УДК 004

В. Б. Барахнин, **А. М. Федотов**

Институт вычислительных технологий СО РАН
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия

Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия

E-mail: bar@ict.nsc.ru; fedotov@sbras.ru

ПОСТРОЕНИЕ МОДЕЛИ ФАКТОГРАФИЧЕСКОГО ПОИСКА *

Рассматриваются теоретические вопросы фактографического поиска, а также разработки технологии извлечения фактографической информации из научных документов достаточно произвольной структуры. Показано, что при создании фактографических информационных систем целесообразно следующее определение факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы. Предложена 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: факт, фактографический поиск, интеллектуальные системы, извлечение фактов.

Введение

В классической монографии [1], изданной ВИНИТИ 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 само́й необходимой научной информации – говорится, что для удовлетворения информационных потребностей первого типа предназначены информационные системы, получившие название *документальных*, второго типа – *фактографических*. В настоящее время наиболее востребованным средством информационного обеспечения научной деятельности становятся *интеллектуальные информационные системы* (ИИС), 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 понятие «фактографические системы» в широком смысле, включающем и ИИС.

Важным этапом процесса функционирования фактографических систем является извлечение из текстов документов содержащихся в них *фактов*, т. е. в наиболее общем смысле, «особого рода предложений, фиксирующих эмпирическое знание» [4]. К сожалению, указанная задача далека не только от сколько-нибудь удовлетворительного решения, но и от четкой формальной постановки. Одна из основных причин этого заключается в том, что с появлени-

* Работа выполнена при частичной поддержке РФФИ (проекты № 11-07-00561, 12-07-00472, 13-07-00258), президентской программы «Ведущие научные школы РФ» (грант НШ 6293.2012.9) и интеграционных проектов СО РАН.

Барахнин В. Б., Федотов А. М. Построение модели фактографического поиска // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. Т. 11, вып. 4. С. 16–27.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2013. Том 11, выпуск 4

ем в конце 1970-х гг. персональных компьютеров возникли мощные средства визуализации информации, вследствие чего резко упал интерес к научным изысканиям в области теории создания информационно-поисковых систем.

Отрицательное воздействие на развитие новых алгоритмов обработки фактографической информации оказал японский проект «компьютеров пятого поколения», стартовавший в начале 1980-х гг. в Японии, который активно подхватили США, СССР, Великобритания и структуры Европейского сообщества. В процессе реализации этого проекта предполагалось, в частности, разработать технологии логических заключений для обработки знаний, способные делать логические выводы из представленных фактов, хранящихся в сверхбольших базах данных и базах знаний, при этом предусматривалась параллельная обработка данных. Доступ к данным должен был осуществляться с помощью языка логического программирования Пролог. Кроме того, планировалось реализовать поиск характерных признаков массивов данных, автоматическое реферирование текстов на естественном языке и т. п. Требуемое для решения поставленных задач резкое увеличение производительности предполагалось достигнуть путем замены программных решений на аппаратные, что означало приостановку теоретических исследований в области фактографического поиска.

Однако в 1992 г. проект завершился, не достигнув цели. Среди множества причин провала проекта мы остановимся лишь на тех, которые связаны с разработкой программного обеспечения. Прежде всего, возможности решения задач в области искусственного интеллекта были переоценены, разработчики питали ничем не обоснованную надежду на то, что возможно создание системы искусственного интеллекта, которая, будучи реализованной на компьютере достаточно большой мощности, будет способна к самоорганизации, проявляющейся, в частности, в самостоятельном (не зависящем от человека) изменении внутренних правил и параметров системы. Эта идея оказалась непродуктивной: система, которой было позволено «самоорганизовываться», быстро утрачивала целостность и начинала проявлять неадекватную реакцию. Наконец, сделанная в процессе реализации проекта ставка на развитие преимущественно аппаратных решений в ущерб программным оказалась ошибочной: аппаратные средства неоправданно усложнялись, а развитие и совершенствование алгоритмов резко затормозилось. Но окончательно похоронило «японский проект компьютеров пятого поколения» появление Интернета, приведшее к возникновению принципиально новой парадигмы распределения и хранения данных. Таким образом, научные изыскания в области теории создания информационно-поисковых систем возобновились лишь в середине 1990-х гг. 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 ИИС 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.

Уточнение понятия «факт»

Прежде чем обсуждать проблемы работы с фактографической информацией, следует уточнить, какое именно содержание мы будет 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 ГОСТе 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, т. е. 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.

К тому же, как уже отмечалось, 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. Например, даже такой, казалось бы, бесспорный факт, как то, что «температура кипения воды равна 100 °С», неявно предполагает указание на условия наблюдения, например химическую чистоту воды и давление в 1 атм, причем последнее условие нельзя заменить на более абстрактное «стандартное атмосферное давление», поскольку в химии таковым согласно решению Международного союза теоретической и прикладной химии (ИЮПАК) считается давление 100 кПа, меньшее 1 атм, и при «стандартном давлении» температура кипения воды несколько меньше 100 °С.

Еще больше проблем возникает в области гуманитарных наук, в частности истории, где некое утверждение, снабженное ссылкой на источник информации, нередко становится новым утверждением, являющимся предметом изучения источниковедения. При этом если исходное высказывание может быть спорным и не являться историческим фактом (например, «Император Александр Первый и старец Фёдор Кузьмич – одно и то же лицо»; о том,

что данное высказывание отнюдь не относится к «лженаучным», а заслуживает, по крайней мере, серьезного обсуждения, см. монографию [8]), то утверждение со ссылкой может являться фактом источниковедения («Князь Н. С. Голицын опубликовал версию о том, что император Александр Первый и старец Фёдор Кузьмич – одно и то же лицо, в журнале "Русская старина", 11 книга, 1880 г.»).

Наконец, рассмотрение в качестве фактов имен собственных предполагает, как показано в [1], наличие связей имен собственных с информацией о конкретных носителях этих имен, так как в противном случае имя несет лишь назывную, но не информационную функцию.

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

Для уточнения смысла, вкладываемого в термин «факт» применительно к той области информатики, которая изучает процессы взаимных преобразований данных, информации и знаний в процессе функционирования ИИС, представляется целесообразным использование семиотического подхода.

Отметим, что на такое понимание факта обратил внимание Л. Витгенштейн [1]. Понятие «факт» является центральным в его «Логико-философском трактате» [9], одним из источников которого, как отметил Витгенштейн предисловии трактата, стали работы Г. Фреге – основателя семиотики. Процитируем основные положения трактата, касающиеся фактов:

«…1.1. Мир есть совокупность фактов, а не вещей.

...

1.2. Мир распадается на факты.

1.21. Любой факт может иметь место или не иметь места, а все остальное останется тем же самым.

....

2. То, что имеет место, что является фактом, – это существование атомарных фактов.

2.01. Атомарный факт есть соединение объектов (вещей, предметов).

2.011. Для предмета существенно то, что он может быть составной частью атомарного факта.

…

2.034. Структура факта состоит из структур атомарных фактов.

2.04. Совокупность всех существующих атомарных фактов есть мир.

2.05. Совокупность всех существующих атомарных фактов определяет также, какие атомарные факты не существуют.

2.06. Существование или несуществование атомарных фактов есть действительность. (Существование атомарных фактов мы также называем положительным фактом, несуществование – отрицательным.)

2.061. Атомарные факты независимы друг от друга.

2.062. Из существования или несуществования какого-либо одного атомарного факта нельзя заключать о существовании или несуществовании другого атомарного факта.

…

4.21. Простейшее предложение, элементарное предложение, утверждает существование атомарного факта.

…

4.22. Элементарное предложение состоит из имен. Оно есть связь, сцепление имен».

[1] Авторы выражают признательность Ю. В. Леоновой, обратившей внимание на определение факта в «Логико-философском трактате» Л. Витгенштейна.

Положения, выдвинутые в «Логико-философском трактате», имеют большое значение для семиотики, в частности, потому, что в нем устанавливается полное соответствие между онтологическими и семантическими понятиями [10]. Кроме того, Витгенштейн не исключает ложные (или, если угодно, представляющиеся на данном уровне познания ложными) утверждения из числа атомарных фактов, а называет такие факты несуществующими.

Нетрудно заметить, что процитированные положения «Логико-философского трактата» (прежде всего, ключевые определения из раздела 2.01: «Атомарный факт есть соединение объектов (вещей, предметов)... Структура факта состоит из структур атомарных фактов») практически полностью воспроизводятся в модели данных «сущность – связь» [11], являющейся основой для унификации различных представлений данных (при этом следует отметить, что в статье [11] для обозначения связи между сущностями не используется термин «факт», а в ее библиографическом списке отсутствует ссылка на «Логико-философский трактат»).

Для единообразия определения понятия «факт» удобно использовать модификацию модели данных «сущность – связь» из той же статьи, называемую моделью множества сущностей. Ее отличительные особенности заключаются в том, что, во-первых, в ней все трактуется как объекты (в том числе, например, цвет, в то время как в модели «сущность – связь» цвет обычно трактуется как «значение», а согласно «Логико-философскому трактату» «2.0251. Пространство, время и цвет (цветность) есть формы объектов»), а во-вторых, все связи в этой модели – бинарные. Связи между объектами в модели множества сущностей также рассматриваются как объекты, связанные, в свою очередь, с объектами – атрибутами связей.

Важно подчеркнуть, что создание фактографических систем подразумевает извлечение фактов не только непосредственно из текста документа, но и из его метаданных. Это следует, например, из традиционного понимания научно-информационного процесса [12], 2-й этап которого (аналитико-синтетическая переработка документальной информации) предусматривает как извлечение сведений о содержании документа (индексирование, аннотирование и т. п.), так и обработку его библиографических данных.

Заметим, что указание источника, из которого извлечен данный факт, в качестве одного из атрибутов факта позволяет с той или иной степенью достоверности отделять «существующие» (в терминологии Витгенштейна) факты от «несуществующих». С этой целью на множестве источников может быть введена шкала их достоверности.

В некоторых случаях целесообразно извлекать и факты, касающиеся не только семантического, но и синтаксического уровня текста. В частности, при анализе поэтических текстов [13] исследуются их метрические, ритмические и фонетические характеристики. При этом они могут представлять не только непосредственный интерес, но и использоваться для установления фактов, касающихся, например, авторства документов. Так, Д. С. Самойлов [14], проанализировав особенности рифм одной из версий продолжения X главы «Евгения Онегина», полностью исключил авторство Пушкина, поскольку в этом тексте процент рифм с совпадающими опорными согласными в несколько раз превышает этот показатель в произведениях Пушкина.

Однако всякий ли факт, содержащийся в тексте или метаданных документа, обрабатываемого ИИС с целью извлечения из него фактов, представляет интерес с точки зрения создателей и пользователей данной ИИС? Чтобы ответить на этот вопрос, формализуем введенное понятие факта подобно тому, как это было сделано в нашей работе [15] 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 ИИС 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 рассматривается некоторая информация (как правило, семантическая; примеры возможных исключений приведены выше), но в справочно-информационный фонд ИИС факт заносится в качестве совокупности элементов дан-

ных, описывающих сущности и связи между ними, что соответствует уже упоминавшемуся соотнесению 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 часть некоторой модели реального мира и связаны с другими данными [16; 17]. Таким образом, применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А. А. Ляпунов (см., например, [18]): «нет модели – нет информации».

В качестве модели предметной области обычно выступает ее *онтология* (какой именно смысл мы вкладываем в это весьма широко трактуемое понятие, будет уточнено далее).

Таким образом, при создании фактографических информационных систем разумно следующее определение факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы.

Отсюда, в частности, вытекает следующее важное замечание: именно онтология фактографической системы определяет, что́ 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 известной беседе с В. Гейзенбергом: «Только теория решает, что можно наблюдать» [19].

Особенности онтологий для фактографических систем

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

В работе [15] установили некоторую определенность в понимании и разграничении использования терминов «тезаурус» и «онтология» (применительно к рассматриваемой предметной области). Более или менее однозначное трактование термина «тезаурус» сложилось еще в конце 1960-х гг. [12]: 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 [20], под онтологией нередко стали понимать широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [21]:

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.

Мы показали [13], 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, [22]). 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 ИИС речь, как правило, идет о создании лишь каркаса онтологии, содержащего только краткие сведения о сущностях, а их более подробное описание происходит в процессе функционирования ИИС 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, «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 [23]. Эта модель, записанная в качестве модели предметной области, имеет вид

$$S = \left\langle K, M, M^j \left\langle K_i, K_{i'} \right\rangle \right\rangle,$$

где K – классы сущностей; M – множество используемых атрибутов сущностей; $M^j \left\langle K_i, K_{i'} \right\rangle$ – 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 может быть представлена как описывающий ее документ

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

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

При создании фактографической информационной системы сущности могут быть представлены в виде описывающих их документов, а в качестве атрибутов сущностей выступают элементы метаданных.

Разумеется, пользуясь знаниями о предметной области, возможно и целесообразно накладывать различные ограничения (морфологические, синтаксические, семантические, структурно-текстовые) на характеристики сущностей, входящих в те или иные классы (подробно принципы установления ограничений описаны в [24]).

Подчеркнем, что предложенная модель фактографической системы отличается от модели документальной информационной системы, описанной нами в работах [3; 23]. Основными структурными элементами документальной информационной системы являются *документы*, понимаемые применительно к данной ситуации как информационные ресурсы, имеющие уникальный идентификатор и обладающие метаданными. В фактографических же системах, основанных, как отмечалось выше, на модели множества сущностей, элементы суть *объекты*, при этом под объектами понимаются и характеристики стержневых сущностей, и связи между объектами. Аналогичный подход встречается в модели RDF [2] консорциума W3, предлагающей рассматривать в качестве элементов системы ресурсы, которые могут представлять и сущности, и их характеристики. Описанный подход, как отмечалось в [3], весьма неудобен для работы с документальной информацией, но полностью соответствует введенному нами пониманию факта, что делает его наиболее пригодным для создания фактографических систем.

Отметим, что применительно к фактографическим информационным системам, создаваемым 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.

[2] Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. URL: http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/.

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 гг. 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 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, [25], 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 конструкциях), 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 совокупное применение методов синтаксического и семантического анализа. Например, общедоступным средством анализа текстов является стеммер (морфологический анализатор) компании «Яндекс» [3], позволяющий извлекать словосочетания заданной структуры, например, (*прилагательное*) + (*существительное*) или (*существительное*) + (*существительное в родительном падеже*), т. е. проводить не только морфологический, но и синтаксический анализ. Для се-

[3] http://company.yandex.ru/technologies/mystem/

мантического анализа текстов может быть применен подробно описанный в [3] алгоритм выявления в тексте терминов, в том числе 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).

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 зачастую чуть ли не непременным атрибутом качественной фактографической системы называют возможность формулировки запроса на естественном языке. Однако из изложенного выше, на наш взгляд, вытекает вывод о том, что такая функция не дает пользователям 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: [26]):

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 конкретной фактографической информации.

Разумеется, возможна и усложненная постановка задачи, когда пользователю требуются не только (или даже не столько) сами факты, но и их анализ, обобщение и т. п. Для решения этой задачи требуются такие компоненты ИИС [2], как рассуждающая информационная система, формализующая правила логического вывода, и интеллектуальный интерфейс (диалог, графика и т. д.).

Таким образом, функционирование фактографических информационных систем как частного случая ИИС основано на двух противоположных процессах: при пополнении фактографической системы новыми фактами происходит преобразование семантической информации в данные, однако непосредственно потребности пользователя удовлетворяет обратный процесс – извлечение из данных нужной пользователю информации и знаний.

Заключение

В данной статье разработана и обоснована концепция построения интеллектуальной системы для поиска фактографической информации, содержащейся в научных документах достаточно произвольной структуры. Показано, что при создании фактографических информа-

ционных систем целесообразно следующее определение факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы. Предложена модель онтологии фактографической системы.

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

Примером практического использования фактографических систем может служить система исследования текстов диссертаций и авторефератов с целью изучения структуры научных связей ученого (научное окружение ученого), структуры и динамики развития научных коллективов (научные школы) и т. п. [27; 28]. Такие исследования дают возможность изучения и оценивания тенденций развития различных научных направлений, взаимосвязей между отдельными сообществами, а также идентификации персон, научных центров и организаций, научных школ. С использованием тезаурусного метода контент-анализа из текстов диссертаций и авторефератов извлекаются именованные сущности (персоны, организации, географические наименования), временны*е* характеристики, ключевые термины, а также выявляются связи между ними, т. е. устанавливаются факты, касающиеся названных сущностей.

Список литературы

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.: ВИНИТИ, 1996.

3. *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010.

4. *Ракитов А.* Факт // Философская энциклопедия. М.: Сов. энциклопедия, 1970. Т. 5. С. 298.

5. *Марчук А. Г.* О распределенных фактографических системах // Тр. X Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2008). Дубна, 2008. С. 93–102.

6. *Марчук А. Г., Марчук П. А.* Архивная фактографическая система // Тр. XI Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL'2009). Петрозаводск, 2009. С. 177–185.

7. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American. 2001. Vol. 284 (5). P. 34–43.

8. *Сахаров А. Н.* Александр I. М.: Наука, 1998.

9. *Витгенштейн Л.* Логико-философский трактат. М.: Изд-во иностр. лит., 1958.

10. *Грязнов А. Ф.* Витгенштейн. Новая философская энциклопедия. М.: Мысль, 2000. Т. 1. С. 406–408.

11. *Чен П. П.-Ш.* Модель «сущность – связь» – шаг к единому представлению данных // СУБД. 1995. № 3. С. 137–158.

12. *Михайлов А. И., Черный А. И., Гиляревский Р. С.* Основы информатики. М.: Наука, 1968.

13. *Барахнин В. Б., Кожемякина О. Ю.* Об автоматизации комплексного анализа русского поэтического текста // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XIV Всерос. науч. конф. Переславль-Залесский, 2012. С. 213–217.

14. *Самойлов Д. С.* Книга о русской рифме. М.: Худож. лит., 1982.

15. *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.

16. *Колмогоров А. Н.* Три подхода к определению понятия «количество информации» // Проблемы передачи информации. 1965. Т. 1, вып. 1. С. 3–11.

17. *Колмогоров А. Н.* Теория информации и теория алгоритмов. М.: Наука, 1987.

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

19. *Heisenberg W.* Der Teil und das Ganze. Gespräche im Umkreis der Atomphysik. München, 1976.

20. *Добров Б. В., Лукашевич Н. В., Синицын М. Н., Шапкин В. Н.* Разработка лингвистической онтологии по естественным наукам для решения задач информационного поиска // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. VII Всерос. науч. конф. Ярославль, 2005. С. 70–79.

21. *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert Systems All Over Again // AAAI-1999 Invited Panel Presentation. 1999.

22. *Нариньяни А. С.* Кентавр по имени ТЕОН: тезаурус + онтология // Диалог'2001 по компьютерной лингвистике и ее приложениям: Тр. междунар. семинара. Аксаково, 2001. Т. 1. С. 184–188.

23. *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. Спец. выпуск. С. 52–58.

24. *Сидорова Е. А.* Онтологический подход к представлению знаний для задачи анализа текстовых ресурсов // Знания – Онтологии – Теории: Материалы Всерос. конф. с междунар. участием. Новосибирск, 2007. Т. 1. С. 221–228.

25. *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.

26. *Горобец Б. С.* Советские физики шутят… Хотя бывало не до шуток. М.: ЛИБРОКОМ, 2010.

27. *Барахнин В. Б., Федотов А. М., Федотова О. А.* Электронная библиотека по научному наследию как фактографическая система // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XV Всерос. науч. конф. Ярославль, 2013. С. 91–97.

28. *Леонова Ю. В., Федотов А. М.* Извлечение знаний и фактов из текстов диссертаций и авторефератов // Системный анализ и информационные технологии: Тр. V Междунар. конф. Красноярск, 2013. Т. 2. С. 85–91.

Материал поступил в редколлегию 02.12.2013

V. B. Barakhnin, A. M. Fedotov

A MODEL OF FACTOGRAPHIC RETRIEVAL

This paper considers the theoretical problems of factographic retrieval, as well as of development of technology for extraction factographic information from scientific documents with rather arbitrary structure. It is shown that when creating information factographic systems, the following definition of fact is advisable: contained in the document's text and metadata, set of relations between the entities, described in the ontology of information system. A model of ontology of factographic system, working with documents of rather arbitrary structure, is proposed. The problems of automated retrieval of facts from documents and organization the interaction between factographic systems and users are discussed.

Keywords: fact, factographic retrieve, intellectual systems, data (facts) mining.

References

1. *Mikhailov A. I., Chernyi A. I., Gilyarevskyi R. S.* Scientific Communications and Informatics. Moscow: Nauka, 1976.

2. *Arsky Yu. M., Gilyarevskyi R. S., Turov I. S., Chernyi A. I.* Infosphere: Information Structures, Systems and Processes in Science and Society. Moscow: VINITI RAS, 1996.

3. *Shokin Yu. I., Fedotov A. M., Barakhnin V. B.* Problems of Information Retrieval. Novosibirsk: Nauka, 2010.

4. *Rakitov A. I.* Fact // Encyclopedia of Philosophy. Moscow: Soviet Encyclopedia, 1970. Vol. 5. P. 298.

5. *Marchuk A. G.* About Distributed Factographic Systems // Proceedings of the Tenth Anniversary of All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (RCDL'2008). Dubna, October 7–11, 2008. P. 93–102.

6. *Marchuk A. G., Marchuk P. A.* Archival Factographic System // Proceedings of the Eleventh Anniversary of All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (RCDL'2009). Petrozavodsk, Russia, September 17–21, 2009. P. 177–185.

7. *Berners-Lee T., Hendler J., Lassila O.* The Semantic Web // Scientific American. 2001. Vol. 284 (5). P. 34–43.

8. *Sakharov A. N.* Alexander I. Moscow: Nauka, 1998.

9. *Wittgenstein L.* Tractatus Logico-Philosophicus. New York: Harcourt Brace, 1922.

10. *Gryaznov A. F.* Wittgenstein. New Encyclopedia of Philosophy. Moscow: Mysl, 2000. Vol. 1. P. 406–408.

11. *Chen P. P.* The Entity-Relationship Model – Toward a Unified View of Data // ACM TODS. 1976. Vol. 1, № 1. P. 9–36.

12. *Barakhnin V. B., Kozhemyakina O. Yu.* About the Automation of the Complex Analysis of Russian Poetic Text // Proceedings of the Fourteenth Anniversary of All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (RCDL'2012). Pereslavl-Zalessky, October 15–18, 2012. P. 213–217.

13. *Samoylov D. S.* Book about Russian Rhyme. Moscow: Khudozhestvennaya Literatura, 1982.

14. *Barakhnin V. B., Fedotov A. M.* Clarification of the Terminology Used in the Description of Intellectual Information Systems, Based on the Semiotic Approach // Izvestiya VUZ: Problems of printing and publishing. 2008. № 6. P. 73–81.

15. *Kolmogorov A. N.* Three Approaches to the Definition of the Concept «Quantity of Information» // Problems of Information Transmission. 1965. Vol. 1, № 1. P. 3–11.

16. *Kolmogorov A. N.* Information Theory and the Theory of Algorithms. Moscow: Nauka, 1987.

17. Lyapunov A. A. On the Ratio of Concepts of Substance, Energy and Information // Lyapunov A. A. Theoretical and Applied Problems of Cybernetics. Novosibirsk: Nauka, 1980. P. 320–323.

18. *Heisenberg W.* Physics and Beyond: Encounters and Conversations. A. J. Pomerans, trans. New York: Harper & Row, 1971.

19. *Mikhailov A. I., Chernyi A. I., Gilyarevskyi R. S.* Fundamentals of Informatics. Moscow: Nauka, 1968.

20. *Dobrov B. V., Loukachevitch N. V., Sinitsyn M. N., Shapkin V. N.* Development of Linguistic Ontology on Natural Sciences for Information Retrieval Purposes // Proceedings of the Seventh Anniversary of All-Russian Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections» (RCDL'2005). Yaroslavl, 2005. P. 70–79.

21. *Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F.* Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. 1999.

22. *Narin'yani A. S.* A centaur by name of «THEON»: Thesaurus + Ontology // Proceedings of the DIALOG'2001 International Workshop. Aksakovo, 2001. Vol. 1. P. 184–188.

23. *Barakhnin V. B., Leonova Y. V., Fedotov A. M.* On the Problem of Formulation of Requirements for Creating Scientific-Organizational Information System // Computational Technologies: Special Issue, 2006. Vol. 11. P. 52–58.

24. *Sidorova E. A.* Ontological Approach to Knowledge Representation for an Analysis of Text Documents // Proceedings of the All-Russian Conference with International Participation «Knowledge-Ontology-Theory» (KONT-07). Novosibirsk, 2007. Vol. 1. P. 221–228.

25. *Bychkov I. V., Ruzhnikov G. M, Hmelnov A. E., Shigarov A. O.* The Heuristic Method for Table Detection in Multi-Format Documents // Computational Technologies, 2009. Vol. 14, № 2. P. 58–73.

26. *Gorobets B. S.* Soviet Physicists Joke... Although there are no Joking. Moscow: LIBROKOM, 2010.

27. *Barakhnin V. B., Fedotov A. M., Fedotova O. A.* Electronic Library on Scientific Heritage as a Factual System // XV Scientific Conference «Digital Libraries: Advanced Methods and Technologies, Digital Collections». Yaroslavl, Russia, 2013. P. 91–97.

28. *Leonova Y. V., Fedotov A. M.* Extraction of Knowledge and Facts from Texts of Theses and Abstracts // Proceedings of the Fifth International Conference «Systems Analysis and Information Technologies» (SAIT-2013). Krasnoyarsk, 2013. Vol. 2. P. 85–91.