Basic level of concepts and formal concept analysis 2: examination of existing basic level metrics

Radim Belohlavek & Martin Trnecka

To cite this article: Radim Belohlavek & Martin Trnecka (2020) Basic level of concepts and formal concept analysis 2: examination of existing basic level metrics, International Journal of General Systems, 49:7, 707-723, DOI: 10.1080/03081079.2020.1837795

To link to this article: https://doi.org/10.1080/03081079.2020.1837795

Published online: 29 Oct 2020.

Submit your article to this journal

Article views: 27

View related articles

View Crossmark data

Citing articles: 1

View citing articles
Basic level of concepts and formal concept analysis 2: examination of existing basic level metrics

Radim Belohlavek and Martin Trnecka

Dept. Computer Science, Palacký University, Olomouc, Czech Republic

ABSTRACT
We exploit the framework developed in the first part of our paper to examine significant approaches to the basic level of concepts proposed in psychological literature. These approaches are mostly presented informally in the literature and we hence provide their formalization. Our formalization and experiments reveal previously unknown relationships between existing approaches to basic level, some of which contradict current psychological beliefs. We argue that from a general perspective, formalization of basic level and related phenomena in the framework of formal concept analysis is beneficial for psychological explorations themselves because it helps put them on a solid ground.

1. Basic level definitions and metrics

Due to subtle nature of concepts, it is not surprising that psychological literature does not contain a single and uniquely interpretable definition of basic level. Even the programmatic study by Rosch et al. (1976) contains several informal, verbal descriptions of basic level. This reflects a variety of existing views regarding the basic level of concepts.

To illustrate this fact, let us present three descriptions by the most influential author on this topic, Eleanor Rosch. For one, Rosch et al. (1976, p. 383) say:

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.

At a different place in the same paper, the authors say (Rosch et al. 1976, 384):

. . . the basic categorization is the most general and inclusive level at which categories can delineate real-world correlational structures.

Finally, Rosch (1978) claims:

“. . . basic level objects are the most inclusive level of classification at which objects have numbers of attributes in common . . .” [given today’s terminology one would say “basic level concepts” instead of “basic level objects”].
Various informal definitions of basic level have been exploited to define various formal and semiformal basic level metrics or indices. Such metrics are functions that assign numbers to concepts in such a way that the larger the number, the stronger it indicates that the concept is a basic level concept. That is, the number $BL(c)$ assigned by the metric $BL$ to the concept $c$ may be interpreted as a truth degree of the proposition “$c$ is a basic level concept”.

A clear benefit of such formalization consists in making the notion of basic level operational. A metric makes it possible to form and test reasonably precise hypotheses regarding basic level. In particular, it makes possible to compare various views and definitions of basic level and also test them against experimental data, which shall be explored in our paper. We contend that formalizing basic level via basic level metrics is thus a methodologically significant step for the psychological research on basic level. In addition, as demonstrated in the first part of our study (Belohlavek and Trnecka 2020), a basic level metric linked with a suitable formal method involving concepts may enhance the method itself.

In the remainder of this paper, we formalize within the framework of formal concept analysis (FCA) selected metrics of basic level and examine their mutual relationships. Our main goal is to study correlations of the various metrics. Doing so, we attempt to explore the question:

Do the various metrics agree in denoting concepts as basic level concepts?

Note at this point that formalization of basic level faces several challenges. For one, like every formal model of psychological phenomena, a formalization of basic level is going to appear simplistic from a psychological standpoint. Furthermore, the formalization has to cope with various issues considered problematic or not yet fully understood from a psychological viewpoint. Examples of these issues include the different choice of basic level concepts by people with extensive domain knowledge compared to people with less knowledge (Rosch 1978; Tanaka and Taylor 1991; Murphy 2002), a question of which attributes to take into consideration to determine basic level (Rosch 1978; Murphy 2002), or the general role of context in determination of basic level (Rosch 1978; Murphy 2002).

2. Formalization of basic level metrics within formal concept analysis

We use FCA as a framework for our formalization; see the first part of our study (Belohlavek and Trnecka 2020) and Ganter and Wille (1999) for preliminaries. We assume that we are given a formal context $⟨X, Y, I⟩$ describing all the available information regarding the objects and attributes. For a given approach $M$ to basic level, we denote the corresponding basic level metric $BL_M$, i.e. $BL_M$ assigns to every formal concept $⟨A, B⟩$ in the concept lattice $B(X, Y, I)$ a degree $BL_M(A, B)$ to which $⟨A, B⟩$ belongs to the basic level.$^1$

2.1. Similarity approach ($S$) – the rudimentary view of Rosch

This is the approach explored in the first part of our paper, i.e. the formalization of the rudimentary view of basic level by Eleanor Rosch. The corresponding metric shall be denoted by $BL_S$. That is, $BL_S(A, B)$ may be regarded as the truth degree of the conjunction of the
following three propositions:

1. the objects of the formal concept \( \langle A, B \rangle \) are similar to each other;
2. the objects of the concepts subordinate to \( \langle A, B \rangle \) are only slightly more similar to each other.

For details we refer to Belohlavek and Trnecka (2020).

### 2.2. Cue validity approach (CV)

The cue validity approach was proposed by Rosch et al. (1976). In this approach, a cue validity of attribute \( y \) for concept \( c \) is defined as the conditional probability \( P(c|y) \) – the probability that an object \( x \) belongs to a concept \( c \), i.e. \( x \in A \), given that the object \( x \) has the attribute \( y \). The total cue validity for \( c \) is defined as the sum of cue validities of each of the attributes of \( c \) Rosch et al. (1976, 384). Basic level concepts are those with a high total cue validity.

Clearly, a key question in formalizing this approach is how to define the probabilities involved. For this purpose we consider the following probability space. The set \( X \), i.e. the set of objects of the given formal context, represents the elementary events. The set \( 2^X \) is the \( \sigma \)-algebra of events, i.e. the subsets \( A \) of \( X \) are the events of the probability space (the \( A \)s are not necessarily the extents of formal concepts). The probability distribution \( P \) is given by

\[
P(\{x\}) = \frac{1}{|X|}
\]

for every object \( x \in X \). Consequently, the probability \( P(A) \) of an event \( A \) is

\[
P(A) = \frac{|A|}{|X|}.
\]

The event corresponding to a set \( B \subseteq Y \) of attributes (\( B \) is not necessarily an intent of a concept) is \( B^\uparrow \), i.e. the set \( \{x \in X \mid \forall y \in B : x \mbox{has} y\} \) of all objects sharing every \( y \in B \). For a formal concept \( c = \langle A, B \rangle \), the informal probability \( P(c|y) \) is then represented by the probability \( P(A|\{y\}^\downarrow) \), which is

\[
P(A|\{y\}^\downarrow) = \frac{|A \cap \{y\}^\downarrow|}{|\{y\}^\downarrow|}.
\]

(1)

Within this framework, the cue validity approach yields the following definition of the degree \( BL_{CV}(A, B) \) for the concept \( c = \langle A, B \rangle \):

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

**Remark 1:** A criticism of the cue validity approach was formulated by Murphy (1982); see also Murphy (2002, 215). This criticism is unwarranted. Namely, Murphy claims that cue validity is monotone with respect to inclusion of categories and, hence, achieves its
maximum for the most general category (i.e. the most general concept). This claim is wrong. Namely, while it is true that \( \langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle \) implies \( P(A_1 | \{ y \}^\perp) \leq P(A_2 | \{ y \}^\perp) \), as the author correctly argues, it does not imply that \( BL_{CV}(A_1, B_1) \leq BL_{CV}(A_2, B_2) \) because the summations run over \( B_1 \) and \( B_2 \) and we have \( B_1 \supseteq B_2 \). A concrete counterexample is easy to obtain.

Note that Murphy (1982, 176) claims:

\[
\ldots \text{By itself, this principle causes category cue validity to increase with category inclusiveness. This effect is compounded by the fact that a more inclusive category will possess more attributes \ldots than its subordinates do. For example, the features possessed by animal include all of the features possessed by bird as well as some that bird does not have: \ldots When category cue validity is calculated by adding up the validities of each cue, this factor also gives animal a higher category cue validity, although bird is the basic category.}
\]

The mistake Murphy makes consists in the assumption that if \( c_1 \) is a subcategory of \( c_2 \), then \( c_2 \) contains all attributes possessed by \( c_1 \). However, this is not true according to a common understanding in which an attribute possessed by a category is an attribute possessed by all objects in the given category. Hence, in this case, the opposite is true: \( c_1 \) contains all the attributes possessed by \( c_2 \). In terms of FCA, if \( c_1 = \langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle = c_2 \), then \( B_1 \supseteq B_2 \).

2.3. Category feature collocation approach (CFC)

The category feature collocation approach is inspired by Jones (1983). It is based on the notion of 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 \( P(y | c) \). The probabilities involved are defined in the probability space defined in Section 2.2. That is, for a formal concept \( c = \langle A, B \rangle \), the informal probabilities \( P(c | y) \) and \( P(y | c) \) are represented by \( P(A | \{ y \}^\perp) \) and \( P(\{ y \}^\perp | A) \), respectively, where \( P(A | \{ y \}^\perp) \) is defined by (1) and

\[
P(\{ y \}^\perp | A) = \frac{|A \cap \{ y \}^\perp|}{|A|}.
\]

The total CFC for \( c = \langle A, B \rangle \) may then be defined as the sum of collocations of \( c \) and each attribute \( y \in Y \). Basic level concepts may then again be understood as concepts with a high total CFC. This leads to

\[
BL_{CFC}(A, B) = \sum_{y \in Y} (P(c | y) \cdot P(y | c)) = \sum_{y \in Y} \left( \frac{|A \cap \{ y \}^\perp|}{|\{ y \}^\perp|} \cdot \frac{|A \cap \{ y \}^\perp|}{|A|} \right).
\]

2.4. Category feature possession (CFP)

This measure is proposed by Jones (1983) as a measure utilizing the notion of category feature collocation. We describe it in terms of FCA. Suppose \( Y = \{ y_1, \ldots, y_m \} \), and suppose that all the categories, i.e. formal concepts considered, are \( c_1 = \langle A_1, B_1 \rangle, \ldots, c_n = \langle A_n, B_n \rangle \). For each attribute \( y_j \) we select the concepts \( c_i \) for which their collocation is maximal by assigning them the score \( S_{ij} = 1 \), and assigning the score \( S_{ij} = 0 \) to the other concepts. We hence define the \( n \times m \) matrix \( S \)

\[
S_{ij} = \begin{cases} 
1 & \text{if } (P(c_i | y_j) \cdot P(y_j | c_i)) = \max_{c_k \in \mathcal{B}(X,Y,I)} (P(c_k | y_j) \cdot P(y_j | c_k)), \\
0 & \text{otherwise}.
\end{cases}
\]
Again, the underlying probability space and the probabilities involved are given as above.

Now, the category feature possession (CFP) for $c_i \in \mathcal{B}(X, Y, I)$, which we denote by $BL_{CFP}(c_i)$, is defined by

$$BL_{CFP}(c_i) = \sum_{j=1}^{m} S_{ij}. $$

The value $BL_{CFP}(c_i)$ is then interpreted as the degree to which $c_i$ is a basic level concept.$^2$

### 2.5. Category utility approach (CU)

Corter and Gluck (1992) introduced the category utility approach, which represents yet another approach to a basic level metric. Their approach is supposed to overcome some shortcomings of those based on cue validity and category collocation. For motivations and further considerations we refer to Corter and Gluck (1992) and Zeigenfuse and Lee (2011). The approach equates the degree to which a concept $c$ belongs to basic level with the so-called category utility of $c$, which is informally defined by

$$cu(c) = p(c) \sum_{y \in Y} [p(y|c)^2 - p(y)^2].$$

Given the probability space as in the previous sections, we therefore obtain the following definition of the degree $BL_{CU}(A, B)$ to which the formal concept $c = \langle A, B \rangle$ is a basic level concept:

$$BL_{CU}(A, B) = P(A) \cdot \sum_{y \in Y} \left[ \left( \frac{P(Y|A)}{P(A)} \right)^2 - P(Y) \right].$$

Note in this context a related approach (Fisher 1987) which is based on the CU approach and performs similarly (Gosselin and Schyns 2001).

### 2.6. Predictability approach (P)

What we call a predictability approach is an approach based on the idea, which is frequently formulated in the literature on the basic level (Murphy 2002, 218ff.), that basic level concepts are highly informative in that they enable good prediction. By this it is meant that they are abstract concepts that still make it possible to predict well the attributes of their objects. Like the similarity approach, this approach has not been formalized in the psychological literature so far.

Our formalization proceeds as follows. We consider a given concept $c$ a basic level concept if it satisfies the following three conditions, which are congruent with the
considerations regarding basic level (Murphy 2002):

1. \( c \) enables good prediction.
2. \( c \) enables prediction much better than its upper neighbors.
3. \( c \) enables prediction only slightly worse than its lower neighbors.

To formalize these conditions, we introduce a graded (fuzzy) predicate \( \text{pred} \) such that \( \text{pred}(c) \in [0,1] \) is naturally interpreted as the truth degree of the proposition "concept \( c = \langle A, B \rangle \) enables good prediction". The above three conditions may then be put in the following form:

1. \( c \) has high \( \text{pred} \);
2. \( c \) has a significantly higher \( \text{pred} \) than its upper neighbors;
3. \( c \) has only a slightly smaller \( \text{pred} \) than its lower neighbors.

We then utilize the principles of fuzzy logic to obtain the truth degrees \( \beta_1, \beta_2, \) and \( \beta_3 \) of these propositions, respectively, in a manner analogous to how we proceeded in the definitions of the truth degrees \( \alpha_1, \alpha_2, \) and \( \alpha_3 \) when formalizing Rosch's approach to basic level (Belohlavek and Trnecka 2020). Finally, we put

\[
BL_P(A, B) = \beta_1 \otimes \beta_2 \otimes \beta_3,
\]

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

We therefore need to define the graded predicate \( \text{pred} \). Consider again the probability space induced by the given formal context \( \langle X, Y, I \rangle \). 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

\[
V_y(x) = \begin{cases} 
1 & \text{if } \langle x, y \rangle \in I \\
0 & \text{if } \langle x, y \rangle \notin I 
\end{cases}
\]

and

\[
V_c(x) = \begin{cases} 
1 & \text{if } x \in A \\
0 & \text{if } x \notin A 
\end{cases}
\]

As is well known (Cover and Thomas 2006), the fact that the value of \( y \) is well predictable for objects in \( c \) corresponds to the fact that the conditional entropy \( E(V_y|V_c = 1) \) is low. A straightforward computation leads to

\[
E(V_y|V_c = 1) = -\sum_{a \in \{0,1\}} P(V_y = a|V_c = 1) \cdot \log P(V_y = a|V_c = 1) =
\]

\[
= -P(V_y = 0|V_c = 1) \cdot \log P(V_y = 0|V_c = 1) +
\]

\[
- P(V_y = 1|V_c = 1) \cdot \log P(V_y = 1|V_c = 1) =
\]

\[
= -\frac{P(V_y = 0 \cap V_c = 1)}{P(V_c = 1)} \cdot \log \frac{P(V_y = 0 \cap V_c = 1)}{P(V_c = 1)} +
\]
\[ P(V_Y = 1 \cap V_c = 1) \cdot \log \frac{P(V_Y = 1 \cap V_c = 1)}{P(V_c = 1)} = \]
\[ = - \frac{|A - \{y\}^\downarrow|}{|A|} \cdot \log \frac{|A - \{y\}^\downarrow|}{|A|} - \frac{|\{y\}^\downarrow \cap A|}{|A|} \cdot \log \frac{|\{y\}^\downarrow \cap A|}{|A|}. \]

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 \( \text{pred}(c) \), and since one may prove that \( p(c) \in [0, 1] \), letting

\[ \text{pred}(c) = 1 - p(c) \]

gives us the desired definition of \( \text{pred} \).

3. Extracting interesting concepts

In the first part (Belohlavek and Trnecka 2020), we demonstrated that for \( BL_S \), the basic level concepts, i.e. the formal concepts \( \langle A, B \rangle \) attaining a high degree \( BL_S(A, B) \), may be regarded as informative, natural, and thus interesting concepts. We performed similar experiments with the other metrics formalized in this paper. Before we proceed to compare these metrics, we therefore describe results of some of these experiments.

In our experiments, we used the following datasets: Drinks, Sports,\(^3\) Zoo, Means of transportation,\(^4\) and the well-known DBLP dataset Miettinen and Vreeken (2011).

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 degree of membership to a basic level include those which may be described as follows: beers, drinks containing magnesium and potassium, energy drinks containing caffeine, liqueurs, milk drinks, mineral waters, energy drinks, sweet vitamin drinks, wines, and some other natural groups of drinks. 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 degree of basic level 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. The Sports, Zoo, and Means of transportation data are analyzed using Rosch’s similarity approach in the first part (Belohlavek and Trnecka 2020).

One may conclude that, by and large, all the metrics tend to select similar concepts as basic level concepts, but there are noticeable differences between the metrics. Particularly similar were the basic level 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 and CFP differ from the other metrics and follow a “different logic”.

To sum up, we observed that the basic level concepts selected by the metrics generally 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.

4. Comparison of basic level metrics

We now provide a comparative analysis of the above metrics. As the metrics represent different quantitative approaches to describe the same phenomenon – the basic level of concepts – the reasons for comparing them are obvious. The basic question, important also from the psychological viewpoint, is: Are the basic levels determined by these metrics related? That is, do the metrics differ significantly and thus describe different notions of basic level, or are they similar and thus essentially describe a single, “objective” notion of basic level? Examination of these questions, which is grossly missing in the literature 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 somewhat ambiguous question of whether two given metrics are similar, in that they determine similar basic levels, in two ways, presented in Section 4.2 and in Section 4.3. Prior to that, however, we consider in Section 4.1 the possible particular metrics formalizing Rosch’s view, which were proposed in the first part of our study (Belohlavek and Trnecka 2020), and select one of them, namely $BL_{SMC}$, for further comparison with the other metrics, due to its properties. In Section 4.2, 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 Section 4.3, we ask the question (less strict and perhaps more natural) of whether the sets $Top^M_r$ and $Top^N_r$ 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.

4.1. Metrics based on Rosch’s view

Of the eight metrics considered by Belohlavek and Trnecka (2020), we select $BL_{SMC}$ for further comparison. One reason, discussed in the first part (Belohlavek and Trnecka 2020) is that the choice of average in the respective formulas seems intuitively most plausible given Rosch’s view. An additional reason is that when observing the rank correlation using the Kendall tau, which is recalled and used in Section 4.2, it appears that $BL_{SMC}$ may be considered as a representative of one of the groups of the metrics, namely the group consisting of $BL_{SMC}$, $BL_{SMC}^m$, $BL_{SMC}^m$, $BL_{SMC}^m$, $BL_{SMC}^m$, and $BL_{SMC}^m$. The second group consists of $BL_{SMC}^m$ and $BL_{SMC}^m$, which are both intuitively less plausible due to the usage of min instead of $\cup$ in the aggregation of similarities. This is illustrated in Table 1 for the Sports data.
Table 1. Kendall tau rank correlation for sports data.

<table>
<thead>
<tr>
<th></th>
<th>$BL^\varnothing_{SMC}$</th>
<th>$BL^{m\varnothing}_{SMC}$</th>
<th>$BL^m_{SMC}$</th>
<th>$BL^\varnothing_j$</th>
<th>$BL^m_j$</th>
<th>$tBL^m_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$BL^\varnothing_{SMC}$</td>
<td>1.000</td>
<td>0.856</td>
<td>0.483</td>
<td>0.533</td>
<td>0.707</td>
<td>0.731</td>
</tr>
<tr>
<td>$BL^{m\varnothing}_{SMC}$</td>
<td>0.856</td>
<td>1.000</td>
<td>0.389</td>
<td>0.471</td>
<td>0.629</td>
<td>0.726</td>
</tr>
<tr>
<td>$BL^m_{SMC}$</td>
<td>0.483</td>
<td>0.389</td>
<td>1.000</td>
<td>0.783</td>
<td>0.551</td>
<td>0.429</td>
</tr>
<tr>
<td>$BL^\varnothing_j$</td>
<td>0.533</td>
<td>0.471</td>
<td>0.783</td>
<td>1.000</td>
<td>0.549</td>
<td>0.457</td>
</tr>
<tr>
<td>$BL^m_j$</td>
<td>0.707</td>
<td>0.629</td>
<td>0.551</td>
<td>1.000</td>
<td>0.823</td>
<td>0.238</td>
</tr>
<tr>
<td>$BL^{m\varnothing}_j$</td>
<td>0.731</td>
<td>0.726</td>
<td>0.429</td>
<td>0.823</td>
<td>1.000</td>
<td>0.105</td>
</tr>
<tr>
<td>$BL^m_{j}$</td>
<td>0.108</td>
<td>0.035</td>
<td>0.572</td>
<td>0.238</td>
<td>0.105</td>
<td>1.000</td>
</tr>
<tr>
<td>$BL^m_j$</td>
<td>0.139</td>
<td>0.069</td>
<td>0.587</td>
<td>0.279</td>
<td>0.166</td>
<td>0.900</td>
</tr>
</tbody>
</table>

4.2. Similarity of basic level metrics

For every input data $(X, Y, I)$, a given metric $BL_M$ (i.e. $M$ being $S$, $CV$, $CFC$, $CFP$, $CU$, and $P$) determines a ranking (with possible ties) of formal concepts in $B(X, Y, I)$, i.e. determines the linear quasiorder $\leq_M$, i.e. a reflexive and transitive relation, defined by

$$\langle A_1, B_1 \rangle \leq_M \langle A_2, B_2 \rangle \iff BL_M(A_1, B_1) \leq BL_M(A_2, B_2).$$

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

We used the datasets described in Section 3. In addition, we employed collections of synthetic datasets of various sizes with tens to hundreds of objects and tens of attributes. The results are depicted in Tables 2−5 for the real data and Tables 6 and 7 for synthetic data.

The table entries describe the rank correlation coefficients. For example, the value 0.754 at row $CV$ and column $CU$ in Table 2 means that the Kendall tau $\tau(\leq_{CV}, \leq_{CU}) = 0.754$, 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 $S$, $CFP$ and

Table 2. Kendall tau rank correlation for Sports data.

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>0.035</td>
<td>0.022</td>
<td>0.170</td>
<td>0.061</td>
<td>0.057</td>
</tr>
<tr>
<td>CV</td>
<td>0.035</td>
<td>1.000</td>
<td>0.637</td>
<td>0.227</td>
<td>0.754</td>
<td>0.170</td>
</tr>
<tr>
<td>CFC</td>
<td>0.022</td>
<td>0.637</td>
<td>1.000</td>
<td>0.302</td>
<td>0.690</td>
<td>0.208</td>
</tr>
<tr>
<td>CFP</td>
<td>0.170</td>
<td>0.227</td>
<td>0.302</td>
<td>1.000</td>
<td>0.231</td>
<td>0.211</td>
</tr>
<tr>
<td>CU</td>
<td>0.061</td>
<td>0.754</td>
<td>0.690</td>
<td>0.231</td>
<td>1.000</td>
<td>0.161</td>
</tr>
<tr>
<td>P</td>
<td>0.057</td>
<td>0.170</td>
<td>0.208</td>
<td>0.211</td>
<td>0.161</td>
<td>1.000</td>
</tr>
</tbody>
</table>
Table 3. Kendall tau rank correlation for Drinks data.

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>-0.181</td>
<td>-0.094</td>
<td>0.055</td>
<td>-0.150</td>
<td>0.055</td>
</tr>
<tr>
<td>CV</td>
<td>-0.181</td>
<td>1.000</td>
<td>0.543</td>
<td>0.020</td>
<td>0.581</td>
<td>-0.232</td>
</tr>
<tr>
<td>CFC</td>
<td>-0.094</td>
<td>0.543</td>
<td>1.000</td>
<td>0.066</td>
<td>0.798</td>
<td>-0.069</td>
</tr>
<tr>
<td>CFP</td>
<td>0.055</td>
<td>0.020</td>
<td>0.066</td>
<td>1.000</td>
<td>0.005</td>
<td>-0.021</td>
</tr>
<tr>
<td>CU</td>
<td>-0.150</td>
<td>0.581</td>
<td>0.798</td>
<td>1.000</td>
<td>1.000</td>
<td>-0.064</td>
</tr>
<tr>
<td>P</td>
<td>0.055</td>
<td>-0.232</td>
<td>-0.069</td>
<td>-0.021</td>
<td>-0.064</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4. Kendall tau rank correlation for Zoo data.

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>-0.156</td>
<td>-0.204</td>
<td>0.013</td>
<td>-0.171</td>
<td>-0.064</td>
</tr>
<tr>
<td>CV</td>
<td>-0.156</td>
<td>1.000</td>
<td>0.722</td>
<td>0.020</td>
<td>0.691</td>
<td>-0.122</td>
</tr>
<tr>
<td>CFC</td>
<td>-0.204</td>
<td>0.722</td>
<td>1.000</td>
<td>0.115</td>
<td>0.665</td>
<td>0.163</td>
</tr>
<tr>
<td>CFP</td>
<td>0.013</td>
<td>0.065</td>
<td>0.115</td>
<td>1.000</td>
<td>0.038</td>
<td>0.023</td>
</tr>
<tr>
<td>CU</td>
<td>-0.171</td>
<td>0.691</td>
<td>0.665</td>
<td>0.038</td>
<td>1.000</td>
<td>0.090</td>
</tr>
<tr>
<td>P</td>
<td>-0.064</td>
<td>0.122</td>
<td>0.163</td>
<td>0.023</td>
<td>0.090</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 5. Kendall tau rank correlation for Means of transportation data.

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>-0.025</td>
<td>-0.045</td>
<td>-0.005</td>
<td>-0.073</td>
<td>-0.022</td>
</tr>
<tr>
<td>CV</td>
<td>-0.025</td>
<td>1.000</td>
<td>0.791</td>
<td>0.020</td>
<td>0.691</td>
<td>-0.013</td>
</tr>
<tr>
<td>CFC</td>
<td>-0.045</td>
<td>0.791</td>
<td>1.000</td>
<td>0.022</td>
<td>0.801</td>
<td>-0.013</td>
</tr>
<tr>
<td>CFP</td>
<td>-0.005</td>
<td>0.020</td>
<td>0.022</td>
<td>1.000</td>
<td>0.016</td>
<td>0.003</td>
</tr>
<tr>
<td>CU</td>
<td>-0.073</td>
<td>0.691</td>
<td>0.665</td>
<td>0.038</td>
<td>1.000</td>
<td>-0.041</td>
</tr>
<tr>
<td>P</td>
<td>-0.022</td>
<td>-0.013</td>
<td>-0.013</td>
<td>0.003</td>
<td>-0.041</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 6. Kendall tau rank correlation for synthetic 75 × 25 datasets (average values for 100 datasets).

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>-0.172</td>
<td>-0.196</td>
<td>-0.025</td>
<td>-0.208</td>
<td>-0.121</td>
</tr>
<tr>
<td>CV</td>
<td>-0.172</td>
<td>1.000</td>
<td>0.677</td>
<td>0.064</td>
<td>0.643</td>
<td>0.123</td>
</tr>
<tr>
<td>CFC</td>
<td>-0.196</td>
<td>0.677</td>
<td>1.000</td>
<td>0.102</td>
<td>0.728</td>
<td>0.103</td>
</tr>
<tr>
<td>CFP</td>
<td>-0.025</td>
<td>0.064</td>
<td>0.102</td>
<td>1.000</td>
<td>0.050</td>
<td>-0.026</td>
</tr>
<tr>
<td>CU</td>
<td>-0.208</td>
<td>0.643</td>
<td>0.728</td>
<td>0.050</td>
<td>1.000</td>
<td>0.228</td>
</tr>
<tr>
<td>P</td>
<td>-0.121</td>
<td>0.123</td>
<td>0.103</td>
<td>-0.026</td>
<td>0.228</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 7. Kendall tau rank correlation for synthetic 100 × 50 datasets (average values for 100 datasets).

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>CV</th>
<th>CFC</th>
<th>CFP</th>
<th>CU</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>1.000</td>
<td>-0.120</td>
<td>-0.094</td>
<td>-0.006</td>
<td>-0.087</td>
<td>-0.009</td>
</tr>
<tr>
<td>CV</td>
<td>-0.120</td>
<td>1.000</td>
<td>0.791</td>
<td>0.020</td>
<td>0.691</td>
<td>-0.013</td>
</tr>
<tr>
<td>CFC</td>
<td>-0.094</td>
<td>0.791</td>
<td>1.000</td>
<td>0.022</td>
<td>0.801</td>
<td>-0.013</td>
</tr>
<tr>
<td>CFP</td>
<td>-0.006</td>
<td>0.020</td>
<td>0.222</td>
<td>1.000</td>
<td>0.016</td>
<td>0.003</td>
</tr>
<tr>
<td>CU</td>
<td>-0.087</td>
<td>0.691</td>
<td>0.801</td>
<td>0.016</td>
<td>1.000</td>
<td>-0.041</td>
</tr>
<tr>
<td>P</td>
<td>-0.009</td>
<td>-0.013</td>
<td>-0.013</td>
<td>0.003</td>
<td>-0.041</td>
<td>1.000</td>
</tr>
</tbody>
</table>

P is significantly 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, CFP, and P represent three different, singleton groups.
4.3. Similarity of sets of top r basic level concepts

Rank correlation may be considered too strict to be taken as a measure of similarity of two basic level metrics. Namely, instead of basic-level ranking, one may arguably be more interested in the set consisting of the top r concepts of \( B(X, Y, I) \) according to the ranking \( \leq_M \) for a given metric M. We denote such set by

\[
\text{Top}_r^M
\]

with the provision that

(a) if the \((r + 1)\)-st, \ldots, \((r + k)\)th concepts are tied with the rth one in the ranking, we add the k concepts to \( \text{Top}_r^M \);
(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 \( \text{Top}_r^M \) and \( \text{Top}_r^N \) similar. We propose the following measure of such similarity.

For formal concepts \( (C, D), (E, F) \in B(X, Y, I) \), denote by \( s((C, D), (E, F)) \) 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; see Section 3.3 of Belohlavek and Trnecka (2020). That is, for formal concepts \( (A_1, B_1) \) and \( (A_2, B_2) \), we define the degree of their similarity by

\[
s((A_1, B_1), (A_2, B_2)) = \text{sim}_{\text{SMC}}(B_1, B_2),
\]

where

\[
\text{sim}_{\text{SMC}}(B_1, B_2) = \frac{|B_1 \cap B_2| + |Y - (B_1 \cup B_2)|}{|Y|}.
\]

Nevertheless, 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

\[
S(\text{Top}_r^M, \text{Top}_r^N) = \min(I_{MN}, I_{NM})
\]

where

\[
I_{MN} = \frac{\sum_{(C,D)\in\text{Top}_r^M} \max_{(E,F)\in\text{Top}_r^N} s((C, D), (E, F))}{|\text{Top}_r^M|}
\]

and

\[
I_{NM} = \frac{\sum_{(E,F)\in\text{Top}_r^N} \max_{(C,D)\in\text{Top}_r^M} s((C, D), (E, F))}{|\text{Top}_r^N|}.
\]

Using basic principles of fuzzy logic (Gottwald 2001), \( S(\text{Top}_r^M, \text{Top}_r^N) \) may naturally be interpreted as the truth degree of the proposition “for most concepts in \( \text{Top}_r^M \) there is a similar concept in \( \text{Top}_r^N \) and vice versa”. Because of this interpretation and because \( S \) is a reflexive and symmetric fuzzy relation with suitable further properties, \( S \) is a good candidate for measuring similarity (Recasens 2011). Clearly, high values of \( S \) indicate high similarity and \( S(\text{Top}_r^M, \text{Top}_r^N) = 1 \) iff \( \text{Top}_r^M = \text{Top}_r^N \).
We inspected the similarity of top $r$ sets of basic level concepts for varying $r$. This is shown in Figures 1–6, each showing 15 functions representing the similarities of the 15 pairs of sets $\text{Top}_r^S$, $\text{Top}_r^\text{CV}$, $\text{Top}_r^{\text{CFC}}$, $\text{Top}_r^{\text{CFP}}$, $\text{Top}_r^{\text{CU}}$, and $\text{Top}_r^P$ and of top $r$ concepts selected by the 6 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 $r$ with an exception of similarities involving CFP. 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 $r$. This 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 until 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 follow a similar pattern, reflecting a certain (weak) form of transitivity; the same may be said of CFP and P. S, P and CFP again seem to represent separate singleton groups. The results thus reveal a similar structure to that observed for similarity of rankings in Section 4.2.

**Figure 1.** Similarities $S$ of sets of top $r$ concepts for Sports data.
**Figure 2.** Similarities $S$ of sets of top $r$ concepts for Drinks data.

**Figure 3.** Similarities $S$ of sets of top $r$ concepts for Transport facilities data.
Figure 4. Similarities $S$ of sets of top $r$ concepts for Zoo data.

Figure 5. Similarities $S$ of sets of top $r$ concepts for $75 \times 25$ datasets (average values for 100 datasets).
5. Conclusions

In the second part of our study, we formalized within FCA five additional, significant approaches to basic level discussed in the psychological literature. We demonstrated that – as with Rosch’s rudimentary view formalized in the first part of our study – the basic level of concepts corresponding to these approaches tends to contain informative, natural concepts. The experiments performed indicate a mutual consistency of the 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, CFP, 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 (Rosch 1978; Murphy 2002). 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 of our findings regarding the relationships between 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, and examination from the present viewpoint of the notion of concept stability and other formal concept indices (Kuznetsov 2007; Klimushkin, Obiedkov, and Roth 2010; Mouakher, Ktyfi, and Ben...
Yahia (2019); computational considerations regarding the computation of basic level; exploration of related psychological phenomena such as the typicality effects (Murphy 2002) for data analysis purposes; for first considerations see Belohlavek and Mikula (2020).

Notes

1. The degree $BLM(A, B)$ in the approaches presented below is not always in the unit interval $[0, 1]$. It may be transformed to $[0, 1]$ using an appropriate scaling. Nevertheless, such transformation is not needed for our considerations.
2. Clearly, $BLCFP(c)$ may exceed 1; the normalized quantity $BLCFP(c_i)/|Y|$ takes values in $[0, 1]$.
4. Both Zoo and Means of transportation are described in Belohlavek and Trnecka (2020).
5. We used a standard implementation of Kendall tau in MATLAB.

Acknowledgments

This two-part paper is an extension of the authors’ conference papers “Basic level of concepts in formal concept analysis” presented at ICFCA 2012 and “Basic level in formal concept analysis: interesting concepts and psychological ramifications” presented at IJCAI 2013. Support by grant no. IGA_PrF_2020_019 of Palacky University is gratefully acknowledged. M. Trnecka acknowledges support by grant no. JG_2020_003 of Palacky University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by IGA of Palacky University [IGA_PrF_2020_019] and Junior research grant of Palacký University [JG_2020_003].

Notes on contributors

Radim Belohlavek received PhD degree in computer science from the Technical University of Ostrava, Czech Republic, in 1998, PhD degree in mathematics from Palacky University, Olomouc, Czech Republic, in 2001, and DSc degree in informatics and cybernetics from the Academy of Sciences of the Czech Republic in 2008. He is professor of computer science at Palacky University. Dr Belohlavek’s academic interests are in discrete mathematics, logic, uncertainty and information, and data analysis. He published three books and over 150 papers in conference proceedings and journals. Dr Belohlavek is a Senior Member of IEEE (Institute of Electrical and Electronics Engineers), and a Member of ACM (Association for Computing Machinery) and AMS (American Mathematical Society), and is a member of editorial boards of several international journals.

Martin Trnecka is assistant professor at the Department of Computer Science, Faculty of Science, Palacky University Olomouc in Czech Republic. He obtained his PhD in Computer Science in 2017 from Palacky University Olomouc. His research interests are in data analysis, particularly in Boolean matrix factorization and formal concept analysis. He published several papers in journals and conferences, including Journal of Computer and System Sciences, Discrete Applied Mathematics, Annals of Mathematics and Artificial Intelligence, IJCAI and CIKM.
References