OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.89:004.4

ПОДХОД К ПОСТРОЕНИЮ ИНТЕЛЛЕКТУАЛЬНЫХ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ СЕМАНТИЧЕСКИХ СЕТЕЙ

Ю.А. Загорулько (*zagor@iis.nsk.su*)
Институт систем информатики им. А.П. Ершова СО РАН,
г. Новосибирск, Россия

Рассматривается подход к построению информационных систем на основе онтологий и семантических сетей. Описываются модели представления данных и знаний в информационной системе, которая имеет гибкую архитектуру и состоит из семантического ядра, обеспечивающего представление и хранение информации в виде сети знаний и данных, и подсистем, поддерживающих разработку и сопровождение онтологии и контента системы, а также представление знаний и данных конечному пользователю.

Ключевые слова: интеллектуальная информационная система, объектно-ориентированная семантическая сеть, онтология, семантическое ядро системы.

Введение

В связи с бурным ростом объемов информации в различных отраслях знаний все более актуальной становится задача эффективного информационного обеспечения научной и производственной деятельности, а также процессов принятия решений. Как правило, эта задача рассматривается в контексте создания хранилищ документов и их систематизации с целью облегчения поиска необходимой информации. Однако, возможностей, предоставляемых построенными в рамках такого подхода информационных систем, оказывается недостаточно для эффективной поддержки научной и производственной деятельности.

Это связано, в первую очередь, с ограниченностью методов и средств представления, поиска и интерпретации данных. В существующих системах данные в основном представляются в виде текстовых документов или формализованных записей баз данных, а интерпретация и представление данных в виде фактов, как правило, не поддерживается и возлагается на пользователя. В то же время для человека, будь то ученый или руководитель, наиболее естественной формой подачи информации является представление ее в виде множества взаимосвязанных фактов. Причем очень востребованы средства для анализа фактов, например, их сопоставления во временном и/или пространственном разрезе, быстрого определения источника данного факта, нахождения множества связанных с ним фактов и т.п. Такие средства могла бы предоставлять информационная система, использующая для представления фактов семантическую сеть, а для их интерпретации как общие знания о мире, так и знания о предметной области, для обслуживания которой она предназначена.

В связи с этим встает задача разработки информационной системы нового поколения, информация в которой представляется в виде сети знаний и данных. (Будем называть такую систему интеллектуальной информационной системой или ИИС). Для решения данной задачи необходимо разработать архитектуру интеллектуальной информационной системы, модели представления данных и знаний, методы «ручного» и автоматического пополнения базы знаний и контента ИИС, методы навигации по информационному пространству системы и содержательного поиска информации в терминах понятий заданной предметной области, а также методы автоматического извлечении знаний и фактов из документов деловой и научной тематики.

В данной работе будет рассмотрена только часть описанных выше задач, а именно – разработка модели представления данных и знаний и архитектуры ИИС.

1. Модели представления данных и знаний в ИИС

Модели представления данных и знаний тесно связаны, в частности, вторая задает интерпретацию первой, поэтому будем их рассматривать в рамках одной главы.

Предлагаемая модель представления данных в ИИС – это объектно-ориентированная семантическая сеть (ОО-сеть) следующего вида:

$$N_{OO} = \langle Ob, R, A_{Ob}, A_R, T, D \rangle . \quad (1)$$

где Ob – множество объектов, каждый из которых включает непустое множество атрибутов из множества A_{Ob}, определенных на типах из T или доменах из D,

$R = Ob \times Ob$ – множество бинарных отношений на объектах из Ob; любое отношение может иметь набор атрибутов из A_R, определенных на типах из T или доменах из D и служащих для специализации (уточнения) связи между объектами (аргументами отношения).

Такая модель является универсальным и достаточно гибким средством для представления структурированных данных (например, фактов) и связей между ними. Однако для того, чтобы эффективно и корректно пользоваться данными, представленными в ОО-сети, необходимо ввести еще один уровень представления – уровень знаний, позволяющий оперировать объектами ОО-сети как экземплярами понятий некоторой предметной или проблемной области. Уровень знаний обеспечивается онтологией представления знаний O_P.

Эта онтология должна обеспечивать представление как понятий предметной и проблемной области ИИС, так и разнообразных семантических связей между ними. Она также должна позволять выстраивать понятия в иерархию «общее–частное» и поддерживать наследование свойств по этой иерархии. Кроме того, она должна предоставлять возможность задания ограничений на значения возможных свойств объектов – экземпляров понятий онтологии.

Онтология представления знаний O_P, удовлетворяющая описанным выше требованиям, описывается следующей семеркой:

$$O_P = \langle C, R, T, D, A, F, Ax \rangle . \quad (2)$$

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

$R = \{R_1, \ldots, R_m\}, R_i \subseteq C \times C, R = \{R_T\} \bigcup \{R_P\} \bigcup \{R_A\}$ – 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 данных;

$D = \{d_1, ..., d_n\}$ – множество доменов $d_i = \{s_1, ..., s_k\}$, где s_i – значение стандартного типа из T;

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

$A = \{a_1, ..., a_w\}, A \subseteq C \times TD \bigcup R_A \times TD$ – конечное множество атрибутов, описывающих свойства понятий C и отношений R_A;

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

Ax – множество аксиом, определяющих семантику классов и отношений онтологии.

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 этому при поиске объектов можно осуществлять транзитивное замыкание по этому отношению. Набор ассоциативных отношений R_A определяется пользователем (разработчиком онтологии конкретной ИИС). Наличие таких отношений позволяет организовать содержательный поиск и навигацию по контенту ИИС. Важной особенностью отношений R_A является то, что они могут иметь собственные атрибуты, специализирующие связь между аргументами.

Онтология представления знаний, вводя формальные описания понятий проблемной области и области знаний ИС в виде классов объектов и отношений между ними, задает структуры для представления контента ИИС в виде реальных объектов и отношений предметной и проблемной области ИИС. Для хранения таких структур и предназначена объектно-ориентированная семантическая сеть, модель которой описана выше. Благодаря онтологии O_P в ИИС появляется формальная спецификация объектов и отношений семантической сети.

Таким образом, введение еще одного уровня представления в ИИС (в виде онтологии O_P) позволяет поднять уровень взаимодействия с ОО-сетью и повысить корректность работы с нею.

2. Архитектура интеллектуальной информационной системы

Интеллектуальная информационная система (рисунок 1), базирующаяся на описанных выше моделях данных и знаний, состоит из семантического ядра, обеспечивающего представление и хранение информации в виде сети знаний и данных, и подсистем, отвечающих (1) за разработку, верификацию и сопровождение системы знаний ИИС, (2) за разработку и развитие контента ИИС и (3) за представление знаний и данных конечному пользователю и другим информационным агентам.

Модули, входящие в семантическое ядро ИИС, представляют в системе уровень хранения (данных). Они обеспечивают все виды взаимодействия с объектно-ориентированной семантической сетью, являющейся основным хранилищем данных в системе. Эти модули являются обязательными компонентами всех ИИС, разрабатываемых в рамках рассматриваемого подхода.

В семантическое ядро ИИС входит менеджер ОО-сети, обслуживающий хранилище данных, и семантический модуль, предоставляющий весь набор операций над ОО-сетью. Отличие семантического модуля от менеджера ОО-сети состоит в том, что он обеспечивают работу с ОО-сетью на уровне системы знаний ИИС (в терминах классов и отношений онтологии), а менеджер ОО-сети – на уровне данных.

Возможны различные реализации модели данных, например, с использованием реляционной СУБД или средств для работы RDF-данными.

Рассмотрим, например, вариант, в котором модель данных реализуется через реляционную БД специального вида. В схеме такой БД каждому типу элементов объектно-ориентированной семантической сети – классу, атрибуту класса, отношению, атрибуту отношения, домену, объекту, значению атрибута объекта – соответствует своя таблица. Между таблицами установлены связи, позволяющие, например, связывать каждый объект ОО-сети с тем или иным классом онтологии, а каждое отношение ОО-сети – с определенным отношением онтологии. Благодаря такой схеме, достигается не только высокая гибкость в представлении знаний и данных, но и возможность динамического расширения наборов классов и типов отношений онтологии, а, соответственно, и видов объектов и отношений, представленных в контенте системы (ОО-сети).

В зависимости от специфики системы, она может включать тот или иной набор модулей других типов. Рассмотрим наиболее типичные из них.

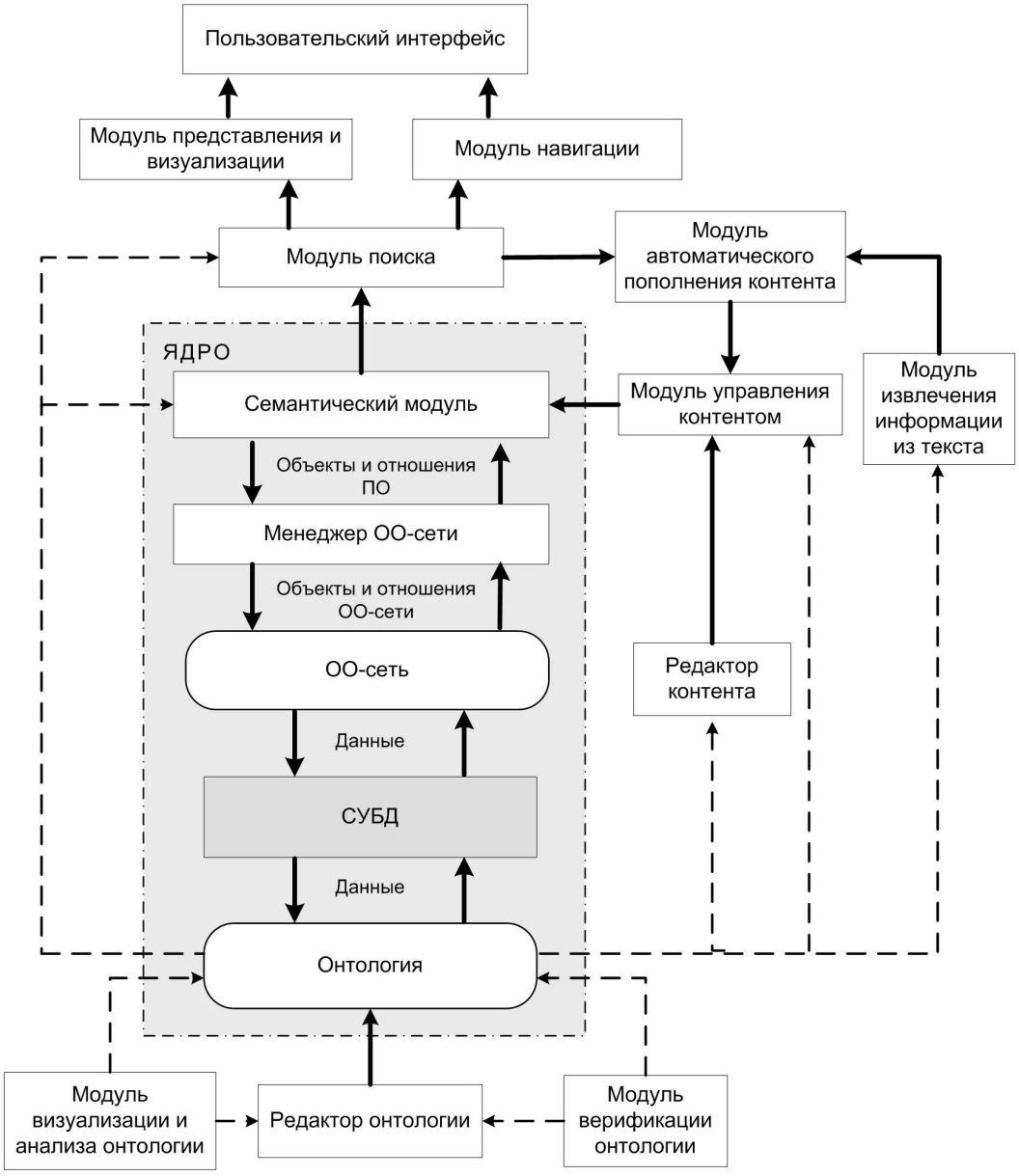

Рисунок 1 – Архитектура типовой ИИС

В интеллектуальную информационную систему должны входить модули, отвечающие за пополнение ее контента знаниями и данными. К ним относятся редакторы онтологий и контента, позволяющие вводить в хранилище данных знания и факты в ручном режиме. При этом редактор контента будет управляться онтологией ИИС. Развитые ИИС могут включать средства автоматического пополнения контента ИИС новыми фактами, а также автоматического извлечения информации из текста.

Для облегчения разработки и сопровождения системы в нее могут включаться модули визуализации, анализа и верификации онтологии и контента ИИС.

Для обеспечения интерфейса с конечными пользователями в систему включаются модули, отвечающие за навигацию по контенту ИИС и содержательный поиск информации в нем в терминах понятий предметной области системы, а также за представление знаний и данных конечному пользователю.

3. Построение онтологии ИИС

В соответствии с предложенной архитектурой каждая интеллектуальная информационная система должна иметь свою онтологию. Рассмотренная в разделе 1 онтология представления служит базисом для построения онтологии конкретной ИСС.

Онтология любой ИИС строится исходя из требований представления и организации знаний и данных в системе и с учетом ее функциональности. В общем случае она включает онтологию проблемной области, онтологию предметной области (области знаний) и онтологию задач (см. рисунок 2).

В зависимости от сложности строящейся системы и проработанности ее области знаний онтология системы может строиться либо непосредственно на основе онтологии представления знаний, либо путем достройки и развития ранее созданных базовых или прикладных онтологий.

Рассмотрим методику построения онтологии ИИС на основе базовых онтологий [Загорулько, 2007]. В качестве базовых выберем три онтологии: онтологию деятельности, которая составляет базис онтологии проблемной области ИИС, онтологию предметного знания, на основе которой строится онтология области знаний ИС, и онтологию базовых задач ИС, которая используется для построения онтологии задач ИС.

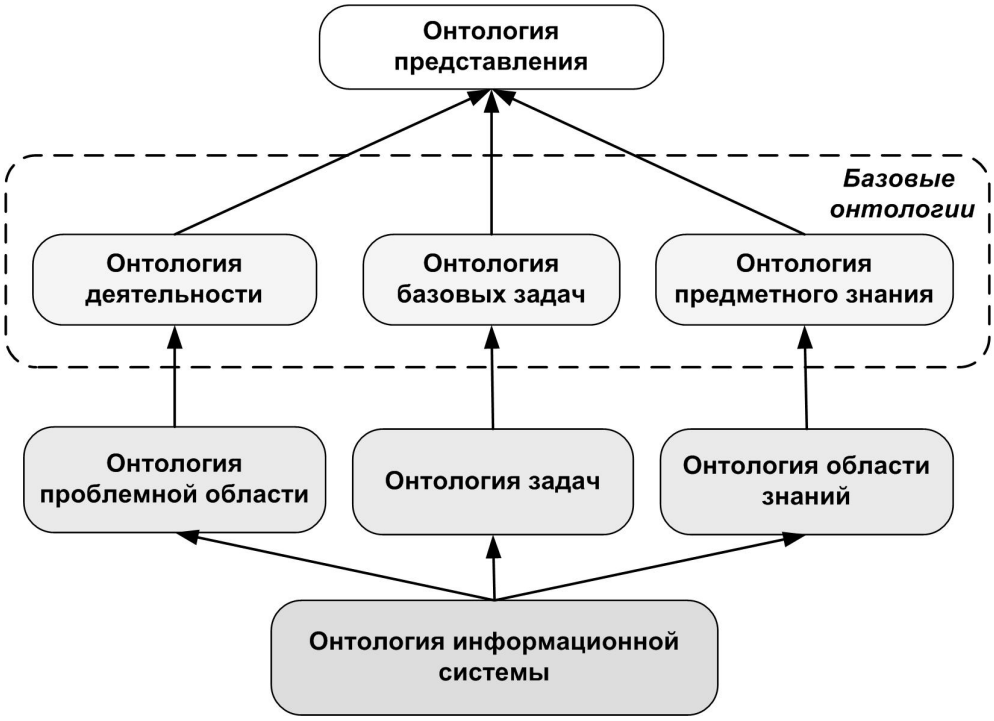

Рисунок 2 – Построение онтологии информационной системы

Первая базовая онтология характеризует проблемную область системы и, фактически, является онтологией верхнего уровня. В качестве такой онтологии может выступать, например, онтология научной и производственной деятельности, которая включает классы понятий, относящиеся к организации научной и производственной деятельности. В нее входят такие классы понятий, как Персона, Организация, Событие, Деятельность, Документ, а также класс Информационный ресурс, который служит для описания информационных ресурсов, представленных в сети Интернет.

Вторая базовая онтология – 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 (научной дисциплины) разделы и подразделы, задать типизацию методов и объектов исследования, описать результаты научной деятельности.

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

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

Онтология, построенная на основе рассмотренных выше методики и базовых онтологий, может использоваться в качестве уровня знаний ИИС, предназначенной для информационной поддержки научной и производственной деятельности, в частности, портала научных знаний [Загорулько, 2008].

Заключение

В докладе рассмотрен подход к построению интеллектуальных информационных систем на основе онтологий и семантических сетей. Описаны модели представления данных и знаний (объектно-ориентированная семантическая сеть и онтология представления знаний) и архитектура интеллектуальной информационной системы. Эта система имеет гибкую архитектуру и состоит из семантического ядра, обеспечивающего представление и хранение информации в виде сети знаний и данных, и подсистем, поддерживающих разработку и сопровождение онтологии и контента ИИС, а также представление знаний и данных конечному пользователю.

Наиболее важные компоненты, разработанные в рамках рассматриваемого подхода (модели данных и знаний, методы построения онтологий, методы и средства организации содержательного доступа) были использованы при построении таких ИИС, как порталы знаний по археологии [Андреева, 2006] и компьютерной лингвистике [Боровикова, 2008], а также прототипа электронного тезауруса по компьютерной лингвистике.

В настоящее время ведется работа по реализации модулей автоматического извлечения информации из текста и автоматического пополнения контента ИИС.

Благодарности

Работа выполняется при финансовой поддержке Президиума РАН (Интеграционный проект СО РАН № 2/12 в рамках программы РАН № 2) и РФФИ (проект № 09-07-00400).

Библиографический список

[Загорулько и др., 2007] Загорулько, Ю.А., Боровикова, О.И. Методологические проблемы построения и использования онтологий в портале научных знаний / Ю.А. Загорулько [и др.] // Тр. IX Международной конференции "Проблемы управления и моделирования в сложных системах". – Самара: Самарский Научный Центр РАН, 2007. – С. 447-454.

[Загорулько и др., 2008] Загорулько, Ю.А., Боровикова, О.И. Подход к построению порталов научных знаний / Ю.А. Загорулько [и др.] // Автометрия. – 2008 – № 1, Т. 44, – С. 100–110.

[Андреева и др., 2006] Андреева, О.А., Боровикова, О.И., Булгаков, С.В., Загорулько, Ю.А., Сидорова, Е.А., Циркин, Б.Г., Холюшкин, Ю.П. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии / О.А. Андреева [и др.] // Труды 10-й национальной конференции по искусственному интеллекту с международным участием (КИИ-2006). – М.: Физматлит, 2006. Т.3, – С. 832-840.

[Боровикова и др., 2008] Боровикова, О.И., Загорулько, Ю.А., Загорулько, Г.Б., Кононенко, И.С., Соколова, Е.Г. Разработка портала знаний по компьютерной лингвистике / О.И. Боровикова [и др.] // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием (КИИ-2008). – М.: ЛЕНАНД, 2008. Т.3, –С.380-388.