РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

2005, том 41, № 5

УДК 519.95

Н. Г. Загоруйко, В. Д. Гусев, А. В. Завертайлов, С. П. Ковалев,
А. М. Налетов, Н. В. Саломатина

(Новосибирск)

СИСТЕМА ONTOGRID ДЛЯ АВТОМАТИЗАЦИИ ПРОЦЕССОВ ПОСТРОЕНИЯ ОНТОЛОГИЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Представлен проект инструментальной системы OntoGrid для автоматизации построения онтологий предметных областей с использованием Grid-технологий и анализа текстов на естественном языке. Рассматривается содержание и текущее состояние разрабатываемых блоков системы OntoGrid.

Введение. Онтологией (О) называется краткое описание структуры предметной области (ПрО), которое включает в себя термины (Т), обозначающие объекты и понятия ПрО, отношения (R) между терминами и определения (D) этих понятий и отношений:

$$O = T, R, D .$$

В графическом представлении онтология имеет вид сети, вершины которой обозначены терминами и отношениями ПрО, а ребра указывают на связи между ними. Начальная вершина, которая содержит название ПрО, связана отношением «часть-целое» с вершинами следующего уровня, представляющими собой базовые категории данной ПрО. Каждая категория связана с вершинами следующего уровня (понятиями) своими отношениями и т. д. Вершины сети могут быть связаны с соответствующими разделами метаинформации, содержащими указание на литературные источники.

Построенная онтология предметной области будет полезна для совершенствования следующих областей деятельности.

1. Системы обучения. Действительно, для первого знакомства с предметной областью было бы полезно иметь в качестве «опорного сигнала» легко воспринимаемую структуру этой области. С помощью онтологии можно быстро находить ссылки на источники информации.

2. Поисковые системы. Наметившийся сейчас переход от поиска информации по ключевым словам к использованию семантически значимых фрагментов текстов существенно облегчается, если используется онтология ПрО.

3. Научные исследования. Большое значение имеет унификация терминологии ПрО. Наличие онтологии ПрО позволит автоматизировать процесс отслеживания полезных данных и знаний в потоке текущей информации.

4. Системный анализ предметной области. Онтология предоставляет структурированную и частично формализованную основу для проведения системного анализа предметной области.

5. Интегрирование данных и знаний. При объединении информационных баз онтология будет помогать устанавливать семантическую эквивалентность одинаковых фактов и понятий, сформулированных в разных терминах.

Почти все известные разработки инструментов для построения онтологий [1] ориентированы на то, что источником знаний, которые нужно отобразить в онтологии, является эксперт в данной прикладной области, освобожденный от программистской работы. Между тем как в процессе разработки, так и в ходе эксплуатации онтологии необходимо постоянно отслеживать новые знания, которые появляются в информационных сетях обычно в виде текстов на естественном языке. Отсюда вытекает необходимость автоматизированного обнаружения нужных знаний в информационных потоках, для чего требуется оснастить инструментальную систему лингвистическим процессором.

Онтология только тогда будет принята научным сообществом, если в ее разработке участвовал широкий круг экспертов данной ПрО. Это требует создания поддержки коллективной деятельности экспертных групп, географически удаленных друг от друга. Удобной технологической средой для реализации такого инструмента является Grid-сеть – распределенная информационно-вычислительная инфраструктура, построенная на основе технологии динамической интеграции вычислительных ресурсов.

В связи с вышеизложенным данный проект нацелен на создание системы автоматизации построения и развития онтологий предметных областей (системы OntoGrid), которая должна быть оснащена лингвистическим процессором, работающим с русскими и английскими текстами, и реализована при помощи Grid-технологии.

В следующих разделах описываются отдельные блоки разрабатываемой системы OntoGrid.

Создание лингвистической базы знаний. Любые работы, связанные с автоматическим анализом текстов, требуют определенного набора лингвистических и алгоритмических ресурсов, основу которых составляют машинные словари (толковые, словообразовательные и другие) и программы морфологического и локального синтаксического анализа, выделения терминологической лексики и т. д.

В настоящее время нами разработаны и реализованы: морфологическая база русского языка; блок морфологического анализа; блок статистического анализа текстов; программа выделения устойчивых словосочетаний в тексте с учетом их морфологической и комбинаторной изменчивости; программа выявления аномалий в позиционном распределении лексических единиц по тексту.

Базой для процедуры морфологического анализа служит электронный словарь Д. Уорта, содержащий свыше 100 тыс. канонических форм [2]. Процедуру индексации (по Зализняку) для большей части словаря удалось автоматизировать, для чего было составлено порядка 200 правил. Полученная таким образом морфологическая база содержит 3,2 млн. словоформ с соответствующими значениями грамматических категорий рода, числа, падежа, времени, лица и т. п. Если искомая словоформа имеется в базовом словаре, то анализ заканчивается ссылкой на ее каноническую форму и морфологиче-

ские характеристики. Иначе выдается сообщение об обнаружении «нового» слова, которое можно нормализовать и назначить ему грамматическую информацию исходя, в частности, из того, что слова, имеющие сходный буквенный состав концов, как правило, имеют аналогичные грамматические характеристики. Эксперименты показали, что такой подход позволяет правильно восстановить каноническую форму новых слов примерно в 80 % случаев.

Основу статистического анализа текстов составляет процедура вычисления их L-граммных характеристик. Термин «L-грамма» здесь означает цепочку из L подряд следующих слов текста. Частотной характеристикой порядка L текста T будем называть совокупность всевозможных представленных в нем L-грамм с указанием частот их встречаемости $_L(T)$. Совокупность частотных характеристик (T) { $_1(T)$, $_2(T), \ldots,$ $_{L_{\max}}(T)$} будем называть полным частотным спектром текста *T*. Здесь $L_{\max}$ – длина максимальной повторяющейся цепочки слов в тексте. Аналогом $_L(T)$ для группы текстов $\overline{T}$ $\{T_1, T_2, \ldots, T_m\}$ является совместная частотная характеристика L-го порядка $_L(\overline{T})$, содержащая частотную информацию только об L-граммах, общих хотя бы для пары текстов из заданной группы. Совокупность общих частотных характеристик со значениями L от 1 до $L_{\max}(\overline{T})$ образует совместный частотный спектр группы текстов $\overline{T}$. Здесь $L_{\max}(\overline{T})$ – длина максимальной цепочки слов, представленной, как минимум, в паре текстов из $\overline{T}$. Совместные частотные характеристики служат основой для вычисления различных теоретико-множественных мер близости для пар и групп текстов [3].

Важную роль при анализе текстов играют устойчивые словосочетания [4]. В основе предложенного нами алгоритма выделения словосочетаний лежит последовательное вычисление частотных характеристик (L $2,3,\ldots$ $\ldots, L_{\max}$) и фильтрация повторяющихся L-грамм в соответствии с критерием устойчивости [5]. Анализ комбинаторной вариативности выделенных «устойчивых» цепочек нацелен на выявление «устойчивых конструкций» типа образцов (или шаблонов): «не только *X*, но и *Y* », «целью … является...», «особенность … заключается …».

Существенное значение при выявлении ключевой лексики имеет информация о распределении слова по длине текста. Слова, демонстрирующие неравномерное распределение, обычно оказываются более значимыми, чем распределенные равномерно (примером последних являются служебные слова). Особо важными считаются повторы, многократно встречающиеся только в одном фрагменте текста. Они могут служить основой для выделения так называемых «сверхфразовых единств» [6]. Нами предложен новый метод выявления в тексте сверхфразовых единств, образуемых сгущениями лексических единиц определенного типа. Апробация метода на литературных и научных текстах позволила сделать следующие выводы [7].

1. Основной механизм кластеризации словоформ в тексте связан с развитием сюжетно-тематической линии.

2. Наиболее информативными оказываются кластеры, представленные среднечастотными словоформами.

3. Большая часть слов, демонстрирующих аномалии в позиционном распределении, относится к категории существительных и прилагательных.

4. При оценивании значимости слов по частоте или функциональному назначению требуется вести контроль позиционного распределения для всех словоформ текста, исключая низкочастотные.

5. Список слов и словосочетаний, демонстрирующих значимые позиционные аномалии, содержит подавляющую часть ключевых слов, указанных самим автором.

Построение семантических сетей текстовых документов. Под системой анализа текста обычно понимается система, для которой определены следующие элементы: формализм для представления смысла текста; база лингвистических знаний (БЛЗ); отображение, переводящее текст в выбранный формализм; набор алгоритмов решения задач анализа текстов, использующих в качестве данных полученное семантическое представление; интерфейс эксперта, если предусмотрено его участие.

Среди классических задач, на решение которых ориентированы такие системы, можно упомянуть классификацию текстов, реферирование, семантически ориентированный поиск текстов по заданным концептам и др. Достаточно широкое распространение получил подход к анализу текста, опирающийся на онтологию как на формальную модель ПрО. При этом система анализа текста проецирует онтологию на текст, выделяет в нем объекты из объема понятий ПрО и связи между ними. Для этого необходимо, чтобы в онтологию входило описание способов реализации понятий и отношений ПрО в текстах. Основной задачей системы анализа текстов при построении онтологии видится как раз автоматизация формирования проекции онтологии (ПроекОнт) на ЕЯ-тексты. Исходя из этого, авторы накладывают следующие требования на БЛЗ и систему в целом.

1. На первом этапе БЛЗ должна представлять собой зачаток ПроекОнт, необходимый для начала функционирования системы. Этот зачаток вносится экспертом.

2. В системе должны быть реализованы механизмы развития БЛЗ в ходе анализа потока текстов ПрО, а также возможность контроля этого развития экспертом. На каждом этапе развития БЛЗ должна являться некоторым приближением к ПроекОнт, на основе которого можно решать задачи анализа текста. На некотором уровне развития БЛЗ должна содержать в себе ПроекОнт.

3. Структура и содержание БЛЗ системы должны быть удобны как при построении семантических представлений текстов, так и при дальнейшем анализе этих представлений.

Пирамидальные семантические сети. В соответствии с изложенными требованиями авторами разрабатывается и реализуется система анализа текста (САТ). В качестве формализма для представления смысла текста удобно использовать семантические сети, удовлетворяющие следующим требованиям.

1. Однородность. Внутренний язык сети должен позволять единообразно описывать элементы разных типов (объекты, отношения, ситуации).

2. Иерархичность. Сеть должна обладать развитыми ассоциативными свойствами и отражать иерархичность реальных объектов.

3. Функциональность. В сети должны быть реализованы процессы формирования связей между семантическими объектами, выделения классов объектов и ситуаций и формирования обобщенных определений этих классов.

4. Полнота. Все части текста, описывающие существенные единицы предметной области, должны быть отражены в сети соответствующими вершинами.

5. Прозрачность. Сеть должна иметь по возможности простой вид и позволять делать удаления несущественных вершин и ребер.

В системе CAT мы используем семантические Q-сети [8], в основе которых лежит аппарат пирамидальных сетей (ПС) В. П. Гладуна [9] и семантические представления И. П. Кузнецова [10].

Элементы ПрО описываются в ЕЯ-текстах элементарными и составными словосочетаниями. Первые обычно состоят из двух слов (анализ данных) и являются реализациями элементарных отношений (r свойство – объект). Вторые (интеллектуальный анализ данных) можно представить в виде комбинации элементарных словосочетаний (интеллектуальный анализ, анализ данных). Они, в свою очередь, являются реализациями составных отношений. Понятия ПрО также выражаются словосочетаниями (одним словом – наименованием понятия, элементарным словосочетанием или комбинацией элементарных словосочетаний).

Пусть D – словарь ПрО, а P – набор отношений, реализации которых мы собираемся искать в текстах. $P \quad R_1 \quad R_2$, где R_1 – множество элементарных отношений с числом аргументов, равным 2 (их реализациями являются словосочетания из двух значимых слов); R_2 – множество элементарных отношений с числом аргументов меньше 2 (их реализации состоят более чем из двух слов). В R_2 входят, например, такие отношения, как родовидовое, часть-целое, синонимия и т. д.

По способу образования фрагменты Q-сети делятся на четыре типа (рис. 1):

1) $\_, r, \_, a, b \quad a \quad _r \; b$ – словосочетание из двух значимых слов $a, b \quad D$, связанных отношением r (например, $a \quad _r \; b$ (анализ данных));

2) $\_, r, s, A, b \quad Aa \quad _r \; b$ – расширение фрагмента A за счет присоединения знаменательного слова b через связь $s \quad a \quad _r \; b$, где $a \quad D$ (например,

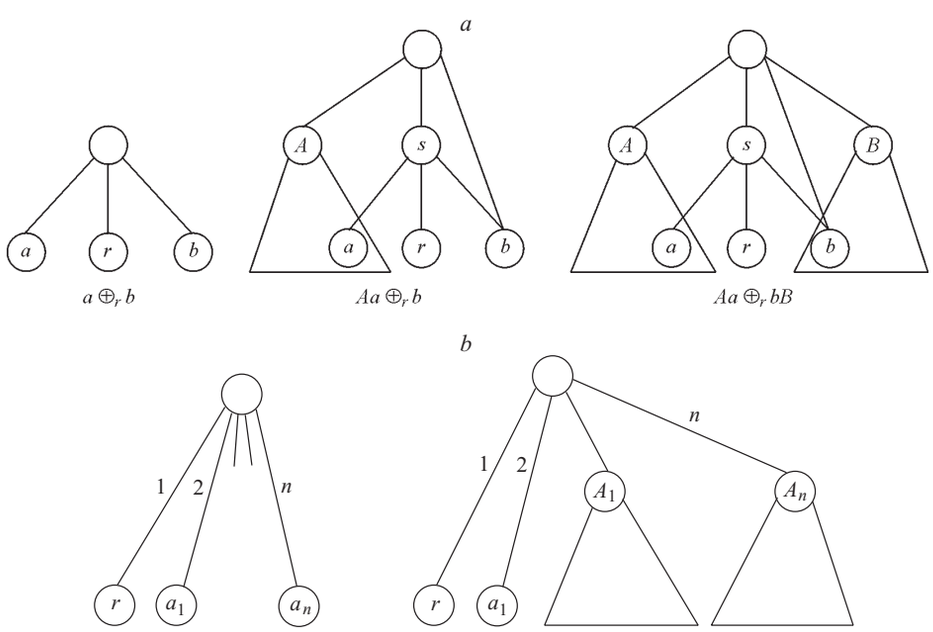


Рис. 1. Способы распознавания Q-сети: фрагменты 1–3-го типов (*a*), фрагменты 4-го типа

$Aa_r b$ (интеллектуальный (анализ данных)), здесь A (анализ данных), а s (интеллектуальный анализ));

3) $\_, r, s, A, B$ $Aa_r bB$ – объединение двух фрагментов А и В через связь $a_r b$, где a D, b D (например, $Aa_r bB$ ((процесс таксономии) (начинается) с (нормировки признаков)), здесь A ((процесс таксономии) начинается), B (нормировка признаков), s (начинается с нормировки)).

4) $d, r, \_, a_1, ..., a_n$ – фрагмент, соответствующий отношению r R_2; $a_1, ..., a_n$ – аргументы этого отношения; d – имя фрагмента. Например, если r – родовидовое отношение, a_1 (задача интеллектуального анализа данных), a_2 (задача таксономии), a_3 (задача распознавания образов), то фрагмент $\_, r, \_, a_1, a_2, a_3$ будет означать, что задачи таксономии и распознавания образов являются задачами интеллектуального анализа данных.

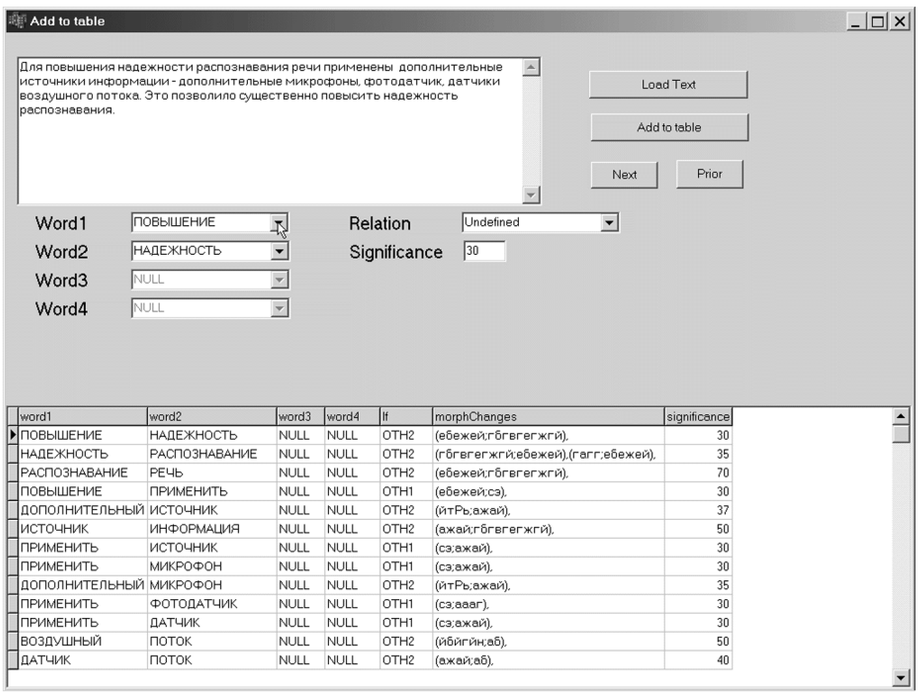
Вершины низшего уровня иерархии сети, не имеющие заходящих дуг, называются рецепторами, которые соответствуют словам или наименованиям отношений. Остальные вершины называются концепторами и соответствуют элементарным или составным словосочетаниям. Рецепторы сети можно рассматривать как фрагменты «нулевого» типа. Таким образом, каждый фрагмент Q-сети соответствует реализации некоторого отношения (т. е. некоторому словосочетанию). Если реализация отношения (словосочетание A) включает в свой состав реализацию другого отношения (словосочетание B), то фрагмент сети А включает в себя фрагмент сети B.

Среди достоинств ПС следует упомянуть развитые ассоциативные свойства, иерархичность, а также, что особенно важно, в них реализованы процессы формирования связей между семантическими объектами, выделения классов объектов и ситуаций и процессы формирования обобщенных определений этих классов. В семантических представлениях И. П. Кузнецова, в свою очередь, все части текста, соответствующие существенным единицам ПрО, вне зависимости от частоты их появления в текстах отражены в сети своими фрагментами. Принадлежность Q-сетей к обоим упомянутым классам дает возможность реализовать на них удобные алгоритмы анализа текста, предложенные в [9, 10].

БЛЗ САТ представляет собой набор элементарных и составных словосочетаний предметной области. БЛЗ удобно использовать в том же виде, который имеют семантические представления текстов, т. е. в виде Q-сети. Условно БЛЗ САТ можно разделить на базу реализаций элементарных отношений (БРО) и набор критичных фрагментов (НКФ), по которым можно определить, какие элементы онтологии затрагиваются в данном тексте.

Формирование БРО. Начальный объем знаний в виде реализаций элементарных отношений ПрО вносится в БРО экспертом непосредственно либо в ходе интерактивного анализа текстов предметной области (рис. 2).

В первом случае эксперт имеет возможность, переходя от одного предложения текста к другому, выбирать наименование элементарного отношения (Relation) и его аргументы, число которых может варьироваться (Word1,…) в выпадающих списках. Каждое словосочетание (реализация отношения) характеризуется значениями набора признаков. В текущей версии в него кроме лексем (аргументов и наименования отношения) входит набор сочетаемости аргументов по морфологическим признакам (заполняется системой с помощью компоненты морфологического анализа [11, 12]) и экспертная оценка значимости словосочетания в рамках предметной области. Каждая строчка БРО содержит набор элементарных словосочетаний, соответствующих реализации $r(\text{Word1}, \text{Word2}, ...)$ некоторого отношения r. Во втором



word1	word2	word3	word4	lf	morphChanges	significance
ПОВЫШЕНИЕ	НАДЕЖНОСТЬ	NULL	NULL	ОТН2	(ебежей;гбгвгегжгй),	30
НАДЕЖНОСТЬ	РАСПОЗНАВАНИЕ	NULL	NULL	ОТН2	(гбгвгегжгй;ебежей),(гагг;ебежей),	35
РАСПОЗНАВАНИЕ	РЕЧЬ	NULL	NULL	ОТН2	(ебежей;гбгвгегжгй),	70
ПОВЫШЕНИЕ	ПРИМЕНИТЬ	NULL	NULL	ОТН1	(ебежей;сэ),	30
ДОПОЛНИТЕЛЬНЫЙ	ИСТОЧНИК	NULL	NULL	ОТН2	(йтРь;ажай),	37
ИСТОЧНИК	ИНФОРМАЦИЯ	NULL	NULL	ОТН2	(ажай;гбгвгегжгй),	50
ПРИМЕНИТЬ	ИСТОЧНИК	NULL	NULL	ОТН1	(сэ;ажай),	30
ПРИМЕНИТЬ	МИКРОФОН	NULL	NULL	ОТН1	(сэ;ажай),	30
ДОПОЛНИТЕЛЬНЫЙ	МИКРОФОН	NULL	NULL	ОТН2	(йтРь;ажай),	35
ПРИМЕНИТЬ	ФОТОДАТЧИК	NULL	NULL	ОТН1	(сэ;ааар),	30
ПРИМЕНИТЬ	ДАТЧИК	NULL	NULL	ОТН1	(сэ;ажай),	30
ВОЗДУШНЫЙ	ПОТОК	NULL	NULL	ОТН2	(йбйгйн;аб),	50
ДАТЧИК	ПОТОК	NULL	NULL	ОТН2	(ажай;аб),	40

Рис. 2. Формирование БРО в ходе интерактивного анализа текста

случае в ходе анализа текста в первую очередь следует выделять наиболее значимые для ПрО фрагменты, далее расширяя их менее важными подробностями. Для обеспечения устойчивости экспертных оценок, редактирования, поиска по БРО можно применить интерфейс «Визуализатор отношений» [13,14]. Экспертная оценка значимости реализации отношения может корректироваться по результатам статистических данных о частоте встречаемости соответствующего словосочетания как в целом в текстах данной ПрО, так и, возможно, локально, для конкретного текста при его анализе.

Алгоритм построения Q-сети. Рассмотрим схему алгоритма построения Q-сети с использованием в качестве лингвистической базы БРО, содержащую реализации отношений с двумя аргументами, т. е. словосочетания из двух слов.

Представим предложение в виде линейной последовательности словоформ $a_1, a_2, ..., a_n$. Проверяя все пары (a_1, a_2), (a_1, a_3), ..., (a_2, a_3), ...,(a_{n-1}, a_n) на присутствие в БРО, сформируем набор потенциальных реализаций отношений предложения T $\{A_1 \quad r_1(a_{11}, a_{12}), ..., A_m \quad r_m(a_{m1}, a_{m2})\}$, где a_{ij} – словоформы предложения.

Словоформы предложения a_i, a_j связаны непосредственно, если в набор потенциальных реализаций отношений предложения входит некоторая реализация $A_k \quad r_k(a_i, a_j)$, и называются опосредованно связанными, если в предложении можно выделить цепочку словоформ $a_i \quad b_1, b_2, ..., b_k \quad a_j$, любые два соседних элемента которой связаны непосредственно (положение в цепочке не связано с порядком слов в предложении). Разобьем предложение на минимальное число компонент связности $\{K_j\}$, каждая из которых состоит из опосредованно связанных друг с другом словоформ.

Введем $f(s, l)$ – функцию веса реализации отношения $r_i(a_{i1}, a_{i2})$ в предложении, где s – оценка значимости данной реализации в рамках ПрО; l – расстояние между словоформами a_{i1} и a_{i2} в предложении (для соседних словоформ l 1); 0 f 1 (f монотонно возрастает по первому и монотонно убывает по второму аргументам).

Для каждой компоненты связности K_i с помощью алгоритма Прима строим связывающую сеть максимального веса такую, что:

1) любые две словоформы компоненты могут быть соединены с помощью реализаций этого подмножества простой (без циклов) цепочкой a_i $b_1, b_2, \ldots, b_k$ a_j единственным образом;

2) суммарный вес элементов выбранного подмножества является максимальным среди всех возможных подмножеств, удовлетворяющих первому условию.

Введем параметры 0 $_2$ $_1$ 1. Из элементов множества T_i веса, не меньшего чем $_1$, последовательно, начиная с самого тяжелого из оставшихся, выбираются реализации отношений A_j $r(a_{j1}, a_{j2})$, аргументы которых не пересекаются с аргументами уже выбранных элементов. Это наиболее существенные элементы компоненты K_i, с которых и начнется построение соответствующей этой компоненте сети. Для каждого такого элемента строим фрагмент 1-го типа a_{j1} $_r$ a_{j2}. Далее из оставшихся элементов множества T_i выбираются реализации самых весомых отношений A_j $r(a_{j1}, a_{j2})$. Если A_j пересекается с ранее выбранными элементами по одному из аргументов, например a_{j1}, находим в сети компоненты максимальный по включению фрагмент F, содержащий рецептор, соответствующий a_{j1}, и строим фрагмент 2-го типа Fa_{j1} $_r$ a_{j2}. Если A_j пересекается с ранее выбранными элементами по обоим аргументам, находим в сети компоненты максимальные по включению фрагменты F_1, F_2, содержащие рецепторы, соответствующие a_{j1} и a_{j2}, и строим фрагмент 3-го типа $F_1 a_{j1}$ $_r$ $a_{j2} F_2$.

Рассматривая все компоненты связности, получаем семантическое представление предложения в виде набора фрагментов. Фрагменты, соответствующие разным компонентам связности предложения, не пересекаются. Наличие несвязных компонент говорит либо об их малой значимости в рамках ПрО, либо о неполноте лингвистической базы предметных знаний.

При построении семантического представления текста связь между предложениями фиксируется при наличии общих фрагментов в их сетевых описаниях.

Представленный алгоритм реализован авторами. Работа алгоритма и соответствующая Q-сеть показаны на рис. 3.

Формирование набора критичных фрагментов. Для определения НКФ составляется обучающая выборка текстов, для каждого из которых экспертом указывается, какие элементы онтологии в нем затронуты. По этим текстам строится общая Q-сеть. Далее на основе алгоритма формирования понятий [9] происходит разделение семантических представлений текстов, затрагивающих разные элементы онтологии ПрО. В ходе этого разделения и получается упомянутый набор критичных фрагментов. Таким образом, каждый класс текстов описывается набором фрагментов, которые типичны или, наоборот, нетипичны для семантических представлений текстов данного класса. При этом соблюдается принцип наибольшей краткости описания (используются фрагменты, занимающие минимально возможное по высоте место в иерархии сети). Логично предположить, что сочетания слов, характеризующие класс текстов, затрагивающих определенные элементы онтоло-

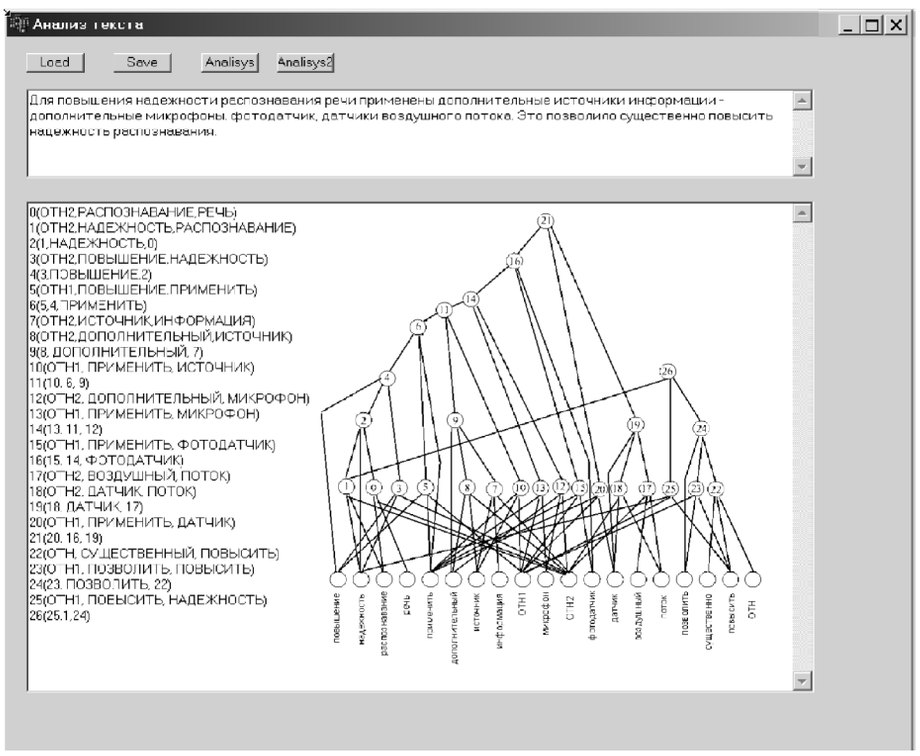


Рис. 3. Анализ фрагмента текста и соответствующая этому фрагменту Q-сеть

гии, и будут отображением этих элементов на текст. Когда онтология еще не построена, лингвистическая база знаний используется САТ как приближение ПроекОнт. Когда онтология построена, такая база становится ее частью.

Система анализа текстов и набор критичных фрагментов могут использоваться и для поддержки существующей онтологии. Соотнесение потока семантических портретов новых текстов с базой значимых фрагментов осуществляет наполнение элементов онтологии ссылками на текстовые документы. По степени наполнения эксперт может принимать решение о разделении «перегруженных» элементов сети и объединении «недогруженных».

При анализе текста часты ситуации, когда вершины фрагментов сети именуются формально разными словами, связанными, однако, отношениями синонимии, гиперонимии/гипонимии (родовидовые), меронимии (часть-целое). Поэтому в системе должен быть предусмотрен выход на тезаурусы WordNet, RussNet (для общеязыковой лексики), а также на специальные тезаурусы для предметной лексики.

Существенную сложность при реализации описанной системы анализа текстов представляет большой объем «ручного» труда при формировании БРО. Этот труд может быть частично автоматизирован. Эксперт может выбирать реализации отношений из предварительно полученного статистическими методами списка устойчивых словосочетаний ПрО [5].

Возможна также автоматизация наполнения БРО за счет формирования определений отношений. С помощью алгоритма формирования понятий [9] или других методов анализа данных, используя БРО как обучающую выборку, можно получить обобщенные правила выделения отношений, не завися-

щие от конкретных лексем. Эти правила формулируются в терминах логических выражений от параметров, входящих в описания отношений, накопленных в БРО. Фактически это соответствует распознаванию образов, где образ есть наименование отношения, а признаковое пространство суть пространство параметров, задействованных при описании реализаций отношений в БРО. Полученные определения должны проверяться экспертом.

Автоматизация процессов создания и развития онтологии в Grid- сети. Структура системы автоматизированного построения онтологий OntoGrid должна отражать специфику трех типов ее клиентов: Эксперта, Пользователя и Администратора [15]. С точки зрения представления создаваемая онтология – это комплект документов определенной структуры. Процесс ее построения состоит из итераций по дополнению и изменению этого комплекта документов. Для изменения онтологии эксперт вносит предложение на проведение определенных модификаций ее содержания. Администратор назначает внесенному предложению рецензентов. После возможного корректирования по решению администратора изменения вносятся в рабочую версию онтологии. На этом итерация процесса эволюции онтологии завершается.

По результатам проведения ряда итераций администратор принимает решение о завершении очередного этапа процесса построения онтологии и публикации ее очередной стабильной версии.

Система, поддерживающая автоматизированное создание и обработку документов онтологии в распределенном режиме, характеризуется набором специфических требований, определяющих ее технологическую организацию. Эти требования обусловливаются тем, что в создании онтологии участвуют многочисленные географически разделенные коллективы экспертов. Топология задействованных узлов сети разработчиков меняется по мере подключения новых коллективов или прекращения работы старых. Участников коллектива экспертов не следует ограничивать жесткими требованиями к архитектуре их рабочих узлов, аппаратных платформ и операционных систем. Адекватную основу для построения систем, удовлетворяющих таким требованиям, предоставляют вычислительные технологии, известные под общим названием Grid [16]. Наиболее развитый инструментарий разработки и развертывания Grid-систем на сегодняшний день предлагается консорциумом “The Globus Alliance”. К числу последних разработок этого консорциума относится архитектура OGSA (Open Grid Services Architecture), основанная на концепции web-сервисов.

При разработке представления структуры онтологии были рассмотрены различные существующие на сегодняшний день подходы и стандарты. Как наиболее обоснованный и перспективный был принят стандарт OWL (Ontology Web Language) [17], разработанный и рекомендованный консорциумом W3C. Язык OWL предназначен для такого представления информации, которое отвечает двум, в некотором смысле противоположным, требованиям. С одной стороны, эта информация содержит знания, а не только представление. С другой стороны, она предназначена для автоматической обработки компьютерными программами в противоположность использованию знаний непосредственно человеком.

OWL обладает большей выразительной силой, чем структурные языки XML, RDF и RDF-S, и может быть представлен в их форме. OWL-документ позволяет, используя лежащую в основе OWL дескриптивную логику, выводить такие факты о сущностях предметной области, которые не содержатся непосредственно в этом документе. Таким образом, OWL-онтология являет-

ся теорией – совокупностью предложений формального языка, замкнутой относительно выводимости. В нашем проекте используется представление онтологии в нотации OWL-RDF.

Для упрощения разработки новых онтологий удобно создавать шаблоны онтологий различных групп предметных областей. При формировании шаблона можно учесть специфику понимания проблематики онтологий, присущую конкретному коллективу экспертов. Данный проект ориентирован на построение шаблона онтологий научно-технических предметных областей, связанных с процессами анализа, синтеза и преобразования информации о произвольных фрагментах реального мира. К числу таких процессов относятся измерение и накопление данных, обнаружение закономерностей (знаний), хранение, обработка и передача данных и знаний, использование знаний для прогнозирования и синтеза. На рис. 4 приведен перечень базовых категорий онтологий проблемных областей такого рода.

В дополнение к основному содержанию онтологии возможно формирование и хранение метаинформации об ее элементах. Содержание метаинформации определяется в соответствии со стандартами Dublin (Guidelines for Implementing Dublin Core in XML) [18]. Согласно этим стандартам метаинформация обеспечивает элементы онтологии такими данными, как реквизиты автора и соавторов, время создания и публикации, источники информации и т. д.

Функциональность системы складывается из двух частей: система должна, с одной стороны, предоставлять удобный полнофункциональный инструмент разработки фрагмента онтологии для эксперта и, с другой стороны, обеспечивать совместную распределенную работу коллектива экспертов над фрагментами общей онтологии. В качестве индивидуального средства работы эксперта с фрагментом онтологии планируется использовать редактор, разработанный группой “Protege Project” [19]. Это средство обеспечивает удобный визуальный контроль за процессом разработки фрагментов онтологии. В дальнейшем предполагается дополнить его средствами доступа к САТ, описанной в предыдущих разделах.

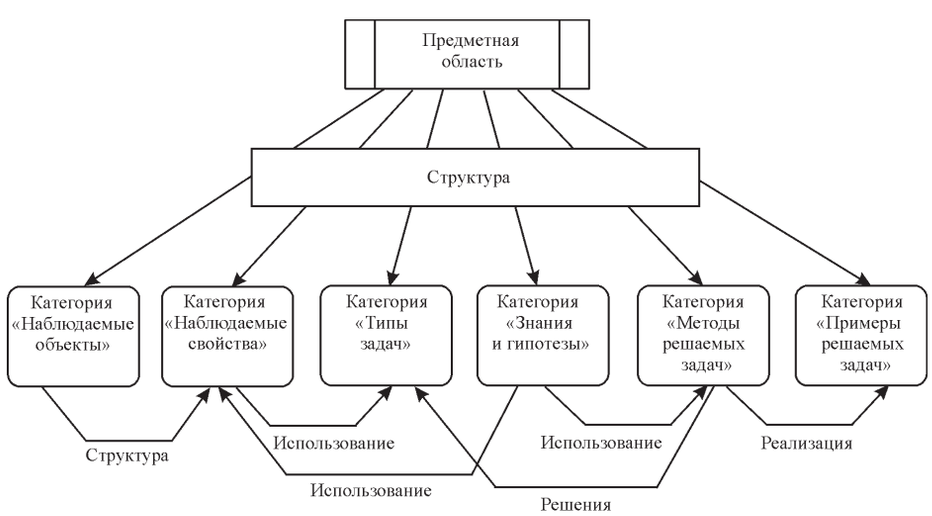


Рис. 4. Базовые категории онтологии

Текущая рабочая версия онтологии хранится в центральном репозитории под управлением сервиса, который обеспечивает выделение фрагментов онтологии экспертам для разработки. Сервис репозитория принимает от эксперта запрос на работу с фрагментом онтологии и составляет OWL-документ, содержащий все утверждения текущей версии онтологии, относящиеся к запрашиваемым классам и объектам. Этот документ с фрагментом онтологии отдается эксперту для разработки. После проведения доработок эксперт возвращает модифицированный фрагмент онтологии репозиторию для внесения изменений в рабочую версию онтологии. В работе системы используется модель оптимистической конкуренции, при которой фрагмент онтологии, выделенный для работы одному эксперту, не блокируется для других экспертов.

По решению администратора сервис репозитория публикует очередную стабильную версию онтологии. Этот документ предоставляется пользователям в качестве текущей версии онтологии.

Предложенная структура представления информации и архитектура ключевых сервисов были успешно апробированы в ходе создания прототипа. В настоящее время ведутся работы по дальнейшей реализации системы OntoGrid.

Заключение. Параллельно с описанными выше исследованиями по созданию инструментальной системы OntoGrid ведется подготовительная работа по организации виртуального коллектива экспертов из различных исследовательских центров, занимающихся проблемой «Интеллектуальный анализ данных» (Data Mining) для совместной разработки онтологии этой предметной области. Редакционной группой из специалистов Новосибирска, Москвы и Мюнхена разрабатывается шаблон онтологий Onto-DM.

В качестве примера, показывающего общую структуру конкретной онтологии, нами совместно с экспертами из Санкт-Петербурга и Минска разработан черновой вариант онтологии предметной области «Распознавание и синтез речевых сигналов», который был обсужден на Международной конференции «Речь и компьютер» (SpeeCom‘2004, Санкт-Петербург, сентябрь 2004) [20].

Авторы выражают благодарность И. А. Борисовой, В. В. Дюбанову, О. А. Кутненко, А. П. Соколовой и В. А. Чуриковой за активное и полезное обсуждение работы.

СПИСОК ЛИТЕРАТУРЫ

1. http://xml.com/2002/11/06/Ontology_Editor_Survey.html
2. Worth D. S., Kozak A. S., Johnson D. B. Russian Derivational Dictionary. N. Y.: American Elsevier Publishing Company, Inc, 1970.
3. Гусев В. Д. Механизмы обнаружения структурных закономерностей в символьных последовательностях // Вычисл. системы. Новосибирск: ИМ СО АН СССР. 1983. Вып. 100. Проблемы обработки информации. С. 47.
4. Белоногов Г. Г., Быстров И. И., Новоселов А. П. и др. Автоматический концептуаль- ный анализ текстов // НТИ. Сер. 2. 2002. № 10. С. 26.
5. Гусев В. Д., Саломатина Н. В. Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Тр. Междунар. конф. «Диалог-2004. Компьютерная лингвистика и интеллектуальные технологии» М.: Наука, 2004. С. 530.

6. Пащенко Н. А., Кнорина Л. В., Молчанова Т. В. и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика. 1983. 7. С. 7.
7. Гусев В. Д., Немытикова Л. А., Саломатина Н. В. Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Вычисл. системы. Новосибирск: ИМ СО РАН, 2002. Вып. 171. Интеллектуальный анализ данных. С. 51.
8. Загоруйко Н. Г., Налетов А. М., Гребенкин И. М. На пути к автоматическому построению онтологии // Тр. Междунар. конф. «Диалог-2003. Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2003. С. 717.
9. Гладун В. П. Планирование решений. Киев: Наук. думка, 1987. С. 17.
10. Кузнецов И. П. Семантические представления. М.: Наука, 1986.
11. Саломатина Н. В. Количественные характеристики вариативности морфемных моделей (на материале словаря канонических форм русского языка) // Вычисл. системы. Новосибирск: ИМ СО РАН, 2001. Вып. 167. Методы обнаружения эмпирических закономерностей. С. 93.
12. Сокирко А. В. Морфологические модули на сайте www.aot.ru // Тр. Междунар. конф. «Диалог-2004. Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2004. С. 559.
13. Загоруйко Н. Г., Налетов А. М., Соколова А. А., Чурикова В. А. Формирование базы лексических функций и других отношений для онтологии предметной области // Тр. Междунар. конф. «Диалог-2004. Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2004. С. 202.
14. Загоруйко Н. Г. Метрологические свойства эксперта // Вычисл. системы. Новосибирск: ИМ СО РАН, 1999. Вып. 166. Обнаружение эмпирических закономерностей. С. 119.
15. Завертайлов А. В., Ковалев С. П. Система поддержки деятельности распределенных экспертных групп по разработке онтологий предметных областей // Тр. Междунар. конф. по вычислительной математике «МКВМ-2004». Новосибирск: ИВМиМГ СО РАН, 2004. С. 56.
16. Grid Computing: Making the Global Infrastructure a Reality. N. Y.: Willey & Sons, 2003.
17. http://www.w3.org/TR/owl-guide/ (Smith M. K., Welty C., McGuinness D. L. OWL Guide. W3 Consortium, 2004).
18. http://dublincore.org/documents/dc-xml-guidelines/ (Powell A., Johnston P. Guidelines for implementing Dublin Core in XML. DCMI, 2003).
19. http://protege.stanford.edu (Protege Project).
20. Galunov V. I., Lobanov B. M., Zagoruiko N. G. Ontology of the subject domain “Speech signals recognition and synthesis” // Proc. 9th Intern. Conf. “Speech and Computer” (SPEECOM’2004). Saint-Petersburg, 2004. P. 448.

Институт математики им. С. Л. Соболева СО РАН,
E-mail: zag@math.nsk.ru

Поступила в редакцию
12 января 2005 г.