Технология создания программных систем информационного обеспечения научной деятельности, работающих со слабоструктурированными документами*

Ю. И. Шокин, А. М. Федотов, В. Б. Барахнин
Институт вычислительных технологий СО РАН, Новосибирск, Россия
Новосибирский государственный университет, Россия
e-mail: shokin@ict.nsc.ru, fedotov@sbras.ru, bar@ict.nsc.ru

Обсуждаются перспективы развития процесса смысловой обработки данных как технологии, при этом в качестве источника данных рассматриваются электронные документы достаточно произвольной структуры.

Ключевые слова: интеллектуальные информационные системы, обработка слабоструктурированных документов.

1. Современные проблемы создания и функционирования информационно-поисковых систем

Развитие высоких технологий в области передачи и обработки информации за последние 10–15 лет привело к созданию принципиально новых возможностей организации практически всех этапов научно-информационного процесса, что, в свою очередь, обусловило качественный рост информационных потребностей научных работников. В настоящее время научные сообщества наиболее развитых стран и регионов мира имеют достаточно мощные информационные системы, в той или иной мере удовлетворяющие потребностям исследователей, однако в процессе их функционирования выявляются весьма значительные проблемы, присущие практически всем программным системам информационного обеспечения научной деятельности.

1.1. Актуализация информации

Существенной проблемой большинства программных систем информационного обеспечения научной деятельности, предназначенных для функционирования в течение неопределенно долгого времени, является недостаточно своевременная актуализация информации (исключение составляют библиотечные системы). Причина этой проблемы состоит в предъявлении к лицам, отслеживающим изменения информации, высоких

*Работа выполнена при финансовой поддержке РФФИ (гранты № 08-07-00229, 09-07-00277 и 10-07-00302), Президентской программы "Ведущие научные школы РФ" (грант № НШ-6068.2010.9), ФЦП "Научные и научно-педагогические кадры инновационной России" на 2009–2013 гг. (госконтракт ГК № П484 от 04.08.2009 г.) и интеграционных проектов СО РАН.

квалификационных требований, возрастающих с усложнением структуры и возможностей поддерживаемой информационной системы, а в нашей стране еще и в недостатке средств для оплаты труда таких сотрудников.

В частности, опыт выполнения интеграционных проектов СО РАН, в рамках которых производилось создание программных систем для разных предметных областей, показал, что рассматриваемые 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 путем 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.

К сожалению, задача автоматизации вовлечения электронных документов в научно-информационный процесс все еще далека от сколько-нибудь удовлетворительного решения. Одна из основных причин сложившейся ситуации заключается в том, что в конце 1970-х годов одновременно с персональными компьютерами появились и мощные средства визуализации информации, вследствие чего были почти остановлены научные изыскания в области теории создания информационно-поисковых систем, которые возобновились только в середине 1990-х в связи с развитием информационных технологий сети Интернет и переходом к распределенному хранению информации. В настоящее время в указанной области получены важные результаты (см., например, монографии [1, 2] и др.), однако эти разработки обычно опираются на неявное предположение о возможности широкого распространения подробной стандартизации представления информации, например, на основе словарей (концепция Semantic Web консорциума W3). К тому же 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. Более того, при свободном характере размещения материалов в сети Интернет требование соблюдения обязательных стандартов представления информации становится всего лишь благим пожеланием (особенно это касается Рунета).

Одно из наиболее неприятных следствий рассматриваемой ситуации — сложность поиска информации, содержащейся в текстовых документах сети Интернет. Это относится и к традиционным методам поиска, характерным для библиотек: поиск по имени автора документа, названию документа или тематический поиск — поскольку *слабоструктурированный* электронный документ (т. е. документ, снабженный метаданными, но при этом имеющий неструктурированные элементы) может не содержать явно заполненных полей метаданных, причем классификационные признаки документа зачастую вообще отсутствуют. Разумеется, обработка слабоструктурированных документов не может быть полностью автоматизирована, и основная задача разработчиков 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 РФ и даже мира, то становится актуальной задача разработки и реализации алгоритмов, автоматизирующих основные этапы научно-информа-

ционного процесса (включая создание тезаурусов и онтологий), посредством интернет-приложений, доступных с любого компьютера сети (разумеется, после аутентификации и авторизации пользователя-эксперта).

1.2. Интероперабельность

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" [3]. 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. д.).

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 [4, 5]. Коротко резюмируя содержание этих работ, можно отметить, что организация поиска в них 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*).

1.3. Взаимодействие с пользователями

При создании информационных систем часто недостаточное внимание уделяется вопросам организации взаимодействия разрабатываемой системы с потребителями информации. Так, А.Н. Колмогоров неоднократно отмечал, что данные представляют информационную ценность лишь тогда, когда они являются составной часть некоторой модели реального мира и связаны с другими данными [6, 7]. Тем самым применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А.А. Ляпунов (см., например, [8]): "Нет модели — нет информации". 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 [9] как структурированная (связанная причинно-следственными и иными отношениями) информация), а не получение каких-либо чисел, гистограмм, отдельных фактов и т. д.

Сказанное, 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 [10].

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, [9]). 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.

Что касается атрибутивного поиска, 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.

Таким образом, 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* [9], куда в качестве составных компонент наряду с традиционной информационной системой входят также рассуждающая информационная система (формализующая правила логического вывода) и интеллектуальный интерфейс (диалог, графика и т. д.), благодаря которому компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека.

При этом следует учитывать, что широта и многогранность информационных потребностей научного сообщества (см., например, [11]) вызывает необходимость массового создания информационных систем, разнообразных как по тематике, так и по целевому назначению, что приводит к необходимости систематического изучения всех этапов процесса разработки информационных систем, включающего стадии создания концептуальной модели, информационной модели и практической реализации системы.

В целом в настоящее время возникла насущная необходимость осмысления процесса обработки компьютерной информации как технологии. Заметим, что аналогичный подход к вычислительному моделированию был осуществлен в начале 1980-х годов в работах Н.Н. Яненко [12] и А.А. Самарского [13] и стал важной вехой в развитии прикладной математики.

2. К вопросу о стадиях переработки информации

В соответствии с [14] под *технологией* будем понимать совокупность методов обработки, изготовления, изменения состояния, свойств и формы сырья, материалов или по-

луфабрикатов 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 [14]). Иными словами, любая технология по своей сути — воспроизводимый инструмент, применяемый для превращения потребляемых факторов в продукцию или, вообще говоря, для достижения планируемых результатов [15].

Приведем еще одно, пожалуй, наиболее краткое из определений технологии: "Технология — способ преобразования данного в необходимое" (см., например, [16]), 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 (подробнее об этом речь пойдет в следующем разделе).

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! Разумеется, человек, владеющий теоретическими 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 с философских, социологических, биологических, физико-математических или кибернетических позиций существует множество подходов к понятию "информация" [9, с. 393], включая так называемую техническую теорию информации, которая по сути является 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 документальной информации и экспертных систем, вкладывают в термин "знания" несколько разный смысл (сравни, в частности, [9] и [17]). При этом в трактовании термина "данные" (понимаемые как факты и идеи, представленные в формализованном виде [18]) 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, позволяющая судить о том, какое из многочисленных определений понятий "информация" или "знание" является "более правильным". Речь идет лишь о 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. На наш взгляд (подробное обоснование см. в [19]), при создании интеллектуальных информационных систем наиболее целесообразно придерживаться многоуровневой модели информации, изложенной, например, в работе В. Гитта [20] (рис. 1). Нижний уровень этой модели соответствует шенноновскому значению термина "информация", три последующих — семиотической триаде (синтактика — семантика — прагматика), а верхний (пятый) уровень носит метафизический характер. При этом наличие в некотором сообщении информации высокого уровня влечет за собой наличие информации всех низших уровней, но, разумеется, не наоборот (еще раз подчеркнем: объем информации зависит, в том числе, от характери-

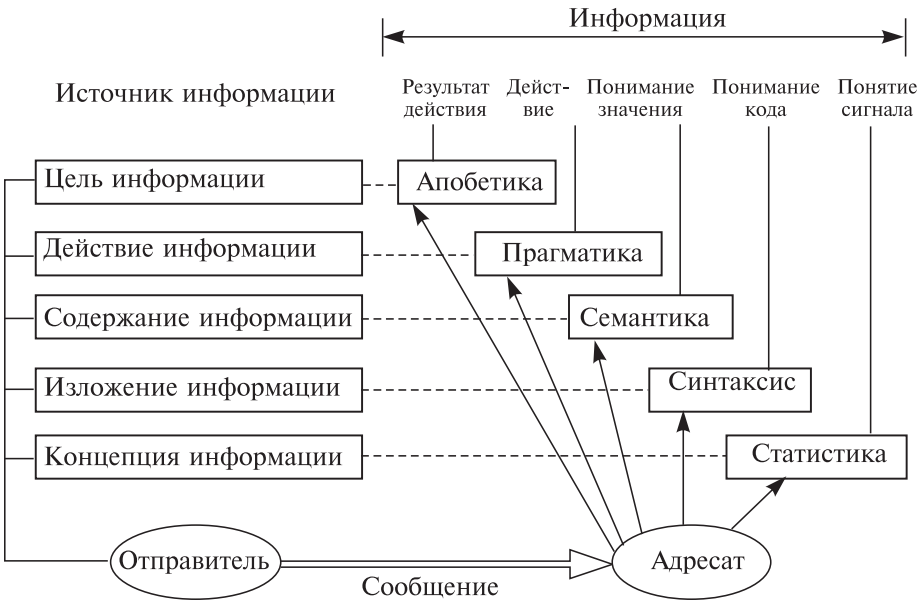

Рис. 1. Пятиуровневая модель информации

стик адресата, причем это касается всех уровней информации, а также от информации, содержащейся в других сообщениях, определяющих контекст данного сообщения).

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). Тем не менее с начала 1980-х годов семиотическая триада заняла прочное место в кибернетике, о чем свидетельствуют соответствующие статьи в "Словаре по кибернетике" [18], хотя в первое время семиотическая терминология применялась, скорее, при описании языка (понимаемого как частный случай знаковой системы) в целом, нежели при анализе отдельных сообщений. Однако к настоящему времени описание непосредственно информации с помощью семиотической терминологии получило широкое распространение в отечественной литературе.

Важно подчеркнуть, что семиотический подход фактически использован при определении базисных понятий в монографии [9]. *Данные* понимаются в ней (в соответствии с традиционным подходом) как факты и идеи, представленные в символьной форме, позволяющей проводить их передачу, обработку и интерпретацию, *информация* — как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная 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.*

Резюмируя, можно сделать вывод о том, что создание технологий компьютерной обработки информации невозможно без анализа стадий процесса ее переработки, т. е. без должного интеллектуального обеспечения технологий, которое основано на всестороннем учете как информационных потребностей научных работников, так и широких возможностей современных аппаратных и программных средств.

3. Системный подход — основа технологии обработки информации

Какие же качественно новые возможности решения указанных выше проблем предоставляют современные компьютеры и языки манипулирования данными? В классических *информационно-поисковых системах* (ИПС) основным элементом (или логической единицей хранения) являлась запись, представлявшая собой поисковый образ документа [18]. При этом важно отметить, что записи не имели непосредственной связи друг с другом, что резко сужало возможности ИПС. В частности, автоматизированные системы, способные строить даже простые категорические силлогизмы (для чего требуется наличие в системе связей между терминами силлогизма), отнесены ([10, с. 149, 150]) к особому классу *информационно-логических систем*. Одной из наиболее очевидных практических проблем, возникающих в силу отсутствия связей между записями, является невозможность установить наличие (или отсутствие) связи между собственным именем и предполагаемым его конкретным носителем, даже если информация о последнем присутствует в ИПС [10, с. 137]. Тем самым ИПС полностью оправдывали свое название — они выдавали в качестве продукта переработки данных именно информацию, но не знания.

Развитие алгоритмических, программных и аппаратных средств информатики привело в 1980-е годы к возможности создания *интеллектуальных информационных систем*, в которых компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека. Интеллектуальные системы (ИнтС) функционируют по следующей схеме [9]:

$$\text{ИнтС} = \text{РИС} + \text{ИПС} + \text{ИнИн},$$

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

Более развитые ИнтС должны иметь также механизм пополнения базы данных, функционируя по схеме

$$\text{ИнтС} = \text{РИС} + \text{ИПС} + \text{ИнИн} + \text{АП},$$

где АП — автоматическое извлечение фактов из текстов и соответствующее пополнение базы данных посредством этих фактов и выводов из них (подробнее см., например, [21]).

Таким образом, интеллектуальная система по сравнению с обычной ИПС обладает новыми возможностями, предоставляющими возможность удовлетворить квалифицированного пользователя в соответствии со схемой документ — факт — рассуждение [9, с. 343], т. е. *интеллектуальные информационные системы позволяют не только извлекать из данных информацию, но и получать новые знания.*

На основании выше сказанного 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 нужных пользователю информации и знаний.*

Для наиболее эффективного функционировании ИнтС в качестве логической единицы хранения целесообразно рассматривать *документ*, понимаемый как информационный ресурс, имеющий (по определению [22]) уникальный идентификатор и обладающий некоторой структурой и содержанием.

Разумеется, документ как информационный ресурс представляет собой поисковый образ исходного документа, причем в некоторых случаях содержание последнего может входить в поисковый образ в качестве одного из элементов (что противоречит ограничению из классической монографии [23], но из контекста следует, что подобное ограничение было вызвано необходимостью уменьшения объема поисковых образов с целью снижения 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". С другой стороны, в фундаментальных работах по информатике и кибернетике [18, 23], вышедших в том числе в конце 1980-х годов, поисковый образ документа не рассматривается даже в качестве вторичного документа.

Для описания документов используются метаданные, как правило, иерархической структуры. Наиболее общий характер имеют метаданные, задающие структуру документа, т. е. описывающие метаданные более низкого уровня (атрибуты документа), которые определяют содержание документа (рис. 2). Наконец, значения этих атрибутов являются фактически метаданными по отношению к исходному документу. Отсюда следует важнейшая отличительная черта рассматриваемого подхода к построению информационных систем: *работа не с данными, а исключительно с метаданными.*

Важно подчеркнуть, что документ может входить в качестве значения некоторого элемента метаданных другого документа. Так, любой документ d_i массива данных 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. Если же 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.

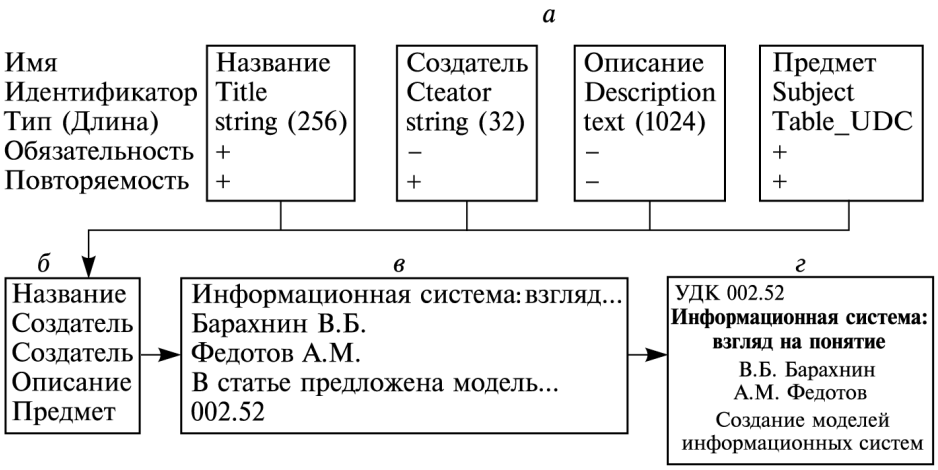

Рис. 2. Иерархия метаданных документа: *а* — структура, *б* — атрибуты, *в* — содержание, *г* — документ

Таким образом, наличие внутренних связей между элементами массива данных позволяет рассматривать его как некоторую *систему* и анализировать с использованием методов общей теории систем (заметим, что классическое определение системы как "множество объектов вместе с отношениями между объектами и между их атрибутами" [24] основано на тех же понятиях, что и, например, реляционная модель данных).

Соответствующий анализ был проведен в [25, 26]. Перечислим основные выводы этих работ, имеющие отношение к технологическим аспектам обработки данных.

Прежде всего отметим, что с использованием системного подхода в [25] удалось дать обоснованную формулировку информационных потребностей научного сообщества и предложить реально выполнимую схему их удовлетворения, учитывающую необходимость компромисса между качеством решения поставленной задачи и разумными сроками ее выполнения. Последний принцип давно является основополагающим в другой отрасли кибернетики — прикладной математике (см., например, [27]), при этом улучшение результата применительно к информационной системе возможно с течением времени и достигается путем расширения массива данных (как путем добавления новых документов, так и расширением структуры уже существующих).

Модель данных в ИнтС 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 [28]. Анализ иерархии метаданных, приведенной на рис. 2, позволяет сделать важный вывод: *описание массива данных посредством метаданных* наделяет их, в том числе, семантикой, воспринимаемой в среде социальных коммуникаций, т. е. *делает данные информацией* (в узком значении этого слова).

Одним из достоинств изложенной модели является простота создания базовой структуры представления информации, отвечающей такой совокупности заранее сформулированных информационных запросов (например, посредством соответствующих гиперссылок), которая в состоянии удовлетворить основные информационные потребности пользователей системы. Эта структура основана на многомерной (т. е. не сводящейся только к предметной) классификация документов, позволяющей включать в метаописание документа некий многомерный набор классификационных признаков, определяющий поисковое предписание, которое соответствует тому или иному информационному запросу из заранее заданного множества (подробнее см. [29]).

Как же добиться возможности реализации следующего технологического шага — получения *новых* (т. е. явно не содержащихся в исходном массиве данных) *знаний*? Для этого, очевидно, необходима, как минимум, хорошая структуризация данных, предусматривающая, в частности, достаточно большое количество поисковых признаков, образующих поисковый образ документа. В свою очередь, объединение поисковых образов однородных документов составляет каталог.

Кроме того, в информационно-поисковом языке, используемом при создании ИнтС, должны присутствовать средства выражения имманентных отношений между предметами, т. е. язык должен обладать парадигматическими отношениями (примером языка, не обладающего этими отношениями, может служить система унитермов — набора одиночных ключевых слов (в редких случаях словосочетаний)). Средством же выражения парадигматических отношений является *онтология* предметной области или ее *тезаурус*, причем граница применения этих терминов весьма размыта (как отмечено

в [30], "...еще недавно сегодняшняя Онтология именовалась Тезаурусом", что иллюстрируют, например, тезаурусы по науковедению и лексикографии [31], которые ввиду своей структурной сложности с сегодняшней точки зрения явно представляются онтологиями). Таким образом, *наличие онтологии (тезауруса)* в качестве составной части информационно-поискового языка, используемого при создании каталога, — *необходимое и достаточное условие* (см. [26]) *возможности получения из данных, уже преобразованных в информацию, новых знаний.*

Заметим, что именно каталог является наиболее естественной формой унификации представления данных, и тем самым, — достаточно простым средством решения отмеченной во введении проблемы синтаксической интероперабельности.

Наконец, рассмотрение массива данных как системы позволяет уделить особое внимание ее динамическим характеристикам, поскольку "...отдельные уровни системы обуславливают определенные аспекты ее поведения, а целостное функционирование оказывается результатом взаимодействия всех ее сторон и уровней" [32].

4. Технология автоматизации обработки слабоструктурированных документов

Важнейшим аспектом работы информационной системы является ее пополнение новыми документами. Опыт создания информационных систем научной направленности показывает, что подобные системы могут успешно развиваться лишь в случае актуализации содержащейся в них информации самими пользователями этих систем. Более того, поскольку в интеллектуальных информационных системах компьютер в диалоговом режиме усиливает комбинаторное мышление и логические возможности человека, то при этом происходит автоматизированное пополнение базы данных. В силу указанных обстоятельств при работе с интеллектуальными информационными системами многих пользователей возможности систем резко возрастают.

Как было отмечено выше, взаимодействие информационных систем с внешними пользователями в плане занесения в них новых данных целесообразно организовывать преимущественно (или даже почти исключительно) через веб-интерфейс, при этом специалисты в предметной области, поддерживающие актуальность информации, могут быть сотрудниками нескольких организаций, расположенных в разных городах и даже странах.

Отметим, что обработка документов, размещенных в сети Интернет, имеет ряд специфических особенностей, отличающих их каталогизацию от каталогизации полиграфических изданий. В частности, каждую публикацию в составе электронного журнала, сборника и т. д. целесообразно представлять как отдельный документ. Это существенно облегчает процесс поиска нужной информации, позволяя вести атрибутивный поиск отдельных статей по авторам, названию, классификационным признакам, ключевым словам и т. д. Разумеется, аналогичный подход весьма желателен и при работе с полиграфическими изданиями (так называемая аналитическая роспись статей), однако данное требование нередко не соблюдается из-за огромных трудозатрат. Как отмечено в [23], один человек за рабочий день способен описать не более 50–70 документов на родном языке и не более 20–30 — на иностранном. При обработке же электронных документов возможна частичная автоматизация процесса каталогизации отдельных публикаций.

Обычно количество организаций, работающих в той или иной конкретной области науки, а также журналов, публикующих статьи соответствующей тематики, сравнительно невелико, поэтому 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. Менее тривиальный характер имеет задача каталогизации множества отдельных документов, размещенных на том или ином сайте (например статей, биографий и т. п.). Так как однородные документы, размещенные на одной сайте, имеют однородную структуру, то наиболее целесообразно использовать алгоритмы, использующие информацию о гипертекстовой разметке обрабатываемых документов. Конечно, такой подход целесообразен лишь для хорошо организованных сайтов с большим объемом однородной информации (что, собственно, и устанавливает рамки применимости рассматриваемой технологии), но именно таковыми являются большинство сайтов, представляющих интерес для создателей систем информационного обеспечения научной деятельности: сайты журналов, содержащие научные статьи, сайты организаций, содержащие описания персон и проектов, и т. п.

Один из возможных алгоритмов решения задачи частичной автоматизации процесса извлечения метаданных разработан и изложен в [33, 34]. Алгоритм, основанный на типичном для интеллектуальных информационных систем человеко-машинном взаимодействии, сводится к выполнению следующих операций:

1) создание шаблона для обрабатываемого сайта;

2) создание списка адресов, где расположены документы;

3) обработка документов;

4) поддержание актуальности информации.

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 в описанной ситуации 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 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. Так, в некоторых отечественных математических журналах используется классификатор УДК, в то время как в международном математическом сообществе более распространен классификатор MSC2000.

Разумеется, наиболее качественно решить задачу классификации может эксперт-человек, поэтому прежде всего следует проверить, не внесена ли информация о полиграфической версии статьи в ту или иную электронную библиографическую базу данных удаленного доступа, в которой документы классифицированы в соответствии с нужным классификатором. Так, в среде математиков весьма популярна база данных журнала “Zentralblatt MATH” (http://www.zentralblatt-math.org/zmath/en). Статью в этой базе можно однозначно идентифицировать по ISSN журнала, его номеру и страницам, на которых она расположена. Однако не все электронные версии журналов содержат номера страниц полиграфических версий статей, поэтому при отсутствии сведений о страницах в процессе идентификации следует опираться на фамилии автора (авторов) в латинской транскрипции.

Подчеркнем, что полная репликация метаданных документа из библиографической базы далеко не всегда может служить эффективной заменой процесса непосредственного извлечения метаданных из слабоструктурированного документа хотя бы потому, что нередко библиографические базы не содержат сведений об url-адресе полной электронной версии документа (например, в базе "Zentralblatt MATH" в описании статьи содержится лишь ссылка на головную страницу сайта соответствующего журнала).

Процесс определения метаданных документа с использованием удаленной библиографической базы также может быть частично автоматизирован [33].

Если же классификационные признаки документа отсутствуют как в нем самом, так и в библиографических базах удаленного доступа, то требуется провести автоматическую классификацию документа исходя непосредственно из его содержания, а также предоставить пользователю возможность осуществления поиска документов "по аналогии" [35]. Для решения этой задачи был разработан и реализован алгоритм автоматической классификации (кластеризации) документов на основании меры их сходства, задаваемой с использованием атрибутов их библиографического описания [36]. Отличительными особенностями этого алгоритма являются, во-первых, использование в процессе координатного индексирования документа не отдельных слов, входящих в словарь предметной области, а терминов-словосочетаний, образующих ее тезаурус; во-вторых, подсчет меры сходства на основании не только координатного индекса документа, но и ключевых слов (в узкобиблиографическом понимании), а также сведений об авторах документа; и, в-третьих, применение продукционных правил, позволяющих изменять весовые коэффициенты, соответствующие тем или иным атрибутам библиографического описания в формуле задания меры сходства на основании апостериорной достоверности значений этих атрибутов.

Для частичной автоматизации процесса создания тезаурусов и онтологий тех или иных разделов науки была разработана и реализована соответствующая методика, основанная на применении предметного указателя специализированных энциклопедий [37], которая обеспечивает высококвалифицированное описание предметной области с использованием надежно выверенных терминов, позволяя провести начальный этап построения онтологии с минимальным привлечением экспертов в данной предметной области.

Заключение

В работе рассмотрены основные направления процесса смысловой обработки данных, содержащихся в слабоструктурированных документах достаточно произвольной структуры, как технологии. Показано, что в основе этой технологии должно быть представление о массиве данных как о системе, описываемой с использованием инфологической модели, благодаря чему между элементами системы (поисковые образы документов) устанавливаются внутренние связи. Описание массива данных посредством метаданных делает *данные информацией*, а наличие онтологии (тезауруса) в качестве составной части информационно-поискового языка, используемого при создании каталога, является обязательным условием возможности получения из данных, преобразованных в информацию, *новых знаний*. Установлено, что применение методов общей теории систем открывает дополнительные возможности исследования технологии смысловой обработки данных. Предложена технология автоматизации извлечения метаданных (в том числе классификационных признаков) из интернет-документов.

Представленные технологии были использованы при создании сайта СО РАН (http://www.sbras.ru), который, по данным рейтинга Webometrics [38], включающего сайты ведущих научно-исследовательских центров всего мира, в течение нескольких лет неизменно занимает наивысшее среди российских сайтов место и входит в первую двадцатку европейских и первую полусотню мировых сайтов, а также ряда связанных с этим сайтом информационных систем.

Список литературы

[1] Krogstie J., Halpin T., Siau K. Information Modeling Methods and Methodologies. Idea Group Publishing, 2005.

[2] Semantic Web and Peer-to-Peer, Decentralized Management and Exchange of Knowledge and Information / Eds. S. Staab, H. Stuckenschmidt. Springer, 2006.

[3] Жижимов О.Л., Турпанов А.А., Федотов А.М. Корпоративный каталог СО РАН // Тр. Восьмой Всероссийской науч. конф. "Электронные библиотеки: Перспективные методы и технологии, электронные коллекции" (RCDL'2006). Ярославль, 2006. С. 226–230.

[4] Фейгин Д. Концепция SOA // Открытые системы. 2004. № 6. http://www.osp.ru/os/ 2004/06/184447/_p1.html

[5] 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. выпуск: Тр. IX рабочего совещ. по электр. публ. (El-Pub2004). С. 29–48.

[6] Колмогоров А.Н. Три подхода к определению понятия "количество информации" // Проблемы передачи информации. 1965. Т. I, вып. 1. С. 3–11.

[7] Колмогоров А.Н. Теория информации и теория алгоритмов. М.: Наука, 1987.

[8] Ляпунов А.А. О соотношении понятий материя, энергия и информация // Проблемы теоретической и прикладной кибернетики. Новосибирск: Наука, 1980. С. 320–323.

[9] Инфосфера: Информационные структуры, системы и процессы в науке и обществе / Ю.М. Арский, Р.С. Гиляревский, И.С. Туров, А.И. Черный М.: ВИНИТИ, 1996.

[10] Михайлов А.И., Черный А.И, Гиляревский Р.С. Научные коммуникации и информатика. М.: Наука, 1976.

[11] Барахнин В.Б., Федотов А.М. Исследование информационных потребностей научного сообщества для построения информационной модели описания его деятельности // Вестник НГУ. Серия: Информационные технологии. 2008. Т. 6, вып. 3. С. 48–59.

[12] Яненко Н.Н. Методологические вопросы современной математики // Вопросы философии. 1981. № 8. С. 60–68.

[13] Самарский А.А. Задачи прикладной математики на современном этапе развития // Коммунист. 1983. № 18. С. 31–42.

[14] Технология // Большой академический словарь. СПб.: Большая Российская энциклопедия, 2003. С. 2000.

[15] Желены М. Управление высокими технологиями // Информационные технологии в бизнесе. Энциклопедия. СПб.: Питер, 2002. С. 81–89.

[16] Технология // Тезаурус по образованию и педагогике / Ин-т информатизации образования в составе Московского гос. гуманитарного ун-та. http://www.mgopu.ru/ininfo/r1_thesaurus.htm#technology

[17] Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. СПб.: Питер, 2000.

[18] Словарь по кибернетике. Киев: Главная редакция Украинской Советской Энциклопедии им. М.П. Бажана, 1989.

[19] Барахнин В.Б., Федотов А.М. Уточнение терминологии, используемой при описании интеллектуальных информационных систем, на основе семиотического подхода // Изв. вузов. Проблемы полиграфии и издательского дела. 2008. № 6. С. 73–81.

[20] Gitt W. Ordnung und information in technik und natur // Am Anfang war die Information. Gräfeling: Resch KG, 1982. S. 171–211.

[21] Осипов Г.С. Лекции по искусственному интеллекту. М.: КРАСАНД, 2009.

[22] Berners-Lee T., Fielding R., Masinter L. Uniform Resource Identifiers (URI). Generic Syntax. RFC 2396. http://www.ietf.org/rfc/rfc2396.txt/

[23] Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. М.: Наука, 1968.

[24] Холл А.Д., Фейджин Р.Е. Определение понятия системы // Исследования по общей теории систем. М.: Прогресс, 1969. С. 252–282.

[25] 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. выпуск: Избр. докл. X Российской конф. "Распределенные информационно-вычислительные ресурсы" (DICR-2005). С. 52–58.

[26] Барахнин В.Б., Федотов А.М. Информационная система: Взгляд на понятие // Вестник НГУ. Сер.: Информационные технологии. 2007. Т. 5, вып. 2. С. 12–19.

[27] Бахвалов Н.С. Численные методы. М.: Наука, 1970.

[28] Langefors B. Infological models and information user views // Information Systems. 1980. No. 5. P. 17–32.

[29] Федотов А.М., Барахнин В.Б. Проблемы поиска информации: История и технологии // Вестник НГУ. Серия: Информационные технологии. 2009. Т. 7, вып. 2. С. 3–17.

[30] Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Тр. междунар. семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Т. 1. Аксаково, 2001. С. 184–188.

[31] Никитина С.Е. Семантический анализ языка науки. М.: Наука, 1987.

[32] Садовский В.Н. Система // Философский энциклопедический словарь. М.: Советская энциклопедия, 1983. С. 610–611.

[33] Барахнин В.Б., Ведерников В.В. Алгоритм автоматической каталогизации статей, опубликованных в электронных версиях научных журналов // Тр. Всероссийской науч. конф. "Научный сервис в сети Интернет: Технологии параллельного программирования". Новороссийск, 2006. С. 277–279.

[34] Барахнин В.Б., Федотов А.М. Ресурсы сети Интернет как объект научного исследования // Изв. вузов. Проблемы полиграфии и издательского дела. 2008. № 1. С. 70–77.

[35] Федотов А.М., Барахнин В.Б. К вопросу о поиске документов "по аналогии" // Вестник НГУ. Серия: Информационные технологии. 2009. Т. 7, вып. 4. С. 3–14.

[36] Барахнин В.Б., Нехаева В.А., Федотов А.М. О задании меры сходства для кластеризации текстовых документов // Там же. 2008. Т. 6, вып. 1. С. 3–9.

[37] Барахнин В.Б., Нехаева В.А. Технология создания тезауруса предметной области на основе предметного указателя энциклопедии // Вычисл. технологии. 2007. Т. 12. Спец. выпуск 2. С. 3–9.

[38] Top 300 R&D European Institutes. http://research.webometrics.info/top300_r&d_europe.asp

Поступила в редакцию 4 октября 2010 г.,
с доработки — 3 ноября 2010 г.