Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications

Radim Belohlávek, Martin Trnecka
Data Analysis and Modeling Lab (DAMOL) 
Department of Computer Science, Palacky University, Olomouc, Czech Republic

Introduction

We present a study regarding basic level of concepts in conceptual categorization. The basic level of concepts is an important phenomenon studied in the psychology of concepts. We propose to utilize this phenomenon in formal concept analysis to select important formal concepts. Such selection is critical because, as is well known, the number of all concepts extracted from data is usually large. We review and formalize the main existing psychological approaches to basic level which are presented only informally and are not related to any particular formal model of concepts in the psychological literature. Interestingly, our formalization and experiments reveal previously unknown relationships between the existing approaches to basic level. Thus, we argue that a formalization of basic level in the framework of formal concept analysis is beneficial for the psychological investigations themselves because it helps put them on a solid, formal ground.

What Do You See?

Why do dogs?

Usually people answer a dog. But why dog? There is a number of other possibilities: Animal, Mammal, Canine beast, Retriever, Golden Retriever, Marley. So why dog? Because “dog” is a basic level concept.

Basic Level Phenomenon

- Extensively studied phenomenon in psychology of concepts.
- When people categorize (or name) objects, they prefer to use certain kind of concepts.
- Such concepts are called the concepts of the basic level.
- Definition of basic level concepts?
  - Are cognitive economic to use; “save the world well”.
  - Several informal definitions proposed.

Basic Level Formalization

- For a given approach M to basic level, we define a function BL_M mapping every concept (A, B) in the concept lattice $IT(X, Y)$ to $[0, \infty)$ or to $[0, 1]$.
- $BL_M(A, B)$ is interpreted as the degree to which (A, B) belongs to the basic level.
- A basic level is thus naturally seen as a graded (fuzzy) set rather than a clear-cut set of concepts.
- We consider the following probability space: X (objects) are the elementary events, $\mathbb{P}$ (sets of objects) are the events, the probability distribution is given by $P(x) = \frac{1}{|X|}$ for every object $x \in X$. For an event $A \subseteq X$, then, $P(A) = |A|/|X|$. The event corresponding to a set $\{y_1, \ldots, y_n\} \subseteq Y$ of attributes is $\{y_1, \ldots, y_n\}$.

Basic Level Metrics

We formalize in formal concept analysis the main existing psychological approaches to basic level.

Cue Validity (CV)

Consider the notion of a cue validity of attribute $x$ for concept $c$, i.e. the conditional probability $p(c|x)$ that an object belongs to $c$ given that it has $x$.

$$BL_{CV}(A, B) = \sum_{p \in P} P(A|p|) = \sum_{p \in P} \frac{|A \cap \{x\}|}{|A|}$$

Similarity approach (S)

- Based on informal (one of the first) definition from Eleanor Rosch (1975).
- Basic level concept satisfies three conditions:
  1. The objects of this concept are similar to each other;
  2. The objects of the superordinate concepts are significantly less similar;
  3. The objects of the subordinate concepts are only slightly more similar.
- Formalized in our previous paper (Belohlávek, Trnecka, IFCA 2012).

Category Feature Collocation Approach (CFC)

- Defined as product $p(c|x) \cdot p(c|y)$ of the cue validity $p(c|x)$ and the so-called category validity $p(c|y)$.

$$BL_{CFC}(A, B) = \sum_{p \in P} P(A|p|) = \sum_{p \in P} \frac{|A \cap \{x\}|}{|A|} \cdot \frac{|A \cap \{y\}|}{|A|}$$

Category Utility Approach (CU)

- Utilizes the notion of category utility $e(c) = p(c) - p(c|y)$.

$$BL_{CU}(A, B) = \sum_{p \in P} P(A|p) = \sum_{p \in P} \frac{|A \cap \{x\}|}{|A|} - \frac{|A \cap \{y\}|}{|A|}$$

Predictability Approach (P)

- Frequently formulated in the literature.
- Basic level concepts are abstract concepts that still make it possible to predict well the attributes of their objects.
- We introduce a graded (fuzzy) predicate $pro_{E}$ such that $pro_{E}(c) \in [0, 1]$ is interpreted as the truth degree of proposition “concept $c \langle A, B \rangle$ enables good prediction”.
- We use the principles of fuzzy logic to obtain the truth degrees $pro_{E}(c)$.

Basic Level in Formal Concept Analysis:

Utilizes the notation of category utility $e(c)$.

$$BL_{P}(A, B) = P(A) \sum_{p \in P} \left( \frac{P(A|p) \cdot P(A|p)}{P(A)} - P(p) \right)^2$$

Comparison of Basic Level Metrics

We provide a comparative analysis of the metrics which represent different quantitative approaches to describe a single phenomenon.

Similarity of Rankings

- For input data $(X, Y, I)$, a given metric $BL_M$ determines a ranking of concepts in $IT(X, Y, I)$, i.e. determines the linear quasorder $\leq_{BL_M}$ defined by $(A, B) \leq_{BL_M} (A', B')$ if $BL_M(A, B) \leq_{BL_M} BL_M(A', B')$.
- We examined the pairwise similarities of the rankings $BL_M$, $CU$, $CFC$, $CV$, and $P$ for various datasets. We used the Kendall tau coefficient to assess the similarities.
- We use several real and synthetic datasets.
- Even though this approach might seem rather strict, significant patterns were obtained.

Comparison of Basic Level Metrics

We provide a comparative analysis of the metrics which represent different quantitative approaches to describe a single phenomenon.

Similarity of Rankings

- For input data $(X, Y, I)$, a given metric $BL_M$ determines a ranking of concepts in $IT(X, Y, I)$, i.e. determines the linear quasorder $\leq_{BL_M}$ defined by $(A, B) \leq_{BL_M} (A', B')$ if $BL_M(A, B) \leq_{BL_M} BL_M(A', B')$.
- We examined the pairwise similarities of the rankings $BL_M$, $CU$, $CFC$, $CV$, and $P$ for various datasets. We used the Kendall tau coefficient to assess the similarities.
- We use several real and synthetic datasets.
- Even though this approach might seem rather strict, significant patterns were obtained.

Comparison of Basic Level Metrics

We provide a comparative analysis of the metrics which represent different quantitative approaches to describe a single phenomenon.

Similarity of Rankings

- For input data $(X, Y, I)$, a given metric $BL_M$ determines a ranking of concepts in $IT(X, Y, I)$, i.e. determines the linear quasorder $\leq_{BL_M}$ defined by $(A, B) \leq_{BL_M} (A', B')$ if $BL_M(A, B) \leq_{BL_M} BL_M(A', B')$.
- We examined the pairwise similarities of the rankings $BL_M$, $CU$, $CFC$, $CV$, and $P$ for various datasets. We used the Kendall tau coefficient to assess the similarities.
- We use several real and synthetic datasets.
- Even though this approach might seem rather strict, significant patterns were obtained.

Similarity of Sets of Top r Basic Level Concepts

- One is arguably more interested in the set consisting of the top r concepts of $IT(X, Y, I)$ according to the ranking $\leq_{BL_M}$ for a given metric M. We denote such set by $Top^{BL_M}r$.
- For formal concepts $(C, D), (E, F) \in IT(X, Y, I)$, denote by $s(C, D), (E, F)$ degree of similarity. For two metrics $M$ and $N$, and a given $r = 1, 2, \ldots$, we define:

$$d_{BL_M}(C, D), (E, F) = \min_{\langle A, B \rangle \in Top^{BL_M}r} s(C, D), (E, F)$$

where

$$I_{MS} = \sum_{(C, D), (E, F)} d_{BL_M}(C, D), (E, F)$$

$$I_{MS} = \sum_{(C, D), (E, F)} \min_{\langle A, B \rangle \in Top^{BL_M}r} s(C, D), (E, F)$$

Similarities $s$ of sets of top r concepts

Simultaneous Approach (P) vs. Subsets Approach (P)

- $s_{Top}^{BL_M}r$ may naturally be interpreted as the truth degree of the proposition “for most concepts in $Top^{BL_M}r$ there is a similar concept in $Top^{BL_M}r$ and vice versa”.

Conclusions

CU, CFC, and CV may naturally be considered as a group of metrics with significantly similar behavior, while S and P represent separate, singleton groups. This observation contradicts the current psychological knowledge. Namely, the (informal) descriptions of S, P, and CU are traditionally considered as essentially equivalent descriptions of the notion of basic level in the psychological literature. On the other hand, CFC has been proposed by psychologists as a supposedly significant improvement of CV and the same can be said of CU versus CFC.