

ПРОБЛЕМЫ РАЗРАБОТКИ ТЕХНОЛОГИИ ФАКТОГРАФИЧЕСКОГО ПОИСКА¹

В.Б.Барахнин, А.М.Федотов

*Институт вычислительных технологий СО РАН, пр. Академика М.А.Лаврентьева, д. 6,
г. Новосибирск, 630090, Россия*

*Новосибирский государственный университет, ул. Пирогова, д. 2,
г. Новосибирск, 630090, Россия,*

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

Аннотация. В статье рассматриваются проблемы разработки технологии извлечения фактографической информации из научных документов достаточно произвольной структуры. Показано, что при создании фактографических информационных целесообразно следующее понимание факта: входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных. Предложена простейшая модель онтологии фактографической системы, работающей с документами достаточно произвольной структуры. Обсуждаются вопросы автоматизированного извлечения фактов из документов и организации взаимодействия фактографических систем с пользователями.

Ключевые слова: фактографические информационные системы, онтология, автоматизированное извлечение фактов

Введение

В классической монографии [1], изданной ВИНТИ и содержащей подробный обзор теоретических проблем фактографического поиска, на основе выделения двух типов информационных потребностей: потребности в сведениях об источниках необходимой научной информации и потребности в самой необходимой научной информации – говорится, что для удовлетворения информационных потребностей первого предназначены информационные системы, получившие название *документальных*, второго типа – *фактографических*. В настоящее время наиболее востребованным средством информационного обеспечения научной деятельности становятся *интеллектуальные системы* (ИнтС), сочетающие возможности информационных систем обоих названных типов и позволяющие удовлетворять информационные потребности квалифицированного пользователя в соответствии со схемой «документ – факт – рассуждение» [2, 3]. В дальнейшем мы будем использовать термин «фактографические системы» в широком смысле, включающем и ИнтС.

Важным этапом процесса функционирования фактографических систем является извлечение из текстов документов содержащихся в них *фактов*, то есть, в наиболее общем

¹ Работа выполнена при частичной поддержке РФФИ (проекты 09-07-00277, 10-07-00302), президентской программы «Ведущие научные школы РФ» (грант НШ 6068.2010.9) и интеграционных проектов СО РАН.

смысле, «особого рода предложений, фиксирующих эмпирич. знание» [4].

К сожалению, указанная задача далека не только от сколько-нибудь удовлетворительного решения, но и от достаточно общей постановки. Одна из основных причин этого заключается в том, что с появлением в конце 1970-х годов персональных компьютеров появились мощные средства визуализации информации, вследствие чего были почти остановлены научные изыскания в области теории создания информационно-поисковых систем, которые возобновились лишь в середине 1990-х в связи с развитием информационных технологий сети Интернет и перехода к распределенному хранению информации. К настоящему моменту в указанной области получены важные теоретические результаты, а также сделан ряд практических шагов по их реализации (см., например, [5, 6]). Эти разработки обычно опираются на неявное предположение о возможности широкого распространения более или менее подробной стандартизации представления информации, например на основе словарей, как это сделано в рамках концепции Semantic Web консорциума W3 [7].

Однако при попытке автоматизировать процесс извлечения фактографической информации из реальных массивов документов, например, размещенных в сети Интернет, использование концепции Semantic Web неизбежно порождает серьезные проблемы, поскольку наработки консорциума W3 носят лишь рекомендательный характер, а объявить их стандартами могут только организации, имеющие соответствующий статус, например ISO, ГОСТ или ANSI. Ввиду этого реальное развитие большинства ресурсов Интернета, в том числе научной направленности, идет без учета подобных необязательных рекомендаций. Более того, свободный характер размещения материалов в сети Интернет превращает требование соблюдения даже обязательных стандартов представления информации всего лишь в благое пожелание (особенно это касается российской части Интернета). Разумеется, сказанное относится еще в большей степени к электронным документам, не размещенным в Интернете и полученным создателями ИнтС для обработки посредством локального доступа.

Таким образом, возникает необходимость разработки технологии извлечения фактографической информации из научных документов достаточно произвольной структуры. Данная статья посвящена обсуждению возникающих при этом проблем.

1. Уточнение понятия «факт»

Прежде чем обсуждать проблемы работы с фактографической информацией, следует уточнить, какое именно содержание мы будем вкладывать в понятие «факт».

К сожалению, в официальных документах: ГОСТ 7.73–96 «Поиск и распространение информации» и ГОСТ 7.74–96 «Информационно-поисковые языки» – этот термин практически не формализован. Так, в ГОСТ 7.74–96 дано лишь косвенное, причем не слишком содержательное, определение факта: «7.7. **фактографическое индексирование:** Индексирование, предусматривающее отражение в поисковом образе документа конкретных сведений (фактов)». Интересно отметить, что иноязычные эквиваленты терминов, относящихся к фактографическому поиску (в отличие от подавляющего большинства прочих терминов), в указанном ГОСТе отсутствуют. Что же касается ГОСТ 7.73–96, то интересующее нас понятие косвенно раскрывается в следующем определении: «3.3.7. **база первичных данных; фактографическая база данных:** База данных, содержащая информацию, относящуюся непосредственно к предметной области».

Подробный анализ значения термина «факт» и его производных, основанный на соответствующих статьях «Философской энциклопедии» и «Словаря современного русского литературного языка», был проведен в монографии [1]. В итоге были выявлены следующие признаки фактов:

1. Факты следует отличать от *данных*, фиксирующих специфику объекта, условия наблюдения и т. п. Понятие же научного факта «предполагает элиминирование такой информации, т. е. требует определенного *обобщения* непосредственных данных». Однако при этом отмечается, что четкого различия между указанными понятиями в «Словаре современного русского литературного языка» не приводится.

2. Фактом можно назвать лишь знание, выдержавшее критическую проверку, то есть полученное в результате обобщения и переработки данных абстрактно-логическим мышлением (разумеется, при этом надо отдавать отчет в том, что достижение абсолютно достоверного знания является лишь идеалом развития науки, практически недостижимым).

3. Любой факт, прежде чем стать объектом научной коммуникации, должен быть преобразован в текст или изображение, получив форму научного документа или его части. Более того, «объектом сбора, хранения, поиска и выдачи в так называемых фактографических информационно-поисковых системах... могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать... любой фрагмент такого текста» [1].

Нетрудно видеть, что сформулированные признаки весьма расплывчаты. Прежде всего, признаки 1 и 2 предполагают обобщение и оценку перерабатываемых данных. Поэтому жесткое соблюдение требований, вытекающих из указанных признаков, выводит работу с фактами за рамки собственно научно-информационной деятельности, поскольку в той или иной степени требует использования теорий и методик конкретных научных дисциплин, к которым относятся данные.

Кроме того, как уже отмечалось выше, очень трудно провести четкую границу между фактами и непосредственными данными. Это касается следующих типов сущностей, описывающих тот или иной объект исследования: имена собственные, хронологические сведения, различные характеристики объектов и т.п. С другой стороны, рассмотрение имен собственных в качестве фактов предполагает, как показано в [1], наличие связей имен собственных с информацией о конкретных носителях этих имен, ибо в противном случае имя несет лишь назывную, но не информационную функцию.

Тенденция стирания граней между понятиями «данные» и «факты» отчетливо проявилась в более современной монографии [2], также изданной ВИНТИ. *Данные* понимаются в ней как факты и идеи, представленные в символической форме, позволяющей проводить их передачу, обработку и интерпретацию, а *информация* – как смысл, приписываемый данным на основании известных правил представления фактов и идей. Структурированная (связанная причинно-следственными и иными отношениями) информация, образующая систему, составляет *знания*.

Для уточнения смысла, вкладываемого в термин «факт» применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний в процессе функционирования ИнтС, представляется целесообразным использование семиотического подхода, подобно тому, как это было сделано в нашей работе [9] для терминов «информация», «знание», «тезаурус», «онтология». В этой работе, в частности, показано, что данные соответствуют синтаксическому уровню сообщения (в том числе документа), информация (в узком смысле!) – семантическому, а знания – прагматическому. Отсюда вытекает, 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.

Следовательно, в качестве «первичного» факта (т.е. содержащегося непосредственно в тексте документа) рассматривается некоторая семантическая информация, но в справочно-информационный фонд ИнтС факт заносится в качестве элемента данных, что соответствует уже упоминавшемуся соотношению данных и фактов из монографии [2].

Но какого рода информация может быть занесена в справочно-информационный фонд системы в виде данных? Ведь сами по себе данные не несут никакой информационной ценности без соответствующих моделей: например, А.Н.Колмогоров неоднократно отмечал, что данные представляют информационную ценность лишь тогда, когда они являются составной частью некоторой модели реального мира и связаны с другими данными [10, 11]. Таким образом, применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А.А.Ляпунов (см., например, [12]): «нет модели – нет информации».

В качестве модели предметной области обычно выступает ее *онтология* (какая именно смысл мы вкладываем в это весьма широко трактуемое понятие – будет уточнено в следующем разделе).

Таким образом, при создании фактографических информационных разумно следующее понимание факта: *входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных.*

Отсюда, в частности, вытекает следующее важное замечание: именно онтология фактографической системы определяет, что будет считаться фактом в рамках этой системы. Здесь мы имеем дело с ситуацией, столь характерной для естественных наук, о которой говорил, например, А.Эйнштейн в своей известной беседе с В.Гейзенбергом: «Только теория решает, что можно наблюдать» [13].

2. Особенности онтологий для фактографических систем

Прежде всего, уточним, какого именно понимания термина онтология мы будем придерживаться в данной работе.

В работе [9] нами было проведено (применительно к рассматриваемой предметной области) установление определенности в понимании и разграничении использования терминов «тезаурус» и «онтология». Более или менее однозначное трактование термина «тезаурус» сложилось еще в конце 1960-х годов [14]: это «словарь-справочник, содержащий все лексические единицы информационно-поискового языка – дескрипторы (вместе с ключевыми словами, которые в пределах данной информационно-поисковой системы считаются синонимами этих дескрипторов), причем дескрипторы в словаре должны быть систематизированы по смыслу, а смысловые связи между ними эксплицитно выражены».

Что же касается термина «онтология», в настоящее время, как отмечено в [15], под онтологией нередко стали понимать широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [16]:

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

Нами было показано [9], что тезаурус становится онтологией тогда, когда связи между дескрипторами не просто эксплицированы (как это предусмотрено в классическом определении тезауруса), но и классифицированы универсальными зависимостями типа «общее – частное», «часть – целое», «причина – следствие» и т.п. (см., например, [17]). Разумеется, это – лишь «нижняя граница» сложности онтологии. Для эффективной работы с фактами следует, чтобы сущности, относящиеся к предметной области, были представлены

не только обозначающими их терминами, но и достаточно широким набором атрибутов, т.е. речь идет об онтологии, обладающей известными признаками модели предметной области.

Разумеется, на первоначальном этапе создания ИнтС речь, как правило, идет о создании лишь каркаса онтологии, содержащего только краткие сведения о сущностях, а их более подробное описание будет происходить в процессе функционирования ИнтС посредством извлечения из документов соответствующих фактов, выступающих в качестве тех или иных атрибутов сущностей. При этом следует хранить и библиографическую ссылку на информационный источник, из которого был извлечен данный факт.

Поскольку, как уже отмечалось выше, «объектом сбора, хранения, поиска и выдачи в так называемых фактографических информационно-поисковых системах... могут быть лишь соответствующие тексты или документы, описывающие некоторые данные или факты, если под документом понимать... любой фрагмент такого текста» [1], постольку в роли онтологии – модели предметной области – может выступать та или иная модель интеллектуальной информационной системы информационной системы, например предложенная нами в работе [18]. Эта модель, записанная в качестве модели предметной области, имеет вид

$$S = \langle K, M, M^j \langle K_i, K_i \rangle \rangle,$$

где K — классы сущностей, M — множество используемых атрибутов сущностей, $M^j \langle K_i, K_i \rangle$ — типы возможных связей между классами сущностей, когда сущность из класса K_i может входить в качестве значения атрибута M^j сущности из класса K_i . Тем самым любая сущность s_i представляется как

$$d_i = \langle m_i^{j,k} \rangle,$$

где $m_i^{j,k}$ — значения атрибутов сущности, k — количество значений (с учетом повторений) j -го атрибута в описании сущности.

При создании информационной системы сущности будут представлены в виде описывающих их документов, а атрибуты сущностей будут представлять собой элементы метаданных.

Предложенная модель онтологии имеет довольно простой вид, что делает ее наиболее пригодной для первоначального этапа создания фактографической системы. Разумеется, она может быть расширена и усложнена, например, путем явного введения необходимых ограничений на атрибуты сущностей (которые можно назвать иначе: *аргументы фактов*) с использованием структуры, называемой *схемой фактов*, описанной в [19, 20].

Схема факта F_k представляется в виде $\langle Arg, Rs, C \rangle$, где Arg – множество дескрипторов аргументов факта (в качестве дескриптора может выступать быть тип словарной единицы, класс информационного объекта и т.п.); $Rs = \langle t, op(t), P \rangle$ – результат применения схемы, где t задает класс объекта, – тип операции, применяемой при условии выполнения ограничения C , P – множество правил для формирования или редактирования объекта, каждое из которых ставит в соответствие атрибуту результирующего объекта либо точное значение атрибута, либо значение атрибута одного из аргументов; наконец, C – множество ограничений (морфологических, синтаксических, семантических, структурно-текстовых), накладываемых на характеристики аргументов факта.

Отметим, что применительно к фактографическим информационным системам, создаваемым систем в рамках концепции Semantic Web, довольно близкий подход был предложен в работе [5]. Речь идет об использовании модели, в которой сущности внешнего мира представляются атрибутированными информационными единицами, а отношения между сущностями реализуются либо в виде прямых ссылок, либо в виде составных

конструкций определенного вида, при этом спецификация такой модели воплощается в виде онтологии.

3. Автоматизированное извлечение фактов из документов

Разработка методик автоматизированного извлечения фактов из документов представляет собой наиболее сложную проблему, возникающую при создании фактографических систем. Это было подчеркнуто еще в [1]: «не существует сколько-нибудь значительных различий в теории и методике построения документальных и фактографических информационно-поисковых систем, если фактографический поиск понимать лишь как процесс отыскания уже готовых данных и фактов, ранее введенных в фактографическую систему... Однако под фактографическим поиском можно понимать и нечто принципиально иное, а именно отыскание машиной требуемых данных и фактов в текстах научных документов, написанных на одном или нескольких разных естественных языках, ...[что] требует оперирования со смыслом текстов его анализа и синтеза, т.е. моделирования достаточно сложных мыслительных процессов».

Собственно говоря, в середине 1970-х годов возможности компьютеров были явно недостаточными для сколько-нибудь полноценного практического решения поставленной задачи. К настоящему моменту рост мощности компьютеров позволил создавать разнообразные алгоритмы для извлечения данных и фактов из документов на естественных языках. Выбор конкретного алгоритма (или, точнее, даже типа алгоритмов) зависит от того, насколько структурированы (и структурированы ли вообще) данные и факты, содержащиеся в конкретном документе.

1. Табличные данные. Они могут выступать, согласно [1], в качестве фактов, если являются, например, характеристиками предметов, географических объектов и т.п. Для их извлечения из документов существуют разнообразные, весьма надежные алгоритмы (см., в частности, [21], включая библиографический обзор).

2. Массивы однородных слабоструктурированных текстовых документов. Нередко первоначальный этап создания онтологий удобно проводить, занося факты, содержащиеся в массивах однородных документов, описывающих предметную область: биографических справочниках, геологических, ботанических или зоологических каталогах и т.п. В таких случаях наиболее целесообразно использовать алгоритмы, учитывающие информацию о закономерностях их текстовой структуры (например, общих для всех документов массива синтаксических и семантических конструкций), а также о гипертекстовой разметке обрабатываемых документов (при наличии таковой). Такой алгоритм, извлекающий факты (метаданные) о библиографии документов, подробно описан, например, в нашей монографии [3]. Он может быть легко адаптирован к фактографической информации произвольного характера, содержащейся в массивах документах, имеющих более или менее однородную текстовую структуру.

3. Тексты произвольного характера. Задача извлечения фактов из произвольных текстов на естественном языке до сих пор, по-видимому, не имеет сколько-нибудь общего решения, поскольку построение такого решения предполагает, в частности, достаточно точное моделирование когнитивной деятельности человека, а также наличие мощных средств как синтаксического, так и семантического анализа текстов, включая подробнейшие онтологии, тезаурусы которых учитывают, например, всё богатство синонимии естественного языка (не столько даже в части научной лексики, сколько в части лексики общеупотребительной).

«Частное решение» этой задачи применительно к той или иной предметной области предполагает, прежде всего, построение онтологии, тезаурус которой включает, наряду с описанием сущностей предметной области, по крайней мере, те пласты

общеупотребительной лексики (разумеется, с учетом синонимии), которые наиболее характерны для данной области.

Непосредственная работа по извлечению фактов из текста может опираться на совокупном применении методов синтаксического и семантического анализа. Например, общедоступным средством синтаксического анализа текстов является стимер компании «Яндекс» (<http://company.yandex.ru/technology/mystem/>), позволяющий извлекать словосочетания заданной структуры, например, *(прилагательное) + (существительное)* или *(существительное) + (существительное в родительном падеже)*. Для семантического анализа текстов может быть применен подробно описанный в [3] алгоритм выявления в тексте терминов, в т.ч. и составных, входящих в словарь онтологии данной предметной области. Само же извлечение факта, относящегося к тому или иному упоминаемому в тексте субъекту, описанному в онтологии, состоит в определении значения предиката, связанного с этим субъектом (описание подробностей конкретной реализации алгоритмов синтаксического и семантического анализа выходит за рамки данной статьи).

4. О взаимодействии фактографических систем с пользователями

Факты, извлеченные из текстов документов, и занесенные в фактографическую информационную систему, могут быть использованы как для дальнейшего получения новых знаний (что, собственно, и характеризует интеллектуальные системы), так и для непосредственного поиска пользователем системы. При этом нередко в качестве чуть ли не постоянного атрибута качественной фактографической системы называют возможность формулировки запроса на естественном языке. Однако из вышеизложенного, на наш взгляд, вытекает вывод о том, что такая функция не дает пользователем специализированных систем каких-то принципиальных удобств. Действительно, коль скоро мы рассматриваем в качестве фактов характеристики сущностей, описанных в онтологии, то весьма несложный интерфейс, позволяющий просматривать онтологию посредством использования последовательности гиперссылок (или даже посредством таблицы), сможет предоставить пользователю возможность без труда найти нужный факт или, по крайней мере, убедиться в том, что этот факт не занесен в систему. С другой стороны, задача «понимания» системой запросов на естественном языке практически эквивалентна задаче извлечения фактов из текстов на естественном языке, о трудностях в решении которой нами сказано выше. При этом следует учесть, что далеко не все пользователи (пусть даже являющиеся высококвалифицированными специалистами в своей предметной области) способны формулировать свой вопрос так четко и недвусмысленно, как, согласно стихотворению проф. А.С.Компанейца, это умел делать на своем знаменитом семинаре в Институте физических проблем АН СССР Л.Д.Ландау (цит. по [22]):

*С первых слов, как Вельзевул во плоти,
Навалился Дау на него:
«Лучше вы скажите, что в работе
Ищется как функция чего?»*

Слишком же расплывчатая постановка вопроса, «не распознанная» информационной системой, может привести к тому, что у пользователя сложится ошибочное мнение, будто бы система не располагает необходимой ему информацией. Таким образом, непосредственный просмотр онтологии представляется наиболее надежным путем получения требуемой фактографической информации.

Еще одним важным примером практического использования фактографических систем в научных издательствах и редакциях журналов может служить проверка

достоверности сведений, содержащихся в рукописях, имеющих биографический, научно-публицистический, обзорный и т.п. характер. Факты, извлекаемые из текста рукописей, подвергаются сравнению с «эталонными» фактами из онтологии информационной системы, и в случае расхождения редакция просит автора уточнить правильность приведенных им сведений.

Заключение

В данной работе намечены основные шаги в направлении разработки технологии обработки фактографической информации, содержащейся в научных документах достаточно произвольной структуры. Показано, что при создании фактографических информационных целесообразно следующее понимание факта: **входящая в текст документа характеристика сущности, описываемой в онтологии информационной системы, представляемая как единичное значение данных**. Предложена простейшая модель онтологии фактографической системы.

Важным этапом практической реализации предлагаемых в статье подходов должна стать реализация алгоритмов синтаксического и семантического анализа текстов с целью извлечения фактов.

Литература

- [1] Михайлов А.И., Черный А.И., Гиляревский Р.С. Научные коммуникации и информатика. – М: Наука, 1976.
- [2] Арский Ю.М., Гиляревский Р.С., Туров И.С., Черный А.И. Инфосфера: Информационные структуры, системы и процессы в науке и обществе. – М.: ВИНТИ, 1996.
- [3] Шокин Ю.И., Федотов А.М., Барахнин В.Б. Проблемы поиска информации. – Новосибирск: Наука, 2010.
- [4] Ракитов А. Факт // Философская энциклопедия. – М: Советская энциклопедия, 1970. – Т. 5. – С. 298.
- [5] Марчук А.Г. О распределенных фактографических системах // Труды Десятой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008). Дубна, 7-11 октября 2008 г. – С.93-102.
- [6] Марчук А.Г., Марчук П.А. Архивная фактографическая система// Труды Одиннадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2009). – Петрозаводск, 17-21 сентября 2009 г. – С. 177-185.
- [7] Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. – 2001. – Vol. 284(5). – P. 34-43.
- [8] Шокин Ю.И., Федотов А.М., Барахнин В.Б. Технология создания программных систем информационного обеспечения научной деятельности, работающих со слабоструктурированными документами // Вычислительные технологии. – 2010 – Т. 15. – № 6. – С. 111-125.
- [9] Барахнин В.Б., Федотов А.М. Уточнение терминологии, используемой при описании интеллектуальных информационных систем, на основе семиотического подхода // Известия вузов. Проблемы полиграфии и издательского дела. – 2008. – № 6 – С.73-81.
- [10] Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. –Т. 1. – Вып. 1. – С.3-11.
- [11] Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987.

- [12] *Ляпунов А.А.* О соотношении понятий материя, энергия и информация // В кн.: А.А. Ляпунов Проблемы теоретической и прикладной кибернетики. – Новосибирск: Наука, 1980. – С. 320-323.
- [13] Heisenberg W. Der Teil und das Ganze. Gespräche im Umkreis der Atomphysik. – München, 1976.
- [14] *Михайлов А.И., Черный А.И., Гиляревский Р.С.* Основы информатики. – М.: Наука, 1968.
- [15] *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.
- [16] *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. – 1999.
- [17] *Нариньяни А.С.* Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. – Аксаково, 2001. – Т. I. – С. 184-188.
- [18] *Баракнин 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.
- [19] *Сидорова Е.А.* Онтологический подход к представлению знаний для задачи анализа текстовых ресурсов // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–07), Новосибирск, 2007. – Т. 1. – С. 221-228.
- [20] *Загорулько Ю.А., Сидорова Е.А., Боровикова О.И.* Онтологический подход к построению систем информационной поддержки научной и производственной деятельности // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–09), Новосибирск, 2009. – Т. 2. – С. 93-102.
- [21] *Бычков И.В., Ружников Г.М., Хмельнов А.Е., Шигаров А.О.* Эвристический метод обнаружения таблиц в разноформатных документах // Вычислительные технологии. – 2009. – Т. 14. – № 2. – С. 58-73
- [22] *Горобец Б.С.* Советские физики шутят... Хотя бывало не до шуток. – М: Книжный дом «ЛИБРОКОМ», 2010.