НА ПУТИ К АВТОМАТИЧЕСКОМУ ПОСТРОЕНИЮ ОНТОЛОГИИ

Загоруйко Н.Г., Налетов А.М., Гребенкин И.М.
Институт Математики СО РАН, НГУ
zag@math.nsc.ru

Аннотация: Обсуждается семантическая сеть (Q-сеть), которая объединяет в себе основные достоинства пирамидальных сетей В.П. Гладуна и семантических сетей Кузнецова И.П. Q-сеть имеет пирамидальную форму и состоит из фрагментов четырех типов. При анализе очередного предложения в нем в первую очередь выделяются фрагменты, которые наиболее часто встречались в ранее рассмотренном корпусе текстов. Анализ заканчивается тогда, когда все составные единицы текста найдут свое отражение в сети. Предусмотрены процедуры упрощения структуры сети и удаления из нее не существенных вершин. Q-сети обладают свойствами однородности и иерархичности. Они позволяют формировать связи между семантическими объектами и создавать обобщенные определения классов объектов. Q-сети содержат специальные вершины для существенных элементов предметной области. Такие сети позволят автоматизировать процесс построения онтологии этой области.

Введение

Для представления текстов Е.Я. хотелось бы использовать семантические сети, обладающие следующими свойствами:

1) Однородность (внутренний язык должен состоять из как можно более однородных частей).

2) Сеть должна обладать развитыми ассоциативными свойствами и должна отражать иерархичность реальных сред.

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

4) Все части текста, отражающие существенные единицы предметной области или цельные комплексные объекты (ФИО, адрес, название организаций и т.п.), вне зависимости от частоты их появления, положения в тексте и других условий, должны быть отражены в сети соответствующими вершинами.

5) Сеть должна иметь по возможности наиболее простой («прозрачный») вид. В ней должны быть предусмотрены механизмы упрощения, удаления случайно попавших вершин, чья существенная роль в дальнейшем не подтвердилась или утратилась со временем.

Применение семантического представления текстов ЕЯ с помощью таких сетей могло бы не ограничиваться только лишь использованием текстов, как данных для работы каких-либо алгоритмов. Они могут быть использованы и для выявления понятий и глубинных связей между классами объектов в данной предметной области (ПО), возможно новых и для экспертов. В конечном итоге такие сети смогут позволить автоматизировать процесс построения онтологии ПО.

Базовые семантические сети

Предлагаемый в данной работе подход является в значительной степени продолжением работ В.П. Гладуна [1] и И.П. Кузнецова [2]. В.П. Гладуном был разработан аппарат построения растущих пирамидальных сетей (ПС - ациклический ориентированный граф, в котором нет вершин с единственной заходящей дугой). В пирамидальной сети информация хранится путем ее отображения в структуре сети. Пирамидой B называется вершина b и все те

вершины, из которых существуют пути в эту вершину b. При построении сети в ней образуются вершины, чьи пирамиды соответствуют отдельным объектам, и вершины, чьи пирамиды соответствуют общим частям нескольких объектов. ПС удобны для выполнения различных операций ассоциативного поиска. Разработанные алгоритмы построения сетей обеспечивают автоматическое установление ассоциативной близости между объектами по общим сочетаниям значений признаков.

При работе с естественно-языковыми текстами Гладун исходит из того, что семантическое представление текста и модель мира являются композициями сведений об объектах, их свойствах, связях между объектами и действиями над объектами. При этом элементарные факты выражаются осмысленными словосочетаниями минимальной длины. В основном это словосочетания, содержащие два знаменательных слова или отдельные слова. При формировании представлений, отображающих смысл текстов, факты считаются реализациями некоторых семантических отношений. В слове или словосочетании, выражающем элементарный факт, аргументы семантического отношения (объекты, действия, состояния, характеристики) кодируются основами словоформ. Остальная часть словосочетания рассматривается как определитель семантического отношения. Каждый аргумент семантического отношения относится к одному из классов, определяемых такими семантическими категориями, как «объект», «свойство», «действие», и т.п. При анализе фразы семантические отношения, кодируемые словосочетаниями, распознаются по выделенным определителям и семантическим категориям знаменательных слов.

Текст естественного языка представляется в виде двухъярусной ПС, являющейся композицией элементарных фактов. Наибольшие пирамиды первого яруса представляют в сети реализации отношений, кодируемые словосочетаниями. В «основании» каждой такой пирамиды находятся рецепторы (вершины, не имеющие заходящих дуг), представляющие знаменательные слова, словосочетания и имя отношения, кодируемого словосочетанием. Пирамиды второго яруса представляют в сети ситуации, описываемые предложениями, абзацами, текстами. При формировании сети возникают вершины, соответствующие пересечениям словосочетаний, предложений, текстов. Одним из основных достоинств ПС является то, что в них реализованы процессы формирования обобщенных определений классов объектов и ситуаций (понятий).

Другой подход к семантическому представлению знаний, на который мы опирались, был предложен И.П. Кузнецовым. Вершины сетей, используемых в этом подходе, могут соответствовать объектам, понятиям, отношениям, логическим составляющим информации, комплексным объектам и др. Кроме того, вводятся вершины другого типа – вершины связи. Они соединяются помеченными ребрами с вершинами, упомянутыми выше (фактически ребра метятся цифрами, определяющими семантический падеж отношения). В результате образуется фрагмент, соответствующий элементарной ситуации, т.е. объектам, связанным отношением. Такой фрагмент называется элементарным (ЭФ). Специального деления вершин на непересекающиеся множества не производится. Каждая из них может играть любую роль. Например, ситуации могут быть связаны своими отношениями, отношения также могут быть объектами другого отношения и т.д. В результате обеспечиваются широкие возможности представления. При этом всему, что может рассматриваться, как самостоятельная единица, существенная для данной ПО, должна быть сопоставлена собственная вершина.

Элементарный фрагмент представляется в виде кортежа $\Phi=<d_1, d_2, d_3, \ldots, d_k>$, где $d_1,\ldots, d_k \in D$ – множество компонент, различаемых в данной предметной области. Под компонентами понимаются объекты, отношения и составленные из них комплексные объекты, рассматриваемые как единое целое. Если указанные объекты с их отношением рассматриваются как единый комплексный объект или ситуация, то последнему сопоставляется собственная вершина d_1, которая занимает в кортеже первое место. Второе место занимает вершина d_2, соответствующая логической составляющей (например, указывающей на истинность или ложность представленного отношения). Вершина d_3 отвечает отношению, а вершины $d_4,\ldots, d_k$ его аргументам. Отношения могут иметь различную местность, что представляется с помощью кортежей элементарных фрагментов различной длины. Для представления ситуаций, состоящих из множеств объектов и отношений, используются множества ЭФ, образующие сеть, части которой называются фрагментами. Семантическая сеть записывается в виде $\Phi=\Phi_1\circ\ldots\circ\Phi_n$.

Q-сети

Каждый из двух рассмотренных выше способов семантического представления удовлетворяет лишь части предъявленных выше требований. Объединение же этих способов покрывает все множество требований. При сравнении этих двух подходов в них обнаруживается много общего. Так, например, ассоциативные вершины ПС во многом подобны вершинам связи, а их пирамиды – фрагментам соответствующей сети. Обратно, учитывая отсутствие специального отделения множества вершин связи от множества вершин понятий и отношений, имеется возможность с помощью введения дополнительных фрагментов придать сети с вершинами связи пирамидальную структуру.

Предлагаемый ниже способ семантического представления текстов в виде Q-сетей использует основные идеи подходов, предложенных Гладуном В.П. и Кузнецовым И.П. Пусть A – словарь знаменательных слов, $R=R_1\cup R_2$ –

множество семантических отношений. R_1 совпадает с множеством семантических отношений, кодируемых словосочетаниями минимальной длины (которое использует В.П. Гладун). R_2 – множество, состоящее из отношений (используемых Кузнецовым И.П.), кодируемых описаниями специально выделенных существенных единиц ПО и цельных комплексных объектов (ФИО, адрес, название организаций и т.п.), а также таких отношений как синонимия, родовидовые отношения и т.д. Текст рассматривается как иерархическая структура фрагментов, каждый из которых представляет некоторую семантическую цельность. При выявлении в тексте отношений из R_1 и их аргументов предполагается использовать правила, основанные на морфологическом анализе знаменательных слов текста в комплексе с сопоставлением каждому аргументу семантического отношения некоторой семантической категории знаменательных слов. Отношения из R_2 предлагается выделять с помощью так называемых опорных конструкций [3]. Например, на основе опорной конструкции: Среди (фрагмент 1) (“имеются”, “рассматриваются”,…) (фрагмент2) (“,”, ”и”) (фрагмент3)…, можно сделать заключение, что фрагмент1 является родовым понятием по отношению к фрагменту2, фрагменту3,…

По способу образования фрагменты делятся на четыре типа:

1) $<\_, r, \_, a, b> \equiv a \oplus_r b$ – словосочетание из двух значимых слов $a,b \in A$, связанных отношением r (например, $a \oplus_r b$= (анализ данных)).

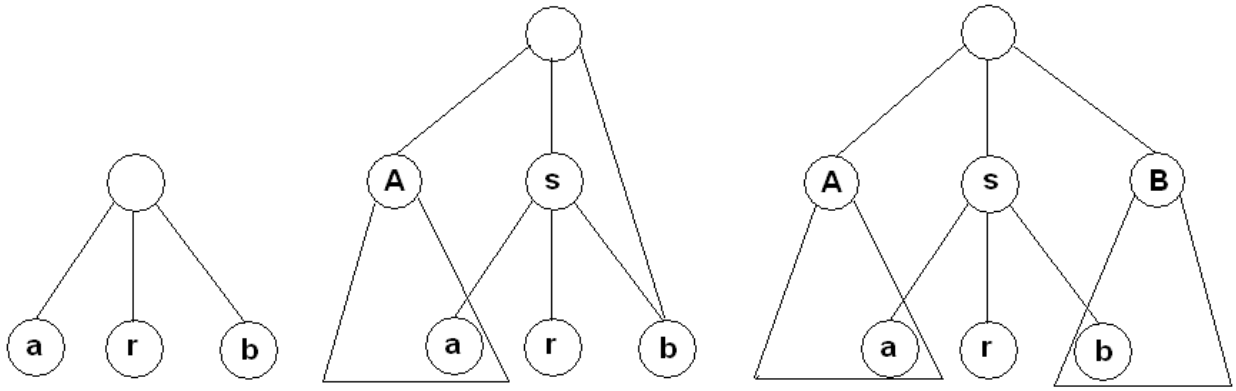
2) $<\_, r, s, A, b> \equiv Aa \oplus_r b$ – расширение фрагмента A за счет присоединения знаменательного слова b через связь $s=a \oplus_r b$, где $a \in A$ (например, $Aa \oplus_r b$=(интеллектуальный (анализ данных)), где A=(анализ данных), s=(интеллектуальный анализ)).

3) $<\_, r, s, A, B> \equiv Aa \oplus_r bB$ – объединение двух фрагментов A и B через связь $a \oplus_r b$, где $a \in A$, $b \in B$ (например, $Aa \oplus_r bB$=((процесс таксономии) начинается) с (нормировки признаков)), где A=((процесс таксономии) начинается), B=(нормировка признаков), s=(начинается с нормировки)).

4) $<d, r, \_, a_1, \ldots, a_n>$ - фрагмент, соответствующий отношению $r \in R_2$, $a_1,\ldots, a_n$ – аргументы этого отношения, d – имя фрагмента. Например, если r-родовидовое отношение, a_1=(«фрукты»), a_2=(«яблоки»), a_3=(«груши»), то фрагмент $<\_, r, \_, a_1,a_2,a_3>$ будет означать, что яблоки и груши являются фруктами.

Часто (но не всегда) фрагменты четвертого типа могут быть полностью (или частично) представлены как фрагмент одного из первых трех типов.

Сопоставим фрагментам каждого типа определенный вид пирамидальной структуры. Фрагмент $a \oplus_r b$ представим в виде пирамиды, состоящей из ассоциативной (не являющейся рецептором) вершины, соединенной заходящими дугами с рецепторами, соответствующими знаменательным словам a, b и отношению r. Фрагмент $Aa \oplus_r b$ представим пирамидой, состоящей из ассоциативной вершины, соединенной заходящими дугами с ассоциативной вершиной, соответствующей фрагменту A, с ассоциативной вершиной, соответствующей фрагменту $s= a \oplus_r b$ и с рецептором, соответствующим слову b. При этом дуга от вершины, соответствующей фрагменту s, помечается как связь между фрагментом A и словом b. Аналогичным образом сопоставляется пирамида фрагменту $Aa \oplus_r bB$, только вместо рецептора, отвечающего слову b, фигурирует ассоциативная вершина, отвечающая фрагменту B.



$a \oplus_r b$ $Aa \oplus_r b$ $Aa \oplus_r bB$,

Рис. 1. Фрагменты 1-го, 2-го и 3-го типов.

Если фрагмент четвертого типа (или его часть) может быть представлена с помощью фрагмента одного из первых трех типов, то ему (или его части) сопоставляется соответствующая пирамида. Если часть из аргументов $a_{i1},\ldots, a_{im}$

не входит в такое представление, то рецепторы, соответствующие им, и рецептор, соответствующий отношению r, непосредственно связываются с ассоциативной вершиной, соответствующей всему фрагменту. Дуги помечаются, указывая на порядок вхождения элементов во фрагмент. В случае, когда $\{a_{i1},\ldots, a_{im}\}=\{a_1,\ldots, a_n\}$ получаем представление, аналогичное элементарному фрагменту сети Кузнецова И.П.

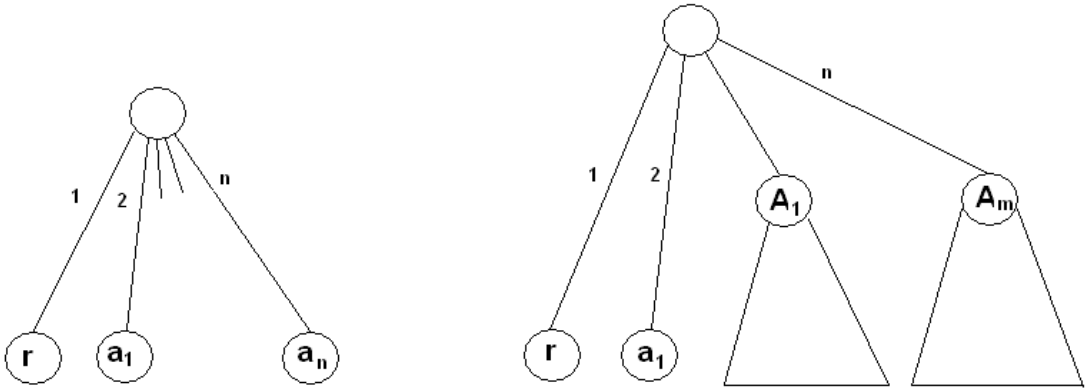


Рис. 2. Фрагменты 4-го типа.

Анализ очередного предложения осуществляется следующим образом:

1) Определяется множество T – множество возможных семантических представлений предложения в виде иерархической структуры фрагментов (и, соответствующей этой структуре сети). Для их построения:

1.1. Выделяются фрагменты 4-го типа (отвечающие отношениям из R_2).

В оставшейся части предложения выполняются следующие действия:

a) Образование фрагмента 1-го типа путем выбора нового опорного словосочетания $a\oplus_r b$.

b) Образование фрагмента 2-го типа $Aa\oplus_r b$, где A-фрагмент из разобранной части предложения, b-знаменательное слово из оставшейся части предложения.

c) Образование фрагмента 3-го типа $Aa\oplus_r bB$, где A,B-фрагменты из разобранной части предложения.

Этот процесс продолжается, пока все предложение не будет представлено одним фрагментом.

Рассматривая различные комбинации действий a,b,c для одного предложения, будем получать разные семантические представления (отличаться будут как наборы фрагментов, так и структура сети). Заметим, что можно сужать множество возможных семантических представлений, добавляя новые отношения в R_2 и, таким образом расширяя множество фрагментов, которые обязательно должны попасть в семантическое представление (например, можно добавить отношение важности, которое может распознаваться по опорным конструкциям вида “особенную важность представляет” (фрагмент 1) и т.д.)

2) Каждый элемент множества T оценивается на основании числа вхождений всех составляющих его фрагментов в уже разобранную часть текста. При этом для каждого фрагмента это число оценивается не по вхождениям его в конечные семантические представления предыдущих предложений, а по тому, существовала ли в этих предложениях альтернатива, в которой такой фрагмент мог бы быть выделен. Для этого, например, можно накапливать статистику вхождений фрагментов 1-го типа (из которых любой фрагмент, в конечном счете, и конструируется) в семантические представления для рассмотренных предложений. Далее, вывод о том, что данный фрагмент A мог быть выделен в некотором предложении L, делается, если в семантическое представление для S входит весь набор фрагментов 1-го типа, необходимый для конструирования фрагмента A.

По мере накопления статистики возможно повторное рассмотрение уже разобранной части текста. При этом в его семантическом представлении фрагменты, чья важность не подтвердилась в дальнейшем, будут заменяться на новые, более перспективные.

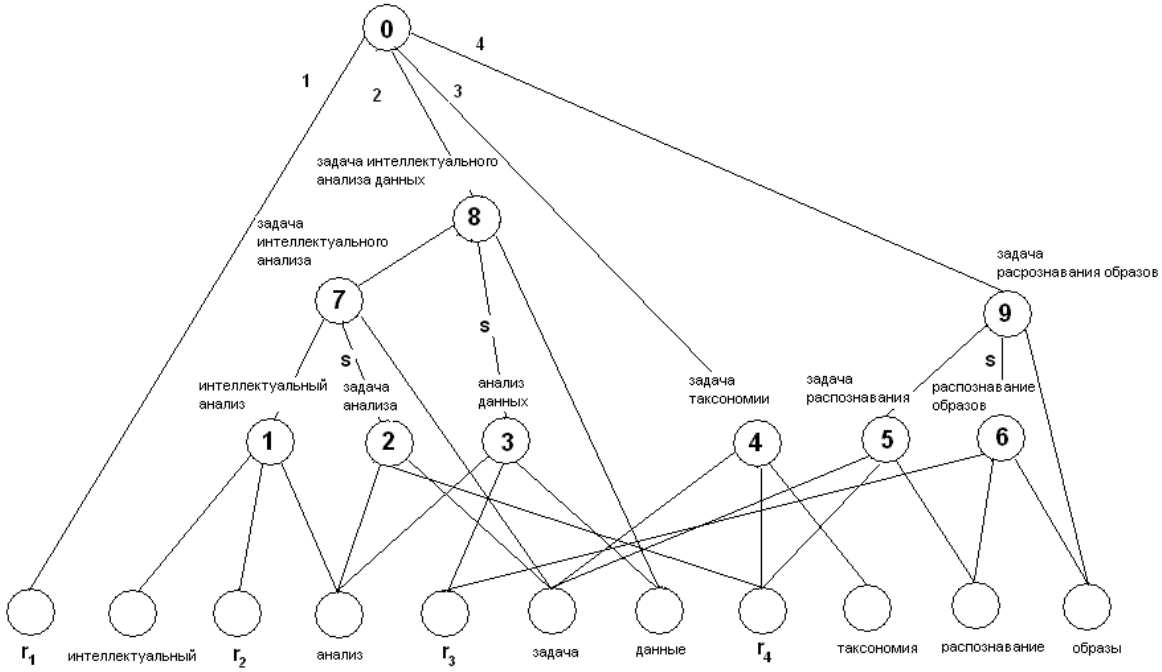
Иллюстрация работы алгоритма

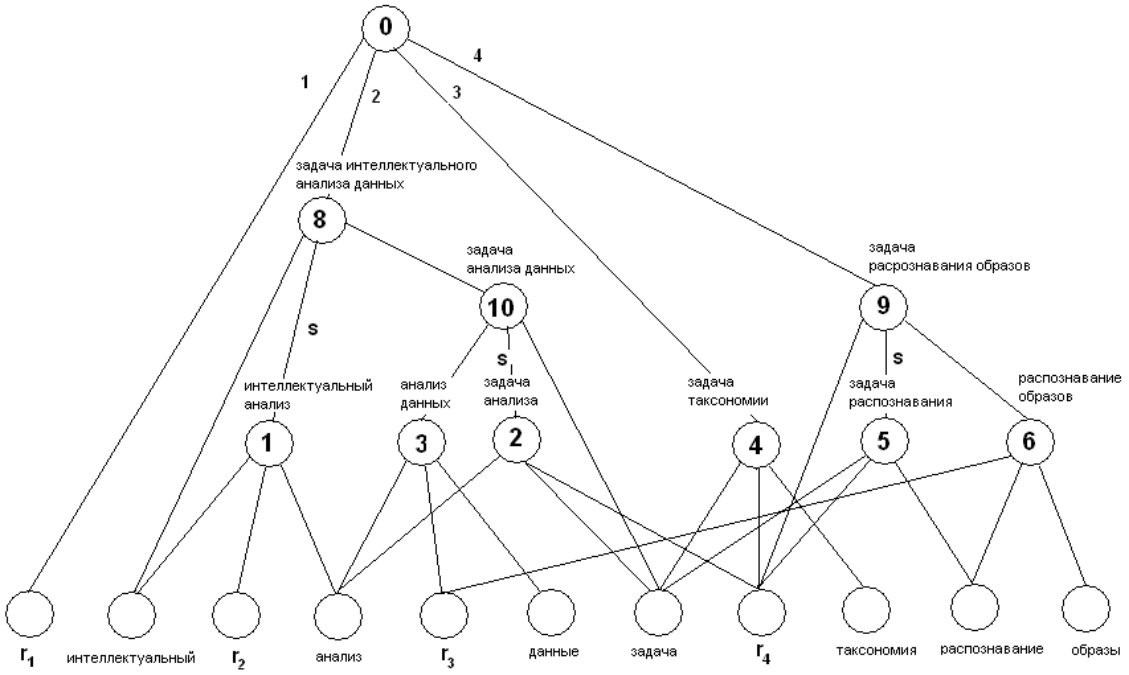
Приведем пример текста и соответствующей ему семантической Q-сети. Представим себе, что программа семантического анализа обработала некоторый корпус текстов из области интеллектуального анализа данных, построила для него соответствующую семантическую сеть и накопила статистику частости встречаемости

фрагментов первого типа и их совместной встречаемости в одном и том же предложении. Затем на анализ поступил новый текст такого содержания: **«Среди задач интеллектуального анализа данных имеются задача таксономии и задача распознавания образов».** Рассмотрим последовательность действий по построению семантической пирамиды для этого предложения.

В соответствие с пунктом 1.1. анализ начинается с поиска ключевых конструкций, которые позволяют установить некоторые виды отношений между понятиями [3]. Обнаруживается ключевая конструкция из слов «среди…имеются», по которой устанавливается родовидовое отношение $r_1 \in R_2$ с родовым понятием в виде фрагмента "задача интеллектуального анализа", и видовых понятий в виде фрагментов "задача таксономии" и "задача распознавания образов". На этом основании строится вершина 0 и четыре нижележащих вершины 1, 2, 3 и 4. (см. рис. 3). В дальнейшем анализ элементов, входящих в состав выделенных фрагментов делается для каждого фрагмента не зависимо от других фрагментов.

Затем в соответствие с пунктом 1.2., переходим к рассмотрению различных комбинаций подпунктов a, b и c. При образовании фрагментов 2-го и 3-го типов следует в первую очередь рассматривать комбинации таких фрагментов первого типа, совместные сочетания которых наиболее часто встречались в ранее рассмотренном корпусе текстов. В результате такой последовательности шагов мы получим несколько вариантов семантического представления данного предложения, изображение двух из которых приводится ниже.





Номера фрагментов указаны внутри соответствующих вершин сети. Дуги, идущие от фрагментов 1-го типа, используемых в качестве связи при образовании фрагментов 2-го и 3-го типов, помечены буквой **s.** Заходящие дуги для вершины 0, соответствующей фрагменту 4-го типа, перенумерованы так, как об этом говорилось выше.

Допустим, что статистика встречаемости различных фрагментов 1-го типа была такой: $n_{1=}10$(«интеллектуальный анализ»), $n_2=10$(«задача анализа»), $n_3=20$(«анализ данных»), $n_4=25$(«задача таксономии»), $n_5=10$(«задача распознавания»), $n_6=15$(«распознавание образов»); и их совместной встречаемости: (1,2)-2, (2,3)-8, (1,2,3)-2, (5,6)-9.

Тогда, можно оценить встречаемость остальных фрагментов величинами $n_7=2$ («задача интеллектуального анализа»), $n_{10}=8$ («задача анализа данных»), $n_8=2$(«задача интеллектуального анализа данных») и $n_9=9$ («задача распознавания образов»).

Для выбора предпочтительного варианта представления воспользуемся мерой качества представления $N=\sum n_i$, где $i \in I$-множество номеров фрагментов, вошедших в это представление. Эта мера будет больше в том случае, если в сети имеются часто встречающиеся вершины, опирающиеся на часто встречающиеся вершины ниже лежащих уровней. Для представленных вариантов получаем $N_1 = 103$ и $N_2=108$, соответственно. Исходя из этого, второе семантическое представление для рассматриваемого предложения считается более предпочтительным.

Заключение

Предлагаемая семантическая Q-сеть имеет пирамидальную структуру и, следовательно, обладает всеми ее достоинствами. Все части текста, отражающие существенные единицы предметной области или цельные комплексные объекты, для выявления которых были введены специальные отношения, всегда будут отражаться в этой сети соответствующими вершинами. Каждая пирамида сети определяет некоторый фрагмент текста одного из четырех типов. Q-сети обладают свойствами однородности и иерархичности, позволяют формировать связи между семантическими объектами. В дальнейшем предполагается, что, представляя с помощью одной Q-сети выборку текстов данной предметной области и используя механизмы формирования обобщенных определений классов объектов и отношений в пирамидальных сетях, можно будет автоматизировать процесс построения онтологии этой предметной области.

Литература

1. Гладун В.П. Планирование решений. Изд. «Наукова думка», Киев, 1987 г.
2. Кузнецов И.П. Семантические представления. Изд. «Наука», М. 1986 г.
3. Харин Н.П. Автоматическое построение тезауруса по текстам документов // Труды Международной конференции по Искусственному Интеллекту КИИ’2002. с. 244-250.