ИНФОРМАТИКА

УДК 004.822

СЕМАНТИЧЕСКАЯ ТЕХНОЛОГИЯ ПРОЕКТИРОВАНИЯ БАЗ ЗНАНИЙ

В. П. ИВАШЕНКО

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 16 мая 2009

Рассматривается технология проектирования баз знаний, основанная на универсальном семантическом способе кодирования знаний. Предлагается подход к построению баз знаний на модульной основе.

Ключевые слова: база знаний, семантическая сеть, онтология, язык представления знаний.

Введение

Для широкого применения интеллектуальных систем, способных повысить качество решения прикладных задач, необходимо большое количество баз знаний. Быстрой разработке достаточного количества баз знаний могло бы способствовать наличие средств разработки интеллектуальных систем, обеспечивающих разработку и проектирование различных компонентов интеллектуальной системы, включая базу знаний. Среди средств, которые могут рассматриваться в качестве основы для разработки баз знаний, можно выделить: оболочки экспертных систем (CLIPS (FuzzyCLIPS, DYNACLIPS, WxCLIPS), SOAR, OPS83, RT-EXPERT, MIKE, BABYLON, WindExS, ES; ACQUARE, Easy Reasoner, ECLIPSE, EXSYS Professional, SIMER+MIR, АТ ТЕХНОЛОГИЯ, CAKE v2.0) [1]; инструментальные пакеты для разработки экспертных систем (G2, ART, KEE, Knowledge KRAFT); системы, ориентированные на обработку онтологий – Protégé, WebOnto, OntoEdit, WebODE, OilEd, OntoLingua.

Достоинствами приведенных инструментальных средств являются: поддержка представления знаний различного вида различными моделями представления знаний в рамках одной системы; наличие средств визуального проектирования баз знаний; наличие средств верификации базы знаний, включая проверку на непротиворечивость; возможность монотонного расширения базы знаний, наличие средств интеграции баз знаний; наличие средств поддержки обмена данными с внешней средой, включая средства обмена данными в реальном времени. Однако для всех указанных средств характерны следующие недостатки: в силу различных ограничений велики сроки разработки баз знаний (отсутствие развитых технологий разработки); узок круг инженеров баз знаний (из-за высоких стартовых требований к разработчику) – от разработчика требуется владение специальными знаниями по моделям и языкам представления знаний; не полностью решен вопрос интеграции баз знаний.

В работе предлагается подход к созданию технологии проектирования баз знаний, в основе которой лежат следующие положения: модульное проектирование баз знаний, интеллектуализация средств поддержки проектирования баз знаний и семантическое представление знаний. Эта технология включает: унифицированную модель баз знаний; библиотеку ip-компонентов баз знаний и инструментальные средства проектирования баз знаний; методику проектирования и интеграции баз знаний.

Унифицированная модель баз знаний

Одной из основных проблем разработки унифицированной модели базы знаний является проблема разработки единых мер, критериев оценки качества представленных знаний. Проблема заключается в том, что, например, существуют критерии, разработанные для отдельных моделей представления знаний, которые не определены для других моделей: критерии непротиворечивости, разработанные для теорий в логических моделях, не всегда просто применить к моделям, в которых теория в явном виде отсутствует, или знания в которых носят императивный характер, а не декларативный; критерии нечеткости представленной информации, разработанные для отдельных нечетких множеств и предикатов, не всегда просто применить в случаях, когда имеются сложноструктурированные знания нечеткого характера; оценки количества информации, разработанные для традиционных способов кодирования, могут иметь различные значения в зависимости от того, как представлять знания такими способами. Поэтому требуется иметь систему мер, которые работали бы для любых видов знаний в рамках разрабатываемой модели и не зависели от особенности ее программной реализации. Тогда уже, исходя из такой системы, можно решать проблему унифицированного представления знаний, которые бы отвечали некоторым общим критериям качества. Унифицированная модель представления знаний использует унифицированное кодирование знаний, унифицированное представление знаний, описание структуры базы знаний, методы оценки качества и сравнения баз знаний.

Основой предлагаемых средств представления знаний является Semantic Code (SC-код) – способ семантического кодирования, обеспечивающий унифицированное кодирование знаний [2]. Особенностями SC-кода являются: простой алфавит, содержащий узлы и дуги, простой синтаксис, базовая теоретико-множественная интерпретация.

Унификация представления знаний обеспечивается семейством совместимых sc-языков, использующих унифицированный способ семантического кодирования (SC-код), которые обеспечивают представление базовых математических абстракций – множеств, чисел, отношений; знаний, которые могут быть представлены различными моделями представления знаний – фреймов, продукций, логических утверждений, нейросетевых моделей; знаний об информации, хранимой на различных медиа-носителях; информации о временных и причинно-следственных взаимосвязях; знаний, описывающих структуру и специфицирующих базы знаний, включая описание целей, вопросов и задач, на решение которых ориентированы интеллектуальные системы. Основным принципом построения sc-языков является представление понятий, соответствующих основным классам объектов, описываемых sc-языком, и отношений между этими объектами – ключевыми узлами такого sc-языка.

В частности, принципами построения гипермедийного sc-языка являются: рассмотрение понятий различных информационных конструкций, рассмотрение отношений между конструкциями одного типа, рассмотрение отношений между конструкциями разных типов (синонимия, трансляция, идентификация).

Принципами логического sc-языка описания нестационарных структур являются: использование нестационарных (ситуативных, временных) дуг принадлежности для представления ситуативных множеств; использование понятия состояния для представления статичных фрагментов предметной области; использование временных отношений между ситуативными множествами и состояниями; использование причинно-следственных отношений между ситуативными множествами и состояниями.

Принципами sc-языка целей, вопросов, задач и обобщенных задач являются: трактовка описаний целей и вопросов как открытых формул со свободными переменными, значения которых надо найти; трактовка вопросов, как целей частного вида, допускающих как минимум два различных целевых состояния – ответа; трактовка понятия задачи как отношения между описанием одного или нескольких допустимых исходных состояний и описанием одного или нескольких целевых состояний.

Множество всевозможных объединений текстов этих языков рассматривается как интегрированный sc-язык представления знаний.

База знаний – связная структурированная информационная конструкция, структура которой состоит не менее чем из одного атомарного раздела, каждому из которых принадлежит свой, описываемый в этом разделе, ключевой элемент этой конструкции, причем в описании

хотя бы одного ключевого элемента присутствует его внешнее обозначение (терм). Структура базы знаний включает:

- декомпозицию разделов базы знаний;
- непустое множество знаков разделов базы знаний;
- непустое множество знаков внешних обозначений;
- множество знаков информационных конструкций, не являющихся текстами sc-кода;
- непустое множество ключевых элементов базы знаний, включая множества знаков предметов, классов, отношений, атрибутивных отношений, утверждений, определений, теорий, задач, доказательств, программ.

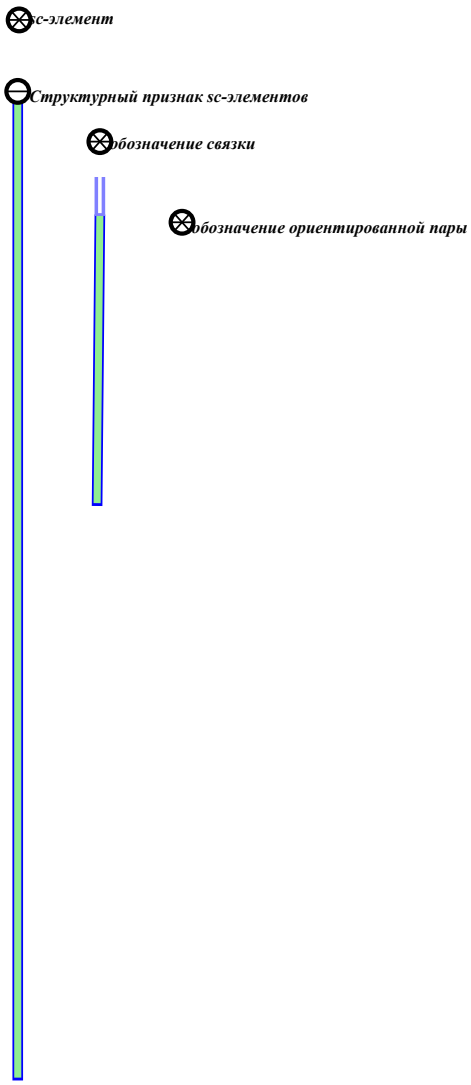

Рис. 1. Онтология SC-кода

К методам оценки качества и сравнения баз знаний относятся следующие: методы оценки объема базы знаний по различным типам элементов, методы проверки связности различных онтологических уровней базы знаний, методы оценки числа пар потенциальных синонимов в базе знаний, методы выявления ошибок в базах знаний.

Библиотека ip-компонентов баз знаний и инструментальные средства проектирования баз знаний

Основная проблема, которую необходимо решить при разработке библиотеки ip-компонентов – это выявление, в рамках разработанной унифицированной модели баз знаний, соответствующих закономерностей между разными оценками качества и построение на их основе спецификаций с целью обеспечения качества и систематизации ip-компонентов по соответствующим критериям.

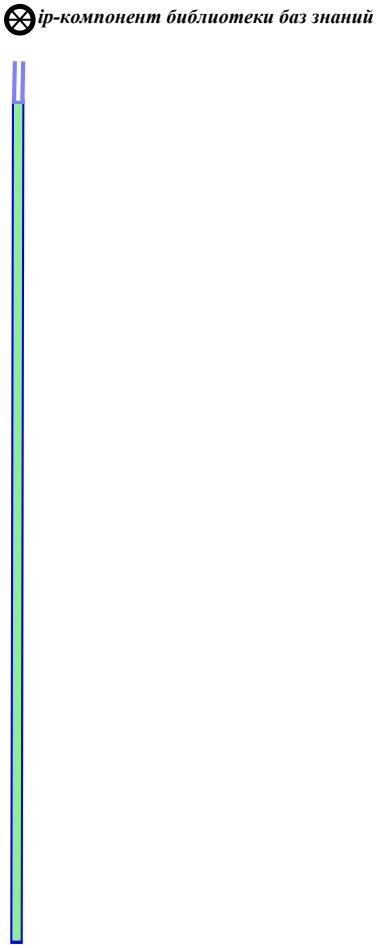

Рис. 2. Онтология ip-компонентов баз знаний

Еще одной проблемой разработки среды проектирования баз знаний является применение и разработка эффективных методов и алгоритмов автоматизации процесса разработки баз знаний, включая верификацию и поиск.

Библиотека ip-компонентов баз знаний рассматривается как sc-система; соответственно, структура библиотеки ip-компонентов баз знаний включает базу знаний, машину обработки знаний и пользовательский интерфейс. В базе знаний библиотеки ip-компонентов баз знаний хранится информация об ip-компонентах разных типов (рис. 2), в соответствии с их типологией, и спецификация ip-компонентов.

В соответствии со структурной типологией ip-компонентов можно выделить: неатомарные ip-компоненты и атомарные ip-компоненты. В состав неатомарных ip-компонентов входят другие ip-компоненты. В состав атомарных – нет. Над ip-компонентами определено отношение декомпозиции, с помощью которого неатомарные ip-компоненты декомпозируются на атомарные ip-компоненты.

Терминологические словари включают множество внешних обозначений ключевых элементов базы знаний, указание синонимии и омонимии терминов, задаваемое через связь терминов с понятиями (ключевыми элементами), которые они выражают. Терминологический

словарь является обязательной, неотъемлемой частью любых других ip-компонентов. В силу этого все ip-компоненты других типов являются атомарными ip-компонентами.

Таксономии включают множество классов основных объектов, рассматриваемых в базе знаний, и систематизацию этих классов на основе теоретико-множественных отношений, в частности отношения включения.

Модели включают множество классов основных объектов и отношений, определенных над этими основными объектами. Модели могут описывать не только отдельные состояния, но и истории, включая ситуативные множества.

Тезаурусы включают онтологию, т.е. множество основных объектов, их определения, пояснения и логическую схему определенных понятий. Среди тезаурусов (справочников) можно выделить следующие типы: справочники по языкам, справочники, не являющиеся справочниками по языкам. Среди справочников по языкам выделяются следующие специальные справочники по всем подъязыкам интегрированного sc-языка. К специальным справочникам также относятся справочник по унифицированному способу кодирования и справочник по графическому способу представления информации, закодированной SC-кодом.

Теории включают аксиоматику, определения, теоремы и другие утверждения, включая возможные задачи и гипотезы относительно основных объектов. Теории могут описывать как стационарные предметные области, так и динамические.

Прикладные базы знаний содержат информацию о предметной области, с конкретными примерами и описаниями, достаточную, чтобы получить исчерпывающие ответы на вопросы об описываемом объекте. Прикладные базы знаний могут включать любые другие вышеперечисленные ip-компоненты.

Таксономии, модели, тезаурусы, теории и справочники являются неатомарными ip-компонентами.

Спецификация ip-компонентов включает указание класса ip-компонента, описание его количественных и качественных характеристик, сертификат, состав, задачно-ориентированный сборник тестовых вопросов для этого ip-компонента, информацию о разработчиках, условиях распространения и информацию для сопровождения.

Операции машины обработки знаний библиотеки ip-компонентов баз знаний разбиваются на: навигационно-поисковые операции, позволяющие находить имеющиеся ip-компоненты по заданной полной или частичной спецификации; операции верификации ip-компонентов; операции добавления ip-компонентов; операции исключения ip-компонентов; другие операции редактирования библиотеки; операции анализа и мониторинга частоты использования ip-компонентов; операции добавления замечаний, отзывов и предложений.

Примерами вопросов, которые можно задать к библиотеке ip-компонентов баз знаний, являются: «найти ip-компонент, в котором есть определение треугольника»; «найти ip-компонент, в котором присутствует описание причинно-следственных закономерностей». Пользовательский интерфейс библиотеки ip-компонентов баз знаний включает знаки классов пользовательских команд редактирования библиотеки, поиска ip-компонентов баз знаний, верификации ip-компонентов баз знаний, обеспечения обратной связи с разработчиками.

Так же, как и библиотека ip-компонентов баз знаний, инструментальные средства проектирования являются sc-системой и их структура аналогична: база знаний, машина обработки знаний, пользовательский интерфейс. База знаний инструментальных средств проектирования баз знаний хранит программы операций редактирования, верификации и отладки, интеграции, а также образцы навигационных вопросов и конструкции гипермедийного языка. К операциям машины обработки знаний инструментальных средств проектирования баз знаний относятся: операции редактирования баз знаний, операции верификации и отладки баз знаний, операции интеграции фрагментов баз знаний. Пользовательский интерфейс инструментальных средств проектирования баз знаний содержит резидентные и нерезидентные узлы пользовательского интерфейса. Последними, в частности, являются: знаки исходных текстов справочной информации, знаки диалогов, знаки атомарных и неатомарных классов пользовательских команд. Знаки классов пользовательских команд включают знаки классов следующих команд: ввода\вывода, редактирования исходных текстов базы знаний, редактирования базы знаний, навигации, верификации и отладки.

Методика проектирования баз знаний

Методика проектирования баз знаний [3] строится на эволюционном подходе и включает описанные ниже этапы:

- разработка тестового сборника вопросов и ответов для проектируемой базы знаний;
- разработка исходных текстов, являющихся формальным представлением ответов на все вопросы тестового сборника;
- формирование набора (сигнатуры) понятий, по которой строится база знаний;
- сопоставление сформированной сигнатуры с имеющимися ip-компонентами с целью поиска имеющихся ip-компонентов;
- формирование списка используемых в базе знаний инструментальной среды ip-компонентов, включая языки представления знаний и готовые фрагменты баз знаний, выделение главной онтологии проектируемой базы знаний, а также – семейства вспомогательных онтологий;
- интеграция найденных ip-компонентов;
- разработка исходного текста отсутствующих онтологий;
- загрузка и отладка исходных текстов разработанных онтологических конструкций для проектируемой базы знаний;
- разработка содержательной структуры проектируемой базы знаний. Декомпозиция проектируемой базы знаний до уровня ее атомарных разделов;
- разработка исходных текстов всех атомарных разделов проектируемой базы знаний;
- поэтапная загрузка и отладка исходных текстов проектируемой базы знаний;
- интеграция автономно отлаженных разделов проектируемой базы знаний как друг с другом, так и с ранее разработанными ip-компонентами баз знаний;
- анализ и повышение качества разработанной базы знаний;
- оформление документации (итогового отчета о результатах проектирования базы знаний).

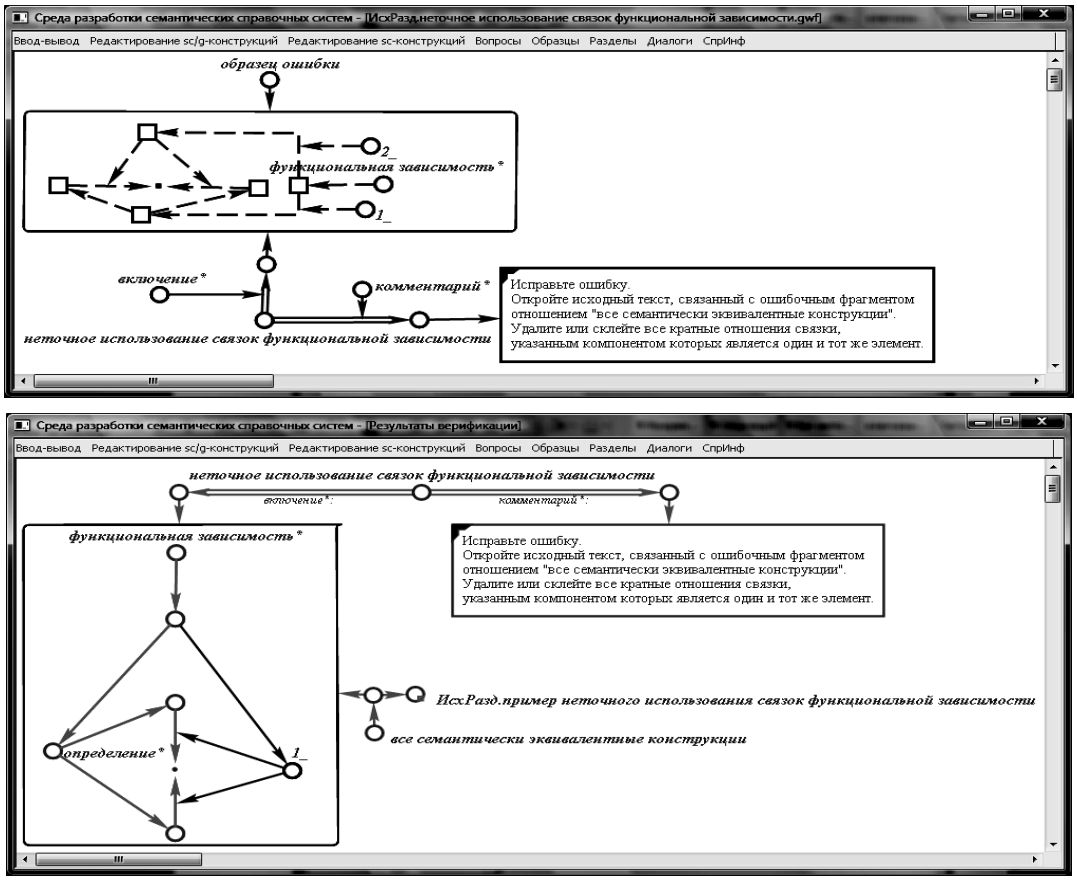

Рис. 3. Примеры описания ошибочной ситуации и диалога системы с пользователем на этапе отладки базы знаний

По всем этим этапам разработаны наборы соответствующих требований и рекомендаций. При формировании тестового сборника вопросов и ответов рекомендуется использовать типовые вопросы (о пояснениях, определениях, теоретико-множественные отношениях и другие) и специальные вопросы для заданной предметной области. При формировании набора понятий требуется обеспечить: связность системы понятий (наличие обобщающих понятий), отсутствие синонимичных (равных) понятий, малое число попарно непересекающихся классов (не больше 8±2), малое число отношений, которые могут быть представлены в виде прямого произведения или соединения других отношений и множеств. При разработке содержательной структуры базы знаний требуется использовать отношение декомпозиции разделов базы знаний и ориентироваться на использование специфицированных типов и шаблонов атомарных разделов.

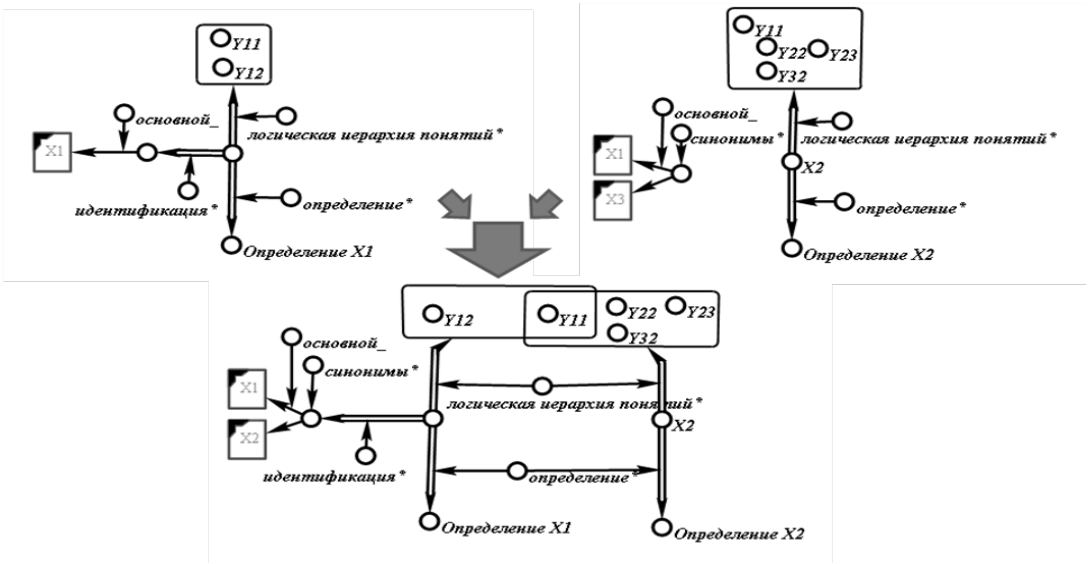

Рис 4. Иллюстрация интеграции фрагментов баз знаний

Рекомендациями по верификации и отладке являются: использование встроенных команд верификации и отладки, использование вопросов из тестового сборника вопросов. Основной проблемой в данной задаче является проблема интеграции фрагментов баз знаний: она включает решение таких задач, как выявление избыточности и синонимии, оптимизацию сигнатуры (набора понятий) базы знаний, другие задачи повышения качества базы знаний. Принципами и подходами для решения этих проблем являются: противопоставление понятий и терминов, выявление потенциальных синонимов в базе знаний, анализ логической схемы (иерархии) понятий базы знаний, использование экспертных знаний для повышения качества баз знаний.

Заключение

Особенностями и достоинствами предложенной технологии проектирования баз знаний являются:

- перенос акцента от традиционной инженерии баз знаний к их интеграции, т.е. к сборке баз знаний из крупных модулей, в том числе ip-компонентов баз знаний;
- унификация ip-компонентов баз знаний;
- обеспечение совместимости разрабатываемых баз знаний;
- широкое использование визуальных методов проектирования.

Работа выполнена при поддержке БРФФИ – РФФИ (грант № Ф08Р-137).

SEMANTIC TECHNOLOGY FOR KNOWLEDGE BASE DESIGN

V. P. IVASHENKO

Abstract

Semantic technology for knowledge base design based on Semantic Code is considered. Module approach for knowledge base engineering is described.

Литература

1. *Гаврилова Т.А., Хорошевский В.Ф.* Базы знаний интеллектуальных систем. СПб, 2000.
2. *Голенков В.В., Елисеева О.Е., Ивашенко В.П. и др.* Представление и обработка знаний в графодинамических ассоциативных машинах / Под ред. В. В. Голенкова. Минск, 2001.
3. *Гулякина Н.А., Ивашенко В.П.* // Докл. БГУИР. 2004. № 6. С. 113-119.