Формирование и кластеризация понятий на основе множества ситуационных контекстов.

Д. В. Михайлов, Г. М. Емельянов

Новгородский Государственный Университет имени Ярослава Мудрого

Цель.

Разработка математической модели процесса формирования и методов качественной оценки понятийной структуры тезауруса на основе множества Семантически Эквивалентных текстов предметно-ограниченной тематики.

Задачи исследования.

1) Разработка и исследование методик формализации контекста для слов Естественного Языка (ЕЯ), обозначающих участников ситуаций в текстовых описаниях последних.
2) Формулирование требований к контексту как базовой структуре для выявления и кластеризации понятий.
3) Разработка и программная реализация метода кластеризации контекстов для существительных русского языка на основе результатов синтаксического разбора предложений.
4) Выработка качественных оценок для понятийной структуры тезауруса как результата кластеризации.
5) Проведение вычислительных экспериментов и выработка практических рекомендаций по используемому множеству исходных текстов.

Ситуационный контекст как основа кластеризации.

Пусть

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

есть последовательность соподчиненных слов, где m_1 — глагол (отглагольное существительное), $\forall m_l \in \{m_2, \ldots, m_{n(k)}\}$ — существительное. Между любой парой слов в (1) имеет место некоторое синтаксическое отношение R_q, где q — его тип, характеризуется падежом зависимого слова и предлогом для связи главного слова с зависимым.

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

$$S_{kl} = \{m_l, g_k\}, \tag{2}$$

в которой $m_l R_q g_k$. При этом обязательно $m_l R_q m_{l+1}$ в (1). Будем называть последовательность S_k вида (1) *ситуационным контекстом* для g_k. В этом случае S_k в совокупности $\left\{S_{kl} \;\middle|\; l = \overline{1, n(k)-1}\right\}$ определяют некоторые ситуации (ассоциируемые с ними понятия) относительно g_k. Причем с любой из S_{kl} ассоциируется более абстрактная ситуация, чем с S_k.

Утверждение 1. *При наличии $S_k = \{m_1, \ldots, m_{n(k)}, g_k\}$ (1) и $S_{k1} = \{m_1, g_k\}$ (2) в анализируемом тексте имеет место частичная СЭ (относительно g_k).*

Пример. *"Характеристика сложности семейства алгоритмов" — "характеристика алгоритмов".* Подобная СЭ может задаваться Генитивной Конструкцией. Сравнение : *"сложность подсемейства модели" — "сложность модели".*

Утверждение 2. *При $m_1 R_q m_2$ возможно существование отношение R_q между m_1 и любым словом последовательности (1).*

Пример. *"Рассматривать на множестве семейств алгоритмов"* — допустимыми являются также словосочетания *"рассматривать на семействах", "рассматривать на алгоритмах".*

Анализ Формальных Понятий и концептуальная кластеризация.

Пусть G — множество объектов, M — множество признаков для объектов из G, $I \subseteq G \times M$. 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 объемом A и содержанием B.*

Определение 3. *Множество всех ФП контекста $K = (G, M, I)$ вместе с заданным на нем отношением порядка $\leq$ обозначают $\Re(G, M, I)$ и называют решеткой ФП.*

Определение 4. *ФП C_1 и C_2 считаются сравнимыми, если либо $C_1 \leq C_2$, либо $C_2 \leq C_1$. Подмножество множества ФП, в котором каждые два элемента сравнимы, называют цепочкой.*

При извлечении из текста $G = \{g_k\}$, M : для $\forall m \in M$ найдется такое g_k, что g_k задает ситуацию СЭ в соответствии с Утверждением 1 и $m \in \{m_1, \ldots, m_{n(k)}\}$ в (1).

$I \subseteq G \times M$ ставит в соответствие каждому $g_k \in G$ соподчиненные слова $m_l \in \{m_1, \ldots, m_{n(k)}\}$ из (1) : для $\forall (g_k, m_l) \in I$ в тексте присутствует $S_{kl} = \{m_l, g_k\}$ (2).

Замечание. *Как следует из (1), в зависимости от наличия/отсутствия предлога p_y между главным и зависимым словом, $\forall m \in M$ может быть представлен как :*

$$m = \begin{cases} x \odot «:» \odot p_y \\ x, \end{cases} \tag{3}$$

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

Формирование пар-кандидатов на включение в отношение I.

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

Вход: $P^S = \{S_k : S_k$ — последовательность вида (1)$\}$;

Выход: $P^K = \{P_k^K : P_k^K = \{(g_k, m) : (g_k, m) \in I\}\}$;

// *В* целях удобства последующего формирования решетки

// *п*ары 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)*

$P^K := P^K \cup \{P_k^K\}$;

$P^S := P^S \backslash \{S_k\}$;

Критерий полезности решетки.

Вводится с целью оценки соответствия формируемого тезауруса требованию иерархичности.

$$F = \max_{j=1}^{J}(\sum_{i=1}^{n_j} |A_i|), \tag{4}$$

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.

Смысл : $\forall C = (A, B) : C \in \Re(G, M, I)$ должно входить в цепочку максимальной длины при $|A| \to \max$.

Определение 5. *Под областью в решетке ФП понимается набор ФП, 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.*

С учетом требований критерия (4) формирование $\Re(G, M, I)$ ведется по областям. С целью минимизации числа спорных ФП 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.

В целях максимизации критерия (4) вводится частота $Cnt(m)$ встречаемости каждого $m \in M$ с различными $g_k \in G$.

$Cnt(m)$ используется для оценки информативности каждого признака из первоначально выявленных для $\{g_k\}$ и подсчитывается как число соответствующих употреблений $m \in M$ в тексте.

Алгоритм 2 Формирование цепочки по масимуму критерия полезности.

Вход: P^C на выходе Алгоритма 3;

Выход: $P^C_{Ch(j)} = \{(g_k, P^C_k) : (g_k, P^C_k) \in P^C | P^C_k$ — набор признаков для $g_k\}$;
P^C_{Rest}; // *Подмножество исходного* P^C, *не вошедшее в* $P^C_{Ch(j)}$
$P^C_{Neigh(j)} \subset P^C_{Ch(j)}$; // *Соседние ФП для тех ФП, относительно которых*
// *рассматривается отношение порядка*

$P^C_{Ch(j)} := \varnothing$; // *Инициализация*
$P^C_{Neigh(j)} := \varnothing$;
Выбрать (g_{max}, P^C_{max}) из $P^C : |P^C_{max}| \to \max$;
$P^C := P^C \backslash \{(g_{max}, P^C_{max})\}$;
$P^C_{Ch(j)} := P^C_{Ch(j)} \cup \{(g_{max}, P^C_{max})\}$;
$P^C_{tmp} := P^C_{max}$;
цикл
Выбрать (g_k, P^C_k) из $P^C : P^C_k \subset P^C_{tmp}$ и $|P^C_{tmp} \cap P^C_k =: Cr| \to \max$;
при $Cr = \varnothing$ **выход**;
$P^C_{tmp} := P^C_k$;
$P^C_{Ch(j)} := P^C_{Ch(j)} \cup \{(g_k, P^C_k)\}$;
$P^C := P^C \backslash \{(g_k, P^C_k)\}$;
Выбрать $\{(g_{Cr}, P^C_{Cr}) | P^C_{Cr} \supseteq Cr\} =: P^{Cr} \subseteq P^C$;
$P^C_{Ch(j)} := P^C_{Ch(j)} \cup P^{Cr}$;
$P^C_{Neigh(j)} := P^C_{Neigh(j)} \cup P^{Cr}$;
$P^C := P^C \backslash P^{Cr}$;
$P^C_{Rest} := P^C$;

Алгоритм 3 (вспомогательный) Формирование наборов признаков для Формальных Понятий.

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 Генерация множества цепочек для "соседних" Формальных Понятий.

Вход: P^C на выходе Алгоритма 3;

Выход: $P_{Ch}^C = \{P_{Ch(j)}^C : P_{Ch(j)}^C$ — цепочка ФП, формируемая Алгоритмом 2.$\}$;

$P_{Ch}^C := \varnothing$; // *Инициализация*
цикл
  Сформировать $P_{Ch(j)}^C$, $P_{Neigh(j)}^C \subset P_{Ch(j)}^C$ и P_{Rest}^C Алгоритмом 2 на основе P^C;
  при $|P_{Ch(j)}^C| \leqslant 1$ **выход**;
  $P_{Ch}^C := P_{Ch}^C \cup \{P_{Ch(j)}^C\}$;
  $P^C := P_{Neigh(j)}^C \cup P_{Rest}^C$;

Алгоритм 5 Формирование частотного словаря для признаков из множества 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
      $P^{Oth} := P^{Oth} \cup \{(g_{k_1}, m_1)\}$;
      $P^U := P^U \backslash \{(g_{k_1}, m_1)\}$;
  $P^{Cnt} := P^{Cnt} \cup \{(m, Cnt(m))\}$;
  $P^U := P^{Oth}$;

Алгоритм 6 Генерация формального контекста.

Вход: $P^S = \{S_k : S_k$ — последовательность вида (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 5 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_{Ch}^C Алгоритмом 4 на основе P^C;

Найти $\max_{j=1}^{J}(|P_{Ch(j)}^C : P_{Ch(j)}^C \in P_{Ch}^C|) =: F_{tmp}$, где J — индексное множество цепочек, (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 $(g_k, P_k^C) \in P^C$

$P_k^C := P_k^C \backslash \{m_c\}$;

$P^{Cnt} := P^{Cnt} \backslash \{(m_c, Cnt(m_c))\}$;

$K := \bigcup_{j=1}^{J} P_{Ch(j)}^C$;

Структура программного комплекса.

Рис. 1. Схема обмена данными между модулями программного комплекса

Здесь :

Cognitive Dwarf — программный пакет синтаксического анализа (ООО "Когнитивные технологии", http://cs.isa.ru:10000/dwarf);

Pairs — модуль извлечения потенциальных пар (g_k, m) из синтаксического дерева;

XML_making — модуль генерации контекста $K = (G, M, I)$ по Алгоритму 6;

Concept Explorer — система анализа данных, реализующая методы Анализа Формальных Понятий, http://conexp.sourceforge.net.

Репрезентативность исходного множества текстов.

Определение 6. *Под репрезентативностью множества текстов понимается способность этого множества отображать все свойства Предметной Области, релевантные для некоторого заданного лингвистического исследования.*

Оценка репрезентативности :

$$F_q = \frac{F_{S_k}}{n_q} = \frac{n_{S_k}}{n \cdot n_q}. \tag{5}$$

Здесь :

F_{S_k} — суммарная частота, с которой последовательности (1), соответствующие условию Утверждения 1, встречаются в анализируемых текстах;

q — тип отношения R_q между словами в (1);

n_q — количество выявленных типов отношений R_q в последовательностях вида (1);

n_{S_k} — количество последовательностей вида (1), извлеченных из анализируемого множества текстов;

n — общее количество слов в анализируемом множестве текстов.

Пример генерации модели понятийной структуры тезауруса в виде решетки Формальных Понятий.

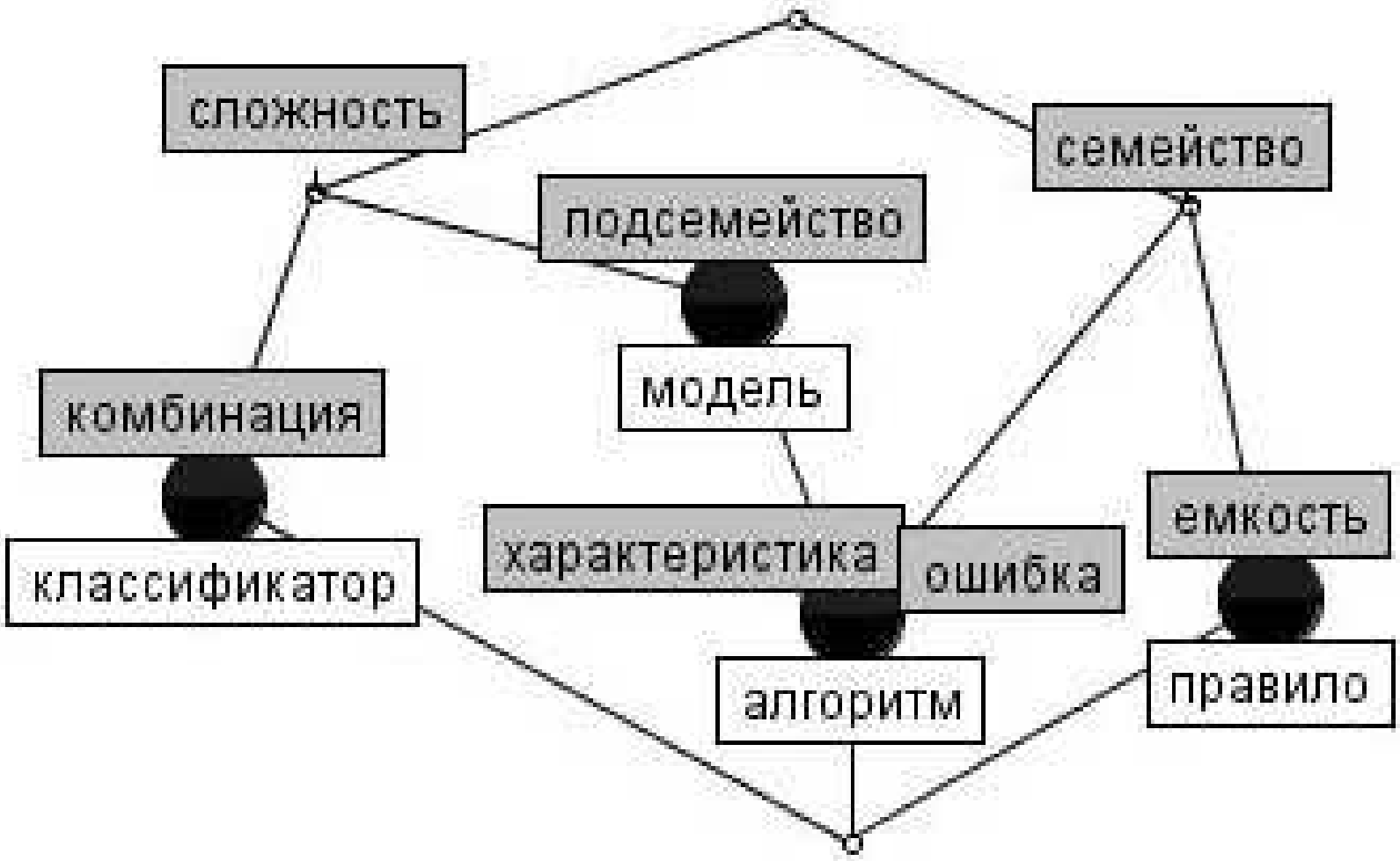

Рис. 2. Пример решетки ФП для множества ситуационных контекстов

Выводы.

- Предложенная модель тезауруса в виде решетки Формальных Понятий позволяет оперировать данными на семантическом уровне без потери или недопустимого упрощения объектов и их признаков.
- Разработанный алгоритм генерации модели понятийной структуры тезауруса в качестве исходных данных для формирования решетки использует результаты синтаксического разбора предложений. Максимум эффективности предложенного метода формирования и кластеризации понятий достигается при совместном его использовании с методами анализа сочестаемости предикатных слов в рамках семантико-синтаксических валентностей последних.
- При использовании последовательностей соподчиненных слов как основы кластеризации введенная оценка репрезентативности для исходного текстовго материала позволяет делать практические выводы о точности используемых алгоритмов синтаксического анализа и направлениях их дальнейшего совершенствования.

Перспективы дальнейших исследований.

- Развитие предложенного метода формирования и кластеризации понятий применительно к Расщепленным Значениям в составе последовательностей соподчиненных слов.
- Разработка и исследование методов предварительной обработки исходного текстового материала с целью максимизации его репрезентативности.