индексации терминов).

2. Нахождение всех потенциально возможных мест вхождения каждого термина в текст (тексты) на основе наличия хотя бы одного

общего слова из лексического словаря. Позиция каждого потенциально возможного вхождения фиксируется.

3. Рассмотрение каждого из возможных мест вхождений с точки зрения соответствия термину в целом. Актуальность вхождения определяется наличием рядом с соответствующей позицией других слов, входящих в термин. Существуют конфигурируемые варианты требований определения актуальности вхождения (точный или неточный порядок слов, минимальное количество слов, входящих в термин, возможность «прерывания» термина посторонними словами и т. п.).

4. Исключение учета вхождений, поглощаемых более длинными вхождениями.

5. Сбор статистики вхождений для каждой пары «текст» — «термин».

На основании изложенного алгоритма реализована программная библиотека, включающая в себя функции для поиска и подсчета количества вхождений в заданный текст (тексты) некоторых последовательностей слов (в данном случае терминов, входящих в словарь онтологии той или иной предметной области). Работа с программной библиотекой осуществляется через веб-интерфейс.

В процессе создания библиотеки встал вопрос о выборе средств, дающих возможность учитывать морфологию слов используемого языка. Выбор пал на свободно распространяемый программный продукт Ispell [233], изначально предназначенный для проверки орфографии на разных языках (язык проверки определяется словарем, который подключает пользователь). Первый русский словарь для Ispell был создан в 1992 г. англоязычными пользователями (см. [157]). Он имел весьма небольшой объем (52 тыс. словоформ) и слабо развитый файл преобразования окончаний слов (affix-файл).

В настоящее время в affix-файл словаря его создателями добавлены правила образования форм существительных, прилагательных, причастий, наречий, изменены правила формирования окончаний глаголов. Основной подход, положенный ныне в основу словаря, заключается в использовании нормализованной формы слова и правил словоизменения, отвечающих грамматике русского языка. Все слова разбиты на флективные классы (типы словоизменения), каждому из которых ставится в соответствие система окончаний всех словоформ слова-представителя. По этой причине сло-

варь одновременно содержит и важную информацию о морфологии слов, которая необходима для современных русскоязычных поисковых систем. В использованной при реализации версии (0.99g2) объем словаря составляет более 137,2 тыс. базовых слов, а полное число образуемых из них словоформ превышает 1,321 млн, при этом словарь постоянно пополняется, в том числе и за счет специальных терминов, предлагаемых пользователями словаря. Таким образом, среднее количество элементов во флективном классе — порядка 10.

Указанный словарь был использован для создания базового лексического словаря программной библиотеки. В процессе работы лексический словарь пополняется в автоматическом режиме без генерации словоформ нового слова, а также в экспертном режиме. Однако работа эксперта по генерации всех словоформ нового слова весьма трудоемка: для существительного с учетом изменения падежа и числа нужно выписать 12 словоформ, для прилагательного с учетом изменения падежа, рода и числа — 24 словоформы, при этом многие словоформы будут повторяться. Такой объем механической работы, помимо больших трудозатрат, чреват неизбежным появлением опечаток.

Для автоматизации работы эксперта построено веб-приложение, автоматически генерирующее все словоформы заданного слова (существительного или прилагательного) русского языка. Мы ограничились только существительными и прилагательными, поскольку именно эти части речи обычно выступают в качестве новых слов, а глаголы в подавляющем большинстве случаев не являются специфическими для той или иной предметной области и включены в основной словарь Ispell.

В основе работы веб-приложения лежит алгоритм Г. Г. Белоногова [38], использующий разбиение слов языка на флективные классы, т. е. типы словоизменения, каждому из которых ставилась в соответствие система окончаний всех словоформ слова-представителя (основа существительных и прилагательных, как правило, остается неизменной). Всего Г. Г. Белоноговым выделено для существительных 66 флективных классов, для прилагательных — 12, каждому из которых поставлен в соответствие полный набор окончаний.

При работе с новым словом эксперт устанавливает при необходимости его начальную форму и указывает его тип: независимое

существительное, прилагательное или зависимое слово-дополнение в родительном падеже. Зависимое слово сразу добавляется в словарь, так как единственной формой слова (применительно к соответствующему контексту) является оно само. При выборе независимого существительного на следующем шаге необходимо указать его род и одушевленность. Для прилагательного дополнительные характеристики не указываются.

Однако размеры *надклассов*, на которые разбиты флективные классы, зачастую слишком велики для эффективной работы пользователя: существительные мужского рода одушевленные — 19, существительные мужского рода неодушевленные — 16, существительные женского рода одушевленные — 8, существительные женского рода неодушевленные — 12, существительные среднего рода — 11, прилагательные — 12. Это затрудняет работу по выбору нужного класса, поскольку специалистами в области когнитивной психологии показано [245], что эффективный выбор возможен, если количество вариантов не превышает 7−9.

Для решения этой проблемы была предложена модификация алгоритма Г. Г. Белоногова, состоящая в автоматическом анализе окончаний нормализованной словоформы внутри каждого надкласса, что приводит к значительному уменьшению количества элементов, из которых предстоит сделать выбор:

— существительные мужского рода одушевленные: 12 + 2 + + 2 + 2 + 1;

— существительные мужского рода неодушевленные: 10 + 3 + 3;

— существительные женского рода одушевленные: 4 + 3 + 1;

— существительные мужского рода одушевленные: 6 + 4 + 2;

— существительные среднего рода: 5 + 5 + 1;

— прилагательные: 4 + 4 + 2 + 1 + 1.

Количество объектов-альтернатив в подавляющем большинстве случаев доведено до рекомендуемого когнитивной психологией. Для сушествительных мужского рода ситуация неулучшаема (например, слова «волос», «голос» и «колос» относятся к разным флективным классам).

Таким образом, программа автоматически проводит предварительный анализ окончания слова, отсеивая те классы, к которым данное слово заведомо принадлежать не может. После этого нужно выбрать флективный класс, которому соответствует слово. Для выбора предоставляется таблица возможных флективных классов,

Вы ввели слово **квазианалитический**, часть речи которого **прилагательное**

Выберите номер флективного класса, слово-представитель которого склоняется так же, как квазианалитический

№ класса	Слово-представитель	Им. п., муж. р., ед. ч.	Им. п., жен. р., ед. ч.	Род. п., муж. р., ед. ч.	Им. п., мн. ч.
104	передний	ий	яя	его	ие
105	хороший	ий	ая	его	ие
106	легкий	ий	ая	ого	ие
111	третий	ий	я	его	и

Рис. 4.8. Выбор флективного класса.

которые определяются словом-представителем и его несколькими характерными словоформами (рис. 4.8).

Мысленно просклоняв данное слово по указанным формам и сравнив полученные окончания с окончаниями из таблицы, можно однозначно определить его флективный класс. После этого программа генерирует все формы слова, отображая их в виде таблицы, в которой они распределены по падежам, числам и родам (если это прилагательное). Выводится список уникальных словоформ, так как обычно слово может иметь одинаковые окончания в разных формах (рис. 4.9). На основании этого списка эксперт принимает решение о занесении словоформ в словарь или, в исключительных случаях, когда сгенерированные словоформы оказываются неверными (например, у слова оказалась изменяемая основа), о переходе в ручной режим работы (разумеется, эксперт, заметив изменяемость основы, может принять решение о переходе в ручной режим и на более раннем этапе).

С помощью системы проведена генерация словоформ математических терминов из Предметного указателя «Математической энциклопедии», отсутствующих в словаре rus-ispell (таких слов около 3 тысяч). Работа системы показала ее удобство для пользователя и высокую эффективность по сравнению с непосредственным склонением слов по правилам русского языка.

Для сравнения отметим, что в алгоритме решения аналогичной задачи из работы [79], классы словоформ определялись без учета теоретических исследований Г. Г. Белоногова путем непосредственного анализа типов окончаний (не в узкоморфологическом, а в «обыденном» смысле) слов. Это приводит к появлению более 10 тыс. классов для существительных и 2,5 тыс. классов для прилагательных (к одному классу отнесены слова, у начальных форм которых

Вы ввели флективный класс номер 106

Слово **квазианалитический** имеет следующие *словоформы*:

Падеж	Муж. р., ед. ч	Жен. р., ед. ч.	Ср. р., ед. ч.	Мн. ч.
Им. п., ед. ч. (Кто? Что?)	квазианалитический	квазианалитическая	квазианалитическое	квазианалитические
Род. п., ед. ч. (Кого? Чего?)	квазианалитического	квазианалитической	квазианалитического	квазианалитических
Дат. п., ед. ч. (Кому? Чему?)	квазианалитическому	квазианалитической	квазианалитическому	квазианалитическим
Вин. п., ед. ч. (Кого? Что?)	квазианалитический	квазианалитическую	квазианалитическое	квазианалитические
Тв. п., ед. ч. (Кем? Чем?)	квазианалитическим	квазианалитической	квазианалитическим	квазианалитическими
Пр. п., ед. ч. (О ком? О чем?)	квазианалитическом	квазианалитической	квазианалитическом	квазианалитических

Уникальные словоформы:

квазианалитический
квазианалитического
квазианалитическому
квазианалитическим
квазианалитическом
квазианалитическая
квазианалитической
квазианалитическую
квазианалитическое
квазианалитические
квазианалитических
квазианалитическими

Добавить в базу!

Рис. 4.9. Генерация словоформ.

совпадают 3 последние буквы), что делает данный алгоритм трудновоспроизводимым. Однако даже столь детальное разбиение не способно дать абсолютно точное различение слов по типу склонения (еще раз напомним: слова «волос», «голос» и «колос» относятся к разным флективным классам), к тому же «эмпирический» характер разбиения вызывает определенные вопросы относительно полноты описания классов.

4.5. Кластеризация текстовых документов на основании меры сходства

> Классификация интегралов:
> 1. *Собственные* — интегралы, которые взял сам, и *несобственные*, которые списал.
> 2. *Определенные* — интегралы, к которым есть ответ, и *неопределенные*, к которым ответа нет.
> 3. *Сходящиеся* — интегралы, которые сходятся с ответом, и *расходящиеся*, которые не сходятся.
>
> *Математический фольклор. Цит. по книге: С. Н. Федин. Математики тоже шутят*

Как уже отмечалось ранее, каталожное описание документа при отсутствии классификационных признаков (т. е. кодов того или иного классификатора) имеет минимальную ценность, поскольку в таком случае процесс поиска документа человеком или его обработка рассуждающей информационной системой может опираться только на простую проверку вхождения тех или иных терминов в текст документа.

К сожалению, даже наиболее структурированные документы — журнальные статьи — далеко не всегда содержат классификационные признаки, к тому же классификатор источника может не совпадать с классификатором, используемым создателями интеллектуальной информационной системы. Изложенные в разд. 4.3 алгоритмы получения классификационных признаков от удаленных источников (библиографических баз) тоже не всегда приводят к желаемому результату, например из-за отсутствия в библиографической базе сведений о рассматриваемом документе.

В этой ситуации требуется провести автоматическую классификацию документа, исходя непосредственно из его содержания.

Один из простейших алгоритмов автоматической классификации документа основан на координатном индексировании и состоит в подсчете количества вхождений ключевых слов в текст классифицируемого документа, после чего документ классифицируется кодами классификатора, относящимися к наиболее часто встречающимся ключевым словам.

Однако эффективность такого алгоритма достаточно низка. Например, термины, наиболее часто встречающиеся в документе, могут относиться к *методу* исследования, в то время как о *предмете* исследования говорится сравнительно кратко в постановке задачи и результатах исследования. Вероятно, такой документ будет классифицирован кодами, относящимися к методу исследования, что вряд ли будет адекватно отражать его содержание. Использование же для координатного индексирования только важнейших метаданных документа: аннотации и ключевых слов (в узкобиблиографическом значении термина) — дает недостаточный материал для статистического анализа.

Решение возникшей проблемы возможно путем анализа содержания документа в контексте других документов, относящихся к той же предметной области, т. е. нахождения для данного документа сходных по содержанию документов, которые уже классифицированы. Этот процесс называется *категоризацией* документов.

Другая важная задача, решаемая создателями интеллектуальных систем обработки информации, — организация поиска документов «по аналогии», который нужен, например, для пополнения каталога информационной системы (либо библиографической базы отдельного исследователя или научного коллектива) документами, сходными в той или иной степени с теми, которые уже занесены в каталог (подробнее о постановке этой задачи см. разд. 1.8). Процесс разбиения документов на классы при отсутствии заранее заданного классификатора называется *кластеризацией* документов. При этом, разумеется, сформированные кластеры могут быть классифицированы апостериорно.

Для решения сформулированных задач требуется разработка способа задания меры сходства документов и выбор эффективного алгоритма кластеризации документов на основании заданной меры сходства. Приводимые ниже результаты соответствующих иссле-

дований изложены нами в [18]. Заметим, что общий обзор моделей и методов кластеризации текстовой информации содержится в статье В. О. Толчеева [165], а подробный анализ алгоритмов кластеризации документов был проведен в работах О. В. Песковой [125] и М. Е. Кондратьева [86], однако мера сходства в них задавалась простейшим образом: с использованием только одной шкалы, так или иначе зависевшей от вхождения терминов в текст документа. Этот подход вполне оправдывает себя при работе с полнотекстовыми документами, однако при кластеризации документов на основании текстов аннотаций содержащейся в них информации может оказаться недостаточно. При этом методы, описанные в работе [125], основаны на извлечении одиночных ключевых слов. Это объясняется как большим ростом вычислительных затрат, связанных с выделением составных терминов из полнотекстовых документов, так и спецификой указанной работы: необходимостью кластеризации библиотечных коллекций, имеющих политематический характер, что затрудняет возможность использования тезаурусов и других внешних словарей, содержащих составные термины. Тем не менее в случае коллекций документов достаточно узкой тематики кластеризация на основе использования составных ключевых терминов дает лучшие результаты даже при отказе от применения тезауруса (см. работу [20]).

В качестве шкал для определения меры сходства документов естественно рассматривать их метаданные (атрибуты библиографического описания; см., например, работу И. В. Некрасова и В. О. Толчеева [110]). Анализ показал, что целесообразно ограничиться следующими атрибутами:

— авторы;

— ключевые слова;

— аннотация.

Так как сравнение аннотаций в явном виде (т. е. как текстовых строк), очевидно, бессмысленно, то из текста аннотаций в соответствии с алгоритмом, изложенным в разд. 4.4, выделяются термины, входящие в словарь онтологии соответствующей предметной области, по которым и ведется измерение сходства.

Опишем алгоритм определения меры сходства двух документов. Количественная характеристика меры сходства определяется на множестве документов D следующим образом:

$$\mu : D \times D \to [0, 1],$$

причем функция μ в случае полного сходства принимает значение 1, в случае полного различия — 0. Вычисление меры сходства между документами d_1 и d_2 осуществляется по формуле вида

$$\mu(d_1, d_2) = \sum \alpha_i \mu_i(d_1, d_2), \qquad (4.1)$$

где i — номер элемента (атрибута) библиографического описания; α_i — весовые коэффициенты, причем $\sum \alpha_i = 1$ (см., например, [49]); $\mu_i(d_1, d_2)$ — мера сходства по i-му элементу (иными словами, по i-й шкале). Так как в описываемой ситуации все шкалы номинальные, то мера сходства по i-й шкале определяется следующим образом: если значения i-х атрибутов документов совпадают, то мера близости равна 1, иначе — 0. При этом необходимо учитывать, что значения атрибутов могут быть составными. В таком случае $\mu_i = n_{i1}/n_{i0}$, где $n_{i0} = \max n_{i0}(d_1), n_{i0}(d_2)$, $n_{i0}(d_j)$ — общее количество элементов, составляющих значение i-го атрибута документа d_j, n_{i1} — количество совпадающих элементов.

Заметим, что изложенный алгоритм измерения меры сходства может быть положен в основу некоторой экспертной системы, обладающей определенными продукционными правилами. Так, значения весовых коэффициентов α_i в формуле (4.1) могут определяться предполагаемой апостериорной достоверностью данных соответствующей шкалы. Например, полное (или даже «почти полное») совпадение значений атрибута «авторы» документа d_1 и документа d_2 более весомо в случае, когда количество значений этого атрибута в документе d_1 достаточно велико (по сравнению со случаем, когда документ d_1 имеет всего одного автора). В такой ситуации мы можем увеличивать значение соответствующего весового коэффициента в формуле (4.1) с одновременным пропорциональным уменьшением других коэффициентов.

Рассмотрим некоторые методы кластеризации документов. В процессе кластеризации происходит разбиение множества документов на классы, при котором элементы, объединяемые в один класс, имеют большее сходство, нежели элементы, принадлежащие разным классам [253] (в нашем случае сравнение документов при формировании кластеров ведется по атрибутам их библиографического описания). Каждый кластер при этом обычно описывается с помощью одного или нескольких идентификаторов, называемых профилем или центроидом. Профиль кластера может быть представлен некоторым формальным объектом, расположенным

в центре кластера, или любым представительным объектом, способным характеризовать остальные объекты этого кластера (подробнее о различных методах определения центроидов будет изложено ниже). С использованием понятия центроида можно искать похожие документы, сравнивая поисковые запросы сначала с профилями кластеров, а затем проверяя записи, входящие в кластеры, имеющие очень близкие профили.

Основная проблема кластеризации документов заключается в таком разнесении документов по группам, при котором элементы каждой группы были бы настолько сходны друг с другом, чтобы в некоторых случаях можно было пренебречь их индивидуальными особенностями. В частности, производить поиск в систематизированном файле гораздо легче, чем в несистематизированном, ибо группы документов, профили которых не имеют сходства с поисковым предписанием, не включаются в углубленный процесс поиска. При кластеризации документов важно прийти к разумному компромиссу относительно размера кластеров, избегая как формирования большого числа очень мелких кластеров (что снижает эффективность кластеризации как выделения множеств сходных документов), так и небольшого количества очень крупных классов (что может вызвать уменьшение точности поиска).

Принято различать две задачи кластеризации: формирование кластеров на основе сведений (свойств и характеристик) о классифицируемых объектах и отнесение объектов к сформированным кластерам (или кластерам, находящимся в процессе формирования.) Собственно формирование классов выполняется обычно на основе сопоставления векторов документов, причем класс определяется как множество всех объектов, имеющих достаточно высокие значения коэффициента подобия. Составление характеристик класса эквивалентно построению профиля; отнесение объектов к классам зависит от степени подобия между идентификаторами объектов и профилями классов.

В настоящей работе мы исследуем методы кластеризации, использующие в качестве критерия для сравнения только заранее заданные элементы библиографического описания документов, не учитывая индивидуальные поисковые возможности данных документов и мнения потребителей об их полезности.

В качестве потенциально пригодных для решения поставленной задачи проанализированы три классических метода кластери-

зации документов: кластеризация путем нахождения клик в полной матрице подобия документов [253], кластеризация по методу Роккио [253] и метод, базирующийся на так называемом жадном алгоритме [89], а также новый алгоритм, основанный на использовании функции конкурентного сходства (FRiS-функции) [42]. Кратко изложим суть перечисленных алгоритмов.

Процесс нахождения клик основан на построении полной матрицы подобия, посредством которой каждой паре документов (d_1, d_2) ставится в соответствие коэффициент подобия $S(d_1, d_2)$. Обычно выбирается пороговое значение O, и матрица подобия приводится к бинарному виду путем замены единицей всех коэффициентов подобия, таких, что $S(d_1, d_2) \geq O$, и нулем — всех остальных. Далее искомые классы определяются как клики, которые могут быть получены из бинарного ряда подобия.

В алгоритме Роккио построение матрицы подобия заменяется проверкой плотности пространства некоторых документов. В качестве возможных центров кластеров выступают только те документы, которые по результатам вычислений оказались расположенными в плотных зонах пространства. Кластеризуемый документ относят к тому классу, подобие с центроидом которого оказалось наиболее высоким.

При использовании жадного алгоритма в матрице подобия находят строку (или столбец — матрица симметрична), сумма компонентов которой будет максимальной. Документ, соответствующий этой строке, объявляют центром первого кластера и включают в кластер все документы, коэффициенты подобия к которым не меньше некоторого наперед заданного порогового значения. Далее выбрасывают все попавшие в кластер документы, вычеркивая из матрицы соответствующие строки и столбцы, после чего процесс повторятся несколько раз, пока все документы не будут кластеризованы.

В методе кластеризации с использованием функции конкурентного сходства при определении меры сходства между двумя документами рассматривается конкурентная ситуация: решение о принадлежности документа d к первому кластеру принимается не в том случае, когда расстояние r_1 до этого кластера «мало», а когда оно меньше расстояния r_2 до конкурирующего кластера. Для вычисления меры конкурентного сходства, измеренной в абсолютной шкале, вводится нормированная величина $F_{12} = (r_2 - r_1)/(r_2 + r_1)$,

называемая функцией конкурентного сходства или FRiS-функцией (от Function of Rival Similarity). Разумеется, на первоначальном этапе кластеризации, когда конкурирующих кластеров еще нет, приходится работать с некоторой модификацией (редукцией) FRiS-функции, использующей виртуальный кластер-конкурент. Суть FRiS-алгоритма состоит в том, что с помощью редуцированной FRiS-функции в качестве центроидов выбираются центры локальных «сгустков» распределения документов, после чего формируются линейно разделимые кластеры.

Теоретически преимущества FRiS-алгоритма заключаются в следующем:

— возможность получения классов произвольной формы;

— описание классов в терминах эталонных образцов (центроидов);

— автоматическое определение наилучшего числа классов.

Тестирование алгоритмов проводилось на электронной базе данных «Сибирского математического журнала», содержавшей библиографические описания статей журнала, вышедших в период с 2000 по 2005 г. (порядка 700 записей). Статьям в указанной базе данных приписаны, помимо стандартных атрибутов (название, автор, год издания и т.п.), еще и соответствующие коды классификатора MSC2000. Это обстоятельство позволило разбить всю работу на два этапа:

1. Нахождение оптимального алгоритма кластеризации. В качестве меры на пространстве документов используется определенная ранее конструкция, однако сравнение ведется только по одному атрибуту — кодам классификатора MSC2000 (обычно документу приписано 3 или более кодов). Поскольку совпадение данных кодов для группы документов является объективным критерием совпадения тематики данных документов, такую меру можно считать образцовой. Если коды классификатора центроида кластера содержались в числе кодов классификатора 2-го уровня данного документа, то мы полагали, что документ отнесен к кластеру правильно.

2. Задание меры на множестве документов, которая после кластеризации базы даст результат, близкий к результату с использованием меры, определенной на этапе 1.

Сравнение трех классических алгоритмов показало, что метод определения кластеров на множестве клик, полученных из матрицы подобия, оказался малопригодным для решения поставленной

задачи, так как имеет тенденцию к образованию большого количества очень мелких групп: по 1–2 документа. Это объясняется тем, что вероятность подобия друг другу всех без исключения элементов в более крупных группах объектов чрезвычайно низка.

Алгоритм Роккио продемонстрировал несколько лучшие результаты: поскольку в данном методе кластеризация происходит вокруг выборочных документов, стало возможным появление достаточно больших классов. Однако, так как алгоритм основан на вычислении плотности пространства документов, то его эффективность зависит от скорости формирования новых кластеров. В нашем случае, за исключением сравнительно небольшого числа документов (порядка 30 % от общего числа), бóльшая часть документов оказывается не включенной ни в один кластер.

Более качественный результат показал жадный алгоритм. Его использование привело к формированию кластерного массива, в котором каждый кластер содержит в среднем порядка 6–10 документов. Таким образом, по сравнению с методом клик и алгоритмом Роккио применение жадного алгоритма имеет ряд преимуществ:

— отсутствует проблема малого количества больших кластеров;

— отсутствует проблема большого количества мелких кластеров;

— нет документов, не попавших ни в один кластер;

— нет проблемы определения профилей документов, т. е. центров, вокруг которых формируются кластеры.

Далее было проведено сравнение FRiS-алгоритма с жадным алгоритмом. Выяснилось, что FRiS-алгоритм дает лучшую точность кластеризации. Ниже, на рис. 4.10, приведены результаты кластеризации базы данных «Сибирского математического журнала» при помощи жадного алгоритма (слева) и FRiS-алгоритма (справа). По горизонтальной оси отмечены номера кластеров, по вертикальной — количество документов в кластере. В качестве критерия принадлежности публикации к кластеру использовался его код классификатора MSC2000 (табл. 4.1). Если коды классификатора центроида кластера содержались в числе кодов классификатора данного документа, то мы полагали, что документ отнесен к кластеру правильно.

Как нетрудно заметить, величина «шума» (верхняя часть столбиков) в кластерах при кластеризации FRiS-алгоритмом сущест-

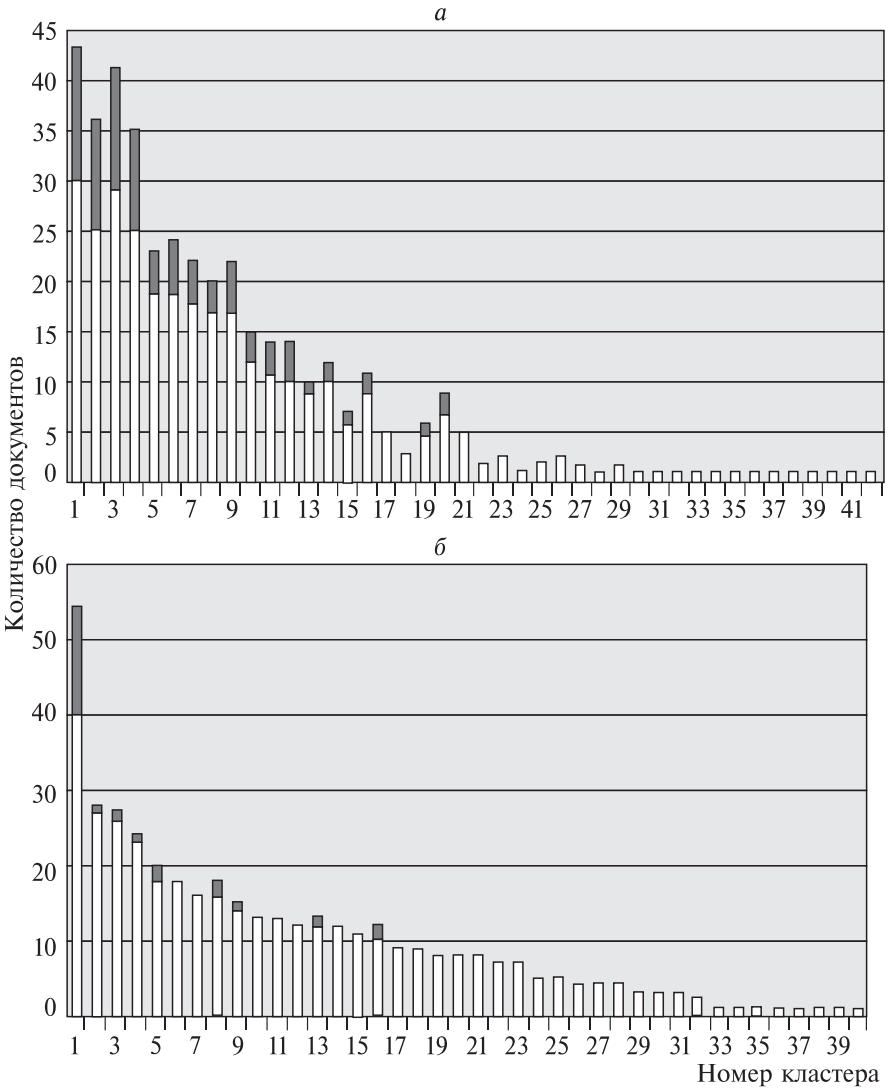

Рис. 4.10. Сравнение жадного (*а*) и FRiS (*б*) алгоритмов.

венно ниже, нежели в случае жадного алгоритма: погрешность классификации для жадного алгоритма составила 12 %, для FRiS-алгоритма 4 %, причем во втором случае разбиение на кластеры более равномерно, а доля одноэлементных кластеров существенно ниже.

Т а б л и ц а 4.1. **Соответствие кластеров разделам классификатора MSC2000**

Номер кластера	Код раздела	Название раздела
1	76Mxx	Основные методы в механике жидкости
2	76Bxx	Несжимаемая невязкая жидкость
3	76Nxx	Сжимаемые жидкости и газовая динамика
4	35Qxx	Дифференциальные уравнения с частными производными: области приложения
5	65Nxx	Численный анализ: дифференциальные уравнения с частными производными, краевые задачи
6	68Uxx	Информационные технологии и их приложения
7	76Dxx	Несжимаемая вязкая жидкость
8	65Fxx	Численный анализ: вычислительные методы линейной алгебры
9	35Kxx	Параболические уравнения и системы
10	74Sxx	Механика деформируемого твердого тела: численные методы
11	68Nxx	Информатика: программное обеспечение
12	65Cxx	Численный анализ: вероятностные методы, моделирование и стохастические дифференциальные уравнения
13	65Lxx	Численный анализ: обыкновенные дифференциальные уравнения
14	35Axx	Дифференциальные уравнения с частными производными: общая теория
15	68Pxx	Теория данных
16	76Rxx	Диффузия и конвекция
17	65Dxx	Численный анализ: численная аппроксимация и вычислительная геометрия
18	90Cxx	Математическое программирование
19	76Fxx	Турбулентность
20	68Mxx	Организация компьютерных систем
21	93Bxx	Теория систем и управление: управляемость, наблюдаемость, структура систем
22	35Cxx	Дифференциальные уравнения с частными производными: представления решений

Окончание табл. 4.1

Номер кластера	Код раздела	Название раздела
23	76Exx	Гидродинамическая устойчивость
24	65Kxx	Численный анализ: математическое программирование, оптимизация и вариационные методы
25	35Bxx	Дифференциальные уравнения с частными производными: качественные свойства решений
26	68Txx	Искусственный интеллект
27	78Axx	Оптика, теория электромагнетизма (общие вопросы)
28	76Axx	Механика жидкости: основы, определяющие уравнения, реология
29	65Rxx	Численный анализ: интегральные уравнения, интегральные преобразования
30	68Wxx	Информатика: алгоритмы
31	34Bxx	Обыкновенные дифференциальные уравнения: краевые задачи
32	49Jxx	Вариационное исчисление и оптимальное управление: теория существования
33	34H05	Обыкновенные дифференциальные уравнения: задачи управления
34	94Axx	Теория информации и коммуникации
35	45Gxx	Нелинейные интегральные уравнения
36	34Axx	Обыкновенные дифференциальные уравнения: общая теория
37	76Rxx	Диффузия и конвекция
38	80Axx	Термодинамика и теплоперенос
39	82Dxx	Статистическая механика, строение вещества: приложения к специальным типам физических систем
40	91Bxx	Теория игр: экономика, общественные науки

К сравнительным недостаткам FRiS-алгоритма следует отнести необходимость вручную задавать число кластеров в разбиении, а также несколько бóльшую вычислительную сложность $O(kN^2)$, где k — задаваемое пользователем число кластеров, по сравнению с $O(N^2)$ у жадного алгоритма. Однако при кластеризации большого

массива документов такое увеличение сложности становится не столь существенным, к тому же этот процесс требуется проводить только единожды. Таким образом, из рассмотренных алгоритмов решения задачи кластеризации документов в качестве оптимального признан FRiS-алгоритм.

При задании меры был принят во внимание тот факт, что значения весовых коэффициентов в формуле (4.1) определяются предполагаемой апостериорной достоверностью данных соответствующей шкалы, и в некоторых случаях один из коэффициентов может быть увеличен с пропорциональным уменьшением остальных.

Для определения весового коэффициента при каждом из атрибутов, была проведена кластеризация выборок из базы данных «Сибирского математического журнала». Рассмотрены выборки различной мощности, а в качестве критерия истинности применялся, как и ранее, результат кластеризации, полученный с мерой, основанной на кодах MSC2000.

Как показал эксперимент, наибольшее сходство с результатом кластеризации по мере, базирующейся на кодах классификатора, достигнуто путем введения следующих продукционных правил:

1) если каждый из документов d_1 и d_2 имеет более двух авторов и, как минимум, $2/3$ из них совпадают, то коэффициент при атрибуте «авторы» равен 1;

2) если каждый из документов d_1 и d_2 содержит более трех ключевых слов и, как минимум, $3/4$ этих слов совпадают, то коэффициент при атрибуте «ключевые слова» равен 1;

3) если каждый из документов d_1 и d_2 содержит более четырех ключевых терминов в аннотации и, как минимум, $3/5$ этих терминов совпадают, то коэффициент при атрибуте «аннотация» равен 1;

4) если условия ни одного из правил 1–3 не выполнены, то коэффициент при атрибуте «авторы» равен 0,2, а при атрибутах «ключевые слова» и «аннотация» равен 0,4.

Последний пункт получен в результате сравнительного анализа ряда наборов коэффициентов (рис. 4.11) при кластеризации выборки из 250 аннотаций «Сибирского математического журнала.

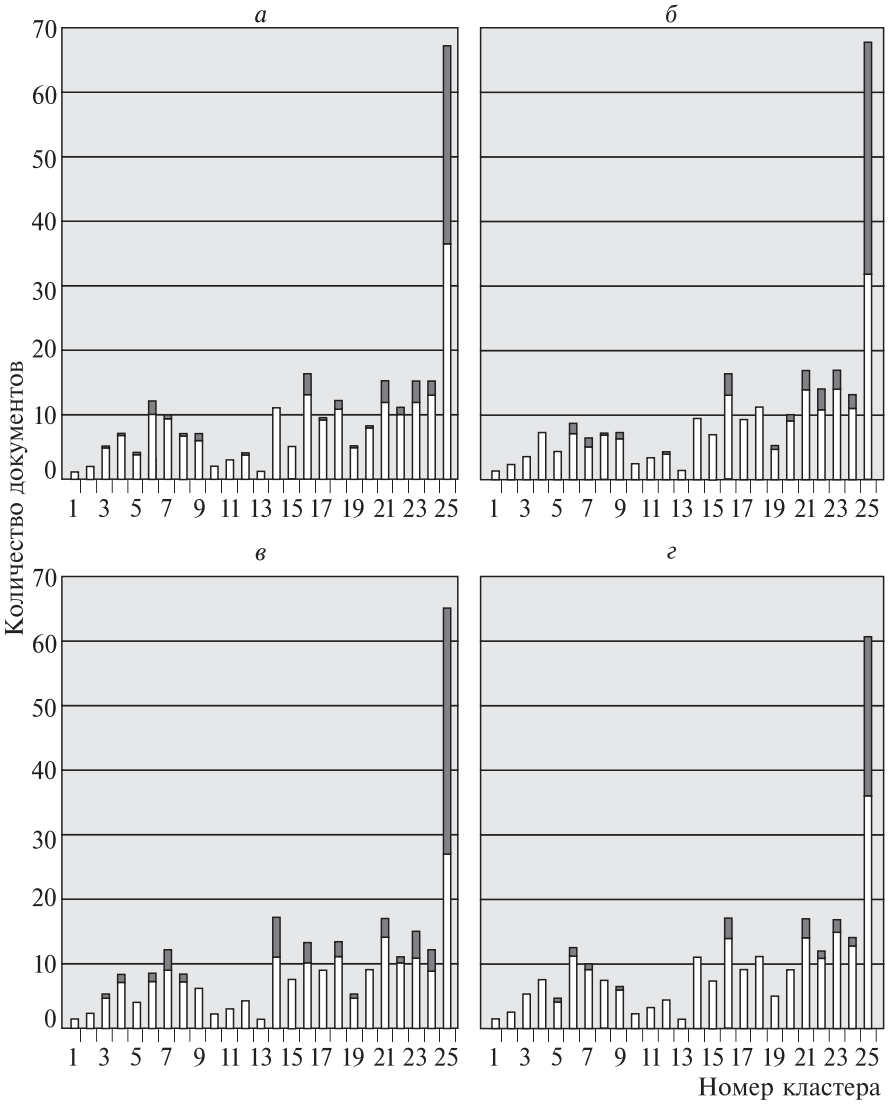

Рис. 4.11. Сравнение весовых коэффициентов.

а — «авторы» — 0,2, «ключевые слова» — 0,4, «аннотация» — 0,4;

б — «авторы» — 0,2, «ключевые слова» — 0,6, «аннотация» — 0,2;

в — «авторы» — 0,4, «ключевые слова» — 0,4, «аннотация» — 0,2;

г — контрольная кластеризация по мере, основанной на кодах MSC2000.

Погрешность кластеризации составляет в случае *а* — 18 %, *б* — 23 %, *в* — 26 %, *г* — 15 %, т. е. набор коэффициентов, соответствующий случаю *а*, дает результат, наиболее близкий к результату контрольной кластеризации.

Сравнительно высокая погрешность объясняется, во-первых, использованием не полных текстов, а аннотаций, и, во-вторых, наличием в выборке документов из близких разделов 2-го уровня MSC2000 (раздел 76Mxx «Основные методы в механике жидкости» частично поглотил разделы 76Bxx «Несжимаемая невязкая жидкость» и 76Nxx «Сжимаемые жидкости и газовая динамика»).

Интересно отметить, что полученные правила применимы как для FRiS-алгоритма, так и для жадного алгоритма.

Изложенный подход к кластеризации документов с использованием FRiS-алгоритма был реализован в виде веб-приложения.

По умолчанию предполагается, что при задании меры сходства на множестве документов будут использованы такие атрибуты библиографического описания, как «авторы», «ключевые слова» и «аннотация». Весовые коэффициенты при этих атрибутах и дополнительные продукционные правила, определяющие апостериорную достоверность данных соответствующей шкалы, выбраны согласно результатам, полученным выше. Однако пользователю предоставляется возможность в случае необходимости задать атрибуты, их весовые коэффициенты и продукционные правила вручную.

Заметим, что веб-приложение может быть, разумеется, использовано и для решения более простой задачи: отнесения новых документов к уже сформированным кластерам, соответствующим, например, разделам используемого классификатора.

* * *

В заключение приведем на рис. 4.12 функциональную схему программной системы, созданной на основе структур и моделей, описанных в гл. 3, с использованием алгоритмов, изложенных в гл. 4. Данная программная система играет важную роль в процессе функционирования как сайта Сибирского отделения РАН (http://www.sbras.ru), который, по данным рейтинга Webometrics [260], включающего сайты ведущих научно-исследовательских центров всего мира, в течение нескольких лет неизменно занимает первое место среди российских сайтов и входит в первую двадцат-

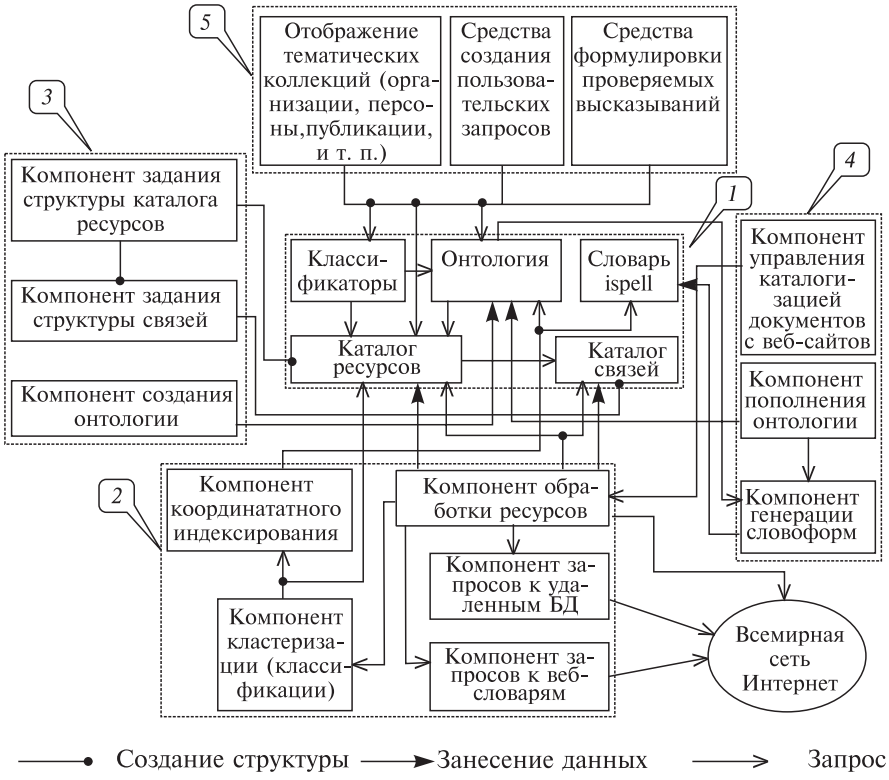

Рис. 4.12. Функциональная схема программной системы.

1 — хранилище данных; *2* — блок извлечения метаданных из интернет-документов; *3* — веб-интерфейс администрирования системы; *4* — веб-интерфейс администрирования данных; *5* — веб-интерфейс пользователя.

ку европейских и первую полусотню мировых сайтов, так и ряда связанных с этим сайтом информационных систем.

Компоненты программной системы реализованы на базе технологии LAMP (платформа Linix, веб-сервер Apache, сервер баз данных MySQL, язык программирования PHP).

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Андреев Ю.Н.* Научно-инновационные комплексы регионов как ресурс развития // Регионология. — 2002. — № 4. — С. 76–87.
2. *Апресян Ю.Д.* Избранные труды. Т. 1. Лексическая семантика. — М.: Школа «Языки русской культуры», Издательская фирма «Восточная литература» РАН, 1995.
3. *Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И.* Инфосфера: Информационные структуры, системы и процессы в науке и обществе. — М.: ВИНИТИ, 1996.
4. *Атре Ш.* Структурный подход к организации баз данных / пер. с англ. — М.: Финансы и статистика, 1983.
5. *Афанасьев К.Е., Шмакова Л.Е.* Компьютерная обработка информации. — Кемерово: Кузбассвузиздат, 2005.
6. *База* данных организаций и сотрудников СО РАН. — http:// www.sbras.ru/ sbras/db/.
7. *Бакалов В.П.* Теория функциональной сложности информационных систем. — Новосибирск: Наука, 2005.
8. *Барахнин В.Б.* Разработка тезауруса предметной области «Математика» // Вычисл. технологии, т. 8. Региональный вестн. Востока, № 3 (19), совм. вып. — 2003. — Ч. 1. — С. 111–115.
9. *Барахнин В.Б., Бычков И.В., Гуськов А.Е. и др.* Распределенный виртуальный музей Сибирского отделения РАН // Тр. Первой междунар. конф. «Системный анализ и информационные технологии». — Переславль-Залесский, 2005. — Т. 1. — С. 41–45.
10. *Барахнин В.Б., Ведерников В.В.* Автоматизированная каталогизация электронных журнальных публикаций // Тр. междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании». — Казахстан, Павлодар, 2006. — Т. 1. — С. 209–214.
11. *Барахнин В.Б., Ведерников В.В.* Алгоритм автоматической каталогизации статей, опубликованных в электронных версиях научных журналов // Тр. Всерос. научн. конф. «Научный сервис в сети Интернет: технологии параллельного программирования». — Новороссийск, 2006. — С. 277–279.
12. *Барахнин В.Б., Григорьева Я.И., Федотов А.М.* Использование тезауруса предметной области для построения информационно-справочных систем по истории науки // Материалы Всерос. конф. с междунар. участием «Знания — Онтологии — Теории» (ЗОНТ-07). — Новосибирск, 2007. — Т. 2. — С. 95–100.
13. *Барахнин В.Б., Куперштох А.А.* Алгоритм координатного индексирования электронных научных документов // Тр. междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании». — Казахстан, Павлодар, 2006. — Т. 1. — С. .228–232.

14. *Барахнин В.Б., Леонова Ю.В.* Информационная модель отношений между документами в информационной системе // Вычисл. технологии. — 2005. — Т. 10. — Спец. вып. — С. 129–137.
15. *Барахнин В.Б., Леонова Ю.В., Федотов А.М.* К вопросу о формулировке требований для построения информационных систем научно-организационной направленности // Вычисл. технологии. — 2006. — Т. 11. — Спец. вып. — С. 52–58.
16. *Барахнин В.Б., Маценко К.С.* Информационная модель системы поддержки инновационной деятельности // Тр. междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании». — Казахстан, Павлодар, 2006. — Т. 1. — С. 233–242.
17. *Барахнин В.Б., Нехаева В.А.* Технология создания тезауруса предметной области на основе предметного указателя энциклопедии // Вычисл. технологии. — 2007. — Т. 12. — Спец. вып. 2. — С. 3–9.
18. *Барахнин В.Б., Нехаева В.А., Федотов А.М.* О задании меры сходства для кластеризации текстовых документов // Вестн. НГУ. Сер. Информ. технологии. — 2008. — Т. 6, вып. 1. — С. 3–9.
19. *Барахнин В.Б., Рубцов Д.Н.* Сравнительные особенности используемых в Рунете информационных моделей описания деятельности крупных организаций и анализ их практической реализации на сайтах научной тематики // Изв. вузов. Проблемы полиграфии и издат. дела. — 2010. — № 4. — С. 97–107.
20. *Барахнин В.Б., Ткачев Д.А.* Кластеризация текстовых документов на основе составных ключевых термов // Вестн. НГУ. Сер. Информ. технологии. — 2010. — Т. 8, вып. 2. — С. 5–14.
21. *Барахнин В.Б., Федотов А.М.* Информационная система: взгляд на понятие // Вестн. НГУ. Сер. Информ. технологии. — 2007. — Т. 5, вып. 2. — С. 12–19.
22. *Барахнин В.Б., Федотов А.М.* Исследование информационных потребностей научного сообщества для построения информационной модели описания его деятельности // Вестн. НГУ. Сер. Информ. технологии. — 2008. — Т. 6, вып. 3. — С. 48–59.
23. *Барахнин В.Б., Федотов А.М.* Методика построения информационно-справочной системы по истории математической науки // Электронные библиотеки. — 2007. — Т. 10, вып. 1. — http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2007/ part1/BF.
24. *Барахнин В.Б., Федотов А.М.* Методологические подходы к построению информационно-справочных систем по истории науки // Тр. Девятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). — Переславль-Залесский, 2007. — С. 84–88.
25. *Барахнин В.Б., Федотов А.М.* О понятии «информационная система» в свете современных информационных технологий // Тр. VI Всерос.

науч.-практ. конф. «Инновационные недра Кузбасса. IT-технологии». — Кемерово, 2007. — С. 139–144.
26. *Барахнин В.Б., Федотов А.М.* Особенности информационно-поисковых систем общего назначения // Тр. Всерос. науч. конф. «Научный сервис в сети Интернет: многоядерный компьютерный мир». — Новороссийск, 2007. — С. 340–344.
27. *Барахнин В.Б., Федотов А.М.* Построение тезауруса для информационно-поисковой системы «Web-ресурсы математического содержания» // Инфокоммуникационные и вычислительные технологии и системы: материалы Всерос. конф. — Улан-Удэ: БурГУ, 2003. — С. 21–23.
28. *Барахнин В.Б., Федотов А.М.* Принципы структурирования сайтов информационной системы научного сообщества (на примере сайта Совета научной молодежи СО РАН) // Вычисл. технологии, т. 9. Вестн. КазНУ им. аль-Фараби, сер.: математика, механика, информатика, № 3 (42), совм. вып. — 2004. — Ч. 1. — С. 254–259.
29. *Барахнин В.Б., Федотов А.М.* Проблемы технологий создания систем смысловой обработки данных // Тр. Десятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008). — Дубна, 2008. — С. 39–44.
30. *Барахнин В.Б., Федотов А.М.* Ресурсы сети Интернет как объект научного исследования // Изв. вузов. Проблемы полиграфии и издат. дела. — 2008. — № 1. — С. 70–77.
31. *Барахнин В.Б., Федотов А.М.* Уточнение терминологии, используемой при описании интеллектуальных информационных систем, на основе семиотического подхода // Изв. вузов. Проблемы полиграфии и издат. дела. — 2008. — № 6. — С. 73–81.
32. *Барахнин В.Б., Федотов А.М., Шокин Ю.И.* Проблемы построения информационно-поисковых систем общего назначения // Тр. VI Всерос. науч.-практ. конф. «Системы автоматизации в образовании, науке и производстве». — Новокузнецк, 2007. — С. 35–39.
33. *Барсегян А.А., Куприянов М.В., Степаненко М.В., Холод И.И.* Методы и модели анализа данных: OLAP и Data Mining. — СПб.: БХВ-Петербург, 2004.
34. *Бахвалов Н.С.* Численные методы. — М.: Наука, 1970.
35. *Бездушный А.А., Бездушный А.Н., Серебряков В.А., Филиппов В.И.* Интеграция метаданных Единого научного информационного пространства РАН. — М.: ВЦ им. А.А. Дородницына РАН, 2006.
36. *Бездушный А.Н., Кулагин М.В., Серебряков В.А. и др.* Предложения по наборам метаданных для научных информационных ресурсов // Вычисл. технологии. — 2005. — Т. 10. — Спец. вып. — С. 29–48.
37. *Белоногов Г.Г., Кузнецов Б.А.* Языковые средства автоматизированных информационных систем. — М.: Наука, 1983.
38. *Белоногов Г.Г., Новоселов А.П.* Автоматизация процессов накопления, поиска и обобщения информации. — М.: Наука, 1979.

39. *Бирюков Б.В.* Кибернетика и методология науки. — М.: Наука, 1974.
40. *Бобров Л.К.* Организация стратегического управления информационной деятельностью библиотек и информационных центров в условиях рынка: автореф. дис. … д-ра техн. наук: 05.25.05. — Новосибирск, 2004.
41. *Богданов А.А.* Тектология: (Всеобщая организационная наука): в 2 кн. — М.: Экономика, 1989.
42. *Борисова И.А., Загоруйко Н.Г.* Функции конкурентного сходства в задаче таксономии // Материалы Всерос. конф. с междунар. участием «Знания — Онтологии — Теории» (ЗОНТ-07). — Новосибирск, 2007. — Т. 2. — С. 67–76.
43. *Босов А.В., Иванов А.В.* Программная инфраструктура информационного web-портала // Информатика и ее применения. — 2007. — Т. 1, вып. 2. — С. 50–64.
44. *Бусленко Н.П., Калашников В.В., Коваленко И.Н.* Лекции по теории сложных систем. — М.: Сов. радио, 1973.
45. *Бэкон Ф.* Новая Атлантида / пер. с англ. // Сочинения в 2 т. — М.: Мысль (сер. Философское наследие), 1978. — Т. 2. — С. 485–518.
46. *Валиев М.К, Китаев Е.Л., Слепенков М.И.* Использование службы директорий LDAP для представления метаинформации в глобальных вычислительных системах. — http://www.keldysh.ru/metacomputing/ism99.html.
47. *Визитная* карточка Сибирского отделения Российской академии наук. — http://www.sbras.ru/cmn/general.html.
48. *Витяев Е.Е., Ковалерчук Б.К., Федотов А.М., Барахнин В.Б. и др.* Обнаружение закономерностей и распознавание аномальных событий в потоке данных сетевого трафика // Вестн. НГУ. Сер. Информ. технологии. — 2008. — Т. 6, вып. 2. — С. 27–68.
49. *Воронин Ю.А.* Начала теории сходства. — Новосибирск: Наука, 1991.
50. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. — СПб.: Питер, 2000.
51. *Гвоздева Е.С., Высоцкий Е.М.* Сегодняшний день будущего российской науки. — Новосибирск: Изд-во СО РАН, 2004.
52. *Гергей Т., Финн В.К.* Об интеллектуальных системах // Экспертные системы: состояние и преспективы. — М.: Наука, 1989. — С. 9–29.
53. *Гиндин С.И.* Семантика текста и различные теории информации // НТИ. Сер. 2. — 1971. — № 10. — С. 10–15.
54. *Голдман С.* Теория информации / пер.с англ. — М.: ИЛ, 1957.
55. *Государственный* НИИ информационных технологий и телекоммуникаций «Информика». — http://www.informika.ru.
56. *Добров Б.В., Лукашевич Н.В.* Вторичное использование лингвистических онтологий: изменения в структуре концептуализации // Тр. Восьмой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2006). — Ярославль, 2006. — С. 56–64.

57. *Добров Б.В., Лукашевич Н.В.* Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе «РОССИЯ» // Тр. Третьей Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2001). — Петрозаводск, 2001. — С. 78–82.
58. *Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н.* Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Тр. Седьмой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2005). — Ярославль, 2005. — С. 70–79.
59. *Доброхотов А.Л.* Универсалии // Философский энциклопедический словарь. — М.: Сов. энцикл., 1983. — С. 702–703.
60. *Елисеев Ю.С., Малинецкий Г.Г., Медведев А.А., Харин А.А.* Инновационный императив // Вестн. нац. комитета «Интеллектуальные ресурсы России». — 2004. — № 2. — С. 61–70.
61. *Единое* научное информационное пространство РАН. — http://www.ras.ru/.
62. *Ермаков Н.Б., Столяров С.В., Федотов А.М.* Модели данных для формирования биологических коллекций // Вестн. НГУ. Сер. Информ. технологии. — 2007. — Т. 5, вып. 2. — С. 35–41.
63. *Ершов Ю.Л., Клименко О.А., Матвеева И.И. и др.* Древовидный каталог математических интернет-ресурсов // Информац. ресурсы России. — 2006. — № 1. — С. 5–8.
64. *Желены М.* Управление высокими технологиями // Информационные технологии в бизнесе: энциклопедия / пер. с англ. — СПб.: Питер, 2002. — С. 81–89.
65. *Жижимов О.Л., Мазов Н.А.* Принципы построения распределенных информационных систем на основе протокола Z39.50. — Новосибирск: Изд-во ИВТ СО РАН, 2004.
66. *Жижимов О.Л., Турпанов А.А., Федотов А.М.* Корпоративный каталог СО РАН // Тр. Восьмой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2006). — Ярославль, 2006. — С. 226–230.
67. *Жмайло С.В.* Исследование и разработка теории и методики построения тезаурусов для информационного поиска в полнотекстовых базах данных: автореф. дис. ... канд. техн. наук: 05.13.17. — М., 2005.
68. *Жукова Е.А., Мелик-Гайказян И.В.* Философские проблемы технологий и феномен Hi-Tech // Философия математики и технических наук. — М.: Академический Проект, 2006. — С. 557–586.
69. *Зацман И.М.* Концептуальный поиск и качество информации. — М.: Наука, 2003.
70. *Зацман И.М.* Семиотические основания и элементарные технологии информатики // Информ. технологии. — 2005. — № 7. — С. 18–31.

71. *Зверев В.С.* Информационное обеспечение инновационной деятельности. — http://sinin.nsc.ru/inf_sys.html.
72. *Зеленков Ю.Г., Сегалович Ю.В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Тр. Девятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). — Переславль-Залесский, 2007. — С. 166-174.
73. *Зыкин С.В.* Разработка и исследование моделей данных и средств организации взаимодействия пользователей с информационными ресурсами: автореф. дис. ... д-ра техн. наук: 05.13.17. — Омск, 2005.
74. *Иконников А.В.* Архитектура // Большая советская энциклопедия. 3-е изд. — М.: Сов. энцикл., 1970. — Т. 2. — С. 296–302.
75. *Интегрированная* система информационных ресурсов (архитектура, реализация, приложения) / отв. ред. В.А. Серебряков. — М.: ВЦ им. А.А. Дородницына РАН, 2004.
76. *Информационная* система «Химия в СО РАН». — http://www.catalysis.nsk.su/chem/.
77. *Информационная* система «Web-ресурсы математического содержания». — http://www.sbras.ru/win/elbib/data/ show_page.dhtml?2+184.
78. *Информационные* бюллетени Яндекса «Контент Рунета». — http://company.yandex.ru/facts/researches/.
79. *Каневский Е.А.* Некоторые вопросы пополнения морфологического словаря терминами предметной области // Тр. междунар. семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. — Аксаково, 2001. — Т. 2. — С. 156–160.
80. *Кант И.* Пролегомены / пер. с нем. — М.; Л.: Соцэкгиз, 1934.
81. *Когаловский М.Р.* Абстракции и модели в системах баз данных // СУБД. — 1998. — № 4–5. — С. 73–81.
82. *Когаловский М.Р.* Технология баз данных на персональных ЭВМ. — М.: Финансы и статистика, 1992.
83. *Колмогоров А.Н.* Теория информации и теория алгоритмов. — М.: Наука, 1987.
84. *Колмогоров А.Н.* Три подхода к определению понятия «количество информации» // Проблемы передачи информации. — 1965. — Т. I, вып. 1. — С. 3–11.
85. *Компьютерра+.* — 12 сент. 2005 г. — С. 6.
86. *Кондратьев М.Е.* Анализ методов кластеризации новостного потока // Тр. Восьмой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2006). — Ярославль, 2006. — С. 108–114.
87. *Контент* Рунета. — http://company.yandex.ru/facts/researches/ya_content_09.xml.
88. *Концепция* открытых систем // Материалы к межотраслевой программе «Развитие и применение открытых систем». — http://www.informika.ru/text/inftech/opensys/3/concept/os_1.html.

89. *Кормен Т., Лейзерсон Ч., Ривест Р.* Алгоритмы: построение и анализ / Пер. с англ. — М.: МЦНМО, 2001.
90. *Кузин Л.Т.* Основы кибернетики. — М.: Энергия, 1973. — Т. 1.
91. *Куценогий К.П., Куценогий П.К., Молородов Ю.И., Федотов А.М.* Разработка структуры метаданных по атмосферным аэрозолям на основе информационной модели // Вычисл. технологии. — 2004. — Т. 9. — Спец. вып. — Ч. 2. — С. 25–33.
92. *Лахути Д.Г.* Проблемы интеллектуализации информационно-поисковых систем: дис. ... д-ра техн. наук: 05.13.17. — Москва, 1999.
93. *Ленин В.И.* Материализм и эмпириокритицизм // Полное собрание сочинений. — 5-е изд. — М.: Политиздат, 1976. — Т. 18.
94. *Лукашевич Н.В.* Описание понятий-ролей в лингвистических и онтологических ресурсах // Сборник тезисов постерных докл. Девятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). — Переславль-Залесский, 2007. — С. 81–89.
95. *Ляпунов А.А.* О соотношении понятий материя, энергия и информация // А.А. Ляпунов Проблемы теоретической и прикладной кибернетики. — Новосибирск: Наука, 1980. — С. 320–323.
96. *Ляпунов А.А., Яблонский С.В.* Теоретические проблемы кибернетики // Проблемы кибернетики. — 1963. — Вып. 9. — С. 5–22.
97. *Ляпунцова Е.В.* Информационно-семиоитческие модели распределенных систем переработки информации: автореф. дис. ... д-ра техн. наук: 05.13.17. — М., 1996.
98. *Мальцева С.В.* Научно-методические основы автоматизации проектирования информационной архитектуры Web-ресурсов Интернет: автореф. дис. ... д-ра техн. наук: 05.13.12. — М., 2004.
99. *Математическая* энциклопедия: в 5 т. — М.: Сов. энцикл., 1977–1985.
100. *Математический* портал. — http://math.ru/history/people/.
101. *Мейер Д.* Теория реляционных баз данных / пер. с англ. — М.: Мир, 1987.
102. *Месарович М., Такахара Я.* Общая теория систем: математические основы / пер. с англ. — М.: Мир, 1978.
103. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Научные коммуникации и информатика. — М.: Наука, 1976.
104. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы информатики. — М.: Наука, 1968.
105. *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы научной информации. — М.: Наука, 1965.
106. *Нариньяни А.С.* Кентавр по имени ТЕОН: Тезаурус + Онтология // Тр. междунар. семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. — Аксаково, 2001. — Т. 1. — С. 184–188.
107. *Нариньяни А.С.* ТЕОН-2: от Тезауруса к Онтологии и обратно // Тр. междунар. семинара Диалог'2002 по компьютерной лингвистике и ее приложениям. — Протвино, 2002. — Т. 1. — С. 307–313.

108. *Народные* русские сказки / под ред. А.Н. Афанасьева: в 3 т. — М: Наука, 1985. — Том 2.
109. *Научная* электронная библиотека eLIBRARY.RU. — http://elibrary.ru/defaultx.asp.
110. *Некрасов И.В., Толчеев В.О.* Построение модели представления библиографического документа // Информ. технологии. — 2005. № 11. — С. 57–63.
111. *Никитина С.Е.* Семантический анализ языка науки. — М.: Наука, 1987.
112. *Новик И.Б., Уёмов А.И.* Моделирование и аналогия // Материалистическая диалектика и методы естественных наук. — М., 1968. — С. 268–293.
113. *Новоселов М.М.* Сходство // Философский энциклопедический словарь. — М.: Сов. энцикл., 1983. — С. 666.
114. *Овдей О.М., Проскудина Г.Ю.* Обзор инструментов инженерии онтологий // Тр. Шестой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2004). — Пущино, 2004. — С. 59–68.
115. *Огурцов А.П., Юдин Э.Г.* Деятельность // Философский энциклопедический словарь. — М.: Сов. энцикл., 1983. — С. 151–152.
116. *Осипов Г.С.* Лекции по искусственному интеллекту. — М.: КРАСАНД, 2009.
117. *Отле П.* Библиотека, библиография, документация: Избранные труды пионера информатики / пер. с англ. и фр. — М.: ФАИР-ПРЕСС, Пашков дом, 2004.
118. *Официальный* сайт Государственной Думы. — http://www.duma.gov.ru.
119. *Официальный* сайт Русской Православной Церкви. — http://www.patriarchia.ru/.
120. *Официальный* сайт Союза писателей России. — http://sp.voskres.ru/prose/.
121. *Панова Н.С., Шрейдер Ю.А.* Принцип двойственности в теории классификации // НТИ. Сер. 2. — 1975. — № 10. — С. 3–10.
122. *Паркер-Роудс А.Ф., Уордли С.* Применение тезаурусного метода при машинном переводе с помощью существующей машинной техники / пер. с англ. // Математическая лингвистика: Сб. переводов. — М.: Мир, 1964. — С. 214–228.
123. *Пахомов Б.Я.* Проблема изменения значений научных понятий // Вопр. философии. — 1973. — № 1. — С. 140–144.
124. *Перечень* зарегистрированных политических партий. — http://www.cikrf.ru/politparty/.
125. *Пескова О.В.* Автоматическое формирование рубрикатора полнотекстовых документов // Тр. Десятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные колекции» (RCDL'2008). — Дубна, 2008. — С. 139–148.

126. *Петров В.М.* Семантика научных терминов. — Новосибирск: Наука, 1982.
127. *Пойа Д.* Как решать задачу / пер. с англ. — М.: Учпедгиз, 1959.
128. *Пойа Д.* Математика и правдоподобные рассуждения / пер. с англ. — М.: Наука, 1975.
129. *Пойа Д.* Математическое открытие / пер. с англ. — М.: Наука, 1970.
130. *Порус В.Н.* Аналогия // Новая философская энциклопедия. — М.: Мысль, 2000. — Т. 1. — С. 103–105.
131. *Постановление* Президиума СО РАН от 13.04.2000 № 137 «Об итогах конкурса интеграционных программ (проектов) СО РАН — 2000 г.» — http://www.sbras.ru/win/anonses/373.html.
132. *Постановление* Президиума СО РАН от 21.02.2003 № 62 «Об итогах конкурса интеграционных проектов СО РАН — 2003 г.» — http://www.sbras.ru/win/anonses/841.html.
133. *Постановление* Президиума СО РАН от 26.01.2006 № 32 «Об интеграционных проектах, выполняемых по заказу Президиума СО РАН». — http://www.sbras.ru/win/anonses/1334.html.
134. *Постановление* Президиума СО РАН от 09.02.2006 № 54 «Об итогах конкурса комплексных интеграционных проектов СО РАН–2006». — http://www.sbras.ru/win/anonses/1341.html.
135. *Постановление* Президиума СО РАН от 09.02.2006 № 55 «Об итогах конкурса междисциплинарных интеграционных проектов СО РАН–2006». — http://www.sbras.ru/win/anonses/1342.html.
136. *Постановление* Президиума СО РАН от 15.01.2009 № 9 «Об итогах конкурса междисциплинарных интеграционных проектов фундаментальных исследований СО РАН на 2009–2011 гг.» — http://www.sbras.ru/win/anonses/1921.html
137. *Постановление* Президиума СО РАН от 15.01.2009 № 10 «Об итогах конкурса проектов, выполняемых совместно со сторонними научными организациями, на 2009–2011 годы». — http://www.sbras.ru/win/anonses/1922.html
138. *Раскина А.А., Солодовник М.П.* Логико-лингвистические аспекты проблемы обработки вопросов в фактографической ИПС // Вопр. информ. теории и практики. — 1979. — № 42.
139. *Резниченко В.А., Проскудина Г.Ю., Овдей О.М.* Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone // Электронные библиотеки. — 2005. — Т. 8, вып. 6. — http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2005/part6/RPO.
140. *Российская* сеть трансфера технологий. — http://www.rttn.ru/.
141. *Рубашкин В.Ш.* Представление и анализ смысла в интеллектуальных информационных системах. — М.: Наука, 1989.
142. *Рубцов Д.Н., Барахнин В.Б.* Выявление дубликатов в разнородных библиографических источниках // Вестн. НГУ. Сер. Информ. технологии. — 2009. — Т. 7, вып. 3. — С. 86–93.

143. *Садовский В.Н.* Основания общей теории систем. — М.: Наука, 1974.
144. *Садовский В.Н.* Система // Философский энциклопедический словарь. — М.: Сов. энцикл., 1983. — С. 610–611.
145. *Сайт* Большого театра. — http://www.bolshoi.ru.
146. *Сайт* Екатеринбургского государственного академического театра оперы и балета. — http://www.uralopera.ru/.
147. *Сайт* Мариинского театра. — http://www.mariinsky.ru/.
148. *Сайт* «Научные сотрудники — математики СО РАН». — http://www.sbras.ru/sbras/math_soran/.
149. *Сайт* Новосибирского государственного академического театра оперы и балета. — http://www.opera-novosibirsk.ru/.
150. *Сайт* «Организации СО РАН». — http://www.sbras.ru/sbras/db/dep.phtml?3++rus.
151. *Сайт* «Перечень важнейших разработок СО РАН, предлагаемых для широкого использования». — http://www.sbras.ru/win/sbras/main-work.html.
152. *Сайт* «Члены Российской академии наук». — http://www.ras.ru/members.aspx.
153. *Самарский А.А.* Задачи прикладной математики на современном этапе развития // Коммунист. — 1983. — № 18. — С. 31–42.
154. *Сеть* передачи данных Сибирского отделения РАН. — http://www.ac-tel.ru/mw/index.php/Введение.
155. *Словарь* «Лингво» компании «Яндекс». — http://lingvo.yandex.ru.
156. *Словарь* по кибернетике. 2-е изд., перераб. и доп. — Киев: Гл. ред. Укр. сов. энцикл. им. М.П. Бажана, 1989.
157. *Словарь* русского языка для Ispell. — http://semiconductors.phys.msu.su/~swan/orthography.html.
158. *Соционет.* — http://socionet.ru/.
159. *Спиркин А.Г.* Знание // В кн.: Философский энциклопедический словарь. — М.: Сов. энцикл., 1983. — С. 192.
160. *Список* крупнейших компаний России журнала «Эксперт». — http://www.raexpert.ru/rankingtable/?table_folder=/expert400/2009/main/.
161. *Стимер* компании «Яндекс». — http://company.yandex.ru/technology/mystem/.
162. *Технология* // Большой академический словарь. — СПб:Бол. рос. энцикл., 2003. — С. 2000.
163. *Технология* // Тезаурус по образованию и педагогике / Институт информатизации образования в составе Московского гос. гуманитарного ун-та им. М.А. Шолохова. — http://www.mgopu.ru/ininfo/r1_thesaurus.htm#technology.
164. *Толковый* словарь русского языка: в 4 т./ под ред. Д.Н. Ушакова. — М.: Сов. энцикл.; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1935–1940.

165. *Толчеев В.О.* Модели и методы классификации текстовой информации // Информ. технологии. — 2004. — № 5. — С. 6–14.
166. *Тюхтин В.С.* Сходство // Большая советская энциклопедия. 3-е изд. — М.: Сов. энцикл., 1976. — Т. 25. — С. 123.
167. *Ульман Дж.* Основы систем баз данных / пер. с англ. — М.: Финансы и статистика, 1983.
168. *Университетская* информационная система РОССИЯ. — http://www.cir.ru/index.jsp.
169. *Устав* Российской академии наук. Утвержден Постановлением Правительства РФ от 19 ноября 2007 г. № 785. — http://www.ras.ru/about/rascharter.aspx.
170. *Федотов А.М.* Методологии построения распределенных систем // Вычисл. технологии. — 2006. — Т. 11. — Спец. вып. — С. 3–16.
171. *Федотов А.М.* Парадоксы информационных технологий // Вестн. НГУ. Сер. Информ. технологии. — 2008. — Т. 6, вып. 2. — С. 3–14.
172. *Федотов А.М., Артемов И.А., Ермаков Н.Б. и др.* Электронный атлас «Биоразнообразие растительного мира Сибири» // Вычисл. технологии. — 1998. — Т. 3, № 5. — С. 68–78.
173. *Федотов А.М., Барахнин В.Б.* К вопросу о поиске документов «по аналогии» // Вестн. НГУ. Сер. Информ. технологии. — 2009. — Т. 7, вып. 4. — С. 3–14.
174. *Федотов А.М., Барахнин В.Б.* Проблемы поиска информации: история и технологии // Вестн. НГУ. Сер. Информ. технологии. — 2009. — Т. 7, вып. 2. — С. 3–17.
175. *Федотов А.М., Барахнин В.Б., Бычков И.В., Жижимов О.Л. и др.* Концепция создания виртуального музея СО РАН // VIII Междунар. конф. по электронным публикациям «EL-Pub2003». — Новосибирск, 2003. — Электронная публикация, № гос. регистрации 3521. — http://www-sbras.nsc.ru/ws/elpub2003/6155/rep6155.pdf
176. *Федотов А.М., Барахнин В.Б., Гуськов А.Е., Жижимов О.Л. и др.* Информационно-справочная система СО РАН // Вычисл. технологии. — 2006. — Т. 11. — Спец. вып. — С. 88–94.
177. *Федотов А.М., Барахнин В.Б., Гуськов А.Е., Леонова Ю.В.* Построение информационной системы научного сообщества на основе интеграции разнородных коллекций ресурсов // Сб. тез. постерных докл. Девятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). — Переславль-Залесский, 2007. — С. 111–117.
178. *Федотов А.М., Барахнин В.Б., Гуськов А.Е., Молородов Ю.И.* Распределенная информационно-аналитическая среда для исследований экологических систем // Вычисл. технологии. — 2006. — Т. 11. — Спец. вып. — С. 113–125.
179. *Фейгин Д.* Концепция SOA / пер. с англ. // Открытые системы. — 2004. — № 6. — http://www.osp.ru/os/2004/06/184447/_p1.html.

180. *Физическая* энциклопедия: в 5 т. — М.: Рос. энцикл., 1998.
181. *Химическая* энциклопедия; в 5 т. — М.: Рос. энцикл., 1998.
182. *Холл А.Д., Фейджин Р.Е.* Определение понятия системы / пер. с англ. // Исследования по общей теории систем. — М.: Прогресс, 1969. — С. 252–282.
183. *Хохлов Ю.Е., Арнаутов С.А.* Обзор форматов метаданных // Российские электронные библиотеки. — http://www.elbib.ru/index.phtml?page=elbib/rus/methodology/md_rev.
184. *Цапенко М.П.* Измерительные информационные системы. — М.: Энергоиздат, 1985.
185. *Цикритзис Д., Лоховски Ф.* Модели данных / пер. с англ. — М.: Финансы и статистика, 1985.
186. *Черняк Л.* От информационно-поисковых систем к корпоративному поиску // Открытые системы. — 2005. — № 11. — http://www.osp.ru/os/2005/11/380532/.
187. *Шокин Ю.И., Барахнин В.Б., Гриншяков Б.Ю.* Методология создания системы информационной поддержки научно-инновационной деятельности региона // Второй форум возрождения китайской северо-восточной старой промышленной базы: научно-техническое сотрудничество Китая и СНГ: сб. докл. — Китай, Харбин, 2006. — С. 179–183 на кит. яз., с. 184–190 на рус. яз.
188. *Шокин Ю.И., Барахнин В.Б., Гуськов А.Е., Клименко О.А. и др.* Единая информационная среда научной организации на примере ИВТ СО РАН // Тр. VII Всерос. науч.-практ. конф. «Инновационные недра Кузбасса. IT-технологии». — Кемерово, 2008. — С. 271–276.
189. *Шокин Ю.И., Белов С.Д., Чубаров Л.Б.* Предварительные результаты тестирования создаваемой системы мониторинга и сбора статистики СПД СО РАН // Вычисл. технологии. — 2007. — Т. 12, № 5. — С. 126–134.
190. *Шокин Ю.И., Ламин В.А., Федотов А.М., Барахнин В.Б. и др.* Распределенная информационная система «Виртуальный музей Науки и техники СО РАН» // Тр. Пятой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2003). — СПб.: НИИ химии СПбГУ, 2003. — С. 112–126.
191. *Шокин Ю.И., Федотов А.М.* Интеграция информационно-телекоммуникационных ресурсов Сибирского отделения РАН // Вычисл. технологии. — 2003. — Т. 8. — Спец. вып. — С. 161–171.
192. *Шокин Ю.И., Федотов А.М.* Информационные ресурсы Сибирского отделения РАН // Информ. ресурсы России. — 1999. — Т. 9, № 4. — С. 12–16.
193. *Шокин Ю.И., Федотов А.М.* Информационные технологии Internet // Вычисл. технологии. — 1997. — Т. 2, № 3. — С. 80–87.
194. *Шокин Ю.И., Федотов А.М.* Развитие распределенных информационно-вычислительных ресурсов в СО РАН // Вычисл. технологии. — 2004. — Т. 9. — Спец. вып. — С. 10–23.

195. *Шокин Ю.И., Федотов А.М.* Распределенные информационные системы // Вычисл. технологии. — 1998. — Т. 3, № 5. — С. 79–93.
196. *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Особенности организации системы управления веб-контентом сайтов информационной поддержки инновационной деятельности // Вычисл. технологии. — 2005. — Т. 10. — Спец. вып. — С. 122–128.
197. *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Технология создания программных систем информационного обеспечения научной деятельности, работающих со слабоструктурированными документами // Вычисл. технологии. — 2010. — Т. 15, № 6. — С. 111–125.
198. *Шокин Ю.И., Федотов А.М., Леонова Ю.В.* Объектная модель документа в электронных коллекциях // VII Междунар. конф. по электронным публикациям «EL-Pub2002». — Электронное изд., № гос. регистрации 0320300063. — http://www.ict.nsc.ru/ws/elpub2002/4488/.
199. *Шрейдер Ю.А.* Информация и метаинформация // НТИ. Сер. 2. — 1974. — № 4. — С. 3–10.
200. *Шрейдер Ю.А.* К определению системы // НТИ. Сер. 2. — 1971. — № 7. — С. 3–8.
201. *Шрейдер Ю.А.* О количественных характеристиках семантической информации // НТИ. Сер. 2. — 1963. — № 10. — С. 35–39.
202. *Шрейдер Ю.А.* О семантических аспектах теории информации // Информация и кибернетика. — М.: Сов. радио, 1967. — С. 15–47.
203. *Шрейдер Ю.А.* Об одной модели семантической информации // Проблемы кибернетики. — М.: Наука, 1965. — Вып. 13. — С. 233–240.
204. *Шрейдер Ю.А.* Равенство, сходство, порядок. — М.: Наука, 1971.
205. *Шрейдер Ю.А., Шаров А.А.* Системы и модели. — М.: Радио и связь, 1982.
206. *Электронная* библиотека MathTree. — http://www.mathtree.ru.
207. *Электронный* атлас биоразнообразия животного и растительного мира Сибири. — http://www.sbras.ru/win/elbib/bio/.
208. *Ядов В.А.* Потребности // Большая советская энциклопедия. 3-е изд. — М.: Сов. энцикл., 1975. — Т. 20. — С. 439–400.
209. *Яненко Н.Н.* Методологические вопросы современной математики // Вопр. философии. — 1981. — № 8. — С. 60–68.
210. *Ackoff R., Emery F.* On purposeful systems. — Ch.; N.Y.: Aldine—Atherton, 1972. / рус. пер. Р. Акофф, Ф. Эмери. О целеустремленных системах — М.: Сов. радио, 1974.
211. *Alexander C. et al.* A pattern language towns, buildings, constructions. — N.Y.: Oxford University Press, 1977.
212. *Barakhnin V., Klimenko O.* Systematization and the search of mathematical web-resources // Proc. Second IASTED Intern. Multi-Conf. on Automation, Control, and Information Technology. Software Engineering. — Novosibirsk: ACTA Press, 2005. — P. 81–84.

213. *Beck K., Cunningham W.* Using pattern languages for object-oriented programs // OOPSLA−87 Workshop on the Specification and Design for Object-Oriented Programming. — http://c2.com/doc/oopsla87.html.
214. *Berners-Lee T., Fielding R., Masinter L.* Uniform Resource Identifiers (URI). Generic Syntax // RFC 2396. — 1999. — http://www.ietf.org/rfc/rfc2396.txt.
215. *Bernier C.L.* Correlative indexes II: Correlative trope indexes // American Documentation. — 1957. — Vol. 8, № 1. — P. 47−50.
216. *Bertalanffy L. von.* Problems of general system theory // Human Biology. — 1951. — № 23. — P. 302−312.
217. *Bertalanffy L. von.* Conclusion // Human Biology. — 1951. — № 23. — P. 336−345.
218. *Brillouin L.* Science and information theory. — N.Y.: Academic Press, 1956. / рус. пер. Л. Бриллюэн. Наука и теория информации. — М.: Физматгиз, 1960.
219. *Bush V.* As we may think // The Atlantic Monthly, July, 1945. — http://www.theatlantic.com/doc/194507/bush.
220. *The CERIF* (Common European Research Information Format) Standard. — http://www.eurocris.org/en/taskgroups/cerif/new_6/new_0/C%3A%5CDocuments+and+Settings%5Ceg53%5CDesktop%5CCERIF_2000_part2.pdf.
221. *Chen P.P.* The entity-relational model. Toward a unified view of data // ACM TODS. — 1976. — № 1. — P. 9−36. / рус. пер. Чен П.П. Модель «сущность—связь» — шаг к единому представлению данных // СУБД. — 1995. — № 3. — С. 137−158.
222. *Codd E.F.* A relational model of data for large shared data banks // Comm. ACM. — 1970. — V. 13. — № 6. — P. 377−387. / рус. пер. Е.Ф.Кодд. Реляционная модель данных для больших совместно используемых банков данных // СУБД. — 1995. — № 1. — С. 145−160.
223. *Community* Research and Development Information Service. — http://cordis.europa.eu/.
224. *The COSINE* and Internet X.500 Schema // RFC 1274. — http://www.networksorcery.com/enp/rfc/rfc1274.txt.
225. *Crescenzi V., Mecca G., Merialdo P.* Roadrunner: Towards automatic data extraction from large web sites // The VLDB Journal. — Rome, 2001. — P. 109−118.
226. *Definition* of the inetOrgPerson LDAP Object Class // RFC 2798. — http://www.faqs.org/rfcs/rfc2798.html.
227. *Dublin* Core Metadata Initiative. — http://dublincore.org/.
228. *European* Research Gateways Online. — http://www.cordis.europa.eu/ergo.
229. *Gitt W.* Ordnung und Information in Technik und Natur // Gitt W. (Hrsg.): Am Anfang war die Information. Gräfeling: Resch KG, 1982. — S. 171–211.

230. *Global* Information Locator Service (GILS). — http://www.gils.net/.
231. *Gruber T.* A translation approach to portable ontology specifications // Knowledge Acquisition Journal. — 1993. — Vol. 5, № 2. — P. 199–220.
232. *ISO/IEC* 11179, Specification and Standardization of Data Elements. — ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/11179.
233. *Ispell* — Spell checker. — http://directory.fsf.org/ispell.html.
234. *Krogstie J., Halpin T., Siau K.* Information modeling methods and methodologies. — Idea group publishing, 2005.
235. *Langefors B.* Infologocal models and information user views // Information Systems. — 1980. — № 5. — P. 17–32.
236. *Langefors B.* Managment information system design // IAG Quart. — 1969. — Vol. 2, № 4. — P. 5–17.
237. *Langefors B.* Some approaches to the theory of information systems // BIT. — 1963. — № 3. — P. 229–254.
238. *Leonova Yu.V., Barakhnin V.B., Fedotov A.M.* On the problem of modeling of the horizontal relations between documents // Вычисл. технологии. — 2007. — Т. 12, № 1. — С. 3–12.
239. *Library* of Congress. — http://www.loc.gov/.
240. *Lightweight* Directory Access Protocol (v3) // RFC 2251. — http://www.faqs.org/rfcs/rfc2251.html.
241. *The MacTutor* History of Mathematics archive. — http://www-history.mcs.st-and.ac.uk.
242. *The Mathematics* Genealogy Project. — http://www.genealogy.ams.org.
243. *Mathematics* Subject Classification. — http://www.ams.org/msc/.
244. *Mayer T.* Our blog is growing up — and so has our index. — http://www.ysearchblog.com/archives/000172.html.
245. *Miller G.F.* The magical number seven, plus or minus two // The Psychol. Rev. — 1956. — Vol. 63. — P. 81–97. / рус. пер. Дж.Миллер. Магическое число семь, плюс или минус два // Инженерная психология. — М.: Прогресс, 1964. — С. 192–225.
246. *Otlet P.* Traite de documentation. — Bruxelles: Ed. Mundaneum, 1934.
247. *Price D.J. de Solla.* Little science, big science. — N.Y., L.: Columbia Univ. Press, 1963. / рус. пер. *Д.Прайс.* Малая наука, большая наука // Наука о науке. — М.: Прогресс, 1966. — С. 281–385.
248. *Resource* Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. — http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.
249. *Resource* Resource Description Framework (RDF) Schema Specification 1.0. W3C Candidate Recommendation 27 March 2000 — http://www.w3.org/TR/2000/CR-rdf-schema-20000327/.
250. *Roget P.M.* Thesaurus of English Words and Phrases classified and arranged so as to facilitate the expression of ideas ahd to assist in literary composition. London, 1852.

251. *Sahuguet A., Azavant F.* Building intelligent web applications using lightweight wrappers // Data Knowledge Engineering. — 2001. — Vol. 36, № 3. — P. 283–316.
252. *Salton G.* Automatic information organization and retrieval. — N.Y.: McGraw-Hill Book Co., 1968. / Рус. пер. *Сэлтон Г.* Автоматическая обработка, хранение и поиск информации. — М.: Сов. радио, 1973.
253. *Salton G.* Dynamic information and library processing. — N.J.: Prentice Hall, 1975. / Рус. пер. *Солтон Дж.* Динамические библиотечно-информационные системы. — М.: Мир, 1979.
254. *Schramm W.* Information theory and mass communication // Communication and Culture. — N.Y.: Holt, Rinehart & Winston, 1966. — P. 521–534.
255. *Segalovich I.* A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search egine // Proc. of the Intern. Conf. on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23-26, 2003, Las Vegas, Nevada, USA. — CSREA Press, 2003. — S. 273–280.
256. *Semantic* Web. — http://www.w3.org/2001/sw/.
257. *Shokin Yu.I., Leonova Yu.V., Barakhnin V.B., Fedotov A.M.* Concerning the problem of work up the model of horizontal relations between the documents in the information systems of scientific community //Proc. 3rd Intern. Conf. on Cybernetics and Information Technologies, Systems and Applications (CITSA 2006). — Orlando, USA, 2006. — Vol. 3. — P. 112–116.
258. *Staab S., Stuckenschmidt H. (Eds.).* Semantic web and peer-to-peer, decentralized management and exchange of knowledge and information. — Springer, 2006.
259. *Task* Force on metadata. Summary report // American Library Association. — 1999. — T. June.
260. *Top* 300 R&D European Institutes. — http://research.webometrics.info/top300_r&d_europe.asp.
261. *Universal* Decimal Classification. — http://www.udcc.org/.
262. *vCard:* The Electronic Business Card. — http://www.imc.org/pdi/.
263. *Web* of Science. — http://wokinfo.com/products_tools/ multidisciplinary/webofscience/.
264. *Webster's* new world dictionary of computer terms. 4th ed. — N.Y.: Prentice Hall, 1992.
265. *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert systems all over again // AAAI–1999 Invited Panel Presentation. — 1999.
266. *Zentralblatt* MATH. — http://www.zentralblatt-math.org/zmath/en/.
267. *Zthes:* a Z39.50 Profile for Thesaurus Navigation. — http:// lcweb.loc.gov/z3950/agency/profiles/zthes-04.html.

Научное издание

Шокин Юрий Иванович
Федотов Анатолий Михайлович
Барахнин Владимир Борисович

ПРОБЛЕМЫ ПОИСКА ИНФОРМАЦИИ

Редактор *М.б. Успенская*
Художественный редактор *Л.В. Матвеева*
Художник *Н.А. Горбунова*
Технический редактор *Н.М. Остроумова*
Корректоры *И.Л. Малышева, Л.А. Анкушева*
Оператор электронной верстки *Р.Г. Усова*

Сдано в набор 10.121.0. Подписано в печать 28.01.11. Бумага ВХИ. Формат 60 □ 90 1/16. Офсетная печать. Гарнитура Journal. Усл. печ. л. 00,0. Уч.-изд. л. 00,0. Тираж 000 экз. Заказ № 000.

Сибирская издательская фирма «Наука» РАН. 630007, Новосибирск, ул. Коммунистическая, 1.
СП «Наука» РАН. 630077, Новосибирск, ул. Станиславского, 25.