Л. В. Найханова

Основные аспекты построения онтологий верхнего уровня и предметной области

Аннотация

Статья посвящена исследованиям в области извлечения и представления знаний в виде онтологий. Рассмотрены виды онтологий, известные средства разработки онтологий. Предложены способы построения онтологий верхнего уровня и предметной области на основе аппарата семиотического моделирования. Онтологии строятся в виде тезаурусов, содержащих терминосистему и категорийно-понятийный аппарат предметной области. При построении онтологии верхнего уровня в качестве источников знаний используются терминологические словари, при построении онтологии предметной области — учебники, учебные пособия и монографии. Онтологии представляются в виде совокупности словарных статей на естественном языке и в виде семантической сети фреймов, имеющей вид иерархии. Создаваемые онтологии предназначены для разработки системы общего доступа к информации.

ВВЕДЕНИЕ

К настоящему времени в самых разнообразных областях человеческой деятельности накоплено большое количество информации, которая используется далеко не в полном объеме. Так, например, информация, организованная в базе данных современными ERP-системами, извлекается посредством плановых и неплановых запросов.

При этом, как правило, использование информации другими системами невозможно. Эта информация характеризуется различными форматами и способами представления, поэтому для того, чтобы накопленные данные приобрели широкую практическую ценность для развития науки и производства, необходимо свести разнообразно представленную информацию к общепонятному виду, что обеспечит возможность ее совместного использования разными системами. Решение этой проблемы связано с задачей извлечения и представления знаний, которая в настоящее время находится в центре внимания многих исследователей. На сегодняшний день в области искусственного интеллекта разработан ряд средств представления знаний, и к наиболее эффективным из них относится онтология.

Создание онтологий осуществляется не только при разработке сред, ориентированных на совместное использование информации несколькими пользователями, но также и при проектировании баз знаний, создании экспертных систем и систем поддержки принятия решений, разработке различных поисковых систем. В связи с тем что экспертные системы принятия решений и во многих случаях поисковые системы используют информацию, накопленную в хранилищах данных, то лучшим решением является создание онтологий уже при проектировании традиционных систем обработки данных на этапе изучения проблемной области и анализа требований. Для решения этой проблемы необходимо создать инструментальные среды, позволяющие осуществлять процесс построения онтологий в интерактивном и автоматическом режимах.

Онтология — это подробная спецификация структуры определенной проблемной области [1]. Основное назначение онтологий — интеграция информации. Онтологии связывают два важных аспекта: во-первых, они определяют формальную семантику информации, позволяя обработку этой информации компьютером, и, во-вторых, определяют семантику реального мира, позволяя на основе общей терминологии связывать информацию, представленную в виде, требуемом для компьютерной обработки, с информацией, представленной в удобной форме для восприятия человеком.

В работе [2] приводится классификация, в которой выделено семь уровней иерархии: онтологии представления, общие онтологии, промежуточные онтологии, онтологии

верхнего уровня, онтологии предметной области, онтологии задач и онтологии приложений.

Онтологии представления определяют концептуализацию, которая лежит в основе формализма представления знаний. Общие онтологии включают фундаментальные аспекты концептуализации, например, такие категории, как «род», «целое», «причина». Промежуточные онтологии содержат общие понятия и отношения, характерные для конкретной предметной области, они могут играть роль интерфейса между различными подобластями предметной области. Онтологии верхнего уровня являются конкретным назначением понятий общих и промежуточных онтологий. Онтологии предметной области содержат понятия определенной области знаний. Онтологии задач описывают определенные задачи области знаний или деятельности, релевантной этой области. Онтологии приложений являются специализацией онтологий предметных областей и задач [1, 2, 3].

Процесс построения онтологий может быть либо восходящим, либо нисходящим. Однако в связи с тем, что восходящий подход чрезвычайно трудоемок и пока не существует средств, которые позволили бы создать полную систему знаний («модель мира»), применяется в основном нисходящий подход к интеграции частных онтологий, ориентированный на конкретные, часто очень ограниченные практические задачи. Таким образом, существует проблема создания онтологий в узкой предметной области, которая ставит вопрос о создании четырех последних уровней иерархии. Неплохо было бы ввести построение этих онтологий как предварительный этап проектирования всех программных систем и особенно традиционных систем обработки данных, так как технологии проектирования современных систем обработки данных (СОД) предполагают разработку их спецификаций. Спецификации СОД строятся на основе изучения предметной области задачи, поэтому логично было бы расширить этот процесс до создания спецификации предметной области, тогда при очередной разработке программной системы расширялся бы объем знаний, что в конечном итоге привело бы к постепенному наращиванию системы знаний за счет соединения или интеграции знаний.

К настоящему времени получили известность средства создания онтологий, такие, как Ontoligua, OntoEdit, OilEd, Protégé, Web-Deso. Среда разработки Ontoligua предназначена для коллективного использования системы базовых знаний при построении собственных онтологий. Она пре-

доставляет разработчику библиотеку модулей, на основе которой осуществляется расширение онтологий. Среда разработки OntoEdit предназначена для проектирования, приспособления и импорта/экспорта моделей знаний в форматах RDF, DAML+OIL, Flogic для/из прикладных систем. Редактор онтологий OilEd в большей степени предназначен для проверки разработанных онтологий на согласованность. Система Protégé является библиотекой, предоставляющей доступ другим приложениям для просмотра баз знаний и позволяющей редактировать и наращивать базы знаний. Система Web-Deso предназначена для создания онтологий некоторой предметной области. Онтологии предметных областей соединяются в одну результирующую онтологию предметной области и помещаются в библиотеку вместе с источниками знаний. Такая же операция выполняется для онтологий задач. Сформированные в библиотеке онтологии интегрируются в онтологию-приложение, которая тоже хранится в библиотеке и предназначена для обеспечения многократного доступа к представленным знаниям. Сравнительные оценки перечисленных средств с точки зрения их внешней и внутренней организации приведены в работе [3]. В настоящей статье приведена сравнительная таблица средств управления онтологиями (см. табл. 1).

В данной работе рассмотрим основные аспекты создания онтологий верхнего уровня и предметной области, необходимые для создания онтологической системы, предназначенной для общего доступа.

1 СПОСОБ ПОСТРОЕНИЯ ОНТОЛОГИЙ ВЕРХНЕГО УРОВНЯ

Данные онтологии должны интегрироваться с уже созданными онтологиями или создаваемыми в перспективе. Это требование определяет то, что понятия и отношения, закладываемые в эти онтологии, носят общеизвестный характер и извлекаются из устоявшихся источников. В связи с этим онтологию верхнего уровня предлагается строить в виде тезауруса, описывающего терминологию предметных областей как терминосистему в виде словаря с концептуальным входом и фиксированными семантическими связями между его единицами с возможностью их редактирования в процессе функционирования.

Таблица 1

Сравнительная таблица средств управления онтологиями

Критерии	Средства				
	Ontoligua	**Protégé**	**OilEd**	**OntoEdit**	**Web-Deso**
Проекты и разработчики	Лаборатория систем знаний Университета Стэнфорда	Проект «Semantic Web», лаборатория медицинской информатики Университета Стэнфорда	Проект «On-To-Knowledge-Project», Университет Манчестера	Проект «Semantic Web», компания Ontoprise GmbH	«Система интеграции знаний», Институт информатики и автоматизации РАН, СПб.
Формат/формализм представления знаний	Логика первого порядка	OKBC-совместимая фреймовая модель знаний	Description Logic	RDF-совместимая фреймовая модель знаний	Объектно ориентированные сети ограничений
Методы моделирования понятий и отношений	Набор аксиоматизированных таксономий и отношений между ними	Сложная таксономия	Сложная таксономия и иерархия	Сложная таксономия и иерархия	Таксономия, иерархия, ассоциативные отношения
Возможность многократного использования онтологий	Включение	Включение, соединение	Включение	—	Соединение
Средства реализации	Архитектура клиент/сервер: Сервер — N/A Клиент — HTML-интерфейс	Самостоятельная Java-программа	Самостоятельная Java-программа	Самостоятельная Java-программа	Архитектура клиент/сервер: Сервер — IIS, ISAPI, MS Visual FoxPro Клиент — HTML-интерфейс, Java-script

Под терминосистемой будем понимать систематизированную совокупность терминов [9]. При построении тезаурусов будем использовать тематическую и иерархическую классификации — членение дисциплины науки на разделы или направления, а внутри раздела единицы тезауруса могут быть связаны между собой иерархическими и неиерархическими отношениями.

В тезаурусе определим два вида словарных статей — «понятие» и «действие» / «операция». Предполагается, что знания для создания терминосистемы будут извлекаться из терминологических словарей. Структура терминосистемы должна определять связи терминов, переходы внутри общей совокупности терминов; описывать семантику, синтактику и прагматику отдельных терминов; включать описание набора семантических предикатов, регулярно связывающих термины в научных текстах.

Для построения компьютерной терминосистемы будем использовать аппарат теории семиотического моделирования, предоставляющий математический базис для построения систем такого типа [4, 5, 6]. Рассмотрим квадрат Д. А. Поспелова [4], показанный на рисунке 1. В этом квадрате первая вершина определяет синтаксис, или способ кодирования знака, вторая — семантику, или понятие о знаке, третья соответствует прагматике — тем процедурам, которые связаны с этим знаком, четвертая — множеству знаков, или фрагменту некоторой структуры на множестве знаков (она играет роль денотата метазнака). Фрагмент структуры на множестве знаков обладает собственным именем, выделяющим его среди остальных. Это имя представлено в вершине 1, понятие о фрагменте дано в вершине 2, а связанные с ним действия — в вершине 3. Стороны квадрата и его диагональ соответствуют различным процедурам, связывающим компоненты знака. Метазнак образует вершины 1, 2 и 3 квадрата.

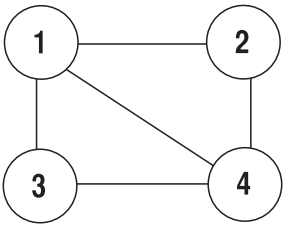

Рис. 1. *Квадрат Поспелова*

В соответствии с данной схемой представления знака в семиотической системе словарные статьи тезауруса можно представить в виде фреймов, а множество фреймов, описывающих термины тезауруса, должно образовывать семантическую сеть иерархического типа [7]. Дуги такой сети соответствуют различным связям между ними, при этом иерархические связи определяются отношениями структуризации, а неиерархические — отношениями иных типов. В связи с тем что построение тезауруса в виде семантической сети фреймов будет выполняться на основе анализа терминологического словаря или нескольких словарей, следует:

1) определить структуры словарной статьи и соответствующего ей фрейма;

2) определить отношения, которые необходимы для построения семантической сети, и способы их выявления;

3) построить иерархическую систему терминов в виде семантической сети понятий.

Так как нам требуется создать тезаурус для общего доступа, он должен иметь два представления: на естественном языке и формальном. Сначала необходимо создать тезаурус на естественном языке, а затем, проанализировав его и придав ему окончательную форму, на его основе можно построить семантическую сеть фреймов.

1.1. ОПРЕДЕЛЕНИЕ СТРУКТУР СЛОВАРНОЙ СТАТЬИ И СООТВЕТСТВУЮЩЕГО ЕЙ ФРЕЙМА

В качестве примера рассмотрим структуру словарной статьи «Понятие». В каждой словарной статье описывается одно понятие/термин. Термин — это знак специальной семиотической системы, обладающий номинативно-дефинитивной функцией или по второму определению: термин — это устоявшееся понятие [8]. Предполагаем, что в терминологических словарях описаны только устоявшиеся понятия рассматриваемой области знаний. Включим в словарную статью следующие элементы: денотат понятия, дефиниции понятия, свойства понятия, синонимы, оппозиты, список терминов, с которыми данное понятие имеет отношения. Тогда структуру словарной статьи «Понятие» можно представить в виде восьмерки:

T = <A, B, C, D, E, F, G, H>.

Рассмотрим элементы структуры.

$A = <A_1, A_2>$ — денотат понятия, где A_1 — уникальное имя понятия или имя словарной статьи; A_2 — знак понятия, соответствующий идентификатору фрейма.

$B= \{B_1, B_2, B_3, ...\}$ — множество дефиниций понятия. Для одного понятия может быть несколько дефиниций. Очевидно, что из общего набора утверждений для дефиниции отбирается только часть. Поскольку термин — это узел в сложной структуре теоретической системы, то различные дефиниции можно рассматривать как «разные пути или способы вхождения нашей мысли фактически в одну и ту же структуру» [8]. Обычно дефиницию определяют как словесно выраженный интенсионал, достаточный для задания экстенсионала, поэтому будем различать содержательную и формальную дефиниции. Множество содержательных дефиниций будет храниться в тезаурусе на естественном языке. Будем считать, что формальная дефиниция определяет понятие через другие понятия, а именно — через собственные свойства и связи с другими понятиями. Формальная дефиниция должна быть представлена в обоих видах тезауруса. Это согласуется с мнением Г. П. Мельникова [9], который выделяет следующие виды дефиниций: субстанциальные, структурные, тезаурусные и системные. Автор отмечает, что системная дефиниция — это такое сочетание субстанциональной (содержательной) и структурной (формальной) форм, при котором из структурной информации вытекает представление о наиболее вероятной субстанциальной информации, а из субстанциальной — наиболее вероятные структурные взаимодействия элементов поля терминосистемы. Таким образом, эти два аспекта обеспечивают представление о ее целостности и функциональной оправданности.

$C = <C_1, C_2, C_3, C_4>$ — множество типов описываемого концептуального объекта, в котором выделены четыре типа согласно Дальбергу:

- сущность C_1: материальные и нематериальные объекты, способы их рассмотрения;
- свойства C_2: количественные, качественные, релятивные (отношения);
- действия C_3: операции, процессы, состояния;
- величины (dimensions) C_4: время, положение, пространство.

$D = <D_1, D_2>$ — пара множеств свойств понятия, где D_1 — множество качественных свойств; D_2 — множество количественных свойств.

$E = \{E_1, E_2, E_3, ...\}$ — множество понятий, описывающих методы/функции, свойственные данному понятию, и отражающих прагматику, связанную с данным понятием.

$F = \{F_1, F_2, F_3, ...\}$ — множество синонимов понятия, или, другими словами, множество понятий, имеющих квантитативные отношения (отношение тождества) с данным понятием.

$G = \{G_1, G_2, G_3, ...\}$ — множество коррелятов, или, другими словами, множество понятий, имеющих отношение оппозиции к данному понятию.

$H = <H_1, H_2>$ — пара множеств понятий, имеющих квалитативные отношения с данным, где:

H_1 — множество понятий, составляющих отношение обобщения с данным $<H_{11}, H_{12}>$, H_{11} — родовое понятий, H_{12} — множество видовых понятий;

H_2 — множество понятий, составляющих отношение агрегации с данным $<H_{21}, H_{22}>$, H_{21} — понятие, являющееся «целым» по отношению к описываемому, H_{22} — множество понятий, являющихся частью описываемого.

Тогда протоструктура фрейма словарной статьи должна иметь следующую структуру:

```
{Идентификатор термина
<имя термина, значение слота>
<тип концептуального объекта, значение>
(качественные свойства понятия
<имя качественного свойства понятия, значение>*)
(количественные свойства понятия
<имя количественного свойства понятия, значение>*)
(действие, связанное с понятием
<имя действия, значение>*)
(синонимы
<имя синонима, ссылка>*)
   (корреляты
<имя коррелята, ссылка>*)
(понятия, имеющие родовидовые отношения с данным
<имя рода, ссылка>*
<имя вида, ссылка>*)
(понятия, имеющие отношения «целое—часть» с данным
<имя целого, ссылка>*
<имя компонента, ссылка>*)
(метазнаки
<имя метазнака верхнего уровня, ссылка>*
<имя метазнака нижнего уровня, ссылка>*)
}
```

Протоструктура фрейма имеет три уровня иерархии. В ней по аналогии с регулярными выражениями символ «*» означает итерацию, круглыми скобками выделены вложенные уровни иерархии.

В терминологических словарях словарная статья начинается с имени термина, а дефиниции, как правило, начинаются сразу после имени термина либо через тире, либо они пронумерованы, поэтому их распознавание не составляет особого труда. Остальные элементы словарных статей тезауруса необходимо искать с помощью диагностирующих конструкций, выявляя отношения между понятиями.

1.2. ОСНОВНЫЕ ВИДЫ ОТНОШЕНИЙ, НЕОБХОДИМЫЕ ДЛЯ ПОСТРОЕНИЯ СЕМАНТИЧЕСКОЙ СЕТИ, И СПОСОБЫ ИХ ВЫЯВЛЕНИЯ

По Дальбергу, к концептуальным отношениям относятся:
- квантитативные (совпадают с логическими отношениями тождества, включения, пересечения, дизъюнкции);
- квалитативные (в большинстве онтологические и включают в себя отношения иерархии, части—целого, оппозитивные и функциональные).

Отношения иерархии (обобщение). Эти отношения принадлежат сфере «быть» и делятся на три подвида:

1) род<—>вид;
2) признак<—>значение признака;
3) инвариант<—>вариант.

Родовидовое или внутрикатегориальное отношение род<—>вид определяет принадлежность к классу, строится на основе комплексного использования параметра смысла и объема номинации и подразделяется на структурные, функциональные, семантические.

Основные диагностирующие конструкции предложения для поиска родового понятия В и видового понятия А:
- А относится к В;
- родом А является В;
- А принадлежит классу/семейству В.

Основные диагностирующие конструкции предложения для поиска видовых понятий В и родового понятия А:
- А имеет следующие виды: B_1, B_2, ...;
- к видам А относятся B_1, B_2, ...

Отношение *признак<—>значение* или отношение *типизация-конкретизация*. Здесь признак или тип — это категория, а значение — это имя конкретной категориальной формы.

Основные диагностирующие конструкции предложения для поиска А — понятия-категории, В — имени категории:

— А — В;
— А имеет имя В;
— А именуется В;
— А называется В.

Отношение *инвариант<—>вариант*. В этом отношении инвариант — это неизменная структурная единица языка науки, а вариант — разновидность структурной единицы языка науки.

Основные диагностирующие конструкции предложения для поиска понятия инварианта А и варианта В:

— к вариантам А относятся B_1, B_2, ...;
— инвариантом В является А.

Отношения агрегации. Они принадлежат сфере «иметь». Лингвисты выделяют четыре подвида:

— целое<—>часть (компонент);
— объект<—>пространство реализации объекта (локализации или позиции);
— объект<—>свойства/признак;
— уровень<—>единица уровня.

В отношении *целое<—>часть (компонент)* целое — это то, что перестает существовать, если отнять любую часть; часть — то, что без целого не существует. Для их поиска будем использовать диагностирующие конструкции, из которых состоит библиотека образцов:

— целое В — часть А:
  А входит в состав В;
  А составляет часть В;
  А является частью;
  А — элементом В;
— часть В — компонент А:
  А состоит из В-ов;
  А включает в себя В как часть;
  А включает в свой состав В как часть;
  А имеет своими частями/элементами B_1, B_2, ..., B_n.

В отношении *объект<—>пространство реализации объекта* пространство реализации — это то пространство, где объект А проявляет свои свойства и функции.

Диагностирующие выражения:
— место А — В;
— А происходит в В;
— А находится в В, А входит в В.

В отношении *объект<—>свойства* свойство — это то, что присуще предметам, что отличает их от других; свойства делятся на существенные и несущественные. В некоторых случаях вместо термина «свойство» удобно использовать термин «признак». Признак — все то, в чем предметы, явления сходны или отличны (показатель, сторона предмета), при этом существуют отличия свойства/признака от части. К ним относятся:

1) логические: свойство не должно иметь субстратного сходства с объектом и свойство не может существовать само по себе, но может проявляться в другой сущности;

2) языковые признаки: свойство имеет номинацию, в которой его «признаковость» выражена морфологическим способом: аффикс «-ость» определяет абстрактное свойство, кроме того, свойство не является главной частью речи.

Основные диагностирующие выражения для поиска следующие:
— А обладает свойством В, имеет В существенным признаком;
— А — В (В — прилагательное или наречие);
— А характеризуется наличием В;
— для А характерно В.

В отношении *уровень<—>единица уровня* выделяют такие единицы, как уровень, единица уровня, тип структуры. Для их диагностики используются следующие выражения:
— А принадлежит уровню В;
— А — единица уровня В;
— А рассматривается на уровне В.

Так как мы используем квадрат Д. А. Поспелова, в котором представлен метазнак, то введем отношения, которые назовем семиотическими.

Семиотические отношения. Данный вид отношений предназначен для выражения соответствий между понятиями знаковых систем. Посредством этих отношений можно создавать иерархии понятий знаковых систем, при этом каждый уровень иерархии будет соответствовать одному метаязыку. Виды семиотических отношений:
— термин<—>способ выражения;
— термин<—>способ представления термина;
— термин одного метаязыка<—>термин второго метаязыка.

Отношение *термин<—>способ выражения* отражает фундаментальные свойства языка как системы знаков, имеющей план выражения и план содержания. Как знаковая система, предназначенная для коммуникации, язык обладает совокупностью средств выражения самых разнообразных значений. Отсюда следует, что в языке должна быть терминология, относящаяся к формальной стороне языка, и терминология, называющая значения и функции. Таким образом, в данном отношении в качестве языкового объекта используется значение или функция. Диагностирующие выражения для прямого отношения:

— А обычно выражает В (пример: флективный класс (А) выражает морфологическую информацию (В));
— А используется для выражения В (пример: номинатив (А) обычно используется для выражения агенса (В));
— А является В.

Таким образом, в данном отношении осуществляется связь двух языковых объектов.

Отношение *термин<—>способ представления термина* фиксирует связь языкового объекта и его представления в модельном языке лингвистики (метаязыке), где ненаблюдаемый теоретический языковый объект может получить наглядный аналог во фрагменте или элементе некоторой модельной записи. Диагностирующие выражения для прямого отношения:

— А обычно представляется посредством В;
— А используется для представления В;
— А используется с помощью В.

Примеры: язык исчисления предикатов (А) применяется в качестве способа представления информационного языка (В); вершины графа (А) используются для представления знаменательных слов (В).

Диагностирующие выражения для обратного отношения: А является представлением В.

Примеры: предложение обычно представляется в виде дерева грамматического разбора; результатом синтаксического анализа является дерево грамматического разбора; результат синтаксического анализа можно представить в виде дерева грамматического разбора. Вокабулы — это способ представления лексической единицы. Толкование — способ представления значения слова. Помета — способ представления грамматических, стилистических и других не собственно семантических характеристик слова.

Таким образом, в данном отношении осуществляется связь языкового и метаязыкового объектов.

Этот тип отношений позволяет строить иерархии, при этом каждый уровень будет соответствовать одному метаязыку, кроме первого, так как на первом уровне определяется отношение между термином, относящимся к плану содержания, и терминами, относящимися к плану выражения.

Отношение *термин одного метаязыка<—>термин второго метаязыка* позволяет установить соответствие между двумя знаковыми системами как в одной предметной области, имеющей разные уровни представления знаний, так и между смежными предметными областями.

Диагностирующие выражения для прямого отношения:
— А представлено в виде/как В;
— А представлено с помощью В;
— способом представления А является В.

Примеры:
— сетевая грамматика (А) имеет своим способом представления языковой ориентированный граф (В);
— значение слова (А) представляется в словарях в виде толкования (В);
— семантическое отношение (А) между терминами может быть представлено в тезаурусе как тезаурусная функция (В).

Различие между способом выражения и способом представления осуществляется при помощи падежа или предлога, а представляется в дереве зависимостей в виде нумерованной стрелки.

Один и тот же объект может иметь множество способов представления. Это зависит от целей описания и разработанности искусственных языков. В кибернетических лингвистических моделях существует целая цепочка способов представления структуры текста и языковых единиц — от чисто лингвистических до чисто машинных.

Семиотическим отношениям соответствует логическая функция — эквиваленция. На основе анализа семиотических отношений будет создано множество связей $Q = <Q_1, Q_2, Q_3>$, описывающих связи между фрагментами (словарными статьями или фреймами) семантической сети тезауруса, где:

Q_1 — способ выражения термина, который отражает свойства языка как системы знаков;

Q_2 — способ представления термина, который фиксирует связь языкового объекта и его представления в модельном языке;

Q_3 — способ представления термина метаязыка высшего/низшего уровня, который устанавливает соответствие между знаковыми системами как в одной предметной области, имеющей разные уровни представления знаний, так и между смежными предметными областями.

При описании словарной статьи «Действие» необходимо рассмотреть функциональные отношения, релевантные квалитативным отношениям.

Функциональные отношения. Отношение, принадлежащее сфере процессуальности, представляет собой многоместный предикат «Операция», в котором выделяют следующие аргументы (места):

A: субъект операции;

B: инструмент или способ/метод/алгоритм выполнения операции;

C: начальный объект или исходные данные;

D: конечный объект или результат/выходные данные;

E: событие, активизирующее операцию.

Субъектом операции может выступать человек или устройство. Диагностирующие выражения для выявления субъекта операции:

— A совершает B;
— A B-ет.

Примеры:

— анализатор (A) отсекает (B) окончания (C);
— алгоритм (A) распознает (B) значение слова (C) и приписывает ему соответствующий индекс (D).

Начальный объект/исходные данные. При анализе научных текстов в предложениях распознается объект, над которым совершается операция. В спецификациях информационных систем это исходные данные некоторой функции или процедуры.

Диагностирующие выражения для выявления начального объекта:

— A осуществляется над B;
— A применяется к B;
— A меняет B;
— A B-ет.

Примеры:

— операция пассивизации (A) применяется к активной конструкции (B);
— умлаут A меняет гласную корня (B);
— начальный объект сегментации (A) — предложение (B).

Конечный объект, или результат/выходные данные. При анализе научных текстов в предложениях распознается объект, являющийся конечным результатом выполнения операции. В спецификациях информационных систем это выходные данные некоторой функции или процедуры.

Диагностирующие выражения для выявления конечного объекта:

— в результате А получается В;
— в результате А образуется В;
— А приводит к В;
— А превращает B_1 в B_2.

Примеры:

— в результате операции номинализации (А) образуется номинализованная конструкция (В);
— насыщение валентностей (А) приводит к образованию насыщенной синтаксической структуры (В);
— операция пассивизации (А) превращает активную конструкцию (B_1) в пассивную (B_2).

Инструмент, или способ, или метод, или алгоритм. Объединяет разнородные термины, часть из них действительно является названием метода или приема, и это зачастую отражается в их дефинициях.

Наличие диагностирующих предложений при описании различных видов позволяет решать эту задачу как задачу диагностики, которая предполагает определение множества А нормальных состояний, в которых искомое отношение может существовать в той или иной логико-синтаксической структуре предложения. Однако некоторые конструкции диагностирующих предложений повторяются у разных типов отношений. Это означает, что может быть сформировано конфликтное множество нераспознанных отношений. Часть из них может быть разрешена за счет анализа лексиса понятий, а часть — путем введения контрпримеров.

Таким образом, решение задачи можно свести к построению правил, определяющих, в каком отношении находится понятие. База правил может быть построена в виде системы продукций *pr*. Каждая продукция *pr* представляет собой пару <*q*, *r*>, где *q* — условие применимости, *r* — программа, которая выполняется, если условие применимости истинно. Так как каждое предложение представляет собой некоторую ситуацию, в которой оказались лексемы, то условие применимости должно описывать одну из возможных ситуаций, которые свойственны тому или иному

отношению, и иметь предикатное представление. При представлении диагностирующих конструкций в предикатном представлении учитываются морфологические и синтаксические характеристики лексем и диагностирующих предложений. В данной статье вид предикатов и их классификация не представлены из-за большого объема данного раздела.

На вход анализатора предложений терминологического словаря подается текущее предложение, которое представляется в виде текущей ситуации, и проверяется условие применимости продукции. Если условие применимости продукции истинно, то выполняется программа *r*, цель которой — заполнение элементов словарных статей.

1.3. ПОСТРОЕНИЕ ИЕРАРХИЧЕСКОЙ СИСТЕМЫ ТЕРМИНОВ

После анализа терминологического словаря будут заполнены не все элементы словарных статей тезауруса, а только та часть, которая имеется в терминологическом словаре в описании терминов. Для дальнейшей систематизации терминов, т. е. построения иерархической системы, необходимо выполнить анализ заполненных словарных статей. Для этого будем использовать следующие типы операций обобщения: по именам, признакам, характеристикам и структуре.

Обобщение по именам. Эта операция связана с установлением отношения «элемент/класс» между некоторой группой знаков, т. е. установление определенного сходства сущностей в процессе сравнения или сопоставления их признаков (свойств), в результате которого может порождаться новое, более общее понятие, которое объединяет целый класс подобных понятий, что соответствует абстракции типизации. Это новое понятие может находиться в другой словарной статье, и тогда необходимо установить связи между этими фрагментами тезауруса. Если такое понятие отсутствует в тезаурусе, то его необходимо ввести. Добавление понятия можно осуществлять в интерактивном режиме или использовать другой терминологический словарь. Таким образом, данная операция является основой способа поиска пары противоположных абстракций *типизация—конкретизация*.

Обобщение по признакам. Эта операция устанавливает отношение «вид/род» между некоторой группой знаков на ниже/вышележащем уровне по отношению к уровню, где располагается анализируемый термин. При выявлении родового понятия анализируются понятия, расположенные на одном уровне с рассматриваемым термином, и осуществляется объединение понятий на основе их сходства или выявленного подобия части признаков. При этом порождается новое (или существующее в тезаурусе) понятие, которое и является обобщением исходных понятий. Как и в предыдущем случае, прописываются связи между фрагментами. При поиске видовых понятий текущего понятия осуществляется анализ понятий нижележащего уровня. Как правило, используется нисходящий подход при поиске родовидовых отношений. Операция *обобщение по признакам* в такой интерпретации является основой способа поиска пары противоположных абстракций *обобщение—специализация.*

Эта операция используется также для поиска отношения «инвариант<—>вариант». В этом случае необходимо полное сходство признаков двух рассматриваемых понятий. Если это условие выполняется, то в тезаурусе прописывается их взаимосвязь.

Обобщение по характеристикам. Эту операцию будем использовать для установления связей между понятиями, находящимися в отношении «часть—целое». Это отношение невозможно определить посредством каких-либо операций, поэтому будем полагать, что они будут прописаны при анализе терминологического словаря. Здесь необходимо только прописать ссылки между ними.

Обобщение по структуре. Обобщение такого типа становится возможным лишь на уровне семиотического моделирования, когда используется язык, ориентированный на знаковую сущность моделей окружающего мира. Эта операция может быть использована для поиска семантических отношений в тезаурусе.

Все операции формализуются в продукционные правила. Применение продукций позволит окончательно построить иерархию понятий в виде семантической сети. На этом этапе должно быть сформировано фреймовое представление словарных статей.

2 СПОСОБ ПОСТРОЕНИЯ ОНТОЛОГИЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Онтологию предметных областей предлагается строить в виде тезауруса, описывающего категорийно-понятийный аппарат предметной области. Структуры тезаурусов идентичны, однако в данном случае роль источника знаний играют учебники, учебные пособия, монографии. Для построения категорийно-понятийного аппарата предметной области необходимо решить следующие задачи:

1) выделение понятий научного текста и их классификация на терминологию и профессионализмы;

2) построение взаимосвязей между понятиями;

3) анализ тезауруса и выявление общенаучной, межнаучной и частнонаучной терминологий.

2.1. ВЫДЕЛЕНИЕ ПОНЯТИЙ НАУЧНОГО ТЕКСТА

Для выделения понятий в научном тексте предлагается использовать статический способ, основанный на частотном принципе. Для этого необходимо в начале выполнить лексический, морфологический и синтаксический анализы текста. В результате анализов осуществляется преобразование текста в поток лексем с характеристиками, отражающими морфологические признаки, а также создается синтаксическая структура предложений текста.

Лексемы делятся на классы. Примерами таких классов являются грамматические классы слов, такие, как существительные, прилагательные, глаголы и т.д. Лексический анализатор должен выдавать следующую информацию: поток основ слов или множество векторов лексем $L = \{l_i \,|i=1..k,\ k$ — общее количество лексем в потоке$\}$, множество $L^S=\{\rho_i^{ls}|i=1..k',$ k' — количество разновидностей лексем в потоке, $k' \leqslant k\}$, вектор ρ_i^{ls} содержит статические характеристики лексемы l_i и множество векторов $L^V=\{\rho_i^{ls}|i=1..k\}$, которые описывают динамические характеристики лексемы l_i, зависящие от контекста. Вектор ρ_i^{ls} содержит значения параметров лексемы l_i, которые характеризуют лексему в общем:

$$\rho_i^{ls} = <n_i, l_i, f_i, m_i, p_i>,$$

где n_i — уникальный номер вектора ρ_i^{ls}; l_i — основа лексемы; f_i — частота встречаемости лексемы в тексте; m_i — класс лексемы (здесь часть речи); p_i — указатель на группу векторов, описывающих динамические параметры лексемы.

Вектор ρ_i^{ls} содержит значения таких параметров, которые отражают морфологические и синтаксические характеристики лексемы, такие, как падеж и число лексемы для существительных и прилагательных, адрес лексемы в тексте E текста d:

$$\rho_i^{ls} = <p_i, n, c_i, a_i>,$$

где p_i — уникальный номер вектора лексемы ρ_i^{ls}; n — уникальный номер вектора лексемы ρ_i^{ls}; c_i — морфологическая информация; a_i — адрес лексемы l_n; $a_i = <n_i^l, n_i^s, n_i^p, n_i^d, n_i^c>$, n_i^l — порядковый номер лексемы в предложении; n_i^s — порядковый номер предложения в документе; n_i^p — номер параграфа; n_i^d — номер раздела; n_i^c — номер главы.

Алгоритм морфологического анализа базируется в большей части на алгоритме, предложенном в [11]. Для выполнения морфологического анализа используются следующие словари S:

- готовых словоформ, выраженных существительными S_1;
- основ существительных S_2;
- окончаний существительных S_3;
- основ прилагательных и причастий S_4;
- окончаний прилагательных S_5;
- основ глаголов S_6;
- окончаний глаголов S_7;
- наречий S_8;
- предлогов S_9;
- союзов S_{10};
- морфологической информации S_{11}.

Словарь готовых (неизменяемых) словоформ — это упорядоченный по алфавиту перечень лексем-существительных, неизменяемых в зависимости от грамматической формы. Словарь наречий — упорядоченный по алфавиту перечень наречий со слитным написанием.

Словарь окончаний существительных — это перечень всех возможных окончаний имен существительных.

Словарь основ существительных, прилагательных, глаголов имеет структуру вида: [*номер основы*] [*номер флективного класса*] [*основа*]. Номер флективного класса — трехзначный: первая цифра означает часть речи (1 — существительное, 2 — прилагательное), две последующие цифры — порядковый номер флективного класса в пределах одного грамматического класса слов.

Для определения необходимой морфологической информации, а именно числа и падежа для имен существительных и прилагательных, используется таблица соответствия флективного класса и окончания S_{12}.

После обработки всех словоформ документа осуществляется «упорядочивание» векторов ρ_i^{ls} множества L^V по элементам n, n_i^c, n_i^d, n_i^p, n_i^s, n_i^l, которое фиксируется значениями элементов p_i векторов ρ_i^{lv}. В этом случае несколько векторов ρ_i^{ls} с подряд идущими значениями p_i описывают все случаи использования лексемы l_n. Их количество представляет собой частоту встречаемости f_n данной лексемы в документе. Наименьший номер p_i векторов ρ_i^{ls}, описывающих лексему l_n, является указателем p_n вектора ρ_n^{ls}. Таким образом, результаты анализов текста будут представлены множествами L, L^S и L^V.

В связи с тем что термин может состоять из нескольких лексем, необходимо рассмотреть вопрос о возможных словосочетаниях, обозначающих термин. Слова или словосочетания выполняют в предложении определенную семантико-синтаксическую функцию. Классификация слов по морфологическим признакам предполагает их разбиение на части речи: существительное, глагол, прилагательное, наречие и т. д. Сформируем множество терминов научного текста, выраженных словосочетаниями, T^D.

Для выделения словосочетаний рассмотрим их лексико-грамматические типы. Различают следующие лексико-грамматические типы словосочетаний: глагольные, именные, наречные [10]. Глагольные словосочетания имеют следующие модели:

1) глагол + существительное или местоимение (с предлогом или без предлога): купить хлеба, обратиться к нему;

2) глагол + инфинитив или деепричастие: просит приехать, сидеть задумавшись;

3) глагол + наречие: поступать правильно, повторять дважды.

Именные словосочетания делятся на субстантивные, адъективные, с главным словом числительным и с главным словом местоимением.

Основные модели субстантивных словосочетаний:

1) согласуемое слово + существительное: ясный день, мой мир;

2) существительное + существительное: город в огнях, отрывок из поэмы;

3) существительное + наречие: шаг вперед, лов зимой;

4) существительное + инфинитив: готовность помочь, повод поговорить.

Основные модели адъективных словосочетаний:

1) прилагательное + наречие: по-праздничному нарядный, едва слышный;

2) прилагательное + существительное (местоимение): широкий в плечах, равнодушный ко всему;

3) прилагательное + инфинитив: способный организовать, готовый сопротивляться.

Последние типы словосочетаний с главным словом числительным и главным словом местоимением являются синтаксически несвободными и не отличаются разнообразием моделей: двое друзей, два товарища, некто в белом, что-нибудь особенное.

Словосочетания наречного типа (с предикативными и непредикативными наречиями) имеют две модели:

1) наречие + наречие: по-летнему жарко, весьма вкусно;

2) наречие + существительное: больно руку, высоко в горы, задолго до праздника.

Из приведенных в предыдущем разделе классификаций видно, что для построения категорийно-понятийного аппарата научно-технического текста необходимо использовать модель субстантивных именных словосочетаний, выражаемую схемой: согласуемое слово + существительное. В этой модели существительное является стержневым словом, а согласуемое слово — зависимым и может выражаться как прилагательным, так и существительным. В общем случае именные словосочетания могут включать в свой состав следующие классы слов: существительные, прилагательные, предлоги, сочинительные союзы и наречия. Количество слов в именных словосочетаниях колеблется от двух до пятнадцати и в среднем составляет три слова [10].

Множество различных структур именных словосочетаний приведено в таблице 2.

Таблица 2

Структурный состав именных словосочетаний

№ п/п	Схема структуры	Примеры
1	Прилагательное + существительное	Информационная система Образовательное сообщество
2	Прилагательное + прилагательное + существительное	Информационная поисковая система Научное образовательное сообщество
3	Прилагательное + прилагательное + прилагательное + существительное	Сетевое научное образовательное сообщество Управляющая цифровая вычислительная машина
4	Существительное + существительное в родительном падеже	Система подготовки Сообщество детей
5	Существительное + прилагательное + существительное в родительном падеже	Система образовательных сообществ Активизация научной деятельности
6	Существительное + прилагательное + прилагательное + существительное в родительном падеже	Система научных образовательных сообществ
7	Существительное + прилагательное + прилагательное + + прилагательное + существительное в родительном падеже	Система сетевых научных образовательных сообществ
8	Прилагательное + существительное + существительное в родительном падеже	Автоматический поиск информации Научная деятельность учащихся
9	Прилагательное + существительное + существительное в родительном падеже + существительное в родительном падеже	Автоматизированная система поиска информации
10	Прилагательное + прилагательное + существительное + прилагательное + существительное в родительном падеже	Распределенная поисковая система научной информации
11	Существительное + существительное в родительном падеже + + существительное в родительном падеже	Система поиска информации Система подготовки кадров

Продолжение

№ п/п	Схема структуры	Примеры
12	Существительное + существительное в родительном падеже + + существительное в родительном падеже + существительное в родительном падеже	Разработка системы поиска информации Структура системы подготовки кадров Система управления базами данных
13	Прилагательное + существительное + существительное в родительном падеже + прилагательное + существительное в творительном падеже	Концептуальные основы управления научной деятельностью
14	Прилагательное + существительное + предлог + прилагательное + существительное в родительном падеже	Образовательная коммуникация для конкретного учащегося
15	Прилагательное + существительное + предлог + существительное в родительном падеже + + существительное в родительном падеже	Информационная система для поиска информации
16	Существительное + существительное в родительном падеже + + сочинительный союз + существительное в родительном падеже + существительное в родительном падеже	Система хранения и поиска информации Система подготовки и переподготовки кадров
17	Существительное + предлог + существительное в предложном падеже + существительное в родительном падеже	Взаимодействие в области науки
18	Прилагательное + существительное + существительное в родительном падеже + сочинительный союз + существительное в родительном падеже	Комплексный подход становления и развития Образовательное сообщество детей и взрослых

Необходимым является выделение словосочетаний всех приведенных в таблице структур. Для поиска вышеперечисленных словосочетаний используются продукционные правила. Для каждого словосочетания или понятия,

выраженного одной лексемой, формируются словарные статьи.

Заполнение элементов словарных статей можно выполнить посредством продукций, созданных для диагностических конструкций, описанных в разделе «Основные виды отношений, необходимых для построения семантической сети, и способы их выявления».

2.2. ПОСТРОЕНИЕ ВЗАИМОСВЯЗЕЙ МЕЖДУ ПОНЯТИЯМИ

Формирование иерархической системы признаков можно выполнить способом, аналогичным описанному способу в разделе «Построение иерархической системы терминов».

2.3. АНАЛИЗ ТЕЗАУРУСА И ВЫЯВЛЕНИЕ ОБЩЕНАУЧНОЙ, МЕЖНАУЧНОЙ И ЧАСТНОНАУЧНОЙ ТЕРМИНОЛОГИЙ

Будем считать, что терминосистема содержит понятия общенаучного характера для данной области знаний, а тезаурус, содержащий категорийно-понятийный аппарат, должен содержать частнонаучную систему понятий. Межнаучная система понятий выявляется в том случае, когда имеется несколько тезаурусов, описывающих смежные области знаний.

При анализе созданного категорийно-понятийного аппарата должны анализироваться два созданных тезауруса. Для этого должна использоваться операция «сходства/различия», которая заключается в проверке денотата, дефиниций и свойств понятия на соответствие. Применение данной операции может привести к следующим результатам:

1) полное совпадение;
2) частичное совпадение;
3) полное различие.

Полное совпадение означает, что понятие носит общенаучный характер и в данной словарной статье необходимо оставить только ссылку на тезаурус, содержащий терминосистему, и все элементы, кроме денотата, следует обнулить.

Частичное совпадение означает, что данное понятие тоже носит общенаучный характер, но имеет конкретное назначение для описываемой предметной области. В этом случае необходимо в денотат понятия добавить ссылку на тезаурус, содержащий терминосистему.

Полное различие означает, что понятие носит частнонаучный характер и принадлежит только описываемой предметной области.

Таким образом, в работе применяются следующие методы исследований. В качестве основного формализма представления знаний используется логика предикатов первого порядка, вспомогательными формализмами являются фреймы, системы продукций, семантические сети. Для построения иерархии на основе исследования семантических и онтологических отношений между понятиями используются методы моделирования понятий и отношений. Семиотическое моделирование для построения иерархии на основе исследования отношений между конкретными предметными областями выполняется на основе методов моделирования онтологических моделей. Средством прототипной программной реализации является комплекс Java-программ. Описание характеристик создаваемых онтологий приведено в таблице 3.

ЗАКЛЮЧЕНИЕ

В статье рассмотрены основные аспекты построения онтологий верхнего уровня и предметной области, позволяющие создавать онтологии в автоматическом режиме. При этом описаны виды онтологий, приведены известные средства разработки онтологий, предложены способы построения онтологий верхнего уровня и предметной области.

Онтологии верхнего уровня строятся в виде терминосистемы определенной области знаний на основе исследования терминологических словарей и представляют собой тезаурус, имеющий вид базы фактов. Онтологии предметных областей строятся как категорийно-понятийный аппарат предметной области на основе исследования учебников, учебных пособий и монографий.

При создании моделей онтологий используется аппарат семиотического моделирования, предложенный Д. А. Поспеловым и развитый его учениками. Для реализации операций моделирования используются системы продукций. Такой подход позволяет обновлять базу продукционных правил, для чего разработана инструментальная среда, которая вводит правила на ограниченном подмножестве естественного языка.

Таблица 3

Основные характеристики создаваемых онтологий

№ п/п	Виды онтологий	Характеристики				
		Содержание	**Источник знаний**	**Структура представленных знаний**	**Формальная структура на языке знаний в электронном виде**	**Декларативный компонент системы**
1	Онтологии верхнего уровня	Терминосистемы определенной области знаний	Терминологический словарь	Иерархия понятий в виде семантической сети	Семантическая сеть фреймов	Совокупность словарных статей
2	Онтологии предметных областей	Категорийно-понятийный аппарат предметной области	Учебники, учебные пособия и монографии	Иерархия понятий в виде семантической сети	То же	То же
3	Онтологии задач	Методы решения конкретных задач или деятельностей	Аналитик, спецификации структур и методов обработки данных	База продукционных правил	Представление методов решения задач как множеств возможных ситуаций	База продукционных правил
4	Онтологии-приложения	Правила вывода	Онтология предметных областей и онтологии задач	Семантические сети и база продукционных правил	Система доступа в виде системы вывода, основанной на знаниях	

В перспективе предполагается строить онтологии задач, содержащие методы и их решения, представленные в виде базы продукционных правил.

Литература

1. *Guriano N.* Understanding, Building, and Using Ontologies / A Commentary to "Using Explicit Ontologies in KBS Development" // International Journal of Human and Computer Studies, 1997. V. 46. № 2/3.

2. *Guriano N., Gangemi A., Pisanelli D.M., Steve G.* An Overview of the ONIONS Project: Applying Ontologies to the Integration of Medical Terminologies // Data & Knowledge Engineering, 1999. V. 31.

3. *Смирнов А.В.* Онтологии в системах искусственного интеллекта: способы построения и организации / А. В. Смирнов, М. П. Пашкин, Н. Г. Шилов, Т. В. Левашова // Новости искусственного интеллекта. — 2002. — № 1.

4. *Поспелов Д.А.* Прикладная семиотика и искусственный интеллект / Д. А. Поспелов // Программные продукты и системы. — 1996. — № 3.

5. *Поспелов Д.А.* Введение в прикладную семиотику / Д. А. Поспелов, Г. С. Осипов // Новости искусственного интеллекта. — 2002. — № 6.

6. *Осипов Г.С.* От ситуационного управления к прикладной семиотике / Г. С. Осипов // Новости искусственного интеллекта. — 2002. — № 6.

7. *Вагин В.Н.* Знание в интеллектуальных системах / В. Н. Вагин // Новости искусственного интеллекта. — 2002. — № 6.

8. *Никитина С.Е.* Семантический анализ языка науки / С. Е. Никитина. — М.: Наука, 1987.

9. *Мельников Г.П.* Основы терминоведения / Г. П. Мельников. — М.: Изд-во ун-та дружбы народов, 1991.

10. *Беловольская Л.А.* Синтаксис словосочетания и простого предложения (ttp://www.philology.ru/linguistics2/belovolskaya-01.htm).

11. *Белоногов Г.Г.* Автоматизированная обработка научно-технической информации. Лингвистические аспекты / Г. Г. Белоногов, Б. А. Кузнецов, А. П. Новоселов // Итоги науки и техники. — М.: ВИНИТИ, 1984. — Т. 8. — (Сер. Информатика).