УДК 002.53:004.89

ПОДХОД К ОБЕСПЕЧЕНИЮ МНОГОЯЗЫЧНОГО ДОСТУПА К СИСТЕМАТИЗИРОВАННЫМ ЗНАНИЯМ И ИНФОРМАЦИОННЫМ РЕСУРСАМ ЗАДАННОЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Ю.А. Загорулько

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск
Новосибирский государственный университет
E-mail: zagor@iis.nsk.su

Рассматривается подход к обеспечению содержательного многоязычного доступа в портале знаний, интегрирующем знания и информационные ресурсы, относящиеся к заданной предметной области, на основе онтологии. Введение в систему многоязычного тезауруса, включающего термины предметной области, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах, делает систему способной «понимать» разноязычные ресурсы и обеспечивать поиск и визуализацию информации на разных языках.

Ключевые слова:
Информационная система, онтология, портал знаний, содержательный доступ, многоязычный тезаурус, информационные ресурсы.

Введение

Постоянный рост объемов информации по различным отраслям знаний делает задачу эффективного информационного обеспечения научных и производственных процессов все более актуальной. Однако, в существующих информационных системах (ИС) данные, как правило, представляются в виде текстовых документов (в корпоративных информационных системах) или информационных ресурсов (в интернет-каталогах и порталах). В то время как для человека, будь то ученый или руководитель, наиболее естественной формой подачи информации является представление ее в виде взаимосвязанных фактов. Обеспечить такое представление информации и эффективный ее поиск может только ИС, 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].

В данной работе рассматривается подход к созданию информационных систем, основанных на онтологиях, в частности интернет-порталов знаний [2], в которых онтология используется как для систематизации и интеграции знаний и информационных ресурсов заданной области знаний в единое информационное пространство, так и для организации содержательного доступа к ним. Такие порталы знаний обеспечивают представление взаимосвязанных фактов (в виде сети знаний и данных), их поиск и управляемую онтологией навигацию.

Другой важной проблемой, которую нужно решать при организации эффективной информационной поддержки описанных выше процессов, является обеспечение эффективного доступа к документам и ресурсам, представленным на разных языках. Упомянутые выше порталы знаний не поддерживают такой возможности, так как используемая в них онтология описывает только знания о моделируемой проблемной и предметной области, но не обеспечивает представления знаний об используемой в данной области знаний лексике для всех требуемых языков.

Таким образом, встает задача поддержки в портале знаний нескольких языков. Эту задачу можно было бы решить, расширив имеющуюся онтологию необходимыми лингвистическими знаниями, однако введение в онтологию дополнительных сущностей и отношений сделало бы ее громоздкой и непрозрачной, затруднило бы ее развитие и сопровождение. В связи с этим было принято решение расширить систему знаний портала еще одним компонентом – многоязычным тезаурусом [3, 4]. Введение такого тезауруса, включающего термины проблемной и предметной области, т. е. слова и словосочетания на нескольких естественных языках, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах, сделает портал способным «понимать» разноязычные ресурсы, поддерживать навигацию по его информационному пространству и воспринимать запросы на разных языках.

Вопросы совместного использования тезауруса и онтологии уже неоднократно обсуждались в других работах, но в них в основном рассматривалось применение тандема «тезаурус-онтология» в задачах обработки текстов [5] и/или информационного поиска [4, 6]. Наш подход призван обеспечить совместное использование онтологии и тезауруса не только для решения этих задач, но и для интеграции знаний и ресурсов, релевантных заданной ПО в единое информационное пространство, а также обеспечения эффективной навигации по нему с использованием удобного для пользователя естественного языка. В связи с этим мы предлагаем использовать в онтологии все отношения, необходимые для представления знаний о данной ПО, оставляя чисто лингвистические отношения (синонимия, эквивалентность и т. п.) для тезауруса.

Описанию нашего подхода к разработке и совместному использованию онтологии и тезауруса для обеспечения многоязычного содержательного доступа в портале знаний и посвящена данная работа.

Структура системы знаний

При создании портала знаний, обеспечивающего содержательный доступ к знаниям и информационным ресурсам определенной тематики, необходимо решить задачи, связанные с обработкой текстовых ресурсов, представлением их содержания в виде взаимосвязанных фактов, организацией поиска нужной информации и её визуализации на разных языках. Для решения этих задач используются знания о структуре и терминологии данной ПО, о структуре и типологии ресурсов (документов), а также знания о свойствах языка (лексике), на котором представлены тексты этих ресурсов.

Технология создания такого портала знаний предполагает организацию его системы знаний на базе интеграции многоязычного тезауруса и онтологии предметной области.

Формально система знаний (*KS*) портала описывается четверкой вида:

$$KS=\langle Os, Th, ICs, Ir\rangle,$$

где *Os* – онтология портала знаний; *Th* – многоязычный тезаурус предметной и проблемной области портала знаний, *ICs* – информационное наполнение (контент) портала знаний, которое строится на основе структур, заданных в онтологии *Os*; *Ir* – информационные ресурсы, интегрированные в информационное пространство портала знаний.

Рассмотрим, как устроены онтология и многоязычный тезаурус.

Онтология портала знаний

Для представления онтологии портала необходим формализм, обеспечивающий гибкие средства описания понятий его проблемной и предметной областей и разнообразных семантических связей между ними. Важным требованием к нему является возможность выстраивания понятий ПО в иерархию «общее-частное» и поддержка наследования свойств по этой иерархии. Этот формализм также должен предоставлять возможность задания ограничений на значения свойств объектов ПО и описания семантики отношений в виде аксиом.

В качестве формализма, удовлетворяющего описанным выше требованиям, была предложена метаонтология следующего вида:

$$O=\left\langle C, R, T, D, A, F, Ax\right\rangle,$$

где

- $C=\{C_1,...,C_n\}$ – конечное непустое множество классов, описывающих понятия данной предметной или проблемной области;
- $R=\{R_1,...,R_m\}$, $R_i\subseteq C\times C$, $R=\{R_T, R_P\}\cup R_A$ – конечное непустое 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;
- $T=\{t_1,...,t_n\}$ – конечное непустое множество стандартных типов;
- $D=\{d_1,...,d_k\}$ – множество доменов $d_i=\{s_1,...,s_k\}$, где s_i – значение стандартного типа *string*;
- $TD=T\cup D$ – обобщенный тип данных, включающий множество стандартных типов и множество доменов;
- $A=A_C\cup A_R=\{a_1,...,a_w\}$ – конечное множество атрибутов, описывающих свойства понятий *C* ($A_C\subseteq C\times TD$) и отношений R_A ($A_R\subseteq R_A\times TD$);
- *F* – множество ограничений на значения атрибутов понятий и отношений, т. е. предикатов вида $p_i=(e_1,...,e_m)$, где e_k – это либо имя атрибута ($e_k\in A$), либо константа ($e_k\in td_j$, где $td_j\in TD$);
- *Ax* – множество аксиом, определяющих семантику классов и отношений онтологии.

Отношение включения «часть-целое» R_P наделено свойством транзитивности, благодаря этому при поиске объектов может выполняться транзитивное замыкание по этому отношению.

Набор ассоциативных отношений R_A определяется пользователем. Наличие таких отношений позволяет организовать содержательный поиск и навигацию по контенту портала. Важной особенностью отношений R_A является то, что они могут иметь собственные атрибуты, специализирующие связь между аргументами.

Онтология портала знаний строится на основе описанной выше метаонтологии. Для упрощения настройки портала на выбранную область знаний и его дальнейшего сопровождения его онтологии выделены базовые онтологии, независимые от предметной области портала, и предметная онтология, описывающая определенную область знаний.

В качестве базовых онтологий были выбраны две онтологии. Первая из них характеризует проблемную область системы. Она не зависит от предметной области системы, фактически, являясь онтологией верхнего уровня. В качестве такой онтологии может выступать онтология научной или производственной деятельности [2], которая включает классы понятий, относящиеся к организации научной и производственной деятельности, такие как Персона, Организация, Событие, Деятельность, Документ (Публикация), используемые для описания участников научной и производственной деятельности, мероприятий, научных программ и проектов, различного типа публикаций. В эту онтологию также включен класс Информационный ресурс, который служит для описания информационных ресурсов, представленных в сети Интернет.

Вторая онтология – онтология предметного знания, задает метапонятия для описания понятий возможных предметных областей. В качестве такой

онтологии выступает онтология научного знания [2], которая включает метапонятия, задающие структуры для описания понятий конкретной области знаний, такие, как Раздел науки, Метод исследования, Объект исследования, Предмет исследования, Научный результат.

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

Так как базовые онтологии являются универсальными, то при создании конкретного портала знаний необходимо разрабатывать только онтологию соответствующей ему области знаний. Такая онтология строится на основе онтологии предметного знания, причем понятия этой онтологии являются реализациями метапонятий онтологии научного знания и организуются в 5 иерархий «общее-частное», соответствующих каждому такому метапонятию. Указанные иерархии связываются между собой ассоциативными отношениями, часть которых наследуется из базовых онтологий, а часть отражает специфику моделируемой области знаний.

Построенная таким образом онтология не только описывает предметную и проблемную область портала знаний, но и задает структуры для представления реальных объектов (в том числе, информационных ресурсов) и связей между ними. В соответствии с этим данные на портале представлены как множество взаимосвязанных информационных объектов (фактов), каждый из которых соответствует некоторому понятию онтологии (является его экземпляром) и имеет заданную им структуру. Семантика связей между информационными объектами определяется отношениями, заданными между соответствующими понятиями онтологии. Совокупность таких информационных объектов и их связей образует информационное содержание или контент портала.

Многоязычный тезаурус

Тезаурус должен обеспечивать возможность взаимодействия с порталом на нескольких языках, в том числе, навигацию, поиск, а также обработку информационных ресурсов, представленных на разных языках.

Тезаурус имеет следующую структуру:

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

где

- $Tr=\{Tr_1,...,Tr_n\}$ – конечное непустое множество терминов, представляющих понятия и отношения некоторой предметной области; из всего множества терминов Tr выделяется подмножество базовых терминов $Trb \subseteq Tr$, которые считаются экспертами наиболее предпочтительными для представления имен понятий и отношений;
- $At=\{at_1,...,at_w\}$ – конечное множество атрибутов, описывающих свойства терминов Tr;
- $Rt = \{Rt_1, ..., Rt_m\}, Rt_i \subseteq Tr \times Tr, Rt = R_{SBT} \cup R_{SNT} \cup R_{AT} \cup \{R_{USE}, R_{UF}, R_{LE}, R_{TO}\}$

  – конечное непустое множество бинарных отношений, заданных на терминах ПО согласно принятым стандартам ГОСТ и ISO [7]:
  - R_{SBT} – множество бинарных отношений, связывающих некоторый термин с термином более общего (в широком смысле) понятия; множество обратных к ним отношений – R_{SNT};
  - R_{AT} – конечное множество ассоциативных отношений между терминами;
  - R_{UF} – бинарное отношение, связывающее наиболее предпочтительный термин с синонимами (менее подходящими терминами) на том же языке; обратное к нему отношение – R_{USE};
  - R_{LE} – отношение лексической эквивалентности между терминами, определяющее одно и тоже понятие на разных языках;
- R_{TO} – отношение, устанавливающее соответствие между термином тезауруса и понятием или отношением онтологии, т. е. $R_{TO} \subseteq Trb \times Eo$, где Trb – множество базовых терминов тезауруса, $Eo=C \cup R$ – множество понятий и отношений онтологии;
- Axt – множество аксиом, определяющих семантику связей между терминами.

Схема представления связей онтологии и тезауруса показана на рисунке.

В качестве примера приведем две аксиомы:

Аксиома A1: if (A R_{UF} B) then (B R_{USE} A), где A и B – термины тезауруса.

Аксиома A2: if (A_{L1} R_{TO} C) & (A_{L2} R_{TO} C) then (A_{L1} R_{LE} A_{L2}), где A_{L1} – термин тезауруса на языке $L1$, A_{L2} – термин тезауруса на языке $L2$, C – понятие онтологии.

Методика построения онтологии и многоязычного тезауруса

В рамках предложенной методики, первой строится онтология предметной области ИС, которая затем дополняется многоязычным тезаурусом, причем знания о новых языках могут добавляться в тезаурус по мере необходимости.

Рассмотрим процесс построения онтологии ПО.

На первом этапе выполняется, так называемая «фиксация» онтологии, которая включает следующие шаги.

Сначала строится «скелет» предметной области, задающий ее самую общую структуру. На этом шаге выявляются наиболее важные понятия области знаний, которые мы будем называть базовыми. Для этого инженеры знаний обращаются к энциклопедическим словарям, учебникам и другим материа-

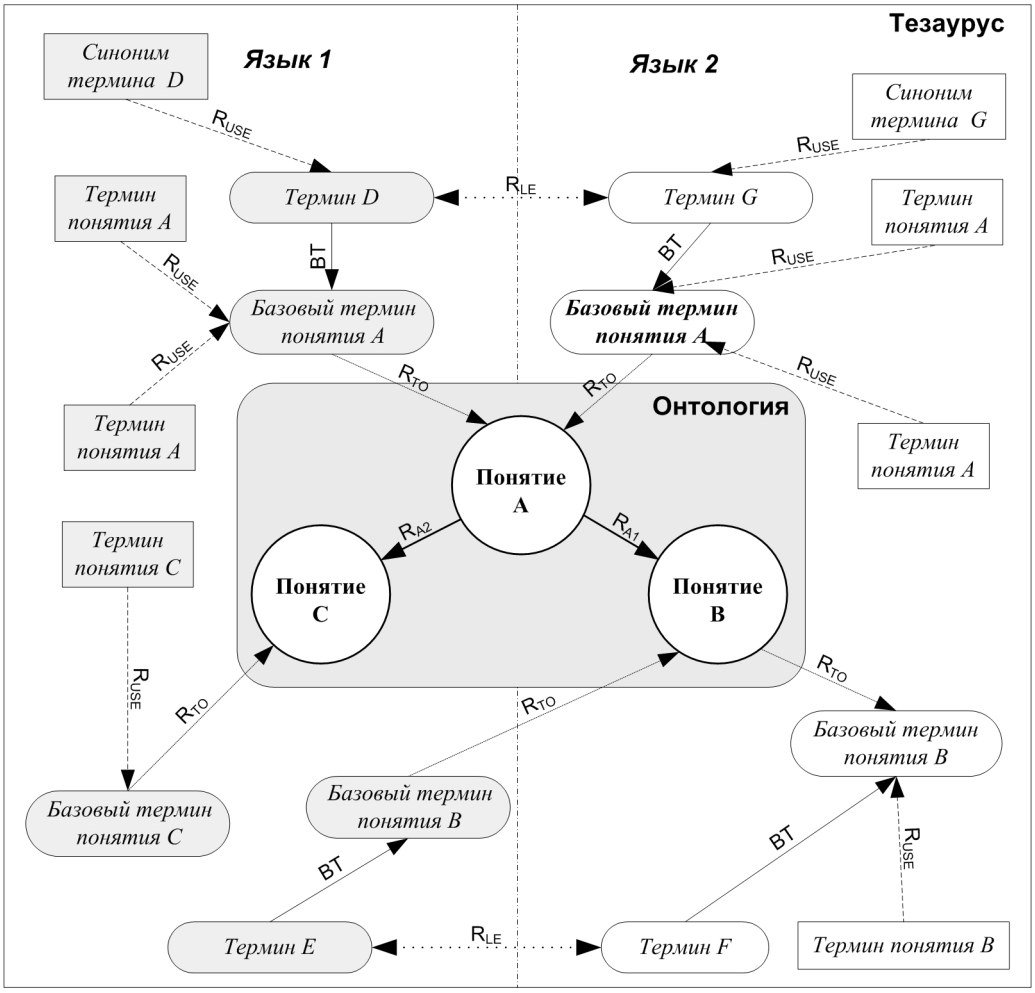

Рисунок. *Схема представления связей онтологии и тезауруса*

лам справочного характера, где уже дана какая-то систематизация понятий данной ПО. Все базовые понятия упорядочиваются в несколько иерархий «общее-частное» и, если необходимо, «часть-целое», вершиной каждой такой иерархии является соответствующее метапонятие онтологии предметного знания.

Затем полученная структура дополняется новыми понятиями и отношениями, существующими в ПО. Для этого собирается достаточно представительный корпус текстов, релевантных предметной области портала, и выполняется автоматическое извлечение из этих текстов значимой лексики, т. е. слов и словосочетаний, наиболее точно характеризующих данную ПО. Из полученного словника совместно с экспертами выбираются наиболее предпочтительные термины, которые будут использоваться в качестве названий понятий и отношений в онтологии ПО.

Для каждого понятия и отношения онтологии разрабатываются точные текстовые определения.

На следующем этапе выполняется кодирование онтологии, т. е. создание формальной спецификации онтологии, включающей:

- иерархии классов, описывающих понятия предметной области;
- множество заданных на классах отношений;
- множество атрибутов, описывающих свойства понятий и отношений;
- множество доменов, описывающих значения атрибутов.
- множества ограничений и аксиом, описывающих свойства понятий и отношений.

Следует заметить, что при построении онтологии могут использоваться справочные материалы и тексты, представленные на разных языках, но в спецификации онтологии должен использоваться только один язык.

Тезаурус строится как лингвистическое дополнение онтологии. Первой строится часть тезауруса, обслуживающая «главный язык» системы, т. е. тот язык, на котором инженеры знаний и эксперты разрабатывали онтологию. Для построения этой части тезауруса используется тот же корпус текстов, по которому строилась онтология, и полученный на этапе построения онтологии словник. Из словника выбираются все значимые термины и связы-

ваются между собой тезаурусными отношениями. Термины, выбранные в качестве названий понятий и отношений онтологии (базовые термины), также связываются отношениями R_{T0} с соответствующими элементами онтологии (понятиями и отношениями).

Построение частей тезауруса, соответствующих другим языкам, выполняется аналогичным образом: для каждого языка собирается достаточно представительный корпус текстов, релевантных ПО портала знаний, выполняется автоматическое извлечение лексики, составляется словник, выбираются термины, между ними устанавливаются тезаурусные отношения. Заключительным шагом включения нового языка в многоязычный тезаурус является связывание его базовых терминов с базовыми терминами «главного языка» отношениями эквивалентности.

Возможен и другой порядок, когда сначала с использованием двуязычных словарей, онтологии и полученного при ее создании словника, определяются термины нового языка, являющиеся эквивалентами базовых терминов «главного языка», а затем уже эта структура дополняется остальными терминами нового языка и тезаурусными отношениями.

Заключение

Предложен подход, обеспечивающий многоязычной содержательный доступ к знаниям и информационным ресурсам заданной предметной области на основе совместного использования онтологии и тезауруса. Связи, существующие между терминами тезауруса и понятиями онтологии, обеспечивают визуализацию представленной в портале знаний информации на разных языках, а также создают предпосылки для их совместного использования при поиске и обработке информации. Подход применяется для обеспечения многоязычного доступа в портале знаний по компьютерной лингвистике [8].

СПИСОК ЛИТЕРАТУРЫ

1. Guarino N. Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6–8 June 1998. Amsterdam, IOS Press. – P. 3–15.
2. Загорулько Ю.А. Построение порталов научных знаний на основе онтологий // Вычислительные технологии. – 2007. – Т. 12. – Спецвып. 2. – С. 169–177.
3. Лукашевич Н.В., Добров Б.В. Двуязычный информационный поиск на основе автоматического концептуального индексирования // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. конф. «Диалог-2003». – Протвино. – 11–16 июня 2003 г. – М.: Наука, 2003. – С. 425–432.
4. Dagobert Soergel. Multilingual thesauri and ontologies in cross-language retrieval // AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. – Stanford: Stanford University, March 24–26, 1997.
5. Нариньяни А.С. Кентавр по имени ТЕОН: Тезаурус + Онтология // Диалог'2001 по компьютерной лингвистике и ее приложениям: Труды Междунар. семин. – Аксаково, 2001. – Т. 1. – С. 184–188.
6. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям: основные принципы разработки и текущее состояние // X Национальная конференция по искусственному интеллекту с международным участием: Труды. – Обнинск, 25–28 сентября 2006 г. – М.: Физматлит, 2006. – С. 489–497.
7. Нгуен М.Х., Аджиев А.С. Описание и использование тезаурусов в информационных системах, подходы и реализация // Электронные библиотеки. – 2004. – Т. 7. – Вып. 1. – режим доступа: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part1/NA. – 17.02.2009.
8. Боровикова О.И., Загорулько Ю.А., Загорулько Г.Б., Кононенко И.С., Соколова Е.Г. Разработка портала знаний по компьютерной лингвистике // XI Национальная конференция по искусственному интеллекту с международным участием (КИИ-2008): Труды. – М.: ЛЕНАНД, 2008. – Т. 3. -С. 380–388.

Поступила 17.02.2009 г.