# Формализация «естественных» понятий[*]

Витяев Е.Е.
*vityaev@math.nsc.ru*
Институт математики им. С.Л.Соболева,
Новосибирский государственный университет (Новосибирск, Россия)

В работах Eleanor Rosch 1973, 1978, на основании большого количества экспериментов, были сформулированы следующие принципы категоризации «естественных» категорий:

«**Cognitive Economy**. *The first principle contains the almost common-sense notion that, as an organism, what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible. To categorize a stimulus means to consider it, for purposes of that categorization, not only equivalent to other stimuli in the same category but also different from stimuli not in that category. On the one hand, it would appear to the organism's advantage to have as many properties as possible predictable from knowing any one property, a principle that would lead to formation of large numbers of categories with as fine discriminations between categories as possible*».

«**Perceived World Structure**. *The second principle of categorization asserts that … perceived world – is not an unstructured total set of equiprobable co-occurring attributes. Rather, the material objects of the world are perceived to possess … high correlational structure … combinations of what we perceive as the attributes of real objects do not occur uniformly. Some pairs, triples, etc., are quite probable, appearing in combination … with one, sometimes another attribute; others are rare; others logically cannot or empirically do not occur*».

Первый принцип невозможен без второго – когнитивная экономия не возможна без структурированности мира. Непосредственно воспринимаемые объекты (basic objects) – информационно богатые связки наблюдаемых свойств, создающую категоризацию. «Categories can be viewed in terms of their clear cases if the perceiver places emphasis on the correlational structure of perceived attributes … By prototypes of categories we have generally meant the clearest cases of category membership» (E. Rosch, 1978).

В дальнейшем теория «естественных» понятий Eleanor Rosch получила название прототипической теории понятий (prototype theory). Основные ее черты описываются в Ross, et al 2008 следующим образом: «The prototype view … keeps the attractive assumption that there is some underlying common set of features for category members but relaxes the requirement that every member have all the features. Instead, it assumes there is a probabilistic matching process: Members of the category have more features, perhaps weighted by importance, in common with the prototype of this category than with prototypes of other categories».

В дальнейших исследованиях было обнаружено, что моделей, основанных на признаках, сходстве и прототипах, недостаточно для описания классов. Необходимо учитывать теоретические, причинные и онтологические знания, относящиеся к объектам классов. Например, люди не только знают, что птицы имеют крылья, могут летать и вить гнезда на деревьях, но также и то, что птицы вьют гнезда на деревьях, потому что могут летать, и летать, потому что они имеют крылья.

Учитывая эти исследования, Bob Rehder 2003 выдвинул теорию причинных моделей (causal-model theory), в соответствии с которой: «people's intuitive theories about categories of objects consist of a model of the category in which both a category's features and the causal mechanisms among those features are explicitly represented. In other words, theories might make certain combinations of features either sensible and coherent … in light of the relations linking

[*] Автор выражает благодарность К.В.Анохину за обращение внимания на данный класс работ и множество полезных замечаний.

them, and the degree of coherence of a set of features might be an important factor determining membership in a category».

В теории причинных моделей отношение объекта к категории основывается уже не на множестве признаков и близости по признакам, а на основании сходства порождающего причинного механизма: «Specifically, a to-be-classified object is considered a category member to the extent that its features were likely to have been generated by the category's causal laws, such that combinations of features that are likely to be produced by a category's causal mechanisms are viewed as good category members and those unlikely to be produced by those mechanisms are viewed as poor category members» (B. Rehder 2003).

Для представления причинного знания в работах Griffiths, T. L., & Tenenbaum, J. B. 2009, B. Rehder 2003, 2011 были использованы Байесовские сети. Однако они не поддерживают циклов и поэтому не могут моделировать циклические причинные связи.

Предлагаемая нами формализация прямо основана на циклических причинных связях, представленных неподвижными точками предсказаний по причинным связям.

Пусть X(a) – множество свойств объекта *a*, заданных некоторым множеством предикатов, $(P_{i_1} \& ... \& P_{i_k} \Rightarrow P_{i_0}) \in MS(X)$ – множество максимально специфических (E.Vityaev 2012) условных связей, выполненных на X, $\{P_{i_1}, ..., P_{i_k}\} \subset X$. Тогда оператор предсказания Pr записывается следующим образом (Витяев Е.Е и др., 2014):

$$Pr(X) = \Phi_{Krit}(X \cup \{P_{i_0} \mid (P_{i_1} \& ... \& P_{i_k} \Rightarrow P_{i_0}) \in MS(X)\} \cup \{\neg P_{i_0} \mid (P_{i_1} \& ... \& P_{i_k} \Rightarrow \neg P_{i_0}) \in MS(X)\}),$$

где $\Phi_{Krit}(X)$ – оператор, модифицирующий множество признаков X путем добавления или удаления некоторого из признаков так, чтобы определенный критерий Krit согласованности причинных связей по взаимному предсказанию признаков X был максимальным. Критерий Krit по-своему измеряет интегрированную информацию системы причинных связей MS(X), определенную в теории G.Tononi (M. Oizumi, L. Albantakis, G. Tononi, 2014). Неподвижная точка достигается тогда, когда $Pr^{n+1}(X(a)) = Pr^n(X(a))$, для некоторого n, где $Pr^n$ есть n кратное применение оператора Pr. Компьютерный эксперимент (Витяев Е.Е и др., 2014) подтвердил возможность формирования понятий в данной формализации.

*Работа выполнена при финансовой поддержке гранта РФФИ, проект 15-07-03410-а.*

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. Psychological Review, 116, 56.

Masafumi Oizumi, Larissa Albantakis, Giulio Tononi. 2014. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0 // PLOS Computational Biology, May 2014, V.10. Issue 5.

Bob Rehder. 2003. Categorization as causal reasoning // Cognitive Science 27, 709–748.

Bob Rehder, Jay B. Martin. 2011. Towards A Generative Model of Causal Cycles // 33rd Annual Meeting of the Cognitive Science Society, (CogSci 2011), Boston, Massachusetts, USA, 20-23 July 2011, V.1 pp. 2944-2949.

Rosch, E.H. 1973 Natural categories // Cognitive Psychology 4. P. 328-350.

Rosch, E. 1978 Principles of Categorization // Rosch, E. & Lloyd, B.B. (eds), Cognition and Categorization, Lawrence Erlbaum Associates, Publishers, (Hillsdale), P. 27–48

B. H. Ross, E. G. Taylor, E. L. Middleton, and T. J. Nokes. 2008. Concept and Category Learning in Humans // H. L. Roediger, III (Ed.), Cognitive Psychology of Memory. Vol. [2] of Learning and Memory: A Comprehensive Reference, 4 vols. (J.Byrne Editor), Oxford: Elsevier, P. 535-556.

E.E. Vityaev, A.V. Demin, D. K. 2012. Ponomaryov. Probabilistic Generalization of Formal Concepts // Programming and Computer Software, Vol. 38, No. 5. P. 219–230.

Витяев Е.Е., Неупокоев Н.В. Формальная модель восприятия и образа как неподвижной точки предвосхищений // Подходы к моделированию мышления. УРСС Эдиториал, Москва, 2014г., стр. 155-172.