Проблемы технологий создания систем смысловой обработки данных♣

© В.Б. Барахнин А.М. Федотов

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет

bar@ict.nsc.ru fedotov@sbras.ru

Аннотация

В работе обсуждаются перспективы развития процесса смысловой обработки данных как технологии, при этом в качестве источника данных рассматриваются интернет-документы достаточно произвольной структуры, что принципиально отличает описываемый подход от концепции Semantic Web.

1 Введение

Проблема доступа к информации является одной из основных проблем, возникающих в деятельности научного исследователя. Любой научный процесс порождает огромные объемы данных, и работать с ними становится все сложнее по мере того, как гигабайты данных превращаются в терабайты. Так, еще в начале 1960-х годов американский историк и социолог науки Дерек де Солла Прайс на основании исследований развития науки в течение последних 200 лет выявил следующую эмпирическую закономерность [1]: *любой достаточно большой сегмент науки в нормальных условиях растет экспоненциально, то есть любые параметры науки за определенный промежуток времени удваиваются*. Эта закономерность получила название *закона экспоненциального роста науки*.

Происшедшее за последние 10-15 лет 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 практический всех этапов научно-информационного процесса, что, в свою очередь, обусловило качественный рост информационных потребностей научного сообщества. Отсюда следует необходимость разработки и создания новых инструментальных средств и алгоритмов для сбора и анализа данных, содержащихся в разнообразных интернет-документах научной тематики.

Однако при решении поставленной задачи возникают трудности, зачастую носящие принципиальный характер.

Во-первых, распределенное хранение информации требует *интероперабельности* (т.е. обеспечения взаимодействия) разнородных информационных источников. Различают два уровня интероперабельности: семантический и технический [2], причем в последнем иногда выделяют синтаксический уровень [3]. Семантическая интероперабельность, заключающаяся в использовании согласованных стандартов метаданных (обзор которых применительно к научным информационным системам приведен в [3]), как правило, соблюдается. Проблемы возникают на уровне технической интероперабельности, точнее, согласования моделей данных и форматов их представления (что относится к синтаксической интероперабельности).

При этом очень важно иметь в виду следующее обстоятельство. Основным источником электронных документов в настоящее время является сеть Интернет. Однако ее развитие сети изначально носит децентрализованный характер, поэтому выработка сколько-нибудь сложных стандартов представления информации – не более чем благое пожелание. Разумеется, определенным группам разработчиков сайтов, относящихся к той или иной предметной области, удается договориться об использовании единых стандартов и технологий, позволяющих, в частности, интегрировать ресурсы соответствующих сайтов, однако можно с уверенность утверждать, что имеются сайты, содержащие важную информацию из данной предметной области, но не соответствующие этим стандартам. Поэтому, говоря об интероперабельности, затрагивающей широкий спектр ресурсов Интернет, необходимо априорно рассчитывать лишь на минимальную их стандартизацию.

Во-вторых, в настоящее время наука об обработке информации, особенно в ее прикладном аспекте, несколько отстает от соответствующих аппаратно-программных средств, хотя еще А.Н. Колмогоров показал, что данные представляют информационную ценность лишь тогда, когда они являются составной часть некоторой модели реального мира и

Труды 10-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2008, Дубна, Россия, 2008.

связаны с другими данными.хотя данные представляют информационную ценность лишь тогда, когда они являются составной часть некоторой модели реального мира и связаны с другими данными [4, 5]. Как отмечал А.А. Ляпунов [6], «информация всегда относительна, она зависит от того, какой информационной системой она воспринимается». Аналогичное отставание наблюдается и в прикладных исследованиях, посвященных извлечению из информации *знаний,* понимаемых как структурированная (связанная причинно-следственными и иными отношениями) информация [7]. Разработчикам программных средств обработки данных зачастую недостает понимания того обстоятельства, что конечная цель работы, связанной с применением информационных технологий – *понимание* того или иного явления, а не получение каких-либо чисел, гистограмм, отдельных фактов и т.п.

Одна из основных причин сложившейся ситуации заключается в том, что с появлением в середине 1970-х годов персональных компьютеров появились мощные средства визуализации информации, вследствие чего были практически остановлены научные изыскания в области интеллектуального поиска, которые возобновились лишь в середине 1990-х с появлением сети Интернет, приведшем к распределенному хранению информации. В настоящее время в этой области получены важные результаты (см., например, монографии [8, 9] и др.), однако эти разработки, лежащие, как правило, в рамках концепции Semantic Web [10], обычно опираются на неявное предположение о возможности широкого распространения более или менее подробной стандартизации представления информации. Проблема заключается в том, что наработки консорциума W3 носят *лишь рекомендательный* характер, а объявить их *стандартами* могут только организации, имеющие соответствующий статус, например ISO или ГОСТ, поэтому реальное развитие большинства ресурсов Интернет, в том числе научной направленности, идет без учета соответствующих рекомендаций.

Особо подчеркнем, что при разработке *собственных* ресурсов разработчикам информационных систем следует придерживаться таких стандартов метаданных, которые обеспечивают ориентированность на семантические технологии описания и использования информации, в т.ч. Semantic Web, однако при обработке *чужих* информационных ресурсов надо быть готовым к худшему.

Ресурсы, разработанные без учета рекомендаций консорциума W3, зачастую не могут быть обработаны с использованием онтологий 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» [11]. Заметим, что это утверждение принадлежит коллективу создателей крупнейшего тезауруса русского языка РуТез, предназначенного для автоматической обработки больших текстовых коллекций, который представляет собой иерархическую сеть 50 тысяч понятий с более чем 125 тысячами слов и выражений (статистика относится к 2007 году [12]).

В итоге разработчики прикладных информационных систем, интегрирующих разнородные интернет-ресурсы, стоят перед дилеммой: или опираться на мощный (хотя и созданный еще в досетевую эпоху) аппарат для манипулирования с *данными*, но не с *информацией*, либо использовать результаты последних теоретических исследований, которые, однако, не всегда позволяют обрабатывать огромные пласты информации, доступной через сеть Интернет (внутренние проблемы концепции Semantic Web, связанные, например, со сложностью создания онтологий верхнего уровня, в данной работе мы не рассматриваем).

Таким образом, возникает третья проблема: разработка методологии изучения интернет-ресурсов достаточно произвольной структуры с целью вовлечения данных, содержащихся в этих ресурсах, в сферу научных исследований.

Резюмируя сказанное выше, можно сделать вывод о том, что комплексное решение указанных проблем возможно лишь при осмыслении процесса обработки компьютерной информации как *технологии.* Заметим, что аналогичное осмысление другой области кибернетики – вычислительного моделирования – было осуществлено в начале 1980-х годов в работах Н.Н. Яненко [13] и А.А. Самарского [14] и стало важной вехой в развитии прикладной математики.

2 Проблема стадий переработки информации

В соответствии с [15] будем понимать под технологией определенную последовательность методов обработки, изготовления, изменения состояний и свойств сырья или материалов в процессе производства продукции. Иными словами, любая технология по своей сути – инструмент, применяемый для превращения потребляемых факторов в продукцию, или, вообще говоря, для достижения планируемых результатов (см. [16]) .

Сошлемся еще на одно, пожалуй, наиболее краткое из определений технологии: «технология – способ преобразования данного в необходимое» (см., например, [17]), 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, о чем шла речь выше.

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 теоретическим основами информатики, по некотором размышлении ответит, что исходным материалом служат данные, а конечным продуктом – знания (или, по крайней мере, семантическая информация). Тем не менее, приведенный пример показывает, что проблемы возникают уже на терминологическом уровне.

Поскольку существует множество подходов к понятию «информация» с философских, социологических, биологических, физико-математических или кибернетических позиций ([7], с. 393), включая так называемую «техническую» теорию информации, которая является, по сути, теорией передачи и хранения данных, постольку можно обнаружить десятки порой противоречащих друг другу определений того, что является информацией или знанием. Вряд ли существует некая «абсолютная» точка зрения, с которой возможно было бы судить о том, какое из многочисленных определений является «более правильным». Поэтому цель проводимого ниже экскурса в историю терминологии 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.

Реальное осознание сложностей в преодолении разрыва между шенноновским понятием информации и концепцией семантической информации как средства социальной коммуникации возникло в середине 1960-х гг. (см., например, работы У. Шрамма [18] и Ю.А. Шрейдера [19], причем Ю.А. Шрейдер показал, что о количестве семантической информации в данном сообщении есть смысл говорить лишь применительно к конкретному приемнику сообщения).

Попытка «телеологического» описания особенностей восприятия сообщения субъектом была предпринята Р. Акоффом и Ф. Эмери [20]. 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).

Тем самым Р. Акофф и Ф. Эмери, по-видимому, одними из первых специалистов в области информатики обратили на многоуровневость восприятия сообщения получателем и сделали попытку описать эти уровни.

Наконец, в начале 1980-х годов немецким исследователем В. Гиттом была предложна пятиуровневая модель информации [21], наиболее полно отражающая различные аспекты термина «информация». Ее уровни (снизу вверх):

– *статистика;*
– *синтаксис;*
– *семантика;*
– *практика;*
– *апобетика* (результат).

Анализируя эту модель, нетрудно видеть, что ее нижний уровень соответствует шенноновскому значению термина «информация», три последующих – семиотической триаде (синтагматика – 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 высоких уровней, но, разумеется, не наоборот (еще раз напомним: объем информации зависит, в том числе, 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). Тем не менее, с начала 1980-х годов семиотическая триада заняла прочное место в кибернетике, о чем свидетельствуют соответствующие статьи в «Словаре по кибернетике» [24], 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 моменту 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 фундаментальной монографии [7], 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, *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*.

Таким образом, проделанный нами анализ показывает, что создание технологий компьютерной обработки информации невозможно без анализа стадий процесса переработки информации, иными словами, без должного интеллектуального обеспечения технологии, которое, разумеется, основано на широких возможностях современных аппаратных и программных средств.

3 Системный подход – основа технологии обработки информации

Какие же качественно новые возможности предоставляют мощности современных компьютеров и языков манипулирования данными? В классических *информационно-поисковых системах* (ИПС) основным элементом (или же логической единицей хранения) являлась *запись* в базе данных, представлявшая собой поисковый образ документа [22]. 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 наличия в системе связей между терминами силлогизма), отнесены ([23, с.149-150]) к особому классу *информационно-логических систем*. Одной из наиболее очевидных практических проблем, порождаемых отсутствием связей между записями, является невозможность установить наличие (или отсутствие) связи между собственным именем и предполагаемым его конкретным носителем, даже если информация о предполагаемом носителе и присутствует в ИПС [23, 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 информатике 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 [7]:

ИнтС = РИС + ИПС + ИнИн+ АП,

где РИС – рассуждающая информационная система (формализующая правила логического вывода), ИнИн- интеллектуальный интерфейс (диалог, графика и т.д.).

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

ИнтС = РИС + ИПС + ИнИн,

где АП – автоматическое извлечение данных из текстов и соответствующее пополнение базы данных посредством этих фактов.

Таким образом, интеллектуальная система обладает по сравнению с обычной ИПС новыми возможностями, позволяя удовлетворить квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [7, с. 343], то есть *интеллектуальные информационные системы позволяют не только извлекать из данных информацию, но и получать новые знания*.

На основании 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 нужной пользователю информации и знаний*.

Таким образом, для наиболее эффективного функционировании ИнтС целесообразно рассматривать в качестве логической единицы хранения *документ,* понимаемый как информационный ресурс, имеющий (по определению [24]) 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 входить в поисковый образ в качестве одного из элементов (это противоречит ограничению из классической монографии [25], 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». С другой стороны, в фундаментальных работах по информатике и кибернетике [22, 25], вышедших, в том числе в конце 1980-х годов, поисковый образ документа не рассматривается даже в качестве вторичного документа.

Важнейшей особенностью данного подхода является использование для описания документов метаданных иерархической структуры. Наиболее общий характер имеют метаданные, задающие структуру документа, т.е. описывающие метаданные более низкого уровня (атрибуты документа), которые определяют содержание документа (см. рис. 1). Наконец, значения этих атрибутов является фактически метаданными по отношению к исходному документу. Отсюда следует *важнейшая отличительная черта предлагаемого 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. Подробнее, любой документ d_i массива данных представляется как $d_i = <m_i^{j,k}>$, где $m_i^{j,k}$ – значения элементов метаданных M^j из набора единого набора метаданных $M = \bigcup M^i$, 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 $m_{i,i'}^{l,k}$ – атрибуты этой связи, являющиеся значениями соответствующих элементов метаданных.

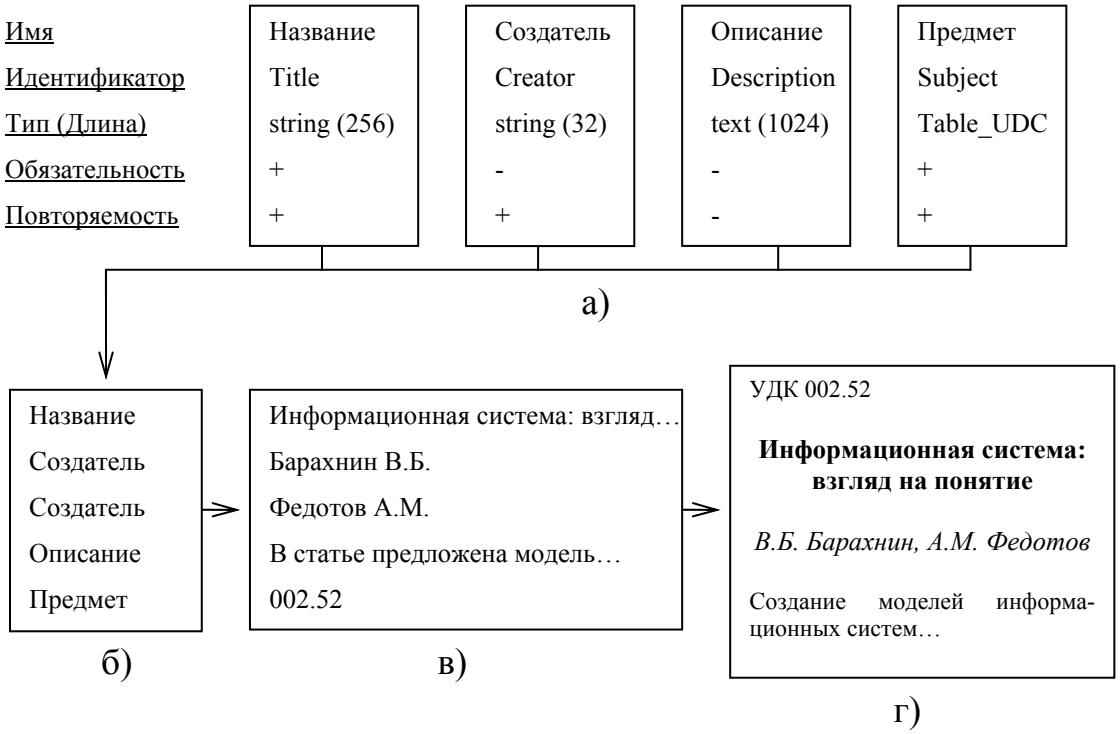

Рис. 1. Иерархия метаданных документа: а) структура, б) атрибуты, в) содержание, г) документ

Таким образом, наличие внутренних связей между элементами массива данных позволяет рассматривать его как некоторую *систему* и анализировать его с использованием методов общей теории систем (заметим, что классическое определение системы «множество объектов вместе с отношениями между объектами и между их атрибутами» [26] основано на тех же понятиях, что и, например, реляционная модель данных).

Соответствующий анализ был проведен нами в [27, 28]. Перечислим основные выводы этих статей, имеющие отношение к технологическим аспектам обработки данных.

Прежде всего, отметим, что с использованием системного подхода в [27] 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, [29]), при этом улучшение результата возможно с течением времени и достигается, применительно к информационной системе, посредством расширения массива данных (как посредством добавления новых документов, так и расширением структуры уже существующих).

Модель данных в системе смысловой обработки данных строится посредством задания классов документов K_i (т.е. множеств документов, описываемых при помощи одних и тех же 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. для построения модели данных используется комбинация иерархической и реляционной моделей, что сближает ее с инфологическими моделями [30]. Анализ иерархии метаданных, приведенной на рис. 1, 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).

Как же добиться возможности реализации следующего технологического шага – получения *новых* (т.е. явно не содержащихся в исходном массиве данных) *знаний*? Очевидно, необходима, как минимум, хорошая структуризация данных, предусматривающая, в частности, достаточно большое количество поисковых признаков, образующих поисковый образ документа, причем соответствующие документы-описания должны быть объединены в *каталог*.

Кроме того, в информационно-поисковом языке, используемом при создании каталога, 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 [31], «…еще недавно сегодняшняя Онтология именовалась Тезау-

русом», что иллюстрируют, например, тезаурусы по науковедению и лексикографии [32], 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, *является необходимым и достаточным условием* (см. [28]) *возможности получения из данных, уже преобразованных в информацию, новых знаний.*

Попутно заметим, что именно каталог является наиболее естественной формой унификации представления данных, тем самым служа 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» [33].

4 Технология автоматизации обработки интернет-документов

Важнейшим вопросом является пополнение информационной системы новыми интернет-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, при этом происходит автоматизированное пополнение базы данных. В силу указанных обстоятельств при работе с интеллектуальными информационными системами многих пользователей, возможности систем резко возрастают.

Так как пользователи, принимающие участие в актуализации информации, могут находиться в разных регионах России и даже мира, то 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.

Отметим, что обработка интернет-документов имеет ряд специфических особенностей, отличающих их каталогизацию от каталогизации полиграфических изданий. В частности, 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 и т.п.

Ввиду того, что информационная система работает не с данными, а исключительно с метаданными, сбор интернет-документов сводится к сбору их метаданных, к тому же непосредственное копирование документов может вызвать серьезные вопросы относительно соблюдения авторских прав.

Однако, как уже было отмечено ранее, значительная доля интернет-документов не содержит метаданных, явно заданных в соответствии с какими-либо рекомендациями, позволяющими проводить автоматическое извлечение, вследствие чего возникает необходимость разработки технологии частичной автоматизации извлечения метаданных из документов произвольной структуры.

Так как однородные документы, размещенные на одной сайте, имеют однородную структуру, то наиболее целесообразно использовать алгоритмы, использующую информацию о гипертекстовой разметке обрабатываемых документов (см., например [34, 35]), при этом надо иметь в виду, что документ может не обладать xml-разметкой и не содержать метаданные в мета-теге, поэтому следует ориентироваться только на html-разметку.

Один из возможных алгоритмов решения задачи частичной автоматизации процесса извлечения метаданных разработан и изложен нами в [36]. Алгоритм, основанный на типичном для интеллектуальных информационных систем человеко-машинном взаимодействии, сводится к выполнению последовательных 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 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 очень популярна база данных журнала «Zentralblatt MATH» (http://www.zentralblatt-math.org/MATH/home), содержащая более 2 миллионов 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 фамилии автора (авторов) в латинской транскрипции.

Подчеркнем, 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 [36].

Наконец, если 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. 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. [37]. Его отличительными особенностями 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 текста документа, но и ключевых слов (в узко-библиографическом 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.

5 Заключение

В данной работе намечены первые шаги в направлении осмысления процесса смысловой обработки данных, содержащихся в интернет-документах достаточно произвольной структуры, как технологии. Показано, что в основе этой технологии должна лежать представление о массиве данных как о системе, описываемой с использованием метаданных посредством комбинации иерархической и реляционной моделей, 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 необходимым и достаточным 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 т.ч. классификационных признаков) из интернет-документов.

Описанные технологии были использованы при создании сайта СО РАН www.sbras.ru, занимающего (по данным за июль 2008 года) согласно рейтингу Webometrics [38] наивысшее среди российских сайтов 54-е место в мире и 18-е – в Европе (всего в рейтинг-лист входят 500 ведущих сайтов университетов и научно-исследовательских центров всего мира), а также ряда связанных с этим сайтом информационных систем.

Литература

[1] Прайс Д. Малая наука, Большая наука // Наука о науке. – М. : Прогресс, 1966. – С. 281–385.

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

[3] Бездушный А.Н., Кулагин М.В., Серебряков В.А., Бездушный А.А., Нестеренко А.К., Сысоев Т.М. Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. – 2005. – Т. 10. – Специальный выпуск. – С. 29–48.

[4] Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т. I. – Вып. 1. – С. 3–11.

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

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

[7] 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.

[8] Krogstie J., Halpin T., Siau K. Information Modeling Methods and Methodologies. – Idea group publishing, 2005.

[9] Staab S., Stuckenschmidt H. (Eds.) Semantic Web and Peer-to-Peer, Decentralized Management and Exchange of Knowledge and Information. – Springer, 2006.

[10] Semantic Web. http://www.w3.org/2001/sw/.

[11] Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н. Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Труды Седьмой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2005). – Ярославль, 2005. – С. 70–79.
[12] Лукашевич Н.В. Описание понятий-ролей в лингвистических и онтологических ресурсах // Сборник тезисов постерных докладов Девятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2007). – Переславль-Залесский, 2007. – С. 81–89.
[13] Яненко Н.Н. Методологические вопросы современной математики // Вопросы философии. – 1981. – № 8. – С.60-68.
[14] Самарский А.А. Задачи прикладной математики на современном этапе развития // Коммунист. – 1983. – № 18. – С. 31–42.
[15] 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.
[16] Желены М. Управление высокими технологиями // Информационные технологии в бизнесе. Энциклопедия. – СПб. : Питер, 2002. – С. 81–89.
[17] Технология // Тезаурус по образованию и педагогике. – Институт информатизации образования в составе Московского государственного гуманитарного университета им. М.А. Шолохова. http://www.mgopu.ru/ininfo/r1_thesaurus.htm#technology.
[18] Schramm W. Information Theory and Mass Communication // In Communication and Culture. N.Y.: Holt, Rinehart & Winston, 1966. P. 521–534.
[19] Шрейдер Ю.А.О семантических аспектах теории информации // Информация и кибернетика. – М. : Советское радио, 1967. – С. 15–47.
[20] Ackoff R., Emery F. On Purposeful Systems. – Ch. – N.Y. : Aldine – Atherton, 1972.
[21] Gitt W. Ordnung und Information in Technik und Natur // In: Gitt W. (Hrsg.): Am Anfang war die Information. Grdfeling: Resch KG, 1982. – S. 171–211.
[22] Словарь по кибернетике. – Киев : Гл. ред. Украинской Советской Энциклопедии им. М.П. Бажана, 1989.
[23] Михайлов А.И., Черный А.И, Гиляревский Р.С. Научные коммуникации и информатика. – М. : Наука, 1976.
[24] Berners-Lee T., Fielding R., Masinter L. Uniform Resource Identifiers (URI). Generic Syntax. RFC 2396. – http://www.ietf.org/rfc/rfc2396.txt/.
[25] Михайлов А.И., Черный А.И, Гиляревский Р.С. Основы информатики. – М. : Наука, 1968.
[26] Холл А.Д., Фейджин Р.Е. Определение понятия системы // Исследования по общей теории систем. – М. : Прогресс, 1969. – С. 252–282.
[27] Барахнин В.Б., Леонова Ю.В., Федотов А.М. К вопросу о формулировке требований для построения информационных систем научно-организационной направленности // Вычислительные технологии. – 2006. – Т. 11. – Специальный выпуск. – С. 52–58.
[28] Барахнин В.Б., Федотов А.М. Информационная система: взгляд на понятие // Вестник НГУ. Сер.: Информационные технологии. – 2007. – Том 5. – Вып.2. – С. 12–19.
[29] Бахвалов Н.С. Численные методы. – М. : Наука, 1970.
[30] Langefors B. Infological models and information user views // Information Systems. – 1980. – № 5. – P. 17–32.
[31] Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям, т. 1. – Аксаково, 2001. – С. 184–188.
[32] Никитина С.Е. Семантический анализ языка науки. – М. : Наука, 1987.
[33] Садовский В.Н. Система // Философский энциклопедический словарь. – М. : Советская Энциклопедия, 1983. – С. 610–611.
[34] Crescenzi V., Mecca G., Merialdo P. Roadrunner: Towards automatic data extraction from large web sites // In: The VLDB Journal. – Rome, 2001. – P. 109–118.
[35] Sahuguet A., Azavant F. Building intelligent web applications using lightweight wrappers // Data Knowledge Engineering. – 2001. – V. 36 – № 3. – P. 283–316.
[36] Барахнин В.Б., Ведерников В.В. Алгоритм автоматической каталогизации статей, опубликованных в электронных версиях научных журналов // Труды Всероссийской научной конференции «Научный сервис в сети Интернет: технологии параллельного программирования». – Новороссийск, 2006. – С. 277–279.
[37] Барахнин В.Б., Нехаева В.А., Федотов А.М. О задании меры сходства для кластеризации текстовых документов // Вестник НГУ. Сер.: Информационные технологии. – 2008. – Т. 6. – Вып. 1. – С. 3–9.
[38] Top 100 R&D European Institutes. – http://www.webometrics.info/top100_r&d_europe.asp

Technologies for construction of intelligence system for data analysis

V.B. Barakhnin, A.M. Fedotov

The paper describes technologies for construction of intelligence system for data analysis.

♣ Работа выполнена при частичной поддержке РФФИ: проекты 06-07-89060, 06-07-89038, 07-07-00271; президентской программы «Ведущие научные школы РФ» (грант № НШ-931.2008.9) и интеграционных проектов СО РАН.