Формирование и кластеризация понятий в задаче распознавания образов в пространстве знаний

Степанова Н. А., Емельянов Г. М.
StepanovaNadya@gmail.com
Великий Новгород, ГОУ ВПО НовГУ им. Ярослава Мудрого

Данная работа представляет метод, основанный на теории решеток, для извлечения лексических знаний из текста и последующего упорядочивания знаний. Извлечение знаний выполняется с целью пополнения лексического ресурса, базовым элементом которого является понятие, объединяющее толкование значения лексем с формами лексем.

Будем использовать расширение теории решеток — теорию Анализа Формальных Понятий (АФП) [1]. АФП является инструментом концептуальной кластеризации, так как Формальные Понятия (ФП) решетки являются классами с заданной в виде содержания понятий интерпретацией. При извлечении из текста лексемы, не содержащейся в лексиконе, требуется отнести данную лексему на основе ее признаков к одному из уже имеющихся в лексиконе ФП (классов) или образовать с помощью лексемы новый класс. Построенный, таким образом, лексикон необходимо обрабатывать с целью извлечения классов более высокого уровня абстракции, состоящих из нескольких ФП. Описанная выше задача кластеризации является классической задачей распознавания образов.

Приведем основные определения АФП. Пусть G и M — множества, называемые соответственно множествами объектов и признаков, а $I \subseteq G \times M$ — бинарное отношение. Если $g \in G$ и $m \in M$, то gIm имеет место, если g обладает признаком m. Тройка $\mathbb{K} = (G, M, I)$ называется формальным контекстом. Для произвольных $A \subseteq G$ и $B \subseteq M$ вводится пара отображений: $A' = \{m \in M | \forall g \in A : gIm\}$, $B' = \{g \in G | \forall m \in B : gIm\}$. Пара множеств (A, B), таких что $A' = B$ и $B' = A$, называется ФП с объемом A и содержанием B. ФП (A_1, B_1) называют подпонятием понятия (A_2, B_2), если $A_1 \subseteq A_2$, при этом (A_2, B_2) называют суперпонятием понятия (A_1, B_1), обозначается $(A_1, B_1) \leqslant (A_2, B_2)$. Множество всех ФП контекста $\mathbb{K}$ вместе с отношением порядка называют решеткой ФП $\underline{\mathfrak{B}}(G, M, I)$. Подмножество ФП, в котором каждые два элемента являются сравнимыми, называют цепочкой, а если каждые два элемента являются несравнимыми, называют антицепочкой.

Предлагается использовать Генитивные Конструкции (ГК) [2] в качестве базовой структуры обработки текста. При синтаксическом разборе текста выделяются отдельные ГК с указанием Опорного Слова (ОС) и существительного Генитивной Группы (ГГ). Сорта — элементы «наивной картины мира», классы, к которым язык относит более конкретные реалии. Если две ГК относятся к одному сорту, то их ОС и/или ГГ также

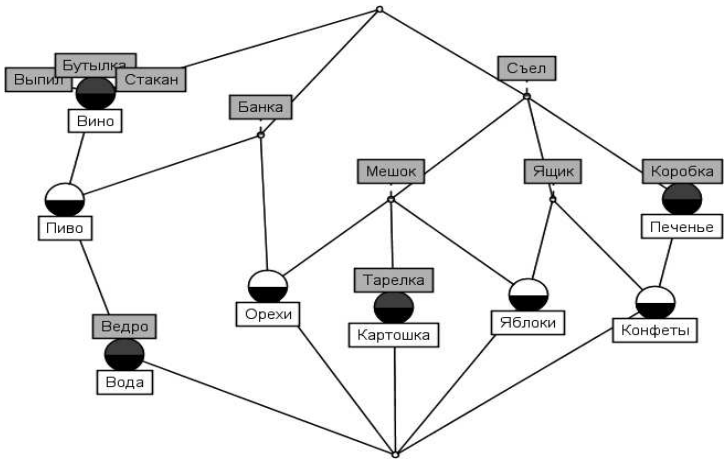

Рис. 1. Формальная решетка генитивных конструкций.

попарно относятся к одному сорту [2], а значит в их толкованиях должны содержаться общие для них свойства, определяемые их сортом. К одному сорту с некоторой вероятностью будем относить ГК при совпадении их форм ОС или ГГ.

Пусть V_s — множество форм ОС и $v_s \in V_s$, V_{gg} — множество ГГ и $v_{gg} \in V_{gg}$. Бинарным отношением $I \subseteq V_{gg} \times V_s$ назовем множество пар (v_{gg}, v_s) правильных генитивных конструкций. В $\mathbb{K} = (V_{gg}, V_s, I)$ ОС рассматриваются как признаки объектов (ГГ), означающие, что объекты имеют общие свойства. По формальному контексту $\mathbb{K}$ построим решетку $\underline{\mathfrak{B}}(V_{gg}, V_s, I)$. Для целей дополнительной спецификации значения ГК будем использовать глаголы, для модели управления которых ГК занимает место одного из актантов. В расширенном контексте в качестве множества признаков будем использовать объединение множеств $V_s \cup V_g$, где V_g — множество форм глаголов. Приведем пример решетки формальных понятий для набора ГК, извлеченных из текста (рис. 1). Все объекты из объема ФП обладают набором общих свойств A', которые описываются признаками из содержания понятия. Набор признаков будем рассматривать как толкования значений соответствующих лексем из объема ФП. Таким образом, через ГК из текста происходит извлечение знаний, представленных формальными понятиями.

Рассмотрим алгоритм сегментации решетки, выделяющий из первоначальной решетки L классы ФП, которые являются более высоким уровнем абстракции, чем отдельные ФП. Любое подмножество ФП будет обязательно иметь уникальное Наименьшее Общее Суперпонятие

(НОСП). Областью в решетке называется набор ФП, связанных отношением порядка с одним НОСП. Задачей алгоритма сегментации решетки является конвертация решетки L в набор формальных решеток $\{L'\}$, где каждой решетке $L_i \in \{L'\}$ частично соответствует область первоначальной решетки L и каждое ФП, кроме вершинного и наименьшего ФП, принадлежит только к одной итоговой решетке из $\{L'\}$. Поскольку области в решетке могут пересекаться, то для выполнения условия принадлежности ФП только к одной итоговой решетке необходимо, чтобы классы L_i лишь частично соответствовали областям первоначальной решетки $\{L\}$. Спорным ФП первоначальной решетки L называют такое ФП C, что C принадлежит более чем к одной области решетки L, у которых НОСП являются несравнимыми ФП.

Для максимизации количества элементов в итоговых классах в качестве НОСП областей решетки L возьмем ФП, являющиеся непосредственными подпонятиями вершинного ФП. Критерием выделения решетки L_i из решетки L является условие, что каждое ФП $C \in L_i$ более схоже с другими ФП из решетки L_i, чем с ФП из решеток $L_j \in \{L'\}$, где $i \neq j$. Мера схожести между ФП вычисляется по формуле

$$\mathrm{spc}(C_i, C_j) = -\log\left(1 - \frac{D_c}{\mathrm{path}_C}\right) \times \frac{|B|}{|B_j \backslash B| + |B_j \backslash B| + |B|}, \qquad (1)$$

где ФП $C = (A, B)$ является НОСП для формальных понятий C_i и C_j; D_c — количество ФП в цепочке, в которой максимальным ФП является вершинное ФП, а минимальным — ФП C; path_C — минимальное количество ФП в цепочке, которой принадлежат вершинное, наименьшее ФП и ФП C. Мера схожести учитывает объем общей ($|B|$) и индивидуальной ($|B_i|, |B_j|$) информации ФП C_i и C_j, специфичность общей информации (D_c), а также неравномерность глубины иерархии решетки (path_C).

Алгоритм сегментации для каждого ФП C_i, являющегося непосредственным подпонятием вершинного ФП, включает C_i в решетку L_i. Далее выполняется поиск в решетке L всех подпонятий ФП C_i, и каждое из этих подпонятий C_j включается в решетку L_i, если все непосредственные суперпонятия ФП C_j являются подпонятиями ФП C_i или совпадают с C_i. В противном случае, ФП C_j является спорным ФП и относится алгоритмом к тому классу, к которому принадлежит то из его непосредственных суперпонятий, для которого значение меры схожести по формуле (1) с ФП C_j будет максимальным.

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

Литература

[1] *Ganter B., Wille R.* Formal Concept Analysis - Mathematical Foundations. — Berlin: Springer-Verlag, 1999. — 284 с.

[2] *Partee B. H.* Formal Semantics, Lectures, RGGU, 2003. — `people.umass.edu/partee/`.