ИНФОРМАЦИОННЫЙ ПОИСК

УДК [025.4.036 : 004] – 047.58

В.Б. Барахнин, А.М. Федотов

Построение моделей документального и фактографического поиска в электронных библиотеках*

Рассматриваются построение моделей документального и фактографического поиска в электронных библиотеках, работающих с документами достаточно произвольной структуры, а также разработка технологии извлечения фактографической информации из научных документов достаточно произвольной структуры. Предложена модель классификации документов электронной библиотеки, основанная на использовании отношения толерантности, учитывающая возможное отсутствие априорно заданных классификаторов. Показано, что при создании фактографических систем целесообразно следующее понимание факта: содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы. Предложена простейшая модель онтологии фактографической системы.

Ключевые слова: *интеллектуальные системы, документальный поиск, факт, фактографический поиск*

ВВЕДЕНИЕ

В классической монографии [1], изданной ВИНИТИ РАН 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 самóй 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 интеллектуальные системы.

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 достаточно общей постановки. Одна из основных причин этого заключается в том, что с появлением в конце 1970-х гг. персональных компьютеров появились мощные средства визуализации информации, вследствие чего были почти остановлены научные изыскания в области теории создания информационно-поисковых систем. Другой причиной приостановки развития новых алгоритмов обработки фактографической информации стало развитие в начале 1980-х гг. в Японии проекта так называемых «компьютеров пятого поколения», 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 в массивах данных автома-

* 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.

тическое 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 имевшихся 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, реализованной на компьютере достаточно большой мощности, способной к самоорганизации, проявляющейся, в частности, в самостоятельном (не зависящим 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. Ошибочным был и выбор языка логического программирования Пролог: программы, написанные на нем, плохо отлаживались и не распараллеливались. 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 ИнтС для обработки посредством локального доступа.

Таким образом, возникает необходимость разработки моделей документального и фактографического поиска в электронных библиотеках (ЭБ), работающих с документами достаточно произвольной структуры. Настоящая статья посвящена построению таких моделей.

МОДЕЛЬ КЛАССИФИКАЦИИ ДОКУМЕНТОВ В ЭЛЕКТРОННЫХ БИБЛИОТЕКАХ

Так как задачи поиска и классификации информации взаимно-обратны, то нам достаточно рассмотреть модель классификации документов, наиболее адекватно отражающую особенности работы с электронной библиотекой, в частности, возможное отсутствие априорно заданных классификаторов.

Наиболее распространенным вариантом классификации библиографических ресурсов является фасетная классификация, теория построения которой формализована индийским библиотековедом Ш.Р. Ранганатаном [8]. Объекты классифицируются одновременно по нескольким независимым друг от друга признакам (фасетам). Применительно к электронным библиотекам (и электронным ресурсам вообще) в качестве фасетов выступают элементы метаданных.

Важно отметить, что при создании научно-образовательных ЭБ, для которых библиографические признаки документов гораздо менее важны по сравнению с обычными электронными библиотеками, подмножества множеств значений библиографических метаданных, образующих значения фасетов, как правило, более широки. Так, ссылки на различные переиздания одного и того же документа с точки зрения научно-образовательных электронных библиотек целесообразно считать эквивалентными.

Простейшая формальная модель классификации документов с использованием структурированных метаданных документов выглядит следующим образом [9]. Пусть в справочно-поисковом аппарате ЭБ хранится информация о документах d_i. При этом любой документ d_i. 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. Рассмотрим подмножество метаданных M_C, определяющее набор классификационных признаков документов, используемых для составления поискового предписания (с учетом заданных логических операций). Для фиксированного элемента метаданных M^j, где $M^j \subset M_C$, заранее определяются подмножества $M_i{}^j$ множества значений этого элемента метаданных (указанные подмножества могут, вообще говоря, пересекаться).

Будем считать два документа *толерантными* (напомним, что толерантность – отношение, которое

обладает свойствами рефлексивности и симметричности, но, вообще говоря, может не обладать, в отличие от отношения эквивалентности, свойством транзитивности; подробно свойства этого отношения исследованы в [10]), если у них значения некоторого элемента метаданных входят в одно и то же подмножество $M_i{}^j$, при этом если значения рассматриваемого элемента метаданных могут повторяться, то документы считаются толерантными при совпадении хотя бы одного из значений. Каждое такое подмножество порождает на множестве документов электронной библиотеки предкласс толерантности, который обозначим $K_i{}^j$.

Более того, в большинстве случаев такие предклассы максимальны, т.е. это –классы толерантности. Предкласс $K_k{}^i$ является классом, если не существует отличного от него (т.е. порожденного другим набором элементов метаданных) предкласса $K_l{}^j$, такого, что $K_k{}^i \subset K_l{}^j$, в противном случае $K_k{}^i$ классом не является.

Выясним в каких случаях предклассы не являются классами (это необходимо, например, для описываемого ниже определения базиса пространства толерантности). Прежде всего, если $M_l{}^i \subset M_k{}^i$, то $K_k{}^i \subseteq K_l{}^i$, и поэтому $K_k{}^i$ классом не является, за исключением конкретного подбора документов, когда $K_k{}^i = K_l{}^i$, но и в этом случае, очевидно, нет смысла рассматривать $K_k{}^i$ в качестве отдельного класса. С содержательной точки зрения этой ситуации соответствует вхождение некоторого раздела классификатора ЭБ в раздел более высокого уровня, когда оба этих раздела учитываются при описании пространства толерантности (разумеется, можно и не учитывать раздел более низкого уровня при определении толерантных элементов, но тогда мы будем иметь дело с пространством толерантности, отличным от первоначального). В описанной ситуации предклассы, не являющиеся классами, определяются априори.

Однако возможна и ситуация, когда $K_k{}^i \subset K_l{}^j$ из-за конкретных особенностей документов ЭБ. Например, в электронной библиотеке по истории математики все документы, имеющие географический признак *Египет*, имеют хронологической признак *до новой эры*, при этом указанный хронологический признаки имеют и документы, относящиеся к другим регионам. Ясно, что в этом случае все документы с признаком *Египет* попарно толерантны не только в силу географического, но и в силу хронологического признака, однако, появление в ЭБ хотя бы одного документа с признаком *Египет*, датируемого *новой эрой*, изменит эту ситуацию. Тем самым в рассматриваемой ситуации предкласс $K_k{}^i$ целесообразно рассматривать (например, при построении базиса) в качестве класса.

Совокупность всех классов толерантности (включая предклассы, рассматриваемые в соответствии со сказанным выше в качестве классов) будем обозначать через H.

Укажем далее, как устроен базис описываемого пространства толерантности (некоторая совокупность H_B классов толерантности называется базисом, если для всякой толерантной пары документов существует класс из H_B, содержащий оба этих документа, а удаление из H_B хотя бы одного класса приводит к потере этого свойства). Очевидно, что множество классов толерантности H_M (включающее по нашему построению, в том числе, и предклассы, рассматриваемые в качестве классов), порожденных всей совокупностью подмножеств $M_i{}^j$, содержит базис. Утверждать, что H_M в точности является базисом нельзя, потому что входящие в него предклассы, не являющиеся классами, могут быть удалены без потери первого свойства из определения базиса. Однако, поскольку добавление в ЭБ даже одного документа может сделать предкласс классом и, стало быть, «полноценным» элементом базиса, постольку рассмотрение таких предклассов в качестве элементов базиса целесообразно с точки зрения организации классификации и поиска документов в электронной библиотеке.

Описание классов толерантности для ЭБ имеет большое практическое значение. Прежде всего, рассмотрим множество всех документов, для которых существует такая совокупность классов (включая предклассы, рассматриваемые в качестве классов) из H, что каждый из этих документов входит в эти и только эти классы. Такое множество представляет собой ядро толерантности, а множество всех ядер толерантности задает отношение эквивалентности на множестве документов ЭБ. При этом для построения ядер толерантности достаточно рассматривать лишь классы (и предклассы) из базиса H_M [10].

Таким образом, поисковое предписание, содержащее подмножество метаданных, определяющее набор классификационных признаков, с указанием сочетаний значений этих метаданных при помощи логических операций, определяет конкретное ядро толерантности на множестве документов, которое и выдается пользователю в качестве ответа на его информационный запрос.

Кроме того, на множестве классов толерантности также можно, в свою очередь, ввести отношение толерантности, при этом толерантными считаются классы, имеющие хотя бы один общий документ. Такая конструкция оказывается полезной, например, для организации поиска документов «по аналогии».

Формализм, основанный на использовании отношения толерантности, оказывается более удобным при создании ЭБ, поскольку в отличие от обычных библиотек, в которых классификаторы заданы априорно, при работе с электронной библиотекой нередко приходится использовать те или иные алгоритмы кластеризации документов (см., например, [3]), а уже потом, исходя из результатов кластеризации, устанавливать подмножества множеств значений элементов метаданных, выступающих в качестве значений фасетов.

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 ГОСТа 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 непосредственно 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: «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 [11]), 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, ибо 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 РАН. *Данные* понимаются в ней как факты и идеи, представленные в символьной форме, позволяющей производить 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. Понятие «факт» является центральным в «Логико-философском трактате» Л.Витгенштейна [12], одним из источников которого, как отметил Витгенштейн в 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».

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 [13]. 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» [14], являющейся основой для унификации различных представлений данных (при этом следует отметить, что в статье [14] 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 всё 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 [15], второй 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.

Более того, в некоторых случаях целесообразно извлекать и факты, касающиеся не только семантического, но и синтаксического уровня сообщения. В частности, при анализе поэтических текстов [16] 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 [17], 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 ИнтС с целью извлечения из него фактов, представляет интерес с точки зрения создателей и пользователей данной системы? Чтобы ответить на этот вопрос, формализуем введенное понятие факта подобно тому, как это было сделано в нашей работе [18] 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 рассматривается некоторая информация (как правило, семантическая; примеры возможных исключений приведены выше), но в справочно-информационный фонд ИнтС факт заносится в качестве совокупности элементов данных, 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 [19, 20]. Таким образом, применение информационных технологий должно основываться на использовании различных моделей (феноменологических, информационных, математических и др.). Как подчеркивал А.А.Ляпунов (см., например, [21]): «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 в следующем разделе).

Таким образом, при создании фактографических информационных систем разумно следующее понимание 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 своей известной беседе с В.Гейзенбергом: «Только теория решает, что можно наблюдать» [22].

ОСОБЕННОСТИ ОНТОЛОГИЙ ДЛЯ ФАКТОГРАФИЧЕСКИХ СИСТЕМ

Прежде всего, уточним, какого именно понимания термина «онтология» мы будем придерживаться в настоящей работе.

В [18] 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 гг. [23]: 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 [24], под онтологией нередко стали понимать широкий спектр структур, представляющих знания о той или иной предметной области с разной степенью формализации [25]:

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 [18], 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, [26]). 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 интеллектуальной системы 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, например предложенная нами в работе [27]. 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 полностью соответствует введенному нами пониманию факта, что делает ее наиболее пригодной для создания фактографической системы. Разумеется, пользуясь знания-

ми 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 [28]).

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.

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. 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, [29], 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 этой 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» (http://company.yandex.ru/technologies/mystem/), 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 семантического 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 нередко в качестве чуть ли непременного атрибута 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. Однако 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 [30]):

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 необходимы такие компоненты ИнтС [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 интеллектуальных систем 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 настоящей статье изложены модели документального и фактографического поиска в электронных библиотеках, работающих с документами достаточно произвольной структуры. Предложена модель классификации документов электронной библиотеки, основанная на использовании отношения толерантности, учитывающая возможное отсутствие априорно заданных классификаторов. Показано, что при создании фактографических информационных систем целесообразно следующее понимание факта: **содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы**. 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.

Примером практического использования фактографических систем может служить проверка в научных издательствах и редакциях журналов 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.

* * *

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

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. Т. 5. – М: Советская энциклопедия, 1970. – 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. Ранганатан Ш.Р. Классификация двоеточием. Основная классификация / пер. с англ. – М.: ГПТНБ СССР, 1970.
9. Федотов А.М., Барахнин В.Б. Проблемы поиска информации: история и технологии // Вестник НГУ. Серия: Информационные технологии. – 2009. – Т. 7, Вып. 2. – С.3-17.
10. Шрейдер Ю.А. Равенство, сходство, порядок. – М.: Наука, 1971.
11. Сахаров А.Н. Александр I. – М: Наука, 1998.
12. Wittgenstein L. Logisch-Philosophische Abhandlung // Annalen der Naturphilosophie. Vol. XIV. Parts 3/4. – Leipzig: Verlag Unesma, 1921. – P..185-262 / пер. Витгенштейн Л. Логико-философский трактат. М.: Издательство иностранной литературы, 1958.
13. Грязнов А.Ф. Витгенштейн // Новая философская энциклопедия. Т.1. – М: Мысль, 2000. – С. 406-408.
14. Chen P.P. The entity-relational model. Toward a unified view of data // ACM TODS. 1976. № 1. P. 9-36. / пер. Чен П. П.-Ш. Модель «сущность–связь» – шаг к единому представлению данных // СУБД. – 1995. – № 3. – С.137-158.
15. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968.
16. Барахнин В.Б., Кожемякина О.Ю. Об автоматизации комплексного анализа русского поэтического текста // Труды Четырнадцатой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» (RCDL’2012). Переславль-Залесский, 15-18 октября 2012 г. – С. 213-217.
17. Самойлов Д. С. Книга о русской рифме. – М.: Художественная литература, 1982.
18. 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.
19. Колмогоров А.Н. Три подхода к определению понятия «количество информации» // Проблемы передачи информации. – 1965. – Т. 1, Вып. 1. – С.3-11.
20. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987.
21. 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.
22. Heisenberg W. Der Teil und das Ganze. Gespräche im Umkreis der Atomphysik. – München, 1976.
23. Михайлов А.И., Черный А.И., Гиляревский Р.С. Основы информатики. – М.: Наука, 1968.
24. 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 ме-тоды и технологии, электронные коллекции» (RCDL'2005). – Ярославль, 2005. – С. 70-79.
25. Welty C., McGuinness D., Uschold M., Gruninger M., Lehmann F. Ontologies: Expert Systems all over again // AAAI-1999 Invited Panel Presentation. – 1999.
26. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Труды международного семинара Диалог'2001 по компьютерной лингвистике и ее приложениям. Т. I. – Аксаково, 2001. – С. 184-188.
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. Сидорова Е.А. Онтологический подход к представлению знаний для задачи анализа текстовых ресурсов // Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ–07). Т. 1. – Новосибирск, 2007. – С. 221-228.
29. Бычков И.В., Ружников Г.М., Хмельнов А.Е., Шигаров А.О. Эвристический метод обнаружения таблиц в разноформатных докумен-тах // Вычислительные технологии. – 2009. – Т. 14, № 2. – С. 58-73.
30. Горобец Б.С. Советские физики шутят… Хотя бывало не до шуток. – М: Книжный дом «ЛИБРОКОМ», 2010.

Материал поступил в редакцию 24.09.14.

Сведения об авторах

БАРАХНИН ВЛАДИМИР БОРИСОВИЧ – доктор технических наук, доцент, старший научный сотрудник Института вычислительных технологий СО РАН, г. Новосибирск.
e-mail: bar@ict.nsc.ru

ФЕДОТОВ АНАТОЛИЙ МИХАЙЛОВИЧ – доктор физико-математических наук, профессор, член-корреспондент РАН, декан Факультета информационных технологий Новосибирского государственного университета, главный научный сотрудник Института вычислительных технологий СО РАН, г. Новосибирск.
e-mail: fedotov@sbras.ru