Проблемы технологий создания систем смысловой обработки данных

В.Б.Барахнин, А.М.Федотов

Институт вычислительных технологий СО РАН,
Новосибирский государственный университет

Актуальность исследования

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.

1. Развитие теории переработки информации

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 связаны с другими данными. Как отмечал А.А.Ляпунов, “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.

2. Синтаксическая интероперабельность

Распределенное хранение информации требует интероперабельности (т.е. обеспечения взаимодействия) разнородных информационных источников. Семантическая интероперабельность, заключающаяся в использовании согласованных стандартов метаданных, 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.

3. Разработка методологии вовлечения в научно-информационный процесс слабоструктурированных документов

В то же время разработки в рамках концепции Semantic Web, 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.

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” (Добров и др., 2005).

Возникает проблема вовлечения в научно-информационный процесс слабоструктурированных документов (т.е. документов, у которых значения атрибутов метаданных, как содержательных, так и структурных, не являются элементами заданных словарей).

Комплексное решение указанных проблем возможно лишь при осмыслении процесса обработки компьютерной информации как технологии

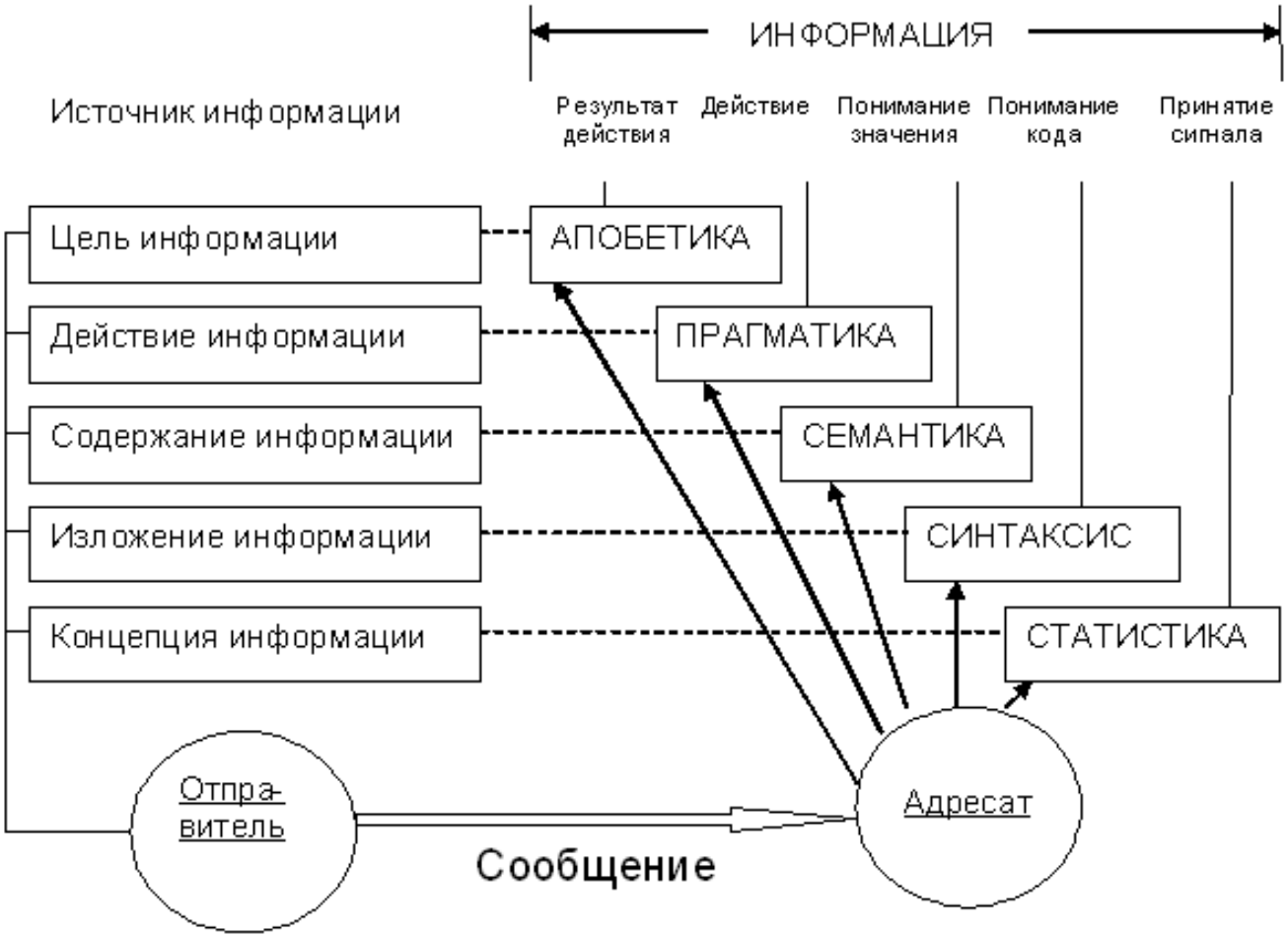
Переработка информации как технология

Будем понимать под технологией определенную последовательность методов обработки, изготовления, изменения состояний и свойств сырья или материалов в процессе производства продукции.

Более кратко: “технология — способ преобразования данного в необходимое”. То есть 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! 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.

Пятиуровневая модель информации (В.Гитт)



Информация: взгляд с точки зрения семиотики

Анализируя пятиуровневую 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).

Важно подчеркнуть, что семиотический подход фактически использован при определении базисных понятий в фундаментальной монографии “Инфосфера” (1996), 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 — прагматическом*.

Интеллектуальные информационные системы как основной инструмент переработки информации

“Инфосфера” (1996):

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

РИС — рассуждающая информационная система,
ИПС — информационно-поисковая система,
ИнИн — интеллектуальный интерфейс,
АП — автоматическое извлечение данных из текстов и пополнение БД.

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

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*.

Выбор основного структурного элемента информационной системы

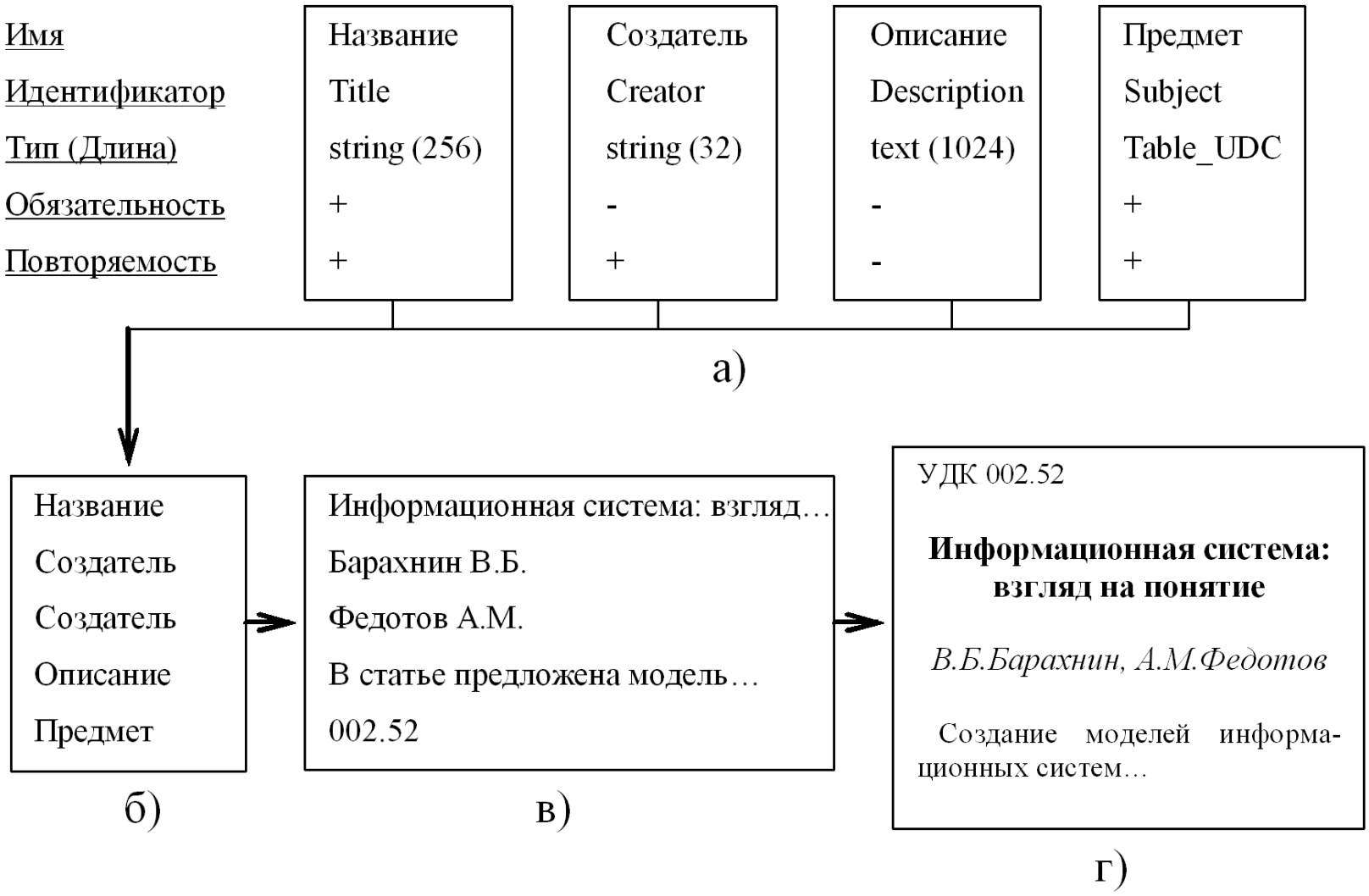
Модель RDF (консорциум W3): элементы суть ресурсы, которые могут представлять и сущности, и их характеристики.
Неудобство: множество равноправных мелких элементов, чрезвычайно много связей, структура далека от естественной.

Модель ИСИР РАН: элементы суть "ресурсы, аналогичные документоподобным объектам".
Неудобство: связи задаются с помощью отношений между типами ресурсов, т.е. связи также имеют внешний характер по отношению к ресурсу.

Модели основаны на концепции Semantic Web и ориентированы 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 (т.е. документов, у которых значения атрибутов метаданных, как содержательных, так и структурных, не являются элементами заданных словарей).

Предлагаемый подход: элементы суть документы (информационные объекты, имеющие, как и всякий ресурс, уникальные идентификаторы, и к тому же обладающие метаданными). Модель позволяет успешно работать со слабоструктурированными документами.

Иерархическая структура метаданных



а) структура, б) атрибуты, в) содержание, г) документ

Совокупность документов как система

Документ может входить в качестве значения некоторого элемента метаданных другого документа. Подробнее, 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.

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 этой связи, являющиеся значениями соответствующих элементов метаданных.

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” основано на тех же понятиях, что и, например, реляционная модель данных).

Использование системного подхода для получение из данных информации

С использованием системного подхода 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, 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 информационной системы 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. для построения модели данных используется комбинация иерархической и реляционной моделей, что сближает ее с инфологическими моделями. Анализ иерархии метаданных, 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).

Получение новых знаний

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. Средством же выражения парадигматических отношений является *онтология* предметной области или ее *тезаурус*, причем граница применения этих терминов весьма размыта. На основании семиотического подхода нами показано, что *тезаурус становится онтологией тогда, когда связи между дескрипторами не просто явно выражены* (как это предусмотрено в классическом определении [**?**]), *но и классифицированы*.

Таким образом, *наличие онтологии (тезауруса)* в качестве составной части информационно-поискового языка, используемого при создании каталога, *является необходимым и достаточным условием возможности получения из данных, уже преобразованных в информацию, новых знаний*.

Наконец, каталог является наиболее естественной формой унификации представления данных, тем самым служа достаточно простым средством решения отмеченной во введении проблемы синтаксической интероперабельности.

Технологии автоматизации обработки слабоструктурированных документов

Были разработаны и реализованы алгоритмы, обеспечивающие автоматизацию основных этапов научно-информационного процесса с участием слабоструктурированных документов:

- автоматизация процесса создания тезаурусов и онтологий, позволяющая проводить начальный этап работы с минимальным привлечением специалистов — экспертов в данной предметной области и вместе с тем обеспечивающая высококвалифицированное описание предметной области с использованием надежно выверенных терминов;
- автоматизация процесса каталогизации слабоструктурированных электронных документов, причем классификационные метаданные, при отсутствии их в самом документе, могут быть получены из удаленной библиографической базы;
- автоматическое координатное индексирование документов с использованием в качестве ключевых слов терминов-словосочетаний (а не отдельных слов);
- автоматизация процесса генерации словоформ для пополнения базовых лексических словарей.
- классификация и кластеризация электронных документов, в т.ч. небольшого объема (аннотаций).

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. Разработаны технологии автоматизации обработки слабоструктурированных документов.

Описанные технологии были использованы при создании Информационно-справочной системе СО РАН www.sbras.ru, занимающей, по данным на июль 2008 г. рейтинга Webometrics, в который входят 500 ведущих сайтов университетов и научно-исследовательских центров всего мира, 1-е место среди российских сайтов (18-е — в Европе, 54-е — в мире).

Спасибо за внимание!