

# Онтологический подход к построению систем информационной поддержки научной и производственной деятельности<sup>1</sup>

Ю.А. Загоруйко, Е.А. Сидорова, О.И. Боровикова

*Институт систем информатики имени А.П. Еришова СО РАН, пр. Лаврентьева, д.6,  
г. Новосибирск, 630090, Россия*

zagor@iis.nsk.su, lena@iis.nsk.su, olesya@iis.nsk.su

**Аннотация.** В докладе рассматривается подход к построению систем информационной поддержки научной и производственной деятельности, обеспечивающих представление и интерпретацию информации в виде фактов. Рассматривается роль, которую может играть онтология в таких системах как на этапе разработки, так и во время функционирования системы. Онтология не только обеспечивает представление знаний о предметной области системы, она используется для описания структуры ее информационного наполнения и при формировании контента хранимых в ней документов, а также служит основой для поддержки содержательного доступа к знаниям и данным, интегрированным в информационное пространство системы. Предлагаемый подход апробирован при разработке специализированных порталов знаний по археологии и компьютерной лингвистике, обеспечивающих эффективный доступ к систематизированным знаниям и информационным ресурсам соответствующей области научных знаний.

**Ключевые слова:** представление знаний, онтология, портал знаний, информационные ресурсы, контент документа, содержательный доступ

## 1 Введение

Большой объем накопленной информации и высокая скорость поступления новой предъявляют все более жесткие требования к современным информационным системам (ИС), особенно к системам, предназначенным для информационного обеспечения научных и производственных процессов. До сих пор эта задача рассматривалась в контексте создания хранилищ документов, обеспечивающих хранение, тематическую рубрикацию и поиск документов по ключевым словам. Однако, возможностей, предоставляемых построенными в рамках такого подхода информационными систем (ИС), оказывается недостаточно для эффективной поддержки научной и производственной деятельности. Это связано, в первую очередь, с тем, что в существующих ИС данные представляются в виде текстовых документов (в корпоративных информационных системах) или информационных ресурсов (в интернет-каталогах и порталах), тогда как для человека наиболее естественной формой подачи информации является представление ее в виде сети взаимосвязанных фактов. Для решения этой задачи необходим переход на качественно новый уровень представления и обработки информации – семантический уровень, что позволит учитывать смысл (содержание) документов, извлекая из них важные для пользователя факты.

Средства представления и интерпретации информации в виде фактов могла бы обеспечивать информационная система, использующая как общие знания о мире, так и знания о предметной области (ПО), которую она обслуживает. В настоящее время такие знания представляются в виде онтологий [1], которые все больше используются не только для описания семантики документов и информационных ресурсов, но и при построении широкого класса

---

<sup>1</sup> Работа выполняется при финансовой поддержке РФФИ (проект № 09-07-00400) и Президиума РАН (Интеграционный проект СО РАН № 2/12 в рамках программы фундаментальных исследований РАН № 2)

информационных систем [2]. Онтология может использоваться в ИС не только для целей удобного для человека представления информации, но и для обеспечения содержательного, в том числе, многоязычного доступа к ней, для поддержки процесса разработки самой ИС и т.п.

Данный доклад посвящен рассмотрению подхода к построению системы информационной поддержки научной и производственной деятельности, центральным компонентом которой является онтология.

## 2 Представление знаний в ИС

Основу системы знаний ИС, предназначенной для информационной поддержки научной и производственной деятельности, составляет онтология. Использование онтологии делает систему знаний ИС легко расширяемой и настраиваемой – в нее могут интегрироваться как новые знания (например, о новых понятиях и отношениях предметной области), так и новые типы документов и информационных ресурсов.

### 2.1 Роль онтологии в ИС

В системе информационной поддержки научной и производственной деятельности онтология выполняет следующие функции:

- Настройка на предметную область (область знаний) информационной системы.
- Описание структуры информационного наполнения (контента) ИС. Организации, персоны, текстовые и мультимедиа документы, как и другие объекты, которые необходимо представить в ИС, могут быть описаны с помощью онтологии. Для этого в нее вводится понятие, характеризующее соответствующий тип объекта, описывается его структура и возможные отношения с объектами (экземплярами) других понятий.
- Описание содержания (контента) документов. На основе онтологии строится содержательная аннотация документа, включающая извлеченные из его текста объекты и связи, соответствующие понятиям и отношениям онтологии.
- Интеллектуальная интеграция информации. На основе онтологии в информационное пространство (ИП) системы может интегрироваться информация из различных информационных источников благодаря тому, что их содержание единообразно отображается в понятия и отношения общей для них онтологии.
- Обеспечение содержательного доступа к знаниям и данным, интегрированным в информационное пространство системы. Пользователь может использовать онтологию в качестве «проводника» для навигации по ИП системы, а также формулировать поисковые запросы, основными элементами которого являются понятия и отношения онтологии. С помощью онтологии также задается удобный для пользователя способ визуализации (представления) информации.
- Поддержка нескольких языков в системе. Онтология, выступая в качестве концептуальной основы многоязычного тезауруса, создает условия для визуализации представленной в ИС информации на разных языках и обеспечения доступа к документам и ресурсам, представленным на разных языках.

### 2.2 Структура и методика построения онтологии ИС

Для спецификации онтологии, которая играет такую важную роль в ИС, предложен формализм, фактически, являющийся онтологией представления.

Формальное описание онтологии представления имеет вид:

$$O_R = \langle C, R, T, D, A, F, Ax \rangle, \quad (1)$$

где  $C = \{C_1, \dots, C_n\}$  – конечное непустое множество классов, описывающих понятия некоторой предметной или проблемной области;

$R = \{R_1, \dots, R_m\}$ ,  $R_i \subseteq C \times C$ ,  $R = R_T \cup R_P \cup R_A$  – конечное множество бинарных отношений, заданных на классах (понятиях):  $R_T$  – антисимметричное, транзитивное, нерефлексивное бинарное отношение наследования, задающее частичный порядок на множестве понятий  $C$ ,  $R_P$  – бинарное транзитивное отношение включения («часть-целое»),  $R_A$  – конечное множество ассоциативных отношений;

$T$  – множество стандартных типов;

$D = \{d_1, \dots, d_n\}$  – множество доменов  $d_i = \{s_1, \dots, s_k\}$ , где  $s_i$  – значение стандартного типа *string*;

$TD = T \cup D$  – обобщенный тип данных, включающий множество стандартных типов и множество доменов;

$A = \{a_1, \dots, a_w\}$  – конечное множество атрибутов, описывающих свойства понятий  $C$  и отношений  $R_A$ ; для каждого класса  $C_i \in C$  выделяется подмножество ключевых атрибутов  $A_i \in A$ , служащих для однозначной идентификации его объектов (экземпляров класса  $C_i$ ); конкретные значения атрибута  $a_i$  должны либо принадлежать одному из доменов  $d_i$  ( $d_i \in D$ ), либо иметь тип  $t_j$  ( $t_j \in T$ );

$F$  – множество ограничений на значения атрибутов понятий и отношений, т.е. предикатов вида  $p_i(e_{i1}, e_{i2})$ , где  $e_{i1}$  – имя атрибута ( $e_{ik} \in A$ ), а  $e_{i2}$  – либо имя атрибута ( $e_{i2} \in A$ ), либо константа ( $e_{i2} \in td_j$ , где  $td_j \in TD$ );

$Ax$  – множество аксиом, определяющих дополнительную семантику классов и отношений онтологии. Важное место среди них занимают аксиомы, позволяющие выводить ассоциативные отношения между объектами:

$if r_{p1}(c_1, c_2) \& \dots \& r_{pn}(c_{m-1}, c_m) then r_c(c_k, c_l),$

где  $r_{pi} \in R$ ,  $pi \in 1 \dots n$ ,  $c_j \in C$ ,  $j \in 1 \dots m$ ,  $r_c \in R_A$ ,  $c_k, c_l \in C$ ,  $k, l \in 1 \dots m$ .

Особенностью отношения  $R_T$  является то, что при наследовании от родительского класса его классу-потомку передаются не только все атрибуты, но и отношения. Отношение включения «часть-целое»  $R_p$  наделено свойством транзитивности, благодаря чему при поиске объектов можно осуществлять транзитивное замыкание по этому отношению. Набор ассоциативных отношений  $R_A$  определяется потребностями конкретной области знаний. Наличие таких отношений позволяет организовать содержательный поиск и навигацию по информационному наполнению (контенту) ИС. Важной особенностью отношений  $R_A$  является то, что они могут иметь собственные атрибуты, специализирующие связь между аргументами.

Таким образом, предложенный формализм не только обеспечивает описание понятий проблемной области и области знаний ИС, но и разнообразных семантических связей между ними. Он позволяет выстраивать понятия в иерархию «общее-частное» и поддерживает наследование свойств по этой иерархии, предоставляет возможность задавать множество аксиом, определяющих семантику классов и отношений онтологии, а также накладывать ограничения на значения возможных свойств объектов – экземпляров понятий онтологии.

Онтология конкретной ИС строится на основе введенного выше формализма согласно предложенной в [3] методологии, главным принципом которой является построение онтологии ИС на основе базовых онтологий путем их достройки и развития, что значительно упрощает создание онтологии ИС и ее дальнейшее сопровождение.

В качестве базовых выбраны три онтологии: *онтология деятельности*, которая составляет базис онтологии проблемной области ИС, *онтология предметного знания*, на основе которой строится онтология области знаний ИС, и *онтология базовых задач ИС*, используемая для построения онтологии задач ИС.

Онтология деятельности характеризует проблемную область системы и, фактически, является онтологией верхнего уровня. В качестве такой онтологии может выступать, например, онтология научной и производственной деятельности, которая включает такие классы понятий, как *Персона*, *Организация*, *Событие*, *Деятельность*, *Документ*, используемые для описания участников научной и производственной деятельности, мероприятий, научных программ и проектов, различного типа публикаций и документов. В эту онтологию также включен класс *Информационный ресурс*, который служит для описания ресурсов, представленных в сети Интернет. Набор атрибутов и связей этого класса основан на стандарте Dublin Core [4].

Онтология предметного знания задает метапонятия для описания понятий возможных областей знаний. В качестве такой онтологии могут выступать, например, онтология научного знания, онтология производства и т.п. Например, онтология научного знания фиксирует основные содержательные структуры, которые могут использоваться для построения онтологий конкретных областей знаний. В частности, эта онтология содержит такие метапонятия, как *Раздел науки*, *Метод исследования*, *Объект исследования*, *Предмет исследования*, *Научный*

*результат*. Используя эти метапонятия, можно выделить и описать значимые для области знаний разделы и подразделы, задать типизацию методов и объектов исследования, описать результаты научной деятельности.

Онтология базовых задач служит для описания базовой функциональности ИС, поэтому она может рассматриваться в качестве спецификации требований к пользовательскому интерфейсу. Эта онтология включает такие базовые понятия, как *Поиск*, *Навигация*, *Просмотр*, *Фильтрация* и т.п., которые могут конкретизироваться при разработке онтологии задач конкретной ИС.

Понятия базовых онтологий связаны между собой ассоциативными отношениями, выбор которых осуществляется не только исходя из полноты представления проблемной и предметной областей ИС, но и с учетом удобства навигации по ее контенту и поиска информации.

### 3 Представление данных в ИС

В рамках рассматриваемого подхода онтология ИС используется в качестве спецификации структуры контента ИС. В самом деле, онтология ИС, вводя формальные описания понятий в виде классов объектов и отношений между ними, задает структуры для представления реальных объектов и связей между ними. В соответствии с этим контент ИС представляется формулой:

$$IC_{IS} = \langle I, R_I, V, A_I, R_{IC}, R_{IR}, R_{IA}, R_{VTD} \rangle, \quad (2)$$

где  $I = \{i_1, \dots, i_n\}$  – конечное множество экземпляров классов онтологии,

$R_I = \{ri_1, \dots, ri_m\}$  – конечное множество экземпляров отношений  $ri_i(i_j, i_k)$ , где  $i_j, i_k \in I$ ,

$V = \{v_1, \dots, v_q\}$  – конечное множество конкретных значений обобщенного типа  $TD$ ,

$A_I = \{ai_1, \dots, ai_w\}$  – конечное множество конкретизированных атрибутов, т.е. бинарных отношений вида  $ai_i(i_x, v_y)$  или  $ai_i(ri_x, v_y)$  между экземпляром класса или отношения и конкретным значением,

$R_{IC} \subseteq I \times C$  – бинарное отношение инцидентности между множеством экземпляров  $I$  и множеством классов  $C$ ,

$R_{IR} \subseteq R_I \times R$  – бинарное отношение инцидентности между множеством экземпляров отношений  $R_I$  и множеством отношений  $R$ ,

$R_{IA} \subseteq A_I \times A$  – бинарное отношение инцидентности между множеством конкретизированных атрибутов  $A_I$  и множеством  $A$ ,

$R_{VTD} \subseteq V \times TD$  – бинарное отношение инцидентности между множеством значений  $V$  и обобщенным типом  $TD$ .

Таким образом, контент ИС представляется в виде множества связанных информационных объектов (ИО). Формально, каждый ИО соответствует некоторому понятию онтологии (является экземпляром какого-либо класса онтологии, т.е. элементом множества  $I$ ) и имеет заданную им структуру. Между конкретными информационными объектами могут существовать связи (элементы класса  $R_I$ ), синтаксис и семантика которых определяется отношениями, заданными между соответствующими понятиями онтологии.

Важную часть контента ИС составляют описания релевантных области знаний системы информационных ресурсов, представленных в Интернет, и описание документов, хранящихся в локальной (текстовой) базе данных системы. В соответствии с формулой (2) описание информационного ресурса включает соответствующий ему экземпляр класса *Информационный ресурс* и набор экземпляров отношений, связывающих данный экземпляр с другими классами онтологии или их экземплярами. Аналогичную структуру имеет описание документа. Содержательно такие описания включают, с одной стороны, общую информацию (мета-информацию) о ресурсе или документе, а с другой – представляют содержание (контент) документа в терминах онтологии области знаний ИС.

### 4 Обеспечение содержательного доступа

Содержательный доступ к знаниям и данным ИС осуществляется через пользовательский интерфейс, вся работа которого базируется на онтологии. Во-первых, пользователь может просматривать онтологию с целью лучшего понимания области знаний ИС и содержания ее контента. Во-вторых, он может использовать ее в качестве «проводника» для навигации по контенту ИС, а также базиса для формулирования запросов к ИС.

#### 4.1 Визуализация знаний и данных в ИС

Настройка визуализации знаний и данных в ИС также выполняется в терминах онтологии. Такая настройка требуется для более удобного представления информации пользователю ИС. При этом для каждого класса онтологии может быть задан шаблон визуализации содержательных ссылок на его объекты. Пользователь может включить в такую ссылку наиболее важные с его точки зрения атрибуты не только данного класса, но и других классов, связанных с ним ассоциативными отношениями.

Различают два вида ссылок – полные и краткие. Полные ссылки используются для представления объектов в визуализируемом списке объектов выбранного класса. Краткие ссылки применяются при визуализации полной информации об объекте: при этом объекты, связанные с визуализируемым объектом ассоциативными отношениями, представляются своими краткими ссылками.

Заметим, что оба вида ссылок на объекты, на самом деле, являются гиперссылками, т.е., щелкнув мышкой по любой из них, пользователь может перейти к просмотру полной информации о представляемом ею объекте.

#### 4.2 Навигация по контенту ИС

Навигация по контенту ИС организуется на основе таксономических отношений  $R_T$ , задающих иерархию понятий, и ассоциативных отношений  $R_A$ , связывающих между собой информационные объекты. При этом обеспечивается возможность выбора информационных объектов определенного класса, фильтрации списков таких объектов, детального просмотра каждого ИО и навигации по его связям, а также просмотра информационных ресурсов, ссылки на которые содержатся в данном ИО.

Навигация по ИС начинается с выбора определенного класса  $C_X$  в дереве понятий онтологии, построенном на основе отношения  $R_T$ . При этом пользователю выдается список информационных объектов выбранного класса  $L(C_X)$ , который отображается в виде html-страницы, содержащей набор ссылок на эти объекты. Для больших списков ИО формируется составная страница, включающая список страниц с элементами навигации по нему.

Формирование списка объектов выбранного класса  $C_X$  выполняется с учетом транзитивного замыкания по отношению  $R_T$ . Вследствие этого результирующий список  $L(C_X)$  будет включать как объекты искомого класса  $C_X$ , так и объекты его классов-наследников, т.е.

$$L(C_X) = \{i_i \mid i_i \in C_X \cup C_T\}, \text{ где } C_X \in C, C_T = \{C_Y \in C \mid \exists R_T(C_X, C_Y)\}.$$

Если на информационных объектах класса  $C_X$  задано отношение включения  $R_P$ , по желанию пользователя множество объектов, включенных в  $L(C_X)$ , может быть представлено либо в виде списка, либо в виде дерева.

Информация о конкретном объекте и его связях отображается в виде html-страницы, формат и наполнение которой зависят от класса объекта и заданных для него отношений. При этом объекты, связанные с данным объектом, представляются на его странице в виде содержательных гиперссылок, по которым можно перейти к их детальному описанию.

Дальнейшая навигация по контенту ИС представляет собой процесс перехода от одних ИО к другим по заданным между ними связям – экземплярам ассоциативных отношений  $R_A$ .

При переходе по конкретной связи любого информационного объекта или в качестве результата поиска мы можем получить достаточно большой список объектов (например, список всех участников крупного проекта или конференции). Для управления такой ситуацией был введен механизм фильтрации списков ИО, который также основывается на онтологии. Отфильтровать текущий список объектов пользователь может, задав конкретные значения атрибутов фильтруемых объектов или ограничения на них, а также на значения атрибутов объектов, связанных с данным объектом ассоциативными отношениями.

### 4.3 Поиск информации

Пользователь может формулировать свои запросы в терминах области знаний ИС. При этом основными элементами запросов являются понятия и отношения онтологии, а также ограничения, которым должны удовлетворять искомые объекты.

Формально, поисковый запрос, построенный в терминах онтологии, выглядит следующим образом:

$$Q = \{ i_X \in C_X \mid P(i_X) \& Rd(i_X) \& Rb(i_X) \}, \text{ где}$$

$C_X \in C$  – класс искомых объектов;

$P(i_X)$  – описание свойств искомого объекта (объектов)  $i_X$  класса  $C_X$  в виде ограничений на значения его атрибутов;

$Rd(i_X)$  – описание свойств объектов, состоящих с объектами класса  $C_X$  в «прямых» отношениях (аналогично  $P(i_X)$ );

$Rb(i_X)$  – описание свойств объектов, состоящих с объектами класса  $C_X$  в «обратных» отношениях.

Результатом выполнения запроса  $Q$  будет множество (список) объектов искомого класса  $C_X$ , удовлетворяющих описанным в запросе свойствам.

Поисковые запросы задаются через специальный графический интерфейс, управляемый онтологией ИС. При выборе пользователем класса искомых информационных объектов автоматически генерируется поисковая форма, в которой можно задать ограничения на значения атрибутов объектов выбранного класса, а также на значения атрибутов объектов, связанных с данным объектом ассоциативными отношениями. При этом также могут быть наложены ограничения и на значения атрибутов таких отношений.

Следует заметить, что при выполнении запроса поиск объектов заданного класса  $C_X$  может выполняться с учетом транзитивного замыкания по отношению  $R_T$ . В этом случае при обработке запроса будут рассматриваться объекты из множества  $\{i_i \mid i_i \in C_X \cup C_T\}$ , где  $C_T = \{C_K \in C \mid \exists R_T(C_X, C_K)\}$ .

## 5 Обеспечение многоязычности в системе

В предлагаемом подходе онтология выступает и в качестве концептуальной основы многоязычного тезауруса. Такой тезаурус строится как лингвистическое дополнение онтологии и включает термины проблемной и предметной области ИС, т.е. слова и словосочетания на нескольких естественных языках, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах. При этом, в отличие от онтологии, где понятия связываются отношениями, характерными для данной области знаний, в тезаурусе термины связываются чисто лингвистическими отношениями, такими как синонимия, лексическая эквивалентность и т.п.

Тезаурус, построенный на основе описанных выше принципов, имеет следующую структуру:

$$Th = \langle Tr, At, Rt, R_{TO}, Axt \rangle, \text{ где}$$

$Tr = \{Tr_1, \dots, Tr_n\}$  – конечное непустое множество терминов, представляющих понятия и отношения некоторой предметной или проблемной области; в множестве  $Tr$  выделяется подмножество базовых терминов  $Trb \subseteq Tr$ , которые считаются экспертами наиболее предпочтительными для представления имен понятий и отношений;

$At = \{at_1, \dots, at_w\}$  – конечное множество атрибутов, описывающих свойства терминов  $Tr$ ;

$Rt = \{Rt_1, \dots, Rt_m\}$ ,  $Rt_i \subseteq Tr \times Tr$  – конечное множество бинарных отношений, заданных на терминах согласно принятым стандартам на тезаурусы (ГОСТ-7.25-2001, ГОСТ-7.24-90 и ISO 5964);

$R_{TO}$  – отношение, устанавливающее соответствие между термином тезауруса и понятием или отношением онтологии, т.е.  $R_{TO} \subseteq Trb \times E_O$ , где  $E_O = C \cup R$  – множество понятий и отношений онтологии;

$Axt$  – множество аксиом, определяющих семантику связей между терминами.

Отношения  $R_{TO}$ , связывающие термины тезауруса  $Tr$  с понятиями онтологии  $C$ , обеспечивают визуализацию представленной в ИС информации на разных языках, делают ее способной «понимать» разноязычные ресурсы, а также поддерживают навигацию по ее контенту и формулирование запросов с использованием удобного для пользователя языка.

## 6 Наполнение контента ИС: основные принципы и проблемы

Наполнение контента ИС может осуществляться как «вручную» – с помощью редактора данных, так и автоматически – подсистемой автоматического извлечения информации из текстов документов и информационных ресурсов.

Редактор данных позволяет создавать, редактировать и удалять информационные объекты, структура которых определяется соответствующими классами онтологии ИС. При создании нового информационного объекта пользователь выбирает в иерархии понятий соответствующий класс. После этого по представленному в онтологии описанию класса автоматически создается форма, включающая поля как для ввода значений атрибутов информационного объекта, так и задания его связей с другими объектами (такие объекты выбираются из списка, выводимого в специальном окне формы). Корректное создание/удаление информационных объектов и отношений между ними, а также редактирование значений их свойств поддерживается средствами редактора на основе заданных в онтологии ограничений.

Наполнение контента ИС «вручную», пусть даже с помощью эргономичного редактора, – весьма трудоемкая задача. В связи с этим была предложена технология автоматического наполнения контента ИС, базирующаяся на онтологии.

В рамках этой технологии для наполнения контента ИС используется информация, представленная в документах, хранящихся в текстовой БД системы, и информационных ресурсах, поставляемых подсистемой сбора онтологической информации. При этом, как отмечалось в главе 3, в контенте ИС фиксируется не только информация, составляющая содержание (или контент) документов и информационных ресурсов, но и сведения о них как об информационных объектах – экземплярах понятий онтологии *Документ* или *Информационный ресурс*, соответственно. При построении контента документов и информационных ресурсов выполняется извлечение из их текстов объектов и связей, соответствующих понятиям и отношениям онтологии. Затем полученные объекты и связи помещаются в контент ИС и становятся доступными через механизмы навигации и поиска.

Если подлежащие обработке текстовые документы просто извлекаются из локальной базы данных системы, то чтобы обработать информационные ресурсы, их сначала необходимо извлечь из сети Интернет. Рассмотрим, как это делается.

### 6.1 Сбор информационных ресурсов по тематике системы

Подсистема сбора онтологической информации [5] осуществляет поиск информационных ресурсов по ссылкам, заданным в специальной базе данных, и определяет их релевантность области знаний системы. Она включает поискового робота, служебную базу данных, а также словарь терминов, характеризующих область знаний системы.

В *служебной базе данных* хранятся ссылки на информационные ресурсы, их мета-описания и другая служебная информация о них, необходимая для поискового робота и дальнейшей работы подсистемы.

*Поисковый робот* последовательно просматривает заданные в БД ссылки и загружает найденные по ним ресурсы и документы для обработки. При этом он осуществляет сбор такой метainформации о них, как название, ссылка, кодировка, формат, размер, дата последнего изменения. Затем, согласно указанным при настройке подсистемы критериям, выполняется анализ загруженных информационных ресурсов и интернет-документов (в дальнейшем – просто *документов*) на предмет их релевантности области знаний системы.

Релевантность документа зависит как от внутренних (текстовых), так и от внешних (ссылочных) характеристик документа.

Текстовые характеристики учитывают наличие терминов словаря в документе, их вес в текстовом содержимом документа и позицию в выделенных фрагментах текста, таких как заголовок документа, аннотация, основной текст, библиографический список и т.п. В зависимости от типа документа каждому из фрагментов назначается определенный вес.

Ссылочные характеристики документа при определении его релевантности учитывают авторитетность ресурса (сайта-родителя), на котором хранится документ, и авторитетность документа-источника, в котором была обнаружена ссылка на документ.

С помощью специальной процедуры анализа гиперссылок из релевантных документов извлекаются и заносятся в служебную базу данных новые ссылки на документы. Таким образом, список ссылок на ресурсы может редактироваться не только вручную настройщиком-экспертом системы, но и пополняться автоматически за счет ссылок, обнаруженных в документах.

Информация о ссылках на ресурсы представляет собой сведения о названии и адресе ссылки, дате ее добавления в БД, статусе обработки и способе добавления (поисковым роботом или экспертом). Эта информация используется в дальнейшем для настройки периодической автоматической проверки сайтов. Кроме того, в служебной базе данных хранится информация об источнике ссылки и сайте-родителе, которому она принадлежит. Источником, в свою очередь, может являться обработанный ранее документ, из которого в процессе анализа была получена новая ссылка. На основе информации о принадлежности ресурсов одному сайту можно установить иерархические отношения между ссылками на ресурсы, которые в дальнейшем учитываются при проверке доступности сайта и обновления его контента.

## 6.2 Общая схема извлечения информации из текста

Для целей автоматического наполнения контента системы информационными объектами нет необходимости делать полный семантический анализ текста документов [6] и извлекать из них всю возможную информацию. Это упрощает процесс анализа документов и позволяет сделать его эффективным и легко настраиваемым на различные области знаний.

Важным свойством онтологии является то, что она задает формат данных в ИС, а, следовательно, определяет какую информацию необходимо извлекать из текста документа, а какую игнорировать. Результат анализа документа представляется в виде семантической сети объектов, которые являются экземплярами понятий и отношений онтологии ПО.

Очевидно, что знаний о предметной области, представленных в онтологии, недостаточно для автоматического извлечения информации из текста – требуются дополнительные знания о языке, на котором эта информация представлена, а также средства, позволяющие, с одной стороны, формировать такие знания экспертам (лингвистам), а с другой стороны, обеспечивать автоматическое применение полученных знаний при обработке документов.

В связи с этим подсистема анализа документов включает следующие компоненты:

- словарный компонент, включающий АРМы (автоматизированные рабочие места) настройки словарей разных типов и исполняемые модули, реализующие основные методы словарной обработки текста;
- конструктор схем или правил извлечения фактов из текста, обеспечивающий настройку процесса постсловарного анализа документа;
- модуль сборки фактов, осуществляющий извлечение фактов по заданным схемам;
- модуль формирования контента документа.

В режиме настройки эксперт с помощью редактора онтологии и соответствующих АРМов формирует базу знаний ИС. При этом ключевая лексика для словарей и онтологий автоматически извлекается из подборки документов, релевантных области знаний ИС [7].

При построении базы знаний решается задача согласования лингвистических знаний, заданных в словаре, со знаниями, представленными онтологией ИС. Следует иметь в виду, что содержащееся в тексте документа описание объекта, соответствующее тому или иному элементу онтологии, редко представляется с помощью одного термина, устойчивого словосочетания или другой единицы словаря. Такое описание обычно включает несколько компонентов, распределенных по тексту. Для того, чтобы понять о каком элементе онтологии идет речь в тексте, необходимо выделить факты, явным образом определяющие класс и/или атрибут понятия или отношения.

Для описания процесса извлечения из текста значимых для ИС фактов, предложено использовать структуры, задающие связи элементов словаря с понятиями и отношениями онтологии [8]. Такие структуры называются *схемами фактов*.

Схема факта  $F_k$  – это тройка вида  $\langle Arg, Rs, C \rangle$ , где

$Arg$  – множество дескрипторов аргументов факта; дескриптором может быть тип словарной единицы, класс информационного объекта или тип факта, играющего вспомогательную роль при описании цепочки, связывающей термины с понятиями и отношениями онтологии.

$Rs = \langle t, op(t), P \rangle$  – результат применения схемы, где  $t$  задает класс объекта,  $op(t)$  – тип операции (создание и/или редактирование объекта класса  $t$ ), применяемой при условии выполнения ограничения  $C$ ,  $P$  – множество правил для формирования/редактирования объекта. Каждое правило ставит в соответствие атрибуту результирующего объекта либо точное значение атрибута, либо значение атрибута одного из аргументов.

$C$  – множество ограничений, накладываемых на характеристики аргументов факта. Выделяются следующие виды ограничений: (1) условия на морфологические и семантические атрибуты аргументов схемы; (2) ограничения синтаксической сочетаемости вершин синтаксических групп, реализующих аргументы схемы; (3) структурно-текстовые ограничения на взаиморасположение аргументов в тексте.

Схемы фактов создаются с помощью конструктора, который предоставляет доступ к онтологии и словарям и обеспечивает корректность построенных схем. Модуль планирования формирует из схем очередь, которая определяет порядок применения схем в процессе анализа с минимальной потерей информации.

В режиме обработки на вход подсистеме анализа поступает текст документа, который передается лексическому процессору словарного компонента. Последний осуществляет сегментацию и извлечение ключевых терминов и передает результаты своей работы модулю сборки фактов, который осуществляет поиск фактов в тексте, используя очередь схем фактов. Результатом работы модуля сборки фактов является множество информационных объектов. Это множество передается модулю формирования контента документов, который идентифицирует и уточняет параметры полученных объектов и, в конечном счете, решает задачу присоединения полученных результатов к контенту ИС.

Остановимся подробнее на последнем компоненте, поскольку именно он обеспечивает взаимодействие подсистемы анализа с контентом ИС и позволяет расширить возможности анализа, связанные с уточнением и корректировкой полученной из документа информации.

### 6.3 Формирование контента документа

Для того чтобы сформировать контент документа необходимо:

- обеспечить корректность объектов, полученных в результате анализа;
- идентифицировать объекты, образующие контент;
- добавить объекты в контент системы и связать их с документом.

Объект считается корректным, если определен его класс и значения всех его определенных (извлеченных из текста) атрибутов удовлетворяют заданным в онтологии ограничениям (в частности, принадлежат соответствующим доменам или типам). Соответствие значений атрибутов доменам обеспечивается на уровне словаря и/или при конструировании схем фактов.

Объект считается идентифицированным, если он корректен и определены все его ключевые атрибуты. Наличие указанных характеристик позволяет однозначно выделить данный объект из множества других объектов, т.е. обеспечить его уникальность в системе.

В процессе идентификации уточняются характеристики полученных при анализе объектов и осуществляется "склеивание" одинаковых объектов на основе анализа локального и глобального контекста.

Под локальным контекстом понимается содержимое анализируемого документа, которое используется для решения таких задач как определение референта для местоимений и неоднозначных именных групп, а также для отождествления объектов на основании частично совпадающего и непротиворечивого набора атрибутов и связей.

Глобальный контекст представлен контентом информационной системы, который содержит как данные о ранее проанализированных документах, так и информацию, вводимую пользователем вручную или поступающую из подсоединенных баз данных. Использование глобального контекста требуется в случае, когда набор ключевых атрибутов определен не полностью. В этом случае осуществляется поиск максимально похожего объекта в контенте системы по известному набору атрибутов и классу объекта, а также его связям. Для этого

строится фокусное множество объекта, найденного в тексте, и сопоставляется с фокусными множествами объектов, совпавших по тому или иному набору атрибутов с идентифицируемым.

Глобальный контекст также может использоваться для уточнения класса объекта по иерархии классов (при этом может уточняться как объект найденный в тексте, так и объект из контента системы) и идентификации и уточнения объектов с помощью восстановления иерархии по отношению "часть-целое" (иерархии вложенности понятий).

## 7 Заключение

Рассмотрен базирующийся на онтологии подход к построению систем информационной поддержки научной и производственной деятельности, обеспечивающих представление и интерпретацию информации в виде фактов. В рамках этого подхода онтология не только обеспечивает представление знаний о предметной области ИС, но и служит для описания структуры ее информационного наполнения и контента хранимых в ней документов, а также поддерживает содержательный доступ к знаниям и данным, интегрированным в информационное пространство системы. Кроме того, использование онтологии делает такие системы легко расширяемыми и настраиваемыми на различные области знаний.

Данный подход был использован в рамках проектов по созданию научных интернет-порталов по археологии [9] и компьютерной лингвистике [10], целью которых было обеспечение содержательного доступа к систематизированным знаниям и информационным ресурсам соответствующей области научных знаний. Благодаря использованию онтологии в качестве центрального компонента этих порталов, пользователи получили удобные средства для представления, поиска и визуализации интересующей их информации, а также управляемой знаниями навигации по их контенту.

## Литература

- [1] Gruber, T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *International Journal of Human-Computer Studies*, Vol. 43, Issues 5-6, November 1995, pages 907–928.
- [2] Guarino, N. Formal Ontology in Information Systems. In N. Guarino (ed.) *Formal Ontology in Information Systems. Proceedings of FOIS'98*, Trento, Italy, June 6-8, 1998. IOS Press, Amsterdam, pages 3–15.
- [3] Загорюлько Ю. А., Боровикова О. И. Технология построения онтологий для порталов научных знаний. *Вестник НГУ. Серия: Информационные технологии*, том 5, выпуск 2, 2007. с. 42–52.
- [4] Using Dublin Core web site, 2009. <http://dublincore.org/documents/usageguide/>
- [5] Боровикова О.И., Загорюлько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний. *Труды международной конференции Диалог'2005 "Компьютерная лингвистика и интеллектуальные технологии" (Звенигород, 1-5 июня 2005)*. М.: Наука, 2005. с. 65–70.
- [6] Хорошевский В.Ф. Управление знаниями и обработка ЕЯ-текстов. *Труды 9-й национальной конференции по искусственному интеллекту с международным участием КИИ'2004*. М.: Физматлит, 2004. Т.2, с.565–572.
- [7] Сидорова Е.А. Многоцелевая словарная подсистема извлечения предметной лексики. *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4-8 июня 2008 г.)*. Вып. 7 (14). М.: РГГУ, 2008. с. 475-481.
- [8] Сидорова Е.А. Онтологический подход к представлению знаний для задачи анализа текстовых ресурсов. *Материалы Всероссийской конференции с международным участием «Знания – Онтологии – Теории» (ЗОНТ-07)*. Новосибирск: Институт математики им. С.Л. Соболева СО РАН, 2007. Т. 1. с. 166-175.
- [9] О.А. Андреева, О.И. Боровикова, С.В. Булгаков, Ю.А. Загорюлько, Е.А. Сидорова, Б.Г. Циркин, Ю.П. Холюшкин. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии. *Труды 10-й нац. конф. по искусственному интеллекту с междунар. участием*. М.: Физматлит, 2006. Т.3. с.832–840.
- [10] Боровикова О.И., Загорюлько Ю.А., Загорюлько Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике. *Труды 11-ой нац. конф. по искусственному интеллекту с международным участием КИИ-2008 (г. Дубна, Россия)*. М.: ЛЕНАНД, 2008. Т.3. с.380–388.