Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

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

Radim Belohlavek and Martin Trnecka
Data Analysis and Modeling Lab (DAMOL)
Department of Computer Science, Palacky University, Olomouc, Czech Republic
radim.belohlavek@acm.org, martin.trnecka@gmail.com

Abstract

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. We argue and demonstrate by examples that basic level concepts may be regarded as interesting, informative formal concepts from a user viewpoint. 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.

1 Motivation, Paper Outline

One of the crucial problems in data mining is the fact that the number of patterns extracted from data may become too large to be reasonably comprehended by a user [Tan *et al.*, 2006]. This is in particular true of formal concept analysis (FCA) which is concerned with extraction of particular groupings, called formal concepts, from object–attribute Boolean data. Typically, a domain expert considers some of the extracted concepts more important (interesting, useful) than others. One possible reason is that the user exploits some background knowledge in judging the concepts' importance. This view resulted in methods that aim at extraction of only certain formal concepts—the important ones according to the background knowledge, see e.g. [Belohlavek and Vychodil, 2006; 2009; Cellier *et al.*, 2008; Dias and Vieira, 2010; Kwuida *et al.*, 2010]. Another possibility is that the user judgment on concepts' importance is due to certain psychological processes and phenomena. In this paper, we examine one such phenomenon—the basic level of concepts.

Our initial study [Belohlavek and Trnecka, 2012] demonstrated that the basic level phenomenon may be utilized to select important formal concepts, making thus a first step in the present research direction. The present paper is considerably more comprehensive. We review key contributions regarding basic level and describe and formalize within FCA five approaches to basic level. By experimental evaluation, we show that the approaches may be used to deal with the problem of a large number of concepts because the corresponding basic levels, i.e. the selected sets of concepts, are reasonably sized and tend to contain significant, useful concepts. Our results support the thesis that basic level is a fuzzy (graded) rather than a clear-cut notion. Moreover, we take advantage of the formalizations of the basic level approaches to assess their mutual relationships. Interestingly, this way we observe strong similarities of some of the approaches—an observation that has not been made in the literature on the psychology of concepts. Last but not least, we discuss some critical issues and challenges in formalizing the basic levels of concepts.

Two main implications of our paper are the following. First, when properly formalized, the psychological approaches to basic level, and possibly other psychological knowledge, may naturally be utilized in data mining. Second, the formalization of psychological knowledge makes possible a more concrete reflection on the psychological theories, as well as a formation and testing of more precise hypotheses regarding the psychological theories, resulting in a mutually beneficial interplay between cognitive psychology on the one hand and formal theories of data and data analysis on the other hand.

2 Basic Level: Issues and Challenges

2.1 The Phenomenon of Basic Level

The basic level phenomenon is, in a sense, encountered in everyday life. When we see a particular dog, we say "This is a dog," rather than "This is a German Shepherd" or "This is a mammal." That is, we prefer to name the object we see by "dog" to naming it by "German Shepherd" or "mammal". Put briefly, basic level concepts are the concepts we prefer in

*Supported by by the ESF project No. CZ.1.07/2.3.00/20.0059, the project is cofinanced by the European Social Fund and the state budget of the Czech Republic.

naming objects.

Such rough definition gives us the first idea but it may hardly be made operational. The first work demonstrating that people consistently use a kind of middle level concepts in speach is [Brown, 1958]. A comprehensive overview of the work on the basic level is provided in [Murphy, 2002]. According to the current view in the psychology of concepts, the basic level of concepts is understood as

> "...the most natural, preferred level at which to conceptually carve up the world. The basic level can be seen as a compromise between the accuracy of classification at a maximally general level and the predictive power of a maximally specific level." [Murphy, 2002, p. 210].

We turn to more particular characterizations of basic level in the next section and then in Section 3. Importantly, since the pioneering work by Rosch [Rosch, 1978; Rosch *et al.*, 1976], many studies confirmed and further developed the observation that basic level concepts have a significant cognitive role. One of the most important features of basic level concepts consists in the fact that they provide us with a lot of information with little cognitive effort [Murphy, 2002; Rosch, 1978]. Related to this fact is the capability to make accurate predictions with small cognitive effort, the capability to categorize quickly, and the fact that basic level concepts are easier to learn, see [Murphy, 2002, pp. 210–214] and references therein. Another important contention, supporting the objective nature of basic levels, is the observation that people across cultures tend to use the same level of concepts in naming animals and plants [Berlin, 1992].

2.2 Basic Level Definitions and Metrics

The psychological literature does not contain a single, robust and uniquely interpretabledefinition of the notion of basic level. Even a single paper, such as the programmatic [Rosch *et al.*, 1976], contains several informal, verbally described definitions. This reflects a variety of existing views regarding basic level and, moreover, is a natural consequence of the many subtleties regarding basic level as well as the nature of psychological research itself. For illustration, let us present three definitions by Rosch, certainly the most influential author in this area. [Rosch *et al.*, 1976, p. 383]: "In general, the basic level of abstraction in a taxonomy is the level at which categories carry the most information, ..., and are, thus, the most differentiated from one another." [Rosch *et al.*, 1976, p. 384]: "...the basic categorization is the most general and inclusive level at which categories can delineate real-world correlational structures." [Rosch, 1978]: "...basic level objects are the most inclusive level of classification at which objects have numbers of attributes in common ..." (note: one would say "basic level concepts" instead of "basic level objects" given today's terminology).

The informal definitions of basic level have been utilized for definitions of (semi)formalized *basic level metrics*, i.e. functions assigning numbers to concepts (categories) in such a way that the larger the number, the stronger it indicates that the concept is a basic level concept. We utilize and formalize within FCA five selected metrics in Section 3.

Clearly, an important benefit of formalization is that it makes the notion of basic level operational. On the one hand, one may form and test reasonably precise hypotheses regarding basic level or compare various views and definitions of basic level and test them against experimental data, benefiting thus the psychological research of basic level. On the other hand—importantly from our standpoint—when linked to a particular formal model involving concepts, such as FCA, the model gets enhanced by a new, potentially powerful tool.

The formalization faces several challenges which result from the many subtleties regarding basic level. Most importantly, like every formal model of psychological phenomena, any formal model of basic level will be simplistic and thus potentially a point of criticism from the psychological standpoint. Furthermore, the formalization has to cope with various issues considered problematic or not yet fully understood from a psychological viewpoint. To give examples, let us recall the fact that basic level concepts of people with less knowledge tend to be more general compared to those of people with extensive domain knowledge [Murphy, 2002; Rosch, 1978; Tanaka and Taylor, 1991]; the problem of which attributes to take into consideration to determine a basic level [Murphy, 2002; Rosch, 1978]; or the role of context in determination of basic level [Murphy, 2002; Rosch, 1978].

3 Formalization of Basic Level Metrics in FCA

3.1 Basic Notions of FCA

We assume familiarity with the basic notions of FCA [Ganter and Wille, 1999]. A *formal context* is denoted by $\langle X, Y, I\rangle$; the binary relation $I \subseteq X \times Y$ tells us whether attribute $y \in Y$ applies to object $x \in X$, i.e. $\langle x, y\rangle \in I$, or not, i.e. $\langle x, y\rangle \notin I$. A *formal concept* of $\langle X, Y, I\rangle$ is any pair $\langle A, B\rangle$ consisting of $A \subseteq X$ and $B \subseteq Y$ satisfying $A^{\uparrow} = B$ and $B^{\downarrow} = A$ where

$$\begin{aligned} A^{\uparrow} &= \{y \in Y \mid \text{for each } x \in X : \langle x, y\rangle \in I\}, \text{ and} \\ B^{\downarrow} &= \{x \in X \mid \text{for each } y \in Y : \langle x, y\rangle \in I\}. \end{aligned}$$

That is, $A^{\uparrow}$ ise the set of all attributes common to all objects from A and $B^{\downarrow}$ is the set of all objects having all the attributes from B, respectively. The set $\mathcal{B}(X, Y, I)$ of all formal concepts of $\langle X, Y, I\rangle$ equipped with the subconcept-superconcept ordering $\leq$ is called the *concept lattice* of $\langle X, Y, I\rangle$. Note that the ordering $\leq$ is defined by $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ iff $A_1 \subseteq A_2$ (equivalently, $B_2 \subseteq B_1$) which models the natural hierarchy of concepts according to which more general concepts have more inclusive extents and less inclusive intents.

3.2 Basic Level Metrics

In this section, we formalize within FCA five selected approaches to basic level. Our common assumption is that we are given a formal context $\langle X, Y, I\rangle$ which describes all the available information regarding the objects and attributes. For a given approach M to basic level, we define a function BL_{M} mapping every concept $\langle A, B\rangle$ in the concept lattice $\mathcal{B}(X, Y, I)$ to $[0, \infty)$ or to $[0, 1]$; $BL_{\mathrm{M}}(A, B)$ is interpreted as the degree to which $\langle A, B\rangle$ 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.

Similarity approach (S)
This approach represents one of Rosch's attempts to operationalize the definition of basic level. It appears in the form of an informal definition in [Rosch, 1978] and, more explicitly, in [Murphy, 2002]. According to this approach, a 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 to each other; 3. the objects of the subordinate concepts are only slightly more similar to each other. This approach was formalized within FCA in [Belohlavek and Trnecka, 2012] to which we refer for details. Due to lack of space, we only note that in this formalization,

$$BL_{\mathrm{S}}(A, B) = \alpha_1 \otimes \alpha_2 \otimes \alpha_3,$$

where α_i represents the degree of validity of the above condition $i = 1, 2, 3$ and $\otimes$ represents a truth function of many-valued conjunction [Gottwald, 2001]. The definition of degrees α_i involves definitions of appropriate similarity measures [Lengnink, 1996; Tversky, 1977] and utilizes basic principles of fuzzy logic.

Cue validity approach (CV)
This approach, proposed in [Rosch *et al.*, 1976], is based on the notion of a cue validity of attribute y for concept c, i.e. the conditional probability $p(c|y)$ that an object belongs to c given that it has y. The *total cue validity for* c is defined the sum of cue validities of each of the attributes of c [Rosch *et al.*, 1976, p. 384]. Basic level concepts are those with a high total cue validity.

Here and below, when speaking of probabilities, we consider the following probability space: X (objects) are the elementary events, 2^X (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 $\{y, \dots\} \subseteq Y$ of attributes is $\{y, \dots\}^{\downarrow}$, i.e. the set of all objects sharing $y, \dots$.

Within FCA, the cue-validity approach thus gives

$$BL_{\mathrm{CV}}(A, B) = \textstyle\sum_{y \in B} P(A|\{y\}^{\downarrow}) = \sum_{y \in B} \frac{|A \cap \{y\}^{\downarrow}|}{|\{y\}^{\downarrow}|}.$$

Remark 1 [Murphy, 2007] (see also [Murphy, 2002, p. 215]) criticizes the cue validity approach on the account that total cue validity is monotone w.r.t. inclusion of categories and, hence, achieves its maximum for the most general category. This claim is wrong. Namely, while it is true that $\langle A_1, B_1\rangle \leq \langle A_2, B_2\rangle$ implies $P(A_1|\{y\}^{\downarrow}) \leq P(A_2|\{y\}^{\downarrow})$, as the author correctly argues, it does not imply that $BL_{\mathrm{CV}}(A_1, B_1) \leq BL_{\mathrm{CV}}(A_2, B_2)$ because the summations run over B_1 and B_2 and we have $B_1 \supseteq B_2$.

Interestingly, one may show that the cue validity approach is congruent to and under certain assumptions agrees with another view of basic level formulated by Rosch, namely the one based on the number of attributes characterizing the categories, which is often dealt with in experiments. Due to space limit, we omit details.

Category feature collocation approach (CFC)
This approach, inspired by [Jones, 1983], utilizes the so-called *collocation of category c and attribute y*, which is the product $P(c|y) \cdot P(y|c)$ of the cue validity $P(c|y)$ and the so-called category validity $P(c|y)$. The total CFC for c may then be defined as the sum of collocations of c and each attribute. Basic level concepts may then again be understood as concepts with a high total CFC. This leads to

$$BL_{\mathrm{CFC}}(A, B) = \textstyle\sum_{y \in Y}(P(c|y) \cdot P(y|c)) =$$
$$= \sum_{y \in Y} \left(\frac{|A \cap \{y\}^{\downarrow}|}{|\{y\}^{\downarrow}|} \cdot \frac{|A \cap \{y\}^{\downarrow}|}{|A|} \right).$$

Category utility approach (CU)
[Corter and Gluck, 1992] introduced yet another approach to basic level. Their approach is supposed to overcome some shortcomings of those based on cue validity and category collocation, cf. also [Zeigenfuse and Lee, 2011]. This approach utilizes the notion of *category utility*, defined by $cu(c) = p(c) \sum_{y \in Y}[p(y|c)^2 - p(y)^2]$. Therefore, in our framework, we obtain that $BL_{\mathrm{CFC}}(A, B)$ equals

$$P(A) \cdot \sum_{y \in Y} \left[\left(\frac{P(\{y\}^{\downarrow} \cap A)}{P(A)} \right)^2 - P(\{y\}^{\downarrow})^2 \right] =$$
$$= \frac{|A|}{|X|} \sum_{y \in Y} \left[\left(\frac{|\{y\}^{\downarrow} \cap A|}{|A|} \right)^2 - \frac{|\{y\}^{\downarrow}|}{|X|})^2 \right].$$

Note that due to lack of space we omit the approach in [Fisher, 1987] (COBWEB) which is based on the CU approach and performs similarly [Gosselin and Schyns, 2001].

Predictability approach (P)
The last approach we present is based on the idea, frequently formulated in the literature [Murphy, 2002], that basic level concepts are abstract concepts that still make it possible to predict well the attributes of their objects (enables good prediction, for short). As with the similarity approach, this approach has not been formalized in the psychological literature. The formalization we propose is based the following idea. We introduce (see below) a graded (fuzzy) predicate *pred* such that $pred(c) \in [0, 1]$ is naturally interpreted as the truth degree of proposition "concept $c = \langle A, B\rangle$ enables good prediction". Then, we use the principles of fuzzy logic to obtain the truth degrees β_1, β_2, and β_3 of the following three propositions, respectively, which are congruent with the considerations regarding basic level [Murphy, 2002]: 1. c has high *pred*; 2. c has a significantly higher *pred* than its upper neighbors; 3. c has only a slightly smaller *pred* than its lower neighbors. This is done in a way analogous to how we approached a similar problem with formalizing the similarity approach to basic level in [Belohlavek and Trnecka, 2012] and we omit details due to lack of space. Finally, we put

$$BL_{\mathrm{P}}(A, B) = \beta_1 \otimes \beta_2 \otimes \beta_3,$$

where $\otimes$ is again an appropriate truth function of many-valued conjunction [Gottwald, 2001], for which we use the product in our experiments.

Therefore, we need to define $pred$. For a given concept $c = \langle A, B\rangle$ and attribute $y \in Y$, consider the random variables $V_y : X \to \{0, 1\}$ and $V_c : X \to \{0, 1\}$ defined by

$V_y(x) = 1$ if $\langle x, y\rangle \in I$ and $V_y(x) = 0$ if $\langle x, y\rangle \notin I$, and
$V_c(x) = 1$ if $x \in A$ and $V_c(x) = 0$ if $x \notin A$.

The fact that the value of y is well predictable for objects in c corresponds to the fact that the conditional entropy [Cover and Thomas, 2006] $E(V_y|V_c = 1)$ is low. One may check that we have

$$E(V_y|V_c = 1) = -\frac{|A - \{y\}^{\downarrow}|}{|A|} \cdot \log \frac{|A - \{y\}^{\downarrow}|}{|A|} - \frac{|\{y\}^{\downarrow} \cap A|}{|A|} \cdot \log \frac{|\{y\}^{\downarrow} \cap A|}{|A|}.$$

Averaging over all the attributes in $Y - B$ (because for $y \in B$ we have $E(V_y|V_c = 1) = 0$), we get an auxiliary quantity

$$p(c) = \sum_{y \in Y - B} \frac{E(V_y|V_c = 1)}{|Y - B|}.$$

Since a low value of p corresponds to a good ability to predict by c, i.e. to a high value of $pred(c)$, and since one may prove that $p(c) \in [0, 1]$, letting

$$pred(c) = 1 - p(c)$$

gives us the desired definition of $pred$.

4 Extracting Interesting Concepts

[Belohlavek and Trnecka, 2012] demonstrated that the basic level concepts corresponding to S, i.e. the formal concepts $\langle A, B\rangle$ with high values of $BL_{\mathrm{S}}(A, B)$, may be regarded as informative, natural concepts. We performed similar experiments with the other metrics proposed in this paper on various datasets, including the Drinks, Sports[1], as well as the well-known DBLP dataset [Miettinen and Vreeken, 2011]. Due to limited scope, we present just a brief summary.

For Drinks (68 drinks as the objects and 25 attributes regarding the composition of drinks), the concept lattice contains 320 formal concepts. The concepts with a high basic-levelness include those which may be described as follows: "beers", "drinks containing magnesium and potassium", "energy drinks containing coffein", "liquers", "milk drinks", "mineral waters", "energy drinks", "sweet vitamin drinks", "wines", and some other natural groups of drinks. Particularly similar were the basic concepts selected by CU and CFC. As a rule, the intents of these concepts contained smaller numbers of attributes compared to the basic level concepts selected by the other metrics. Nevertheless, the concepts selected by CV were considerably similar to those selected by CU and CFC. There was a clear evidence of the fact that P differs from the other metrics and follows a "different logic".

For DBLP (6980 objects—authors of papers in computer science conferences; 19 conferences), the concept lattice contains 2424 concepts. Among the concepts with a high basic-levelness value for most of the metrics were the concept with the intent consisting of SIGMOD and VLDB, which may naturally be described as "authors publishing in top database conferences", the concept with the intent consisting of FOCS and STOCS, naturally described as "authors publishing in top theory conferences", as well as other concepts corresponding to natural, thematically based groupings of conferences.

To sum up, we observed that the basic level concepts selected by the metrics contain natural concepts, informative of the respective domains. Furthermore, we observed that some of the metrics tend to produce the same or intuitively similar basic level concepts. In the next section, we examine this similarity in more detail.

[1] Both are available at http://www.inf.upol.cz/trnecka/datasets.zip

5 Comparison of Basic Level Metrics

In this section, we provide a comparative analysis of the above metrics. Because the metrics represent different quantitative approaches to describe a single phenomenon—the basic level of concepts—the reasons are obvious and include the following ones. Since the metrics are to be used for data analysis purposes such as in Section 4, we wish to describe their relationships, possibly in a quantitative manner. The basic question, important also from the psychological viewpoint, is: are the basic levels determined by these metrics related? Do the metrics differ significantly and thus describe essentially different notions of basic level, or are they similar and thus describe one, "objective" notion of basic level? Examination of these questions, which is grossly missing in the literature (see Section 6) and which is made possible due to our formalization within FCA, may reveal important insight regarding basic level which is significant both from data analysis and psychological viewpoint.

We approach the ambiguous question of whether two given metrics are similar, in that they determine similar basic levels, in two ways. In Sec. 5.1, we ask the (rather strict) question of whether the rankings of concepts according to two given metrics BL_{M} and BL_{N} are similar. In Sec. 5.2, we ask the question (less strict and perhaps more natural) of whether the sets Top_r^{M} and Top_r^{N} consisting of top r concepts according to BL_{M} and BL_{N} are similar. The results reveal some interesting and surprising patterns which are discussed below.

5.1 Similarity of rankings basic level concepts

For every input data $\langle X, Y, I\rangle$, a given metric BL_{M} (i.e. M being S, CV, CFC, CU, or P) determines a ranking (with possible ties) of formal concepts in $\mathcal{B}(X, Y, I)$, i.e. determines the linear quasiorder $\leq_{\mathrm{M}}$ defined by

$$\langle A_1, B_1\rangle \leq_{\mathrm{M}} \langle A_2, B_2\rangle \text{ iff } BL_{\mathrm{M}}(A_1, B_1) \leq BL_{\mathrm{M}}(A_2, B_2).$$

We examined the pairwise similarities of the rankings $\leq_{\mathrm{S}}$, $\leq_{\mathrm{CV}}$, $\leq_{\mathrm{CU}}$, $\leq_{\mathrm{CFC}}$, and $\leq_{\mathrm{P}}$ for various datasets. We used the Kendall tau coefficient [Agresti, 2010; Kendall, 1938] to assess the similarities. Recall that the Kendall tau coefficient $\tau(\leq_{\mathrm{M}}, \leq_{\mathrm{N}})$ of rankings $\leq_{\mathrm{M}}$ and $\leq_{\mathrm{N}}$ is a real number in $[-1, 1]$ based on the numbers of concordant and discordant pairs in the rankings. High values indicate agreement of rankings with 1 in case the rankings coincide; low values indicate disagreement with -1 in case one ranking is the reverse of the other. Even though this approach might seem rather strict, significant patterns were obtained.

We used the Sports and Drinks datasets described in Section 4 as well as collections of synthetic datasets of various sizes with tens to hundreds of objects and tens of attributes. Due to scope limit, we only present the results in Tables 1, 2, 3, and 4, which represent a typical behavior we observed.

	S	CV	CFC	CU	P
S	1.000	-0.093	-0.215	-0.139	-0.175
CV	-0.093	1.000	0.737	0.754	0.170
CFC	-0.215	0.737	1.000	0.789	0.229
CU	-0.139	0.754	0.789	1.000	0.161
P	-0.175	0.170	0.229	0.161	1.000

Table 1: Kendall tau coefficients for Sports data.

	S	CV	CFC	CU	P
S	1.000	0.103	0.151	0.067	-0.036
CV	0.103	1.000	0.555	0.581	-0.232
CFC	0.151	0.555	1.000	0.811	-0.061
CU	0,067	0.581	0.811	1.000	-0.064
P	-0.036	-0.232	-0.061	-0.064	1.000

Table 2: Kendall tau coefficients for Drinks data.

	S	CV	CFC	CU	P
S	1.000	0.051	0.050	0.039	-0.104
CV	0.051	1.000	0.701	0.699	-0.211
CFC	0.050	0.701	1.000	0.791	-0.164
CU	0.039	0.699	0.791	1.000	-0.107
P	-0.104	-0.211	-0.164	-0.107	1.000

Table 3: Kendall tau coefficients for synthetic 75×25 datasets (average values for 100 datasets).

The table entries describe the rank correlation. For example, 0.754 at row CV and column CU in Table 1 means $\tau(\leq_{\mathrm{CV}}, \leq_{\mathrm{CU}}) = 0.754$, indicating a high rank correlation of the lists of concepts sorted according to the CU and CV metrics. As we see from the tables, CV, CFC, and CU tend to be mutually correlated, with CFC and CU being correlated significantly. On the other hand, neither of S and P is correlated with any other metric. Hence, one may conclude that CV, CFC, and CU form a group (with a stronger subgroup consisting of CFC and CU) of metrics that result in considerably similar basic-level rankings of concepts, while S and P represent two different, singleton groups.

5.2 Similarity of sets of top r basic level concepts

Rank correlation may be thought of as representing too strict a criterion. Namely, instead of basic-level ranking, one is arguably more interested in the set consisting of the top r concepts of $\mathcal{B}(X, Y, I)$ according to the ranking $\leq_{\mathrm{M}}$ for a given metric M. We denote such set by

$$Top_r^{\mathrm{M}}$$

with the provision that (a) if the $(r+1)$-st, ..., $(r+k)$-th concepts are tied with the r-th one in the ranking, we add the k concepts to Top_r^{M}; (b) we do not include concepts to which the metric assigns 0. Given metrics M and N, we are interested in whether and to what extent are the sets Top_r^{M} and Top_r^{N} similar. For this purpose, we propose the following measure of similarity.

	S	CV	CFC	CU	P
S	1.000	0.245	0.303	0.350	-0.136
CV	0.245	1.000	0.636	0.631	-0.540
CFC	0.303	0.636	1.000	0.727	-0.645
CU	0.350	0.631	0.727	1.000	-0.458
P	-0.136	-0.540	-0.645	-0.458	1.000

Table 4: Kendall tau coefficients for synthetic 100×50 datasets (average values for 100 datasets).

For formal concepts $\langle C, D\rangle, \langle E, F\rangle \in \mathcal{B}(X, Y, I)$, denote by $s(\langle C, D\rangle, \langle E, F\rangle)$ an appropriately defined degree of similarity, i.e. a number in $[0, 1]$. We present results utilizing the similarity based on the simple matching coefficient [Belohlavek and Trnecka, 2012; Lengnink, 1996] but other options such as the Jaccard coefficient yield similar results. For two metrics M and N, and a given $r = 1, 2, 3, \ldots$, we define

$$S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}}) = \min(I_{\mathrm{MN}}, I_{\mathrm{NM}})$$

where

$$I_{\mathrm{MN}} = \frac{\sum_{\langle C,D\rangle \in Top_r^{\mathrm{M}}} \max_{\langle E,F\rangle \in Top_r^{\mathrm{N}}} s(\langle C, D\rangle, \langle E, F\rangle)}{|\,Top_r^{\mathrm{M}}|}$$

and

$$I_{\mathrm{NM}} = \frac{\sum_{\langle E,F\rangle \in Top_r^{\mathrm{N}}} \max_{\langle C,D\rangle \in Top_r^{\mathrm{M}}} s(\langle C, D\rangle, \langle E, F\rangle)}{|\,Top_r^{\mathrm{N}}|}.$$

Using basic principles of fuzzy logic [Gottwald, 2001], $S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}})$ may naturally be interpreted as the truth degree of the proposition "for most concepts in Top_r^{M} there is a similar concept in Top_r^{N} and vice versa". Because of this interpretation and because S is a reflexive and symmetric fuzzy relation with suitable furhter properties (omitted due to lack of space), S is a good candidate for measuring similarity [Recasens, 2011]. Clearly, high values of S indicate high similarity and $S(Top_r^{\mathrm{M}}, Top_r^{\mathrm{N}}) = 1$ iff $Top_r^{\mathrm{M}} = Top_r^{\mathrm{N}}$.

We inspected the similarity of top r sets of basic level concepts for varying r. This is shown in Figs. 1, 2, and 3, each showing 10 functions representing the similarities of the 10 pairs of sets Top_r^{S}, Top_r^{CV}, Top_r^{CFC}, Top_r^{CU}, and Top_r^{P} of top r concepts selected by the 5 metrics, for varying r.

The figures, which are representative for our experiments with other data as well, indicate the following. First, the similarities tend to increase with increasing r. We regard this monotony a natural property, possibly psychologically relevant, which indicates a kind of mutual consistency of the metrics. Second, the functions tend to be convex, with a faster growth in small values of r, which may be regarded as a certain kind of stability—for relatively small r, the sets of top r concepts may not be similar but their similarity grows fast till a point after which the growth is only small and naturally reflects the property that larger sets are more similar. Third, the degrees of similarity for any two pairs of metrics are reasonably high, indicating that the metrics tend to describe an "objectively existing" basic level. Fourth, one may clearly see a high mutual similarity among CV, CFC, and CU, particularly the similarity of CFC and CU. The mutual similarities of S to any of CV, CFC, and CU are considerably smaller and

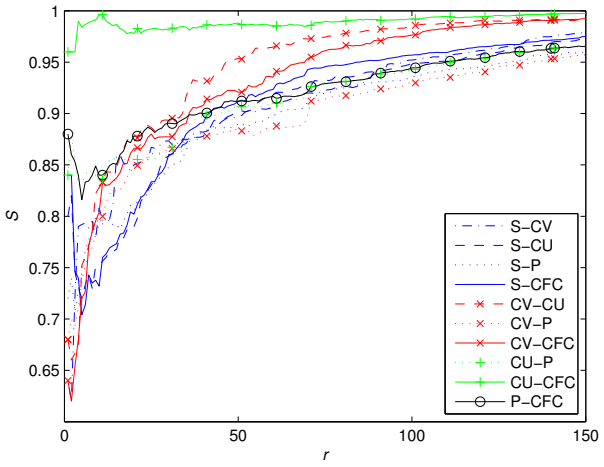

Figure 1: Similarities S of sets of top r concepts for Drinks data.

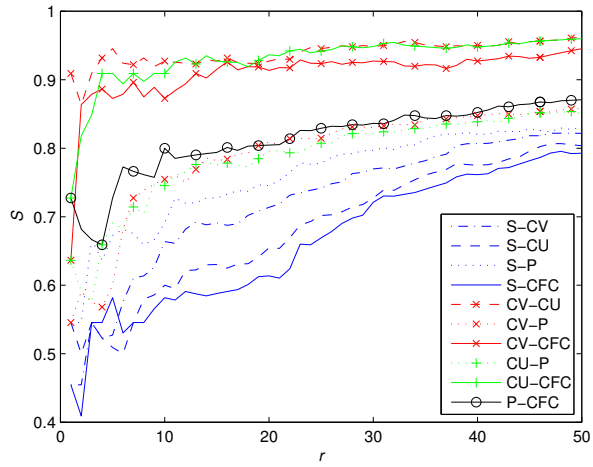

Figure 2: Similarities S of sets of top r concepts for 75×25 datasets (average values for 100 datasets).

follow a similar pattern, reflecting a certain (weak) form of transitivity; the same may be said of P. S and P again seem to represent separate singleton groups. The results thus reveal a similar structure to that observed in Section 5.1.

6 Conclusions

In this paper, we formalized within FCA five approaches to basic level discussed in the psychological literature. We demonstrated that the basic level of concepts corresponding to these approaches tend to contain informative, natural concepts. These concepts may be supplied to a user, preserving important information while reducing significantly the user's overload due to the usually large number of all concepts in the data supplied to the user by ordinary FCA. The experiments performed indicate a mutual consistency of the

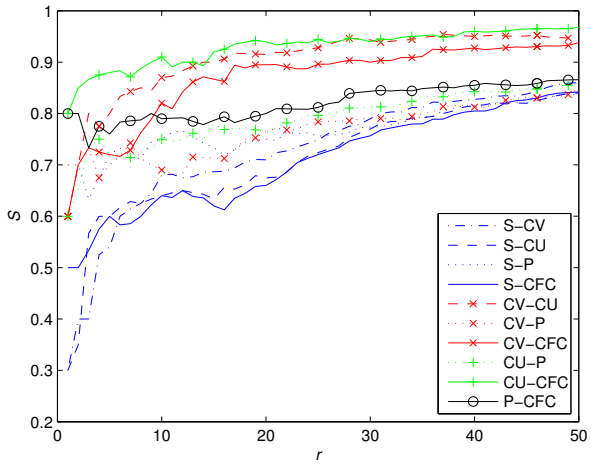

Figure 3: Similarities S of sets of top r concepts for 100×50 datasets (average values for 100 datasets).

proposed basic level metrics but also an interesting pattern. 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 [Murphy, 2002; Rosch, 1978]. 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 [Murphy, 2002].

A future research shall include the following topics: examination from the present viewpoint of the notion of concept stability and other formal concept indices [Klimushkin *et al.*, 2010; Kuznetsov, 2007]; examination of our findings regarding the relationships bewteen the metrics from the viewpoint of the psychology of concepts and reconsideration of some of the above-mentioned psychological views regarding the basic level phenomenon; (psychological) experimental testing of the proposed approaches; comprehensive evaluation of the method selecting basic level concepts as important concepts from data analysis viewpoint; computational considerations regarding the computation of basic level; exploration of related psychological phenomena such as the typicality effects [Murphy, 2002] for data analysis purposes.

References

[Agresti, 2010] A. Agresti. *Analysis of Ordinal Categorical Data, Second Edition*. Wiley, 2010.

[Belohlavek and Trnecka, 2012] R. Belohlavek and M. Trnecka. Basic level of concepts in formal concept analysis. In *Proceedings of ICFCA 2012*, ICFCA'12, pages 28–44, Berlin, Heidelberg, 2012. Springer-Verlag.

[Belohlavek and Vychodil, 2006] R. Belohlavek and V. Vychodil. Formal concept analysis with constraints by closure operators. In *Proceedings of ICCS 2006*, ICCS'06,

pages 131–143, Berlin, Heidelberg, 2006. Springer-Verlag.

[Belohlavek and Vychodil, 2009] R. Belohlavek and V. Vychodil. Formal concept analysis with background knowledge: attribute priorities. *Trans. Sys. Man Cyber Part C*, 39(4):399–409, July 2009.

[Berlin, 1992] B. Berlin. *Ethnobiological classification: principles of categorization of plants and animals in traditional societies*. Princeton University Press, 1992.

[Brown, 1958] R. Brown. How shall a thing be called? *Psychological Review*, 65(1):14–21, January 1958.

[Cellier *et al.*, 2008] P. Cellier, S. Ferré, O Ridoux, and M. Ducassé. A parameterized algorithm to explore formal contexts with a taxonomy. *Int. J. Found. Comput. Sci.*, 19(2):319–343, 2008.

[Corter and Gluck, 1992] J. E. Corter and M. A. Gluck. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2):291–303, 1992.

[Cover and Thomas, 2006] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2006.

[Dias and Vieira, 2010] S. M. Dias and N. Vieira. Reducing the size of concept lattices: The jbos approach. In M. Kryszkiewicz and S. A. Obiedkov, editors, *CLA*, volume 672 of *CEUR Workshop*, pages 80–91. CEUR, 2010.

[Fisher, 1987] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, September 1987.

[Ganter and Wille, 1999] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.

[Gosselin and Schyns, 2001] F. Gosselin and P. G. Schyns. Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological review*, 108(4):735–758, October 2001.

[Gottwald, 2001] S. Gottwald. *A Treatise on Many-Valued Logics*. Studies in Logic and Computation. Research Studies Press, 2001.

[Jones, 1983] G. V. Jones. Identifying basic categories. *Psychological Bulletin*, 94:423–428, 1983.

[Kendall, 1938] M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30:81–89, June 1938.

[Klimushkin *et al.*, 2010] M. Klimushkin, S. Obiedkov, and C. Roth. Approaches to the selection of relevant concepts in the case of noisy data. In *Proceedings of ICFCA 2010*, ICFCA'10, pages 255–266, Berlin, Heidelberg, 2010. Springer-Verlag.

[Kuznetsov, 2007] S. O. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):101–115, April 2007.

[Kwuida *et al.*, 2010] L. Kwuida, R. Missaoui, B. B. Amor, L. Boumedjout, and J. Vaillancourt. Restrictions on concept lattices for pattern management. In M. Kryszkiewicz and S. A. Obiedkov, editors, *CLA*, volume 672 of *CEUR Workshop Proceedings*, pages 235–246. CEUR WS, 2010.

[Lengnink, 1996] K. Lengnink. *Formalisierungen von Ähnlichkeit aus Sicht der Formalen Begriffsanalyse*. PhD thesis, TH Darmstadt, 1996. Verlag Shaker, Aachen, 1996.

[Miettinen and Vreeken, 2011] P. Miettinen and J. Vreeken. Model order selection for boolean matrix factorization. In *Proceedings of the 17th ACM SIGKDD*, KDD '11, pages 51–59. ACM, 2011.

[Murphy, 2002] G. L. Murphy. *The Big Book of Concepts*. The MIT Press, ambridge, MA, 2002.

[Murphy, 2007] G. L. Murphy. Cue validity and levels of categorization. *Psychological Bulletin*, 91:174–177, 2007.

[Recasens, 2011] J. Recasens. *Indistinguishability Operators: Modelling Fuzzy Equalities and Fuzzy Equivalence Relations*. Studies in Fuzziness and Soft Computing. Springer, 2011.

[Rosch *et al.*, 1976] E. Rosch, B. C. Mervis, W. Gray, and D. Johson. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[Rosch, 1978] E. Rosch. Principles of categorization. In E. Rosch and B. B. Lloyd, editors, *Cognition and Categorization*, pages 27–48, Hillsdale (NJ), USA, 1978. Lawrence Erlbaum Associates.

[Tan *et al.*, 2006] P-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Boston, MA, USA, 2006.

[Tanaka and Taylor, 1991] J. W. Tanaka and M. Taylor. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, July 1991.

[Tversky, 1977] A. Tversky. Features of Similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.

[Zeigenfuse and Lee, 2011] M. D. Zeigenfuse and M. D. Lee. A comparison of three measures of the association between a feature and a concept. In L. Carlson, C. Holscher, and T. F. Shipley, editors, *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pages 243–248, Austin, TX, 2011. Cognitive Science Society.