УДК 681.3.06

ФОРМИРОВАНИЕ И КЛАСТЕРИЗАЦИЯ ПОНЯТИЙ НА ОСНОВЕ МНОЖЕСТВА СИТУАЦИОННЫХ КОНТЕКСТОВ

© Михайлов Д.В., Емельянов Г.М., Степанова Н.А.

Новгородский государственный университет им. Ярослава Мудрого,
Институт электронных и информационных систем
ул. Б.С.-Петербургская,41, г.Великий Новгород, Россия, 173003

e-mail: *Dmitry.Mikhaylov@novsu.ru*

Abstract. The approach to the decision of a task of automatic formation of conceptual structure of the thesaurus of the Subject Area on the base of Natural Language's texts is represented. Authors consider (on a material of Russian) the method of revealing of Semantic Classes of concepts on the basis of syntactic contexts of words designating them.

Введение

Центральной задачей анализа смысла высказывания Естественного Языка (ЕЯ) является выделение класса Семантической Эквивалентности (СЭ) [1]. Наиболее известная система классов СЭ в ЕЯ определяется множеством правил синонимических преобразований ЕЯ-высказываний в рамках стандартных Лексических Функций [2]. В общем случае указанная система знаний строится на основе независимых текстовых описаний ситуаций (явлений) действительности выделением различающихся описаний одной и той же ситуации (явления). Далее ставится задача установления степени близости между синонимичными описаниями, но уже различных ситуаций (явлений) и формирования знаний о синонимии в виде прецедентов СЭ [3]. Такая постановка задачи приобретения и систематизации знаний о СЭ делает принципиально возможным ее решение на ЭВМ как задачи таксономии знаний. Тем не менее, при независимом ЕЯ-описании одной и той же ситуации *возникает проблема* оценки адеквантности как формируемых знаний о синонимии, так и самих текстовых описаний ситуаций. Данная *проблема* особенно *актуальна* при распознавании частичных СЭ ЕЯ-высказываний. Наиболее естественным путем ее решения является автоматическое выделение и систематизация понятий заданной Предметной Области, значимых в описываемых ситуациях, непосредственно на основе описывающих текстов. Разработка математической модели процесса формирования понятийной структуры тезауруса на основе текстов предметно-ограниченной тематики для задач анализа СЭ сформулирована как *основная цель* настоящей работы.

1. Семантика синтаксиса как основа кластеризации

Идея предлагаемого решения основана на зависимости лексической сочетаемости слова от его Семантического Класса (СК) [1] в заданном ЕЯ. С СК отождествляется обозначаемое словом понятие (сущность, предмет, явление) реального мира. Поэтому справедливым будет предположить, что и сам СК слова может быть выявлен путем анализа сочетаний слова с другими словами в ЕЯ-текстах по тематике заданной Предметной Области.

В работе [1] нами рассматривалась сочетаемость предикатных слов — глаголов и их производных. Следует отметить, что для извлечения СК слова из набора текстов заданной тематики первостепенную роль играет контекст целевого слова. Наибольшую точность, как показывает практика, дают модели контекста на основе синтаксических связей в предложении [4, 8]. В работе [1] мы фактически рассматривали контекст предикатного слова, который определяется, в первую очередь, синтаксическими связями между предикатом и его семантическими актантами. Для формализации понятий Предметной Области, обозначающих участников тех или иных ситуаций, необходимо ввести в рассмотрение сочетаемость соответствующих существительных со словами, являющимися синтаксически главными по отношению к ним. Причем наряду с сочетаниями «актант–предикат» требуется учитывать произвольные сочетания существительных в тексте между собой (в том числе посредством предлогов).

Каждое выявляемое из текста понятие идентифицируется (в первую очередь) относительно заданного множества ситуаций. Поскольку сами ситуации обозначаются глаголами (либо их производными), наиболее преемлемым вариантом контекста для существительного g_k, обозначающего некоторое выявляемое понятие, будет последовательность соподчиненных слов :

$$S_k = \{m_1, \ldots, m_l, m_{l+1}, \ldots, m_{n(k)}, g_k\}, \tag{1.1}$$

где m_1 — глагол (отглагольное существительное), $\forall m_l \in \{m_2, \ldots, m_{n(k)}\}$ — некоторое существительное. Причем между любой парой слов в (1.1) имеет место некоторое отношение синтаксического подчинения R_q (q — тип отношения, характеризуется падежом зависимого слова и предлогом для связи синтаксически главного слова с зависимым).

Введение в рассмотрение контекста (1.1) дает основание предположить наличие в тексте для $\forall m_l \in \{m_1, \ldots, m_{n(k)-1}\}$ последовательности $S_{kl} \neq S_k$:

$$S_{kl} = \{m_l, g_k\}, \tag{1.2}$$

в которой $m_l R_q g_k$. При этом обязательным является $m_l R_q m_{l+1}$ в (1.1). Будем называть последовательность S_k вида (1.1) ситуационным контекстом для g_k. В этом случае S_k в совокупности с множеством $\left\{S_{kl} \;\middle|\; l = \overline{1, n(k)-1}\right\}$ последовательностей вида (1.2) определяют некоторые ситуации (либо ассоциируемые с ними понятия) относительно g_k. Причем с любой из S_{kl} всегда связывается более абстрактное понятие (ситуация), чем с S_k.

Утверждение 6. *При наличии $S_k = \{m_1, \ldots, m_{n(k)}, g_k\}$ (1.1) и $S_{k1} = \{m_1, g_k\}$ (1.2) в анализируемом тексте имеет место частичная СЭ (относительно g_k).*

Пример. *«Характеристика сложности семейства алгоритмов» — «характеристика алгоритмов».* Подобная СЭ может задаваться, в частности, Генитивной Конструкцией [4, 8]. Для сравнения : *«сложность подсемейства модели» — «сложность модели».*

Утверждение 7. *При $m_1 R_q m_2$ возможно существование отношение R_q между m_1 и любым словом последовательности (1.1).*

Действительно, реальные тексты Естественных Языков, в частности, русского, обладают тем свойством, что при наличии $R_q : m_1 R_q m_2$ в последовательности вида (1.1) возможно установление данного отношения между m_1 и $\forall m_l \in \{m_3, \ldots, m_{n(k)}\}$, а также между m_1 и g_k вне зависимости от уже существующих отношений между словами этой последовательности. Данное свойство следует из соотношения смыслов соподчиненных слов. При этом для установления $m_1 R_q m_i$ (либо $m_1 R_q g_k$) каждое зависимое слово должно быть приведено в соответствующую морфологическую форму.

Пример. *«Рассматривать на множестве семейств алгоритмов»* — допустимыми с точки зрения синтаксиса и семантики русского языка являются также словосочетания *«рассматривать на семействах»* и *«рассматривать на алгоритмах»*.

В настоящей работе в качестве базовой структуры для выявления и кластеризации понятий предлагается использовать ситуационные контексты (1.1), которые участвуют в описании частичных СЭ в соответствии с Утверждением 6. Ставится *задача* : путем синтаксического разбора предложений выявить указанные контексты в анализируемом тексте и на их основе выполнить концептуальную кластеризацию.

2. Концептуальная кластеризация понятий на основе результатов синтаксического разбора предложений

Результатом синтаксического анализа текста является набор деревьев разбора предложений. В настоящей работе синтаксический анализ осуществляется программой «Cognitive Dwarf» (ООО «Когнитивные технологии», http://cs.isa.ru:10000/dwarf). При тестировании данная программа показала самые точные результаты разбора.

На основе полученного набора деревьеа формируются ситуационные контексты (1.1). При этом с каждого дерева последовательно считываются пары (x, y), где x — синтаксически главное слово, y — зависимое слово. Дальнейшая обработка считанных пар направлена на выявление последовательностей (1.1) и (1.2) в соответствии с Утверждением 6. Как результат формируется $\{S_k : S_k$ — последовательность вида (1.1)$\} =: P^S$.

В качестве инструмента концептуальной кластеризации выявленных ситуационных контекстов (1.1) как основы выделения понятий в настоящей работе используются методы теории Анализа Формальных Понятий (АФП) [4, 5, 8] — расширения теории решеток.

Приведем используемые далее основные определения из теории АФП.

Пусть G — множество объектов, M — множество признаков для объектов из G. Имеем также бинарное отношение $I \subseteq G \times M$. Если $g \in G$ и $m \in M$, то gIm имеет место тогда и только тогда, когда g обладает признаком m.

Определение 1. Тройка $K = (G, M, I)$ называется формальным контекстом. При этом для произвольных $A \subseteq G$ и $B \subseteq M$ вводится пара отображений : $A' = \{m \in M | \forall g \in A : gIm\}$ и $B' = \{g \in G | \forall m \in B : gIm\}$.

Определение 2. Пара множеств (A, B), таких что $A \subseteq G$, $B \subseteq M$ и $A' = B$, $B' = A$, называется Формальным Понятием (ФП) с формальным объемом A и формальным содержанием B.

Определение 3. ФП (A_1, B_1) называют подпонятием ФП (A_2, B_2), если $A_1 \subseteq A_2$. При этом (A_2, B_2) называют суперпонятием для ФП (A_1, B_1), обозначается как $(A_1, B_1) \leq (A_2, B_2)$. Отношение $\leq$ будем называть отношением порядка для ФП.

Определение 4. Формальные Понятия C_1 и C_2 считаются сравнимыми, если либо $C_1 \leq C_2$, либо $C_2 \leq C_1$. В противном случае эти ФП называют несравнимыми.

Определение 5. Множество всех ФП контекста $K = (G, M, I)$ вместе с заданным на нем отношением $\leq$ обозначают $\Re(G, M, I)$ и называют решеткой Формальных Понятий.

Определение 6. Подмножество множества Формальных Понятий, в котором каждые два элемента являются сравнимыми, называют цепочкой, а если каждые два элемента являются несравнимыми, называют антицепочкой.

Определение 7. Под областью в решетке Формальных Понятий понимается набор Формальных Понятий, связанных отношением порядка с одним Наибольшим Общим Подпонятием (НОПП) и/или одним Наименьшим Общим Суперпонятием (НОСП). В роли НОПП может выступать наименьшее ФП в решетке, а в роли НОСП — вершинное ФП.

Определение 8. ФП C_2 называется соседним по отношению к ФП C_1 в решетке $\Re$, если они имеют НОСП, отличное от вершинного ФП в этой решетке.

Замечание. АФП по определению есть инструмент концептуальной кластеризации, так как $\forall(A,B) \in \Re$ есть класс с заданной интерпретацией в виде содержания — B.

При извлечении из текста $\{g_k\}$ соответствует множеству объектов G. В множество признаков M включаются существительные и глаголы : для $\forall m \in M$ найдется такое g_k, что g_k задает ситуацию СЭ в соответствии с Утверждением 6 и $m \in \{m_1, \ldots, m_{n(k)}\}$ в (1.1). Отношение $I \subseteq G \times M$ в этом случае ставит в соответствие каждому $g_k \in G$ соподчиненные слова $m_l \in \{m_1, \ldots, m_{n(k)}\}$ последовательности (1.1) : для $\forall(g_k, m_l) \in I$ в анализируемом тексте присутствует $S_{kl} = \{m_l, g_k\}$ (1.2).

Замечание. Как следует из (1.1), в зависимости от наличия/отсутствия предлога между главным и зависимым словом, $\forall m \in M$ может быть представлен как :

$$m = \begin{cases} x \odot «:» \odot p_y \\ x, \end{cases} \tag{2.1}$$

где x — синтаксически главное слово, y — зависимое слово, p_y — предлог, $\odot$ — операция конкатенации.

В процессе генерации $K = (G, M, I)$ пары (g_k, m) выбираются таким образом, чтобы $\forall C = (A, B) : C \in \Re(G, M, I)$ входило в цепочку максимальной длины при $|A| \to \max$. В целях соответствия формируемого тезауруса требованию иерархичности в настоящей работе используется введенный в [4] критерий полезности решетки $\Re$:

$$F = \max_{j=1}^{J}(\sum_{i=1}^{n_j} |A_i|), \tag{2.2}$$

где J — индексное множество цепочек, $j \in J$ — номер цепочки, n_j — количество ФП в цепочке с номером j, i — порядковый номер ФП в цепочке.

Рассмотрим последовательность шагов по построению $\Re(G, M, I)$ с максимальным значением критерия полезности (1.4).

Как следует из Утверждения 7, первоначальное формирование пар-кандидатов (g_k, m) для включения в $I \subseteq G \times M$ может быть выполнено на основе P^S согласно Алгоритму 1.

Следующий этап — сбор информации о частоте $Cnt(m)$ встречаемости каждого $m \in M$ с различными $g_k \in G$. $Cnt(m)$ используется для оценки информативности [4, 8] каждого признака из первоначально выявленных для $\{g_k\}$ и подсчитывается в соответствии с Алгоритмом 2 как число соответствующих употреблений $m \in M$ в тексте [6, с. 203].

С учетом требований критерия (1.4) формирование $\Re(G, M, I)$ в настоящей работе ведется по областям, исходя из соображений максимизации длины цепочек при максимизации объема каждого ФП из входящих в цепочку. С целью минимизации числа спорных ФП [4] каждое следующее ФП в цепочке выбирается по принципу постепенного уменьшения $|B|$ и максимизации количества общих признаков с потенциальным подпонятием при минимальном количестве общих признаков с любым ФП, не входящим в цепочку.

Алгоритм 1 Формирование пар-кандидатов на включение в отношение I.

Вход: $P^S = \{S_k : S_k$ — последовательность вида (1.1)$\}$;
Выход: $P^K = \{P_k^K : P_k^K = \{(g_k, m) : (g_k, m) \in I\}\}$;
// *В целях удобства последующего формирования решетки*
// *пары группируются для каждого* g_k.
$P^K := \varnothing$; // *Инициализация*
пока $P^S \neq \varnothing$
Выбрать S_k из P^S;
$P_k^K := \varnothing$;
для $l = 1, \ldots, n(k)$
$P_k^K := P_k^K \cup \{(g_k, m_l)\}$; // $S_k = \{m_1, \ldots, m_{n(k)}, g_k\}$ *в соответствии с (1.1)*
$P^K := P^K \cup \{P_k^K\}$;
$P^S := P^S \backslash \{S_k\}$;

Алгоритм 2 Формирование частотного словаря для M.

Вход: $P^K = \{P_k^K\}$ на выходе Алгоритма 1;
Выход: $P^{Cnt} = \{(m, Cnt(m)) | m \in M\}$;
$P^{Cnt} := \varnothing$; // *Инициализация*
$P^U := \varnothing$; // *Вспомогательный цикл — объединение списков пар* (g_k, m)
пока $P^K \neq \varnothing$
Выбрать P_k^K из P^K;
$P^U := P^U \cup P_k^K$;
$P^K := P^K \backslash \{P_k^K\}$;
пока $P^U \neq \varnothing$
$P^{Oth} := \varnothing$;
Выбрать (g_k, m) из P^U;
$P^U := P^U \backslash \{(g_k, m)\}$;
$Cnt(m) := 1$;
пока $P^U \neq \varnothing$
Выбрать (g_{k_1}, m_1) из P^U;
если $m = x \odot «:» \odot p_y$ и $m_1 = x \odot «:» \odot p_{y_1}$ **то**
$Cnt(m) := Cnt(m) + 1$;
$P^U := P^U \backslash \{(g_{k_1}, m_1)\}$;
иначе
$P^{Oth} := P^{Oth} \cup \{(g_{k_1}, m_1)\}$;
$P^U := P^U \backslash \{(g_{k_1}, m_1)\}$;
$P^{Cnt} := P^{Cnt} \cup \{(m, Cnt(m))\}$;
$P^U := P^{Oth}$;

Формирование отдельной цепочки $P^C_{Ch(j)}$ на основе множества P^C объектов с заданными наборами признаков ведется согласно Алгоритму 4. Алгоритмом 5 строится множество цепочек для множества $P^C_{Neigh(j)} \subset P^C_{Ch(j)}$. Множество $P^C_{Neigh(j)}$ есть в соответствии с Определением 8 множество Формальных Понятий, соседних по отношению к тем Формальным

Понятиям $C = (A, B) : A = \{g_k\}, B = P_k^C$, между которыми устанавливается отношение порядка при формировании цепочки $P_{Ch(j)}^C$ Алгоритмом 4.

Алгоритм 3 (вспомогательный) Формирование наборов признаков для ФП.

Вход: $P^K = \{P_k^K : P_k^K = \{(g_k, m) : (g_k, m) \in I\}\}$ на выходе Алгоритма 1;
Выход: P^C; // *Множество объектов с наборами признаков*
$P^C := \varnothing$; // *Инициализация*
пока $P^K \neq \varnothing$
$P_k^C := \varnothing$;
Выбрать P_k^K из P^K;
пока $P_k^K \neq \varnothing$
Выбрать (g_k, m) из P_k^K;
$P_k^C := P_k^C \cup \{m\}$;
$P_k^K := P_k^K \backslash \{(g_k, m)\}$;
$P^C := P^C \cup \{(g_k, P_k^C)\}$;
$P^K := P^K \backslash \{P_k^K\}$;

Алгоритм 4 Формирование цепочки по масимуму критерия (1.4).

Вход: P^C на выходе Алгоритма 3;
Выход: $P_{Ch(j)}^C = \{(g_k, P_k^C) : (g_k, P_k^C) \in P^C | P_k^C$ — набор признаков для $g_k\}$;
P_{Rest}^C; // *Подмножество исходного* P^C, *не вошедшее в* $P_{Ch(j)}^C$
$P_{Neigh(j)}^C \subset P_{Ch(j)}^C$; // *Соседние ФП для тех ФП, относительно которых*
// *рассматривается отношение порядка*
$P_{Ch(j)}^C := \varnothing$; // *Инициализация*
$P_{Neigh(j)}^C := \varnothing$;
Выбрать (g_{max}, P_{max}^C) из $P^C : |P_{max}^C| \to \max$;
$P^C := P^C \backslash \{(g_{max}, P_{max}^C)\}$;
$P_{Ch(j)}^C := P_{Ch(j)}^C \cup \{(g_{max}, P_{max}^C)\}$;
$P_{tmp}^C := P_{max}^C$;
цикл
Выбрать (g_k, P_k^C) из $P^C : P_k^C \subset P_{tmp}^C$ и $|P_{tmp}^C \cap P_k^C =: Cr| \to \max$;
при $Cr = \varnothing$ **выход**;
$P_{tmp}^C := P_k^C$;
$P_{Ch(j)}^C := P_{Ch(j)}^C \cup \{(g_k, P_k^C)\}$;
$P^C := P^C \backslash \{(g_k, P_k^C)\}$;
Выбрать $\{(g_{Cr}, P_{Cr}^C) | P_{Cr}^C \supseteq Cr\} =: P^{Cr} \subseteq P^C$;
$P_{Ch(j)}^C := P_{Ch(j)}^C \cup P^{Cr}$;
$P_{Neigh(j)}^C := P_{Neigh(j)}^C \cup P^{Cr}$;
$P^C := P^C \backslash P^{Cr}$;
$P_{Rest}^C := P^C$;

Алгоритм 5 Генерация множества цепочек для «соседних» ФП.

Вход: P^C на выходе Алгоритма 3;
Выход: $P^C_{Ch} = \{P^C_{Ch(j)} : P^C_{Ch(j)}$ — цепочка ФП, формируемая Алгоритмом 4.$\}$;
  $P^C_{Ch} := \varnothing$; // *Инициализация*
  цикл
    Сформировать $P^C_{Ch(j)}$, $P^C_{Neigh(j)} \subset P^C_{Ch(j)}$ и P^C_{Rest} Алгоритмом 4 на основе P^C;
    при $|P^C_{Ch(j)}| \leqslant 1$ **выход**;
    $P^C_{Ch} := P^C_{Ch} \cup \{P^C_{Ch(j)}\}$;
    $P^C := P^C_{Neigh(j)} \cup P^C_{Rest}$;

Максимум значения полезности (1.4) для $\Re(G, M, I)$ в настоящей работе достигается удалением наименее информативных признаков $m \in M$ с наибольшими значениями $Cnt(m)$ из содержания всех ФП во всех цепочках на выходе Алгоритма 5. Максимизация полезности решетки и окончательное формирование $K = (G, M, I)$ осуществляется Алгоритмом 6.

Алгоритм 6 Генерация формального контекста.

Вход: $P^S = \{S_k : S_k$ — последовательность вида (1.1)$\}$;
Выход: $K = (G, M, I)$;
  Сформировать $P^K = \{P^K_k : P^K_k = \{(g_k, m) : (g_k, m) \in I\}\}$ Алгоритмом 1 на основе P^S;
  Сформировать $P^{Cnt} = \{(m, Cnt(m)) | m \in M\}$ Алгоритмом 2 на основе P^K;
  Сформировать P^C Алгоритмом 3 на основе P^K;
  $\Delta_F := 0$;
  пока $\Delta_F \leqslant 0$
    $\Delta_F := |\Delta_F|$;
    Сформировать P^C_{Ch} Алгоритмом 5 на основе P^C;
    Найти $\max^J_{j=1}(|P^C_{Ch(j)} : P^C_{Ch(j)} \in P^C_{Ch}|) =: F_{tmp}$, где J — индексное множество цепочек, (1.4);
    $\Delta_F := \Delta_F - F_{tmp}$;
    Найти $m_c \in M : (m_c, Cnt(m_c)) \in P^{Cnt}$ и $Cnt(m_c)$ — максимально;
    для всех $(g_k, P^C_k) \in P^C$
      $P^C_k := P^C_k \backslash \{m_c\}$;
    $P^{Cnt} := P^{Cnt} \backslash \{(m_c, Cnt(m_c))\}$;
  $K := \bigcup^J_{j=1} P^C_{Ch(j)}$;

3. Экспериментальная проверка полученных алгоритмов

Исходными текстовыми данными для формирования решетки понятий являются рефераты научных статей по тематике Предметной Области, для которой строится тезаурус. Используемое множество статей представляет собой тематическое подмножество того корпуса текстов, который по жанровому разнообразию представленного в нем рода словесности [7] следует отнести к научной прозе. При этом основным требованием к используемому множеству статей является репрезентативность [8].

Определение 9. Под репрезентативностью множества текстов в настоящей работе понимается способность этого множества отображать все свойства Предметной Области, релевантные для некоторого заданного лингвистического исследования.

В настоящей работе в качестве естественной оценки репрезентативности следовало бы взять суммарную частоту F_{S_k}, с которой последовательности (1.1), соответствующие условию Утверждения 6, встречаются в анализируемых текстах. Но с учетом отсутствия ограничений на тип q синтаксического отношения R_q между словами в (1.1) за оценку репрезентативности в настоящей работе принимается отношение частоты F_{S_k} к количеству n_q выявленных типов отношений R_q в последовательностях (1.1) :

$$F_q = \frac{F_{S_k}}{n_q} = \frac{n_{S_k}}{n \cdot n_q}, \tag{3.1}$$

где n_{S_k} — количество последовательностей (1.1), извлеченных из анализируемого множества текстов, n — общее количество слов в анализируемом множестве текстов.

Для апробации предложенных в работе алгоритмов был разработан программный комплекс, схема обмена данными между модулями которого представлена на рис.1.

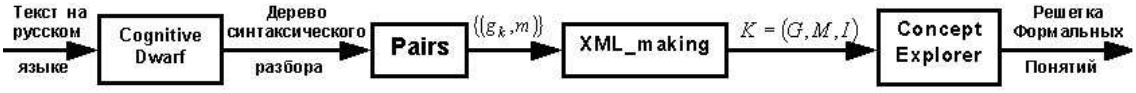

Рис. 1. Схема обмена данными между модулями программного комплекса

Извлечение потенциальных пар (g_k, m) из синтаксического дерева выполняет модуль Pairs. За основу при его реализации авторами была взята программа «Dwarfprint» в составе «Cognitive Dwarf». Генерацию контекста $K = (G, M, I)$ в соответствии с Алгоритмом 6 осуществляет разработанная авторами программа XML_making, которая представляет контекст K на выходе Алгоритма 6 в виде XML-файла. С этой целью в программе XML_making реализована процедура индексирования признаков из M. Визуализацию решетки $\Re = (G, M, I)$ диаграммой линий [5] выполняет ПО Concept Explorer (http://conexp.sourceforge.net/), реализующее методы АФП.

В качестве экспериментального текстового материала была взята работа [9]. Обзорная статья К. В. Воронцова является хорошим примером репрезентативного текста в соответствии с критерием (1.5) с характерной минимизацией n_q при максимизации F_{S_k}. Полученная для [9] решетка ФП представлена на рис.2.

Заключение

Основным *результатом* настоящей работы *является* разработанный авторами *алгоритм формирования понятийной структуры тезауруса заданной Предметной Области на основе описывающих ее текстов русского языка.* Предложенная в работе *модель тезауруса в виде решетки Формальных Понятий позволяет* оперировать данными на семантическом уровне без потери или недопустимого упрощения объектов и их признаков.

Основная *сфера применения результатов* настоящей работы — автоматизация пополнения лингвистических информационных ресурсов. Здесь следует в первую очередь отметить специализированные тезаурусы. Примером может послужить тезаурус по анализу изображений, разрабатываемый исследовательским коллективом Вычислительного центра им. А. А. Дородницына Российской академии наук

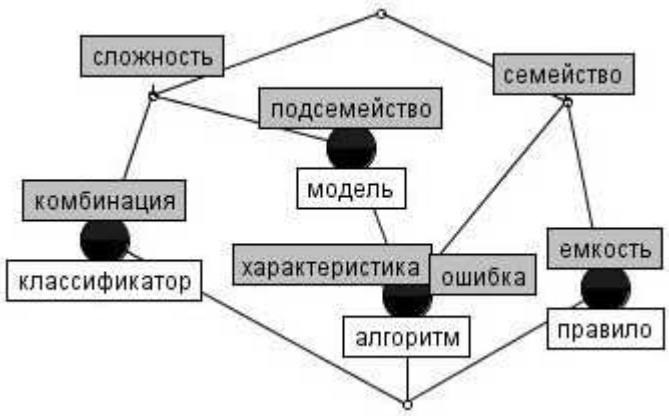

Рис. 2. Пример решетки ФП для множества ситуационных контекстов

Наибольшая эффективность *предложенного метода формирования и кластеризации понятий* может быть достигнута при совместном его использовании с представленным в [1] подходом к анализу сочетаемости предикатных слов. Здесь *в качестве перспективного направления дальнейших исследований* следует отметить развитие предложенного в настоящей работе метода применительно к Расщепленным Значениям [1] в составе последовательностей (1.1).

С учетом результатов машинного эксперимента *отдельного рассмотрения заслуживает* разработка методов предварительной обработки исходного текстового материала с целью максимизации его репрезентативности. *Перспективным* здесь *является* выделение и замена анафор, в первую очередь — анафорических личных местоимений.

Работа выполнена при поддержке РФФИ (проект №06-01-00028).

Список литературы

1. Михайлов Д. В., Емельянов Г. М. Кластеризация семантических знаний в задаче распознавания ситуаций смысловой эквивалентности // Всеросс. конф. ММРО-13. — Москва: Макс Пресс, 2007. — С. 500–503.

2. Мельчук И. А. Опыт теории лингвистических моделей «Смысл⇔текст» : Семантика, синтаксис. — Москва: Школа «Языки русской культуры», 1999. — 345 с.

3. Емельянов Г. М., Корнышов А. Н., Михайлов Д. В. Концептуально-ситуационное моделирование процесса перифразирования высказываний Естественного Языка как обучение на основе прецедентов // Искусственный интеллект. — Донецк, 2006 № 2. — С. 72–75.

4. Степанова Н. А., Емельянов Г. М. Формирование и кластеризация понятий в задаче распознавания образов в пространстве знаний // Всеросс. конф. ММРО-13. — Москва: Макс Пресс, 2007. — С. 206–209.

5. B. Ganter and R. Wille Formal Concept Analysis – Mathematical Foundations. — Berlin : Springer-Verlag, 1999. - 284 pp, 105 figs.

6. Вирт, Никлаус Алгоритмы + структуры данных = программы / Пер. с англ. Л. Ю. Иоффе. — Москва : Мир, 1985. - 406 с.: ил.

7. Рыков В. В. Корпус текстов как семиотическая система и онтология речевой деятельности // Компьютерная лингвистика и интеллектуальные технологии. Международная конференция «Диалог'2004». http://www.dialog-21.ru/conference.

8. Nadezhda Stepanova, Gennady Emelyanov Knowledge acquisition process modeling for question answering systems // Когнитивное моделирование в лингвистике : Труды IX международной конференции. — Казань : Казанский государственный университет, 2007. — С. 344–354.

9. Воронцов К. В. Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. – 2004. – №1. – С.5-24.

Статья поступила в редакцию 27.04.2008