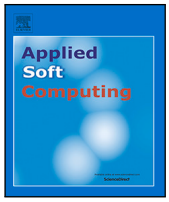
Applied Soft Computing Journal 153 (2024) 111270

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc



Similarity metrics vs human judgment of similarity for binary data: Which is best to predict typicality?

Radim Belohlavek, Tomas Mikula *

Department of Computer Science, Palacký University Olomouc, Czech Republic

ARTICLE INFO

Keywords:
Similarity
Similarity measure
Typicality
Binary data

ABSTRACT

Similarity measures for binary data have been subject to a number of comparative studies. In contrast to these studies, we provide a comparison of similarity measures with human judgment of similarity. For this purpose, we utilize the phenomenon of typicality, whose definition is based on similarity. We observe how well the similarity of objects – either computed by a similarity measure or provided by human judgment – enables the prediction of typicality of these objects in various human categories. In doing so, we examine a large variety of existing similarity measures, and utilize recently available extensive data involving binary data as well as data on human judgment of similarity and typicality.

1. Problem description

Measuring similarity of binary data plays a crucial role in many tasks and has been subject to extensive research. Since the first formulas to measure similarity appeared more than a hundred years ago, a multitude of similarity measures, as well as the dual dissimilarity measures, have been proposed in various areas. Given the number of existing similarity measures for binary data, exploration of the proposed similarity measures has naturally become the subject of a number of studies.

The existing works study various properties of the proposed similarity measures, mutual relationships of the measures, and examine the performance of particular similarity measures; see, e.g., [1–7] for some influential as well as recent studies.[1] The comparative studies usually involve tens of similarity measures (see Section 2.1 for details) and the comparison is typically based on evaluating these measures on data from a particular domain of interest, such as biology or chemistry, and also on randomly generated data.

The primary purpose of our paper is different, namely to compare the large variety of the existing similarity measures using extensive psychological data. Such exploration has not been done before and constitutes the main novelty of our contribution. In particular, we compare the available similarity measures on the one hand with a human judgment of similarity on the other hand. We explore this question indirectly via the important phenomenon of typicality, which – according to a common psychological view – is based on the concept of similarity [8]. For this purpose, we utilize our recent results on typicality and its prediction [9]. In particular, we consider the capability of pairwise similarity ratings – those computed by similarity measures and those provided by human judgment – to predict typicality. In addition to relating similarity measures with a human judgment of similarity, our comparison also provides a view on the relationship between the involved similarity measures themselves. Our study is possible due to the now available high-quality psychological data regarding human categories and related phenomena [10], which involves binary data and data on human judgment of similarity and typicality.

In Section 2.1, we present preliminaries on similarity measures; a list of formulas for all the similarity measures involved in our study along with additional information is supplied in the appendix. The phenomenon of typicality and the formula for computing typicality are the subjects of Section 2.2. In Section 3, we describe the data we use in the present study. Our experimental evaluation is the content of Section 4. Section 5 concludes the paper with observations drawn from the experiments.

2. Similarity and typicality

2.1. Similarity measures

For the purpose of our paper, we follow a general understanding according to which a similarity measure on a set X of objects is a binary

* Corresponding author.
E-mail addresses: radim.belohlavek@acm.org (R. Belohlavek), mail@tomasmikula.cz (T. Mikula).

[1] As works on similarity measures for binary data are rather numerous, we only include selected papers, directly related to our purpose, and refer to these papers for further references.

https://doi.org/10.1016/j.asoc.2024.111270
Received 4 August 2023; Received in revised form 5 January 2024; Accepted 13 January 2024
Available online 17 January 2024

function

$$sim : X \times X \to \mathbb{R};$$

the value $sim(x, y)$ is interpreted as the extent to which x is similar to y. This general approach subsumes a variety of particular similarity measures proposed in the literature. That is, we do not impose possible additional constraints, such as $sim(x, y) = sim(y, x)$, $sim(x, y) \leq sim(x, x)$, or various dual forms of the triangle inequality, which are sometimes considered.

When the similarity of binary data is considered, the set X consists of all possible objects described by n binary attributes, and may hence be conveniently identified with the set $\{0, 1\}^n$ of all n-dimensional binary vectors. Thus, for instance,

$$x = \langle 1, 0, 0, 1, 1 \rangle$$

represents an object described by 5 binary attributes, i.e., $x \in \{0, 1\}^5$, and one has $x_1 = 1$, $x_2 = 0$, $x_3 = 0$, $x_4 = 1$, and $x_5 = 1$. That is, the object has the first, the fourth, and the fifth attribute, but not the second, nor the third.

The similarity measures considered in the literature can conveniently be defined in terms of the values a, b, c, and d, defined as follows. Consider n attributes and two binary vectors $x, y \in \{0, 1\}^n$, and let

$$\begin{aligned} a &= \#\{i \mid x_i = 1 \text{ and } y_i = 1\}, \\ b &= \#\{i \mid x_i = 1 \text{ and } y_i = 0\}, \\ c &= \#\{i \mid x_i = 0 \text{ and } y_i = 1\}, \\ d &= \#\{i \mid x_i = 0 \text{ and } y_i = 0\}. \end{aligned}$$

That is, a is the number of common presences and d is the number of common absences of the attributes $i = 1, \ldots, n$. On the other hand, b is the number of attributes present on x but not on y, and c is the number of attributes absent on x but present on y. While a and d indicate similarity of x and y, b and d indicate dissimilarity. Clearly, $a + b + c + d = n$.

For example, for the vectors

$$\begin{aligned} x &= \langle 0, 1, 1, 0, 0, 1, 0, 1, 1, 0 \rangle, \\ y &= \langle 1, 1, 1, 0, 0, 1, 0, 1, 0, 1 \rangle \end{aligned}$$

in $\{0, 1\}^{10}$, one has

$$\begin{array}{ll} a = 4 & b = 1 \\ c = 2 & d = 3 \end{array}.$$

Now, a similarity measure may be defined by a formula involving the coefficients a, b, c, and d, corresponding to $x, y \in \{0, 1\}^n$, such as

$$sim(x, y) = \frac{a + d}{a + b + c + d} \quad \text{and} \quad sim(x, y) = \frac{a}{a + b + c}. \tag{1}$$

The formulas in (1) actually represent two well-known similarity measures, the simple matching coefficient (SMC) and the Jaccard measure (Jac), respectively.

Measures of similarity for binary data have a long history; see, e.g., [1,2,4]. The first measures were proposed at the end of the 19th century to facilitate the analysis of biological species, which were often described in terms of binary attributes. Since then, numerous other measures have been proposed in areas as diverse as biology, ecology, geology, psychology, chemistry, medicine, information retrieval, machine learning, and bioinformatics. A principal reason for the continuing interest in these measures is the omnipresence of data describing various kinds of items, such as biological species, chemical compounds, performance tests, or documents, in terms of binary attributes, and the need to analyze such data.

Even though no definite categorization or grouping of similarity measures for binary data has been established in the literature, a few classification criteria have been considered. The following two seem best known. The first one attempts to classify the measures into statistically based and co-occurrence based. The statistically based, also called association measures, are often interpretable as correlation coefficients and have usually their values in the interval $[-1, 1]$. Their formulas may seem less intuitive and often contain $ad - bc$ in the numerator. The co-occurrence-based measures are based on the frequencies a and d of co-occurrence of the involved binary attributes, have their values in $[0, 1]$, and are usually defined by intuitive formulas such as (1), which contain a or $a + d$ in the numerator. The second widely used criterion consists in whether the measure takes into account, ignores, or takes into account partially the number d of common absences (negative matches) of the attributes. For instance, while the above SMC measure takes d into account in that it increases similarity, the Jaccard measure ignores d.

In our study, we employ 69 similarity measures, which we selected from a large variety of similarity measures described in the literature, particularly in [1,2,4–6]. The employed measures, along with comments on the logic of our selection and further information about these measures are described in the appendix of this paper. In particular, a list of the employed measures is provided by Table 4, which contains an abbreviation and a name for each measure as well as a formula for computing the values of a given measure. The list is sorted alphabetically by the abbreviations so that a reader may quickly find details about the measures when assessing our experimental results.

2.2. Typicality

The phenomenon of typicality is well known from everyday life: Intuitively, a sparrow is a typical bird, an ostrich is not. Typicality is one of the most important phenomena accompanying human concepts and plays a significant role in a variety of cognitive tasks including categorization and classification. Since being typical is a matter of degree, typicality manifests a graded structure of concepts. Both typicality and the graded structure of concepts have been among the central topics of research in the psychology of concepts since the 1970s. For a comprehensive exposition of typicality and its role in the psychology of concepts, we refer to [8].

According to a mainstream psychological view, which goes back to the seminal work by Eleanor Rosch and her colleagues [11–13], the notion of typicality of an object in a concept (category) is based on the notion of similarity: An object is considered typical in a given concept if the object is similar to the objects to which the concept applies. In [9], we formalized this view of Rosch as follows[2]:

Definition 1. Given a similarity $sim : X \times X \to \mathbb{R}$, an object $x \in X$, and a nonempty set $A \subseteq X$ representing a concept, a *degree of typicality of x in A* is defined by

$$typ(x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}. \tag{2}$$

Formula (2) for typicality results as a straightforward formalization of a verbal description of the psychological view available in the literature and represents the average similarity of the object x to all the objects in A. As demonstrated in [9], the degrees of typicality computed by this formula are highly correlated with human judgment of typicality, i.e., with degrees of typicality provided by humans.

Example 1. Table 1 presents a part of the Zoo data [14], restricted to 9 exemplars of the category "bird" (sparrow, ..., penguin) and some of their binary attributes (feathers, ..., legs 2). The column labeled

[2] In fact, in [9] we used formula (2) to define typicality of x in A, for A being an extent of a so-called formal concept. The definition in the present paper simply gets rid of the constraint and allows A to be a general subset of X, i.e., allows A to represent an arbitrary category.

Table 1
Values of typicality of exemplars of the category "bird" from example 1.

	feathers	eggs	airborne	aquatic	predator	backbone	breathes	tail	domestic	catsize	legs 2	*typ*(J)
sparrow	1	1	1	0	0	1	1	1	0	0	1	0.809
crow	1	1	1	0	1	1	1	1	0	0	1	0.807
vulture	1	1	1	0	1	1	1	1	0	1	1	0.802
duck	1	1	1	1	0	1	1	1	0	0	1	0.784
swan	1	1	1	1	0	1	1	1	0	1	1	0.783
kiwi	1	1	0	0	1	1	1	1	0	0	1	0.763
ostrich	1	1	0	0	0	1	1	1	0	1	1	0.759
chicken	1	1	1	0	0	1	1	1	1	0	1	0.745
penguin	1	1	0	1	1	1	1	1	0	1	1	0.745

$typ(\mathrm{J})$ provides the values of typicality of the exemplars, i.e., the values $typ(x, A)$ computed according to (2), based on the Jaccard similarity (cf. Section 2.1). Note that in (2), x denotes the exemplar whose typicality is being computed, A represents the 9-element set of exemplars of "bird," and $sim(x, x_1)$ denotes the Jaccard similarity of exemplars x and x_1 calculated from the binary descriptions of the two exemplars provided by the corresponding table rows.

The ordering of the exemplars in the table by the values of typicality corresponds to intuition despite the limited number of attributes used in our illustrative example; see [9] for a more comprehensive study of typicality in the context of the Zoo data. Note also that the relatively low dispersion of typicality values results from the limited number of the exemplars and attributes involved in this illustrative example.

3. Data

The availability of high-quality data is essential for any kind of experiment that aims to be psychologically relevant. For our purpose, the Dutch data [10] is unique in this regard, as it provides perhaps the most comprehensive data regarding common human categories and their numerous characteristics, including similarity and typicality. Moreover, the data is considerably larger than the previously available psychological data of similar nature. In this section, we provide a brief description of the data, particularly the parts we use, and our comments regarding usability in experiments along with our technical modifications in this regard.

The Dutch data has been gathered by psychologists at the University of Leuven in a thorough, carefully designed study involving hundreds of human respondents. It basically provides information regarding common language concepts (categories), binary attributes (features) relevant to these categories, objects (exemplars) in these categories, and various psychologically relevant characteristics.

In particular, the data involves 16 linguistic categories. These include both the so-called natural kind and artifact categories, as these two kinds are commonly believed to have distinct properties. Each category is represented by a number of objects (exemplars), such as a robin for the category "bird." There are 10 natural kind categories: "fruit" (30 exemplars); "vegetables" (30); "professions" (30); "sports" (30); the animal categories "amphibians"(5),[3] "birds" (30), "fish" (23), "insects" (26), "mammals" (30), and "reptiles" (22).[4] In addition, there are 6 artifact categories: "clothing" (29), "kitchen utensils" (33), "musical instruments" (27), "tools" (30), "vehicles" (30), and "weapons" (20).[5]

These categories comprise 249 exemplars for the natural kind and 166 exemplars for the artifact categories, which were obtained from humans and are representative of the respective categories.[6] Coverage by these categories is considerable; for instance, the animal categories cover a rather large part of the known animal domain. The objects (exemplars) and attributes (features) were obtained by processes described in [10]. In particular, the attributes were generated by 1003 respondents in two ways: First, respondents were asked to list relevant attributes for a given category (these are called category attributes). Second, they were asked to list relevant attributes for each object involved in the data (these are called exemplar attributes). Furthermore, unions of all exemplar features listed for all the objects in a given category were considered, as well as the union of all exemplar features of all the objects in the animal domain, and an analogous union of exemplar features for the artifact domain.

An essential part of the data are the so-called exemplar-by-feature applicability matrices. These are various matrices in which the rows and columns correspond to some of the objects and attributes, respectively, and the entries contain information about whether a particular object has or does not have a particular attribute. Each of the matrices was filled separately by four respondents. The data also contains the corresponding aggregated matrices, in which the values, viz. 0, 1, 2, 3, and 4, indicate the number of respondents who agreed on that the respective object has the respective attribute. To obtain binary matrices (and thus data with binary attributes) from these aggregated matrices, one naturally thresholds the matrix entries. We present our experiments for a threshold equal to 2. Hence, our binary matrices contain 1 in the entry corresponding to the object x and the attribute y if at least two respondents agreed that x has y.

In particular, we use the binary matrices described in Tables 2 and 3. For instance, the first row in Table 2 refers to two binary matrices: The first one, a 30×28 matrix, describes which of the 30 exemplars of the category "bird" has which of the 28 category attributes of this category (i.e., attributes listed as category attributes for this category by respondents); the second one, a 30×225 matrix, describes which of the 30 exemplars of the category "bird" has which of the 225 exemplar attributes for this category (i.e., all attributes listed as exemplar attributes for some exemplar of "bird"). Similarly, the 129×225 binary matrix referred to by the first row in Table 3 describes which of the 129 objects in the animal domain have which of the corresponding 225 category attributes; the 129 objects are all the objects of the categories "amphibian", "bird", "fish", "insect", "mammal", and "reptile", and the 225 category attributes are all attributes listed as category attributes for these six categories. Likewise, the 129×764 matrix describes which of the objects in the animal domain have which of the corresponding 764 exemplar attributes, i.e., all the attributes listed as exemplar attributes for some of the 129 exemplars in the animal domain.

Typicality ratings, which are present in the Dutch data, were obtained from 112 respondents. For each of the 16 categories and each object in the respective category, the data contains a typicality rating on the scale 1 (very atypical) to 20 (very typical).

The pairwise similarity ratings of the Dutch data come partly from the previous study [15], in which the ratings were obtained for ten of the present categories from 42 participants. The ratings for the other categories were obtained from 92 respondents in [10], who also

[3] Since the category "amphibians" only contains 5 exemplars, and since these exemplars are included in the category "reptiles," we omit it in most of our considerations below; see [10] for reasons to include the exemplars of "amphibians" in "reptiles."

[4] The exemplar-by-feature applicability matrices, which we describe below and use in our experiments, contain only 20 exemplars of the category "reptiles," because the respondents who were to fill in these matrices turned out to not to be familiar with two exemplars (komodo and iguanodon). We hence exclude these two exemplars from our experiments.

[5] Here, we use plural in category names, as the authors do [10]; below, we use singular, i.e., "bird" rather than "birds" to be consistent with our previous writings.

[6] In addition to the 5 amphibians included in reptiles and two omitted exemplars of reptiles (see above), note that three exemplars of artifact categories are included in two distinct categories.

Table 2
Category-based binary matrices used in our experiments.

Category	Objects	Category attributes	Exemplar attributes
bird	30	28	225
clothing	29	38	258
fruit	30	32	233
fish	23	32	156
insect	26	37	214
kitchen utensil	33	39	328
mammal	30	34	288
musical instrument	27	39	218
profession	30	21	370
reptile	20	35	179
sport	30	33	382
tool	30	37	285
vegetable	30	30	291
vehicle	30	34	322
weapon	20	32	181

Table 3
Domain-based binary matrices used in our experiments.

Domain	Objects	Category attributes	Exemplar attributes
animal	129	225	764
artifact	166	301	1,295

provided additional ratings for the ten categories involved in [15] to improve reliability. For every category – except for "amphibians," whose five exemplars are included in "reptiles" – and each pair of objects, the data contains a similarity rating on the scale 1 (totally dissimilar) and 20 (totally similar).

Since the original data contains some minor semantic and technical faults, as well as inconveniences as regards a possible machine processing of the data, we modified the data as follows. For one, since the original data contains some wrongly formatted comma-separated files, we transformed them into a valid format. In addition, the names of some objects and attributes are spelled differently across multiple files in the original data; we therefore unified these names. We also converted all names to lowercase to unify them. No changes were made to the data itself. The result is easily machine-processable data. The corrected version of Dutch data, along with a convenient Python wrapper, is publicly available on GitHub [16].

4. Experiments

4.1. Rationale

Comparing similarities via the ability to predict typicality. The rationale of our experiments may be described as follows. Formula (2) for computing degrees of typicality involves degrees $sim(x, y)$ of similarity. Hence, for a given similarity function sim, the function typ may be regarded as a function $typ(sim)$ parameterized by sim, which assigns to each x in a given universe X of objects and a non-empty subset A of X the degree

$$[typ(sim)](x, A) = \frac{\sum_{x_1 \in A} sim(x, x_1)}{|A|}$$

to which the object x is typical for the concept (category) represented by A.

As explained in Section 3, the Dutch data contains information regarding the objects (exemplars) of a variety of categories, including descriptions of these objects by binary attributes. The descriptions of objects by binary attributes enable one to compute the values $sim(x, y)$ of similarity measures sim for pairs of objects x and y. Consequently, one may compute, for any given category A, the degrees $[typ(sim)](x, A)$ of typicality determined by each particular similarity measure sim. In addition, since the Dutch data also contains information on human judgment of similarity, i.e., contains similarity degrees $\mathrm{HJ}(x, y)$ obtained from humans for pairs of the involved objects x and y, one may also compute the degrees $[typ(\mathrm{HJ})](x, A)$ of typicality determined by human judgment of similarity HJ. From this perspective, different similarities shall generally lead to different predictions of typicality.

Now, since the Dutch data also contains degrees of typicality assessed by humans for the involved categories, one may explore, for a given category A and for each similarity measure sim, a correlation of the computed typicality degrees $[typ(sim)](x, A)$ for the objects x in A on the one hand, and the degrees of typicality obtained for the category A from humans on the other hand. The same kind of correlation may be explored for the typicality degrees $[typ(\mathrm{HJ})](x, A)$ computed using human similarity in place of $[typ(sim)](x, A)$. High correlation implies that the particular similarity (represented by a similarity measure or by human judgment) is capable of predicting well the human judgment of typicality.

One may then explore various questions; most importantly:

- How do the various similarity measures compare in their ability to predict typicality?
- How do the similarity measures compare to a human similarity in the same regard, i.e., in their ability to predict typicality?

It is basically these questions that we examine using the experiments presented below. Note that while various comparisons of selected similarity measures are available in the literature (cf. Section 2.1), comparing similarity measures with human judgment of similarity has never been explored in the literature.

Assessment of correlation. The design of our experiments implies a need to assess correlation in the following scenario. For a given category A and a given similarity function sim (either a similarity measure or a similarity obtained from human judgment), we need to assess the correlation between a typicality rating of objects (exemplars) x of the given category, computed by the above formula for $[typ(sim)](x, A)$, and a typicality rating given by a human judgment. To assess correlation of these two typicality ratings, we use the well-known Kendall tau rank-order correlation coefficient.

Recall that the Kendall-tau coefficient measures agreement between two linear orderings (rank orderings), $<_1$ and $<_2$, on a given set of objects. Its basic version is defined by

$$\frac{\#\text{ concordant pairs} - \#\text{ discordant pairs}}{\#\text{ all pairs}};$$

here, a pair of objects x and y is concordant if $x <_1 y$ and $x <_2 y$, or $x >_1 y$ and $x >_2 y$, and is discordant if $x <_1 y$ and $x >_2 y$, or $x >_1 y$ and $x <_2 y$.

In our scenario, the first ordering of the objects, $<_1$, is determined by the computed typicality $typ(sim)$, while the second one, $<_2$, is given by the human rating of typicality, and the Kendall tau is applied to these orderings. In this sense, Kendall tau measures the extent to which the typicality rating determined by the chosen similarity sim agrees with the typicality rating given by human judgment.

Note also that we chose the τ_b variant of the Kendall coefficient since it properly accounts for ties, i.e., situations in which the same degree of typicality is assigned to two or more objects. The coefficient τ_b ranges from 1 (same ordering) to −1 (inverse, i.e., opposite ordering). We used the implementation of τ_b in a Python library [17].

4.2. Results

Our first set of experiments involves the category-based matrices described in Table 2. As described in Section 3, each of these 30 matrices corresponds to a single category and one of the two kinds of attributes (category and exemplar). For each such matrix and each considered similarity sim (i.e., each considered similarity measure of Table 4 and the human similarity obtained from the Dutch data), we

computed the degrees $[typ(sim)](x, A)$ of typicality for all objects x of the respective category (i.e., for all matrix rows).[7] We then computed the Kendal τ_b correlation coefficient of the computed degrees of typicality and the human-assessed typicality degrees for the given category. The results for all the natural kind categories and their category attributes are displayed in Fig. 1. Fig. 2 shows analogous results for the exemplar attributes. The results for all the artifact categories and their category and exemplar attributes are shown in Figs. 3 and 4, respectively.

In this and the other graphs, we use the abbreviations introduced in the appendix (Table 4) to denote the respective similarity measures. Thus, for instance, typ(Di2) denotes the typicality computed by means of the Di2 (Dice 2) similarity measure. In the same spirit, typ(HJ) denotes the typicality computed by means of the human judgment of similarity. The $typ(sim)$ on the horizontal axis are ordered by the mean value of the correlation coefficients across the involved categories.

The second set of experiments involves the four domain-based matrices of Table 3. We performed analogous computations as in the first set of experiments. First, for each category in the animal domain, we computed the degrees of typicality using all the category attributes of the domain matrix for each object of the category. Then a Kendall τ_b coefficient of the computed typicality degrees and the human-assessed degrees of typicality was computed for each particular category. The results are displayed in Fig. 5. The results of the same computation with all the exemplar attributes of the animal domain replacing the category attributes are shown in Fig. 6. Analogous results for the artifact domain and its categories are presented in Figs. 7 and 8. Notice that the categories "fruit", "profession", "sport", and "vegetable" are not included in the second set of experiments because these categories are not part of the two domains.

To provide a summarized view of the results, we also include Figs. 9, 10, and 11, which display the average correlation coefficients over all the categories in the animal domain, the artifact domain, and in both of these domains, respectively. In each graph, the mean correlation coefficients are presented for the four sets of attributes: the category-based category attributes, the category-based exemplar attributes, the domain-based category attributes, and the domain-based exemplar attributes; cf. Tables 2 and 3.

4.3. Discussion

Both the detailed graphs (Figs. 1–8) and the averaged summary views (Figs. 9–11) reveal notable patterns as regards the ability to predict human judgment of typicality by various similarity functions, as well as regards a comparison of the explored similarity measures and human judgments of similarity. Note first that according to a commonly accepted interpretation, the values of τ_b of rank-order correlation may be interpreted as follows: $\tau_b \geq 0.3$, $0.2 \leq \tau_b < 0.3$, $0.1 \leq \tau_b < 0.2$, and $0.0 \leq \tau_b < 0.1$ indicate strong, moderate, weak, and very weak correlation, respectively; the negative values of τ_b are interpreted analogously.

Consider first the human similarity HJ. Overall, HJ enables rather good predictions of typicality and is among the best similarities in this regard. Not only ranks the human similarity as the sixth best as regards average of correlations across all the categories and all the sets of attributes (Fig. 11) with a rather strong $\tau_b = 0.42$, but performs best as regards prediction of typicality in the animal domain (Fig. 9).

The slightly worse performance of human similarity on the artifact categories and also on the three natural categories outside the animal domain may, in our view, be due to the fact that a judgment of similarity of exemplars of these categories is somewhat problematic (consider, e.g.: What is the similarity degree of sailing and sport fishing, of being an accountant and a postman, sled and bicycle?) and the calculated similarity may hence yield better predictions of typicality.[8]

As regards the performance of all the involved similarities, the averaged summary graph (Fig. 11) indicates that there is a group of similarities with an overall strong correlation of human judgment of typicality. Naturally, this group does not have a sharp boundary, but among its core members are, except for the human similarity HJ discussed above, the similarity measures Co1, RR, int, Di2, and CT3, which all have higher average correlation compared to HJ across all categories and across the artifact domain (Fig. 10). In addition, there is a group of other highly correlated similarity measures, which include Fai, FM, CT4, Fos, Ku2, McC, Sor, SS1, cos, Jac, Maa, and Gle.

Observe that some of the similarity measures display a high average correlation except for predictions in the category-based data with category attributes. We contend that the latter drop in correlation is mainly due to the fact that the category attributes of the smaller, category-based matrices provide less information about the exemplars—a significant phenomenon to which we turn below.

One can also identify a group of similarity measures with a low average correlation and with values around 0, and varying considerably in prediction of typicality over the domain-based and the artifact-based data and the two respective kinds of attributes. These include Den, Co2, Col, Di1, Twd, Fo1, and Gow. From this point of view, Gow seems particularly peculiar as its correlation attains significant negative values in several cases but not in others, which is apparent in all figures except Figs. 1 and 3.

Worth noting is also the good prediction of typicality by Co1 and Di2, and the poor performance of their symmetric counterparts, Co2 and Di1. In both cases, good prediction results when the value of c (see Section 2.1) increases the value of the denominator in the respective similarity formula; hence, if the value $sim(x, x_1)$ involved in formula (2) for typicality gets smaller when x does not have an attribute possessed by x_1 but does not get smaller when x has an attribute not possessed by x_1.

Note at this point that as may be observed in the graphs, certain groups of similarity measures displayed a perfect correlation τ_b in that the correlation coefficient with a human judgment of typicality is the same for all data we explored. This pertains to the pairs BU1 and BU2, Gle and Maa, Ku2 and McC, RG and Sco, and to the triplet Ham, ip, and SMC. In all these cases the respective pairs of similarity measures yield different values, i.e., are distinct functions. Their formulas are, nevertheless, closely related.

Another conclusion which may be drawn from the experiments pertains the quality of attributes. It is well known in the psychology

[7] The similarity measures with undefined values are not included; see Remark 1 in the appendix.

[8] See Section 5 for more details. Human similarity HJ was assessed by the respondents with no context, in that each respondent was asked to judge the similarity for a number of exemplar pairs selected across various categories. We hypothesize that such assessment yields different, likely smaller and less consistent, degrees of similarity compared to an alternative scenario, in which a category name and a list of all exemplars of the category are given, and the respondent is to assess similarity of all exemplar pairs in this category. The name and the list of all objects of the category provide a context for the assessment. In the presence of this context, the assessed similarity degree of, e.g., sled and bicycle, is likely to be higher compared to when no context is present (the context helps one realize, so to say, the similarity because relevant attributes become more apparent in the presence of the context). When assessing typicality, respondents implicitly utilize their context-based judgment of similarity (because then, the category name and the lists of exemplars are available). Now, we hypothesize that the similarity computed using a reasonably good similarity measure sim is likely to be better correlated with the context-based human similarity rather than with the without-context similarity HJ. Hence, the correlation of the human typicality rating with typ(sim) is likely to be higher than the correlation with typ(HJ). This hypothesis would hence explain the slightly worse correlation of the computed typicality based on human similarity compared to computed typicality based on a reasonably good similarity measure.

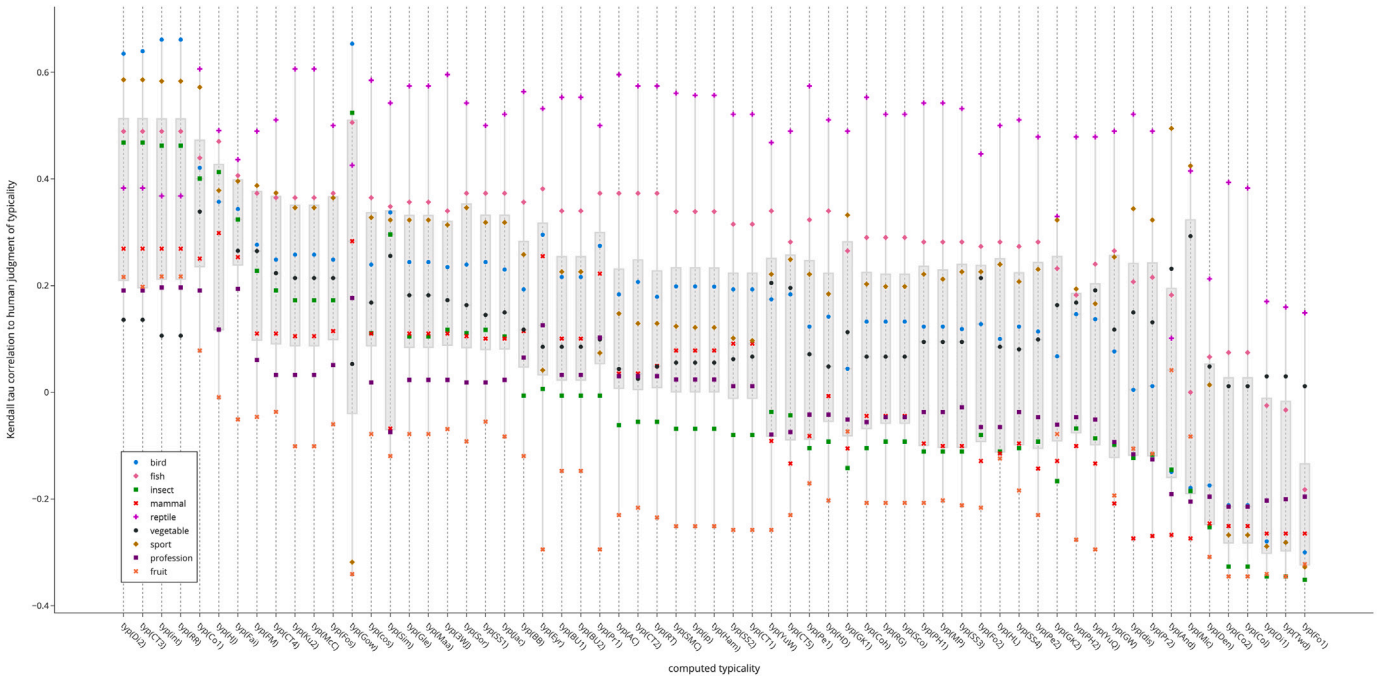


Fig. 1. Correlations of computed typicality to human judgment of typicality across natural categories with category attributes (horizontal axis ordered by mean value).

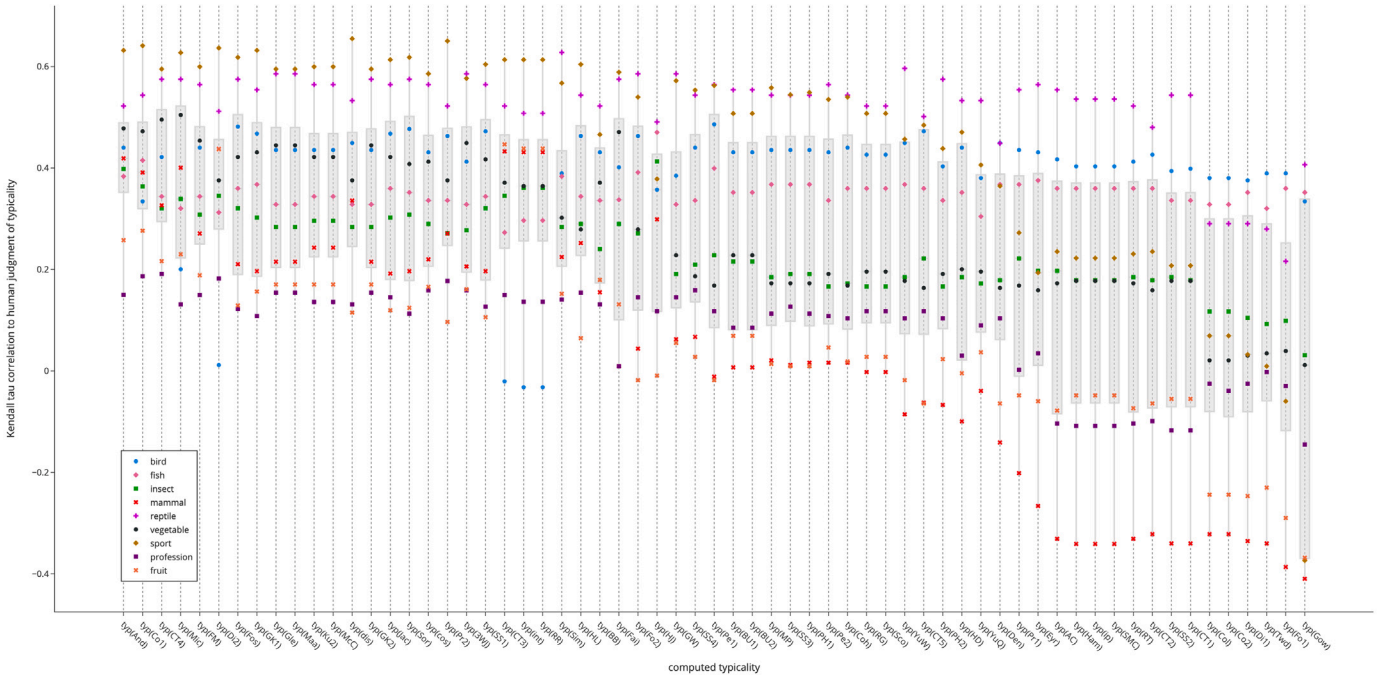


Fig. 2. Correlations of computed typicality to human judgment of typicality across natural categories with exemplar attributes (horizontal axis ordered by mean value).

of concepts that the quality of attributes used to assess typicality and similarity is essential [8]; see also [10] and the references therein. In order to enable good predictions, the attributes need to represent well the aspects people naturally take into account in their judgments on typicality and similarity. This intuitive knowledge has, nevertheless, not been confirmed by any extensive experimentation. Our results provide confirmation of this knowledge. Namely, as is apparent from all the graphs, the exemplar attributes generally result in a better prediction of human judgment of typicality than the category attributes, which are considerably less numerous and provide less distinctive information about the exemplars due to how these kinds of attributes have been collected (see Section 3). This is particularly apparent for the category-based data with the category attributes because, for this data, the numbers of attributes are considerably smaller than for the corresponding data with the exemplar attributes and also much smaller than the numbers of the exemplar and category attributes for the domain-based data. For the domain-based data, the numbers of both kinds of attributes are rather high, resulting in a comparable performance of prediction in this case.

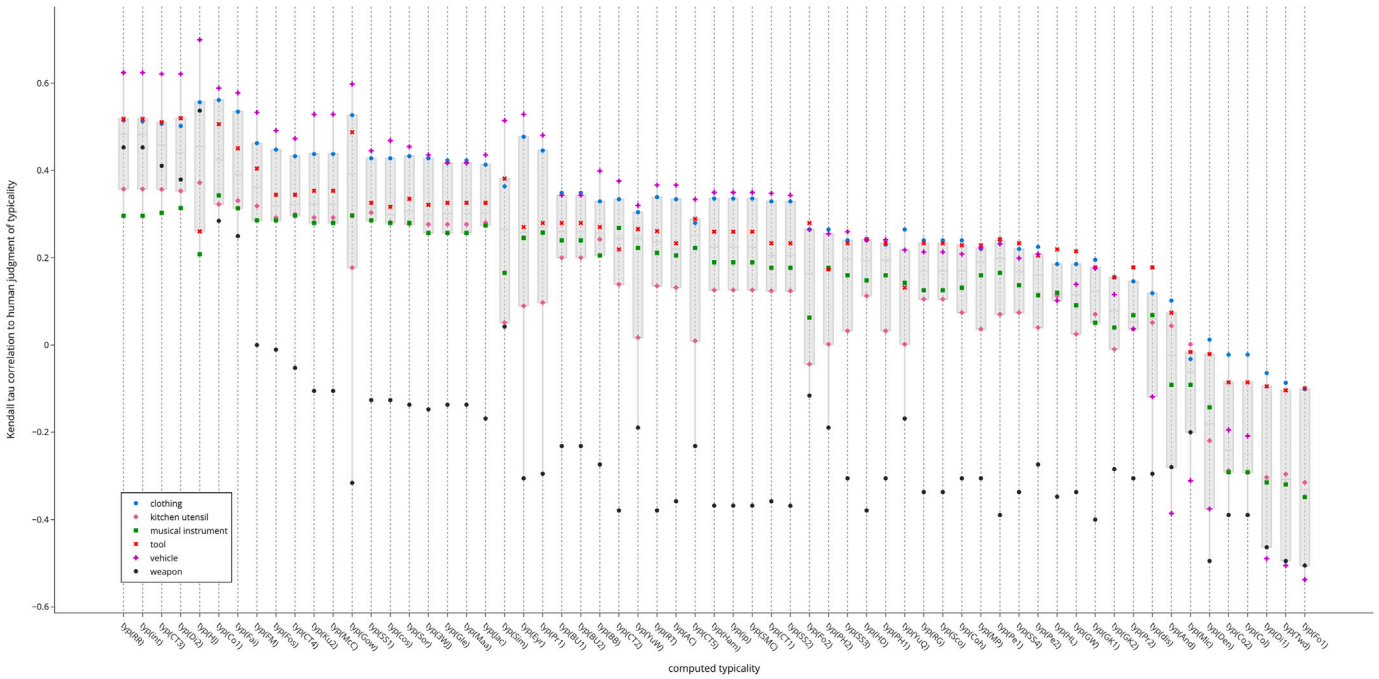


Fig. 3. Correlations of computed typicality to human judgment of typicality across artifact categories with category attributes (horizontal axis ordered by mean value).

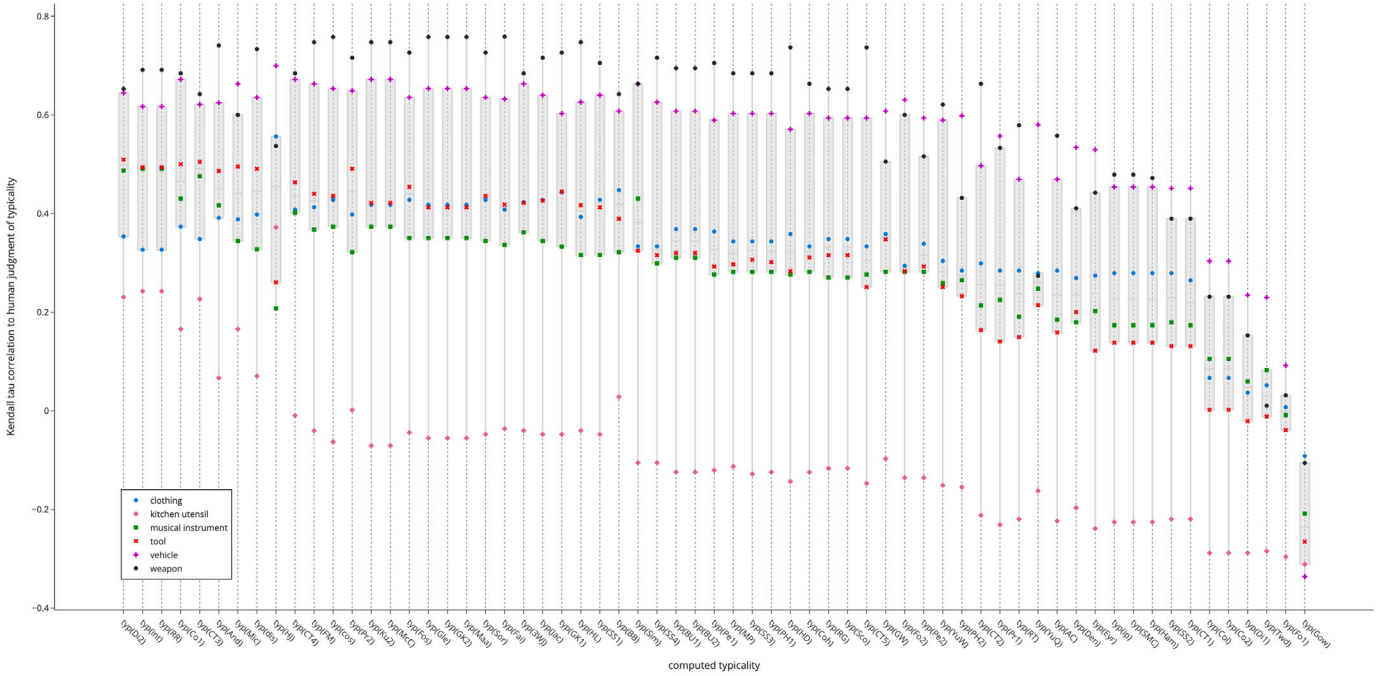


Fig. 4. Correlations of computed typicality to human judgment of typicality across artifact categories with exemplar attributes (horizontal axis ordered by mean value).

As regards a possible answer to the question in the title of our paper, i.e., which similarity is best to predict typicality, it comes as no surprise that there is no clear winner. This seems to result from the fact that all the similarity measures have been carefully designed to serve in certain real situations and have been proven through the test of time. In addition, several measures have been proposed for each particular purpose in the past. It is hence to be expected that groups of similarities, albeit vaguely delineated, rather than a single similarity, might be identified as the best predictors of typicality. In this regard, the group consisting of Co1, RR, int, Di2, CT3, and HJ may be identified as representing the best predictors. It is significant that this group includes the human similarity HJ, which not only confirms an intuitive expectation (human similarity is expected to come out among the best similarities) but also justifies the adequacy of formula (2) for computing degrees of typicality (the formula provides a verified relationship between a human judgment of similarity and a human judgment of typicality). As regards possible common properties of Co1, RR, int, Di2, and CT3, except for Co1, they are examples of the co-occurrence similarity measures defined by intuitive formulas. Moreover, the number d of negative matches (see Section 2.1) does not increase the value of similarity for these measures. We do not have an intuitive explanation for the good performance of the statistically

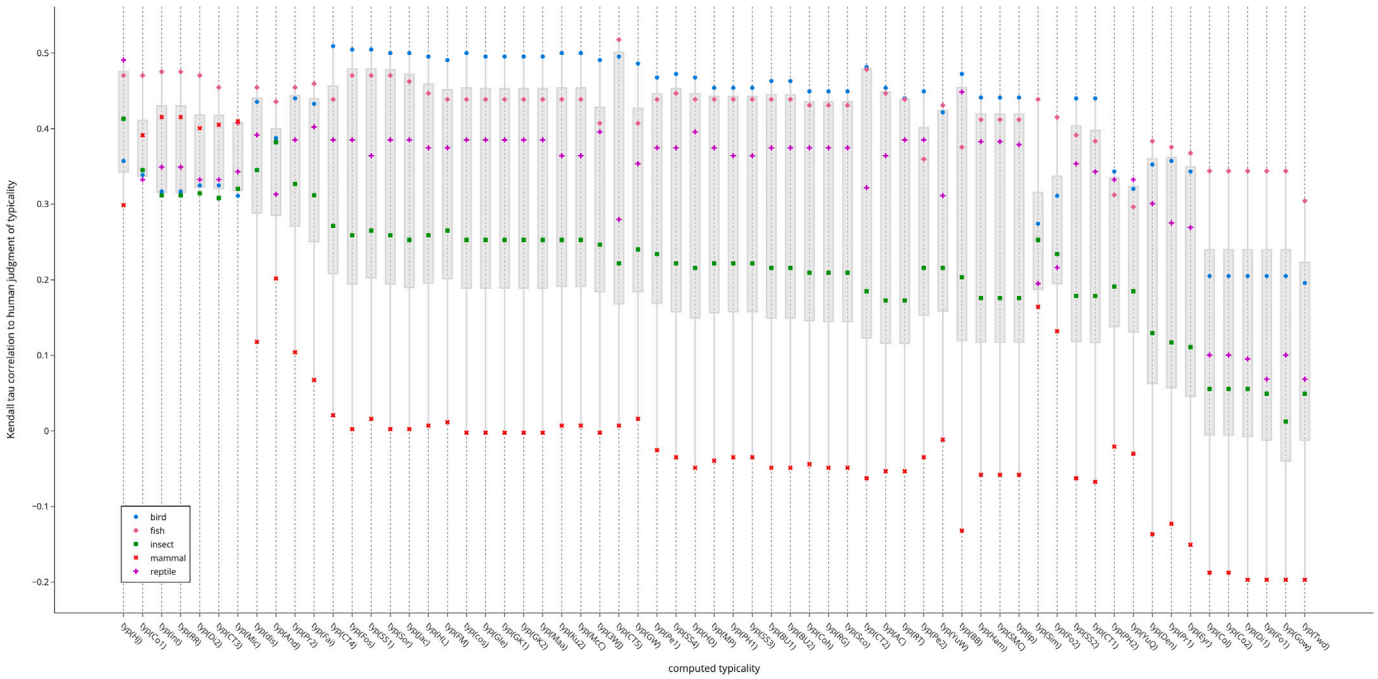


Fig. 5. Correlations of computed typicality to human judgment of typicality across animal domain with category attributes (horizontal axis ordered by mean value).

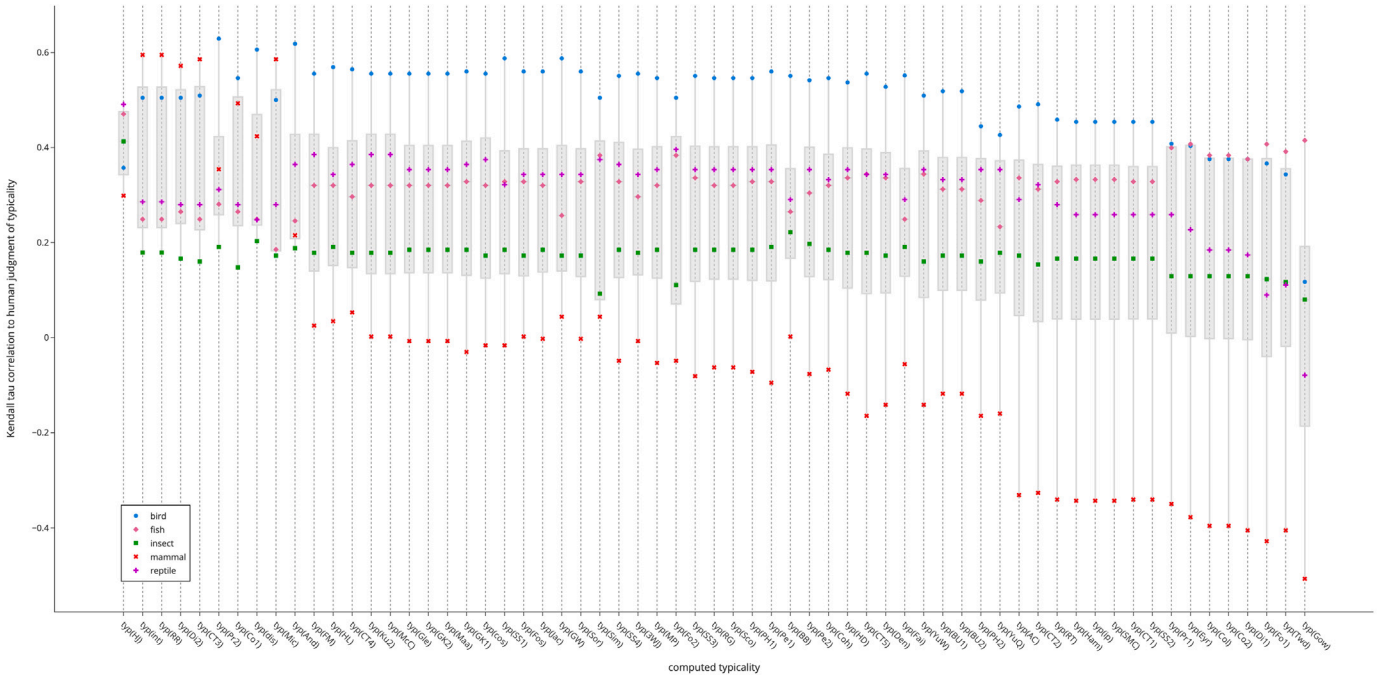


Fig. 6. Correlations of computed typicality to human judgment of typicality across animal domain with exemplar attributes (horizontal axis ordered by mean value).

motivated Co1. The second group that still provides very good predictions of typicality consists of Fai, FM, CT4, Fos, Ku2, McC, Sor, SS1, cos, Jac, Maa, and Gle. A majority of these measures are also co-occurrence based and for all of them, except for Fai, the negative matches (d) do not increase the value of similarity. On the other hand, similarities Den, Gow, Co2, Col, Di1, Twd, and Fo1 lead to poor predictions of typicality. Except for Di1, these are statistically motivated measures and for most of them, the negative matches (d) do increase the similarity value.

The graphs also reveal a few interesting particular observations. For instance, Figs. 3 and 4 display that for the category "weapon," the exemplar attributes result in the best predictions of typicality across all the artifact categories (with correlation values around 0.7), while the category attributes for "weapon" result in the worst prediction, and this holds true for most of the similarity measures. This is likely to be attributed to the small number of category attributes for this category, which turn out poorly informative for the prediction of typicality with most of the measures. We refrain from a detailed exposition of such particular observations, however interesting they may be and leave them for possible future examination due to lack of space.

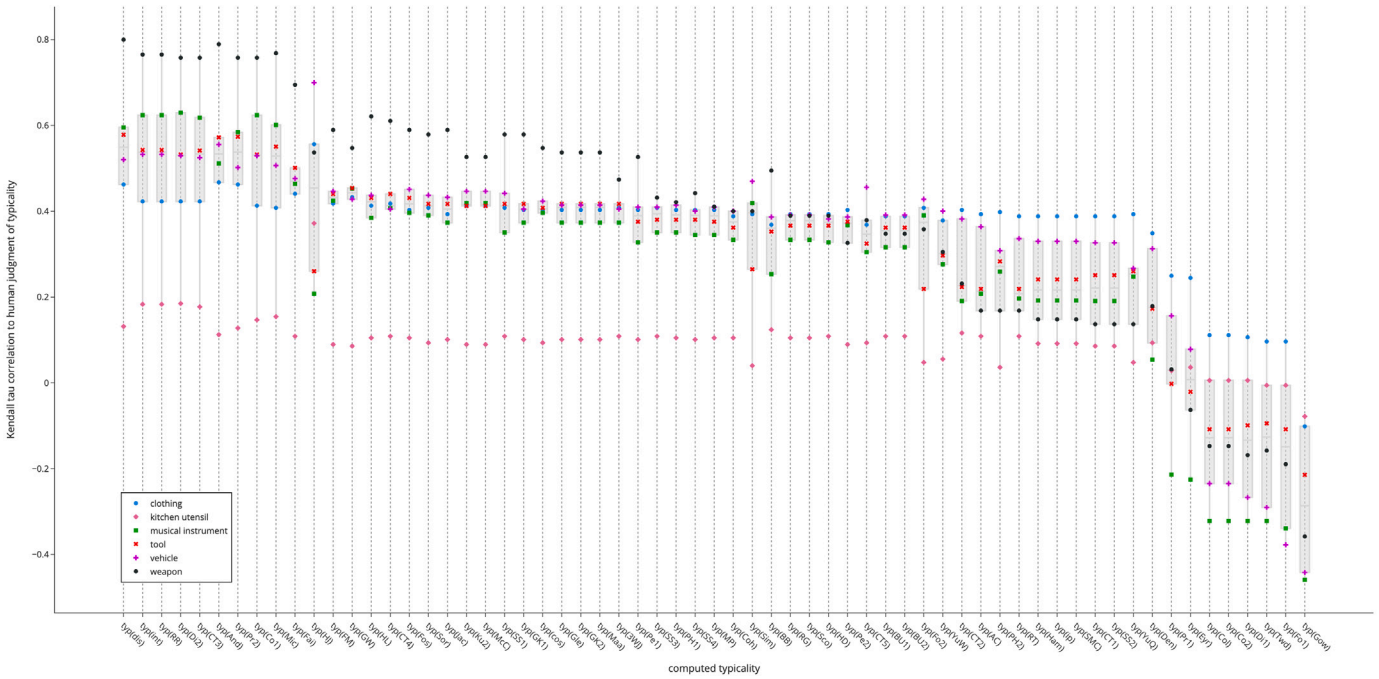


Fig. 7. Correlations of computed typicality to human judgment of typicality across artifact domain with category attributes (horizontal axis ordered by mean value).

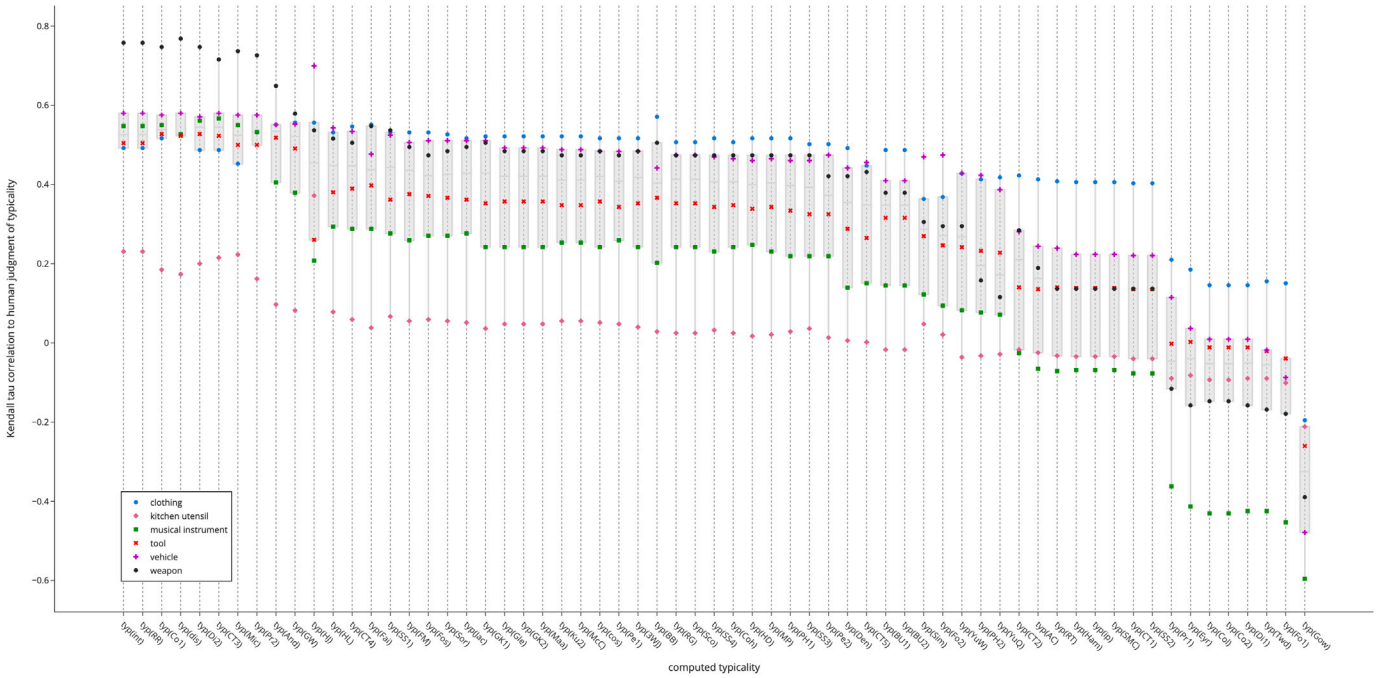


Fig. 8. Correlations of computed typicality to human judgment of typicality across artifact domain with exemplar attributes (horizontal axis ordered by mean value).

As regards possible limitations of the conclusions drawn from our experiments, they are implied, for the most part in our view, by the nature of the test data we use. For one, even though the Dutch data we utilized is rather extensive and involves several binary matrices, which we used, the validity of our conclusions would be improved if supported on yet another data, i.e., data obtained within an independent psychological study. Lack of such data presents a limitation not only to our study but for other possible explorations of a similar kind. Moreover, even though reliability was observed when gathering the Dutch data, both similarity and typicality may still be regarded as considerably subjective phenomena, and hence, a human judgment of both similarity and typicality may suffer from additional forms of possible unreliability compared to when data is obtained by an ordinary physical measurement. The latter problem, however, represents an unavoidable aspect of experimentation with psychological data.

5. Conclusions

Our experiments comparing 62 similarity measures for binary data with human judgments of similarity via their ability to predict human assessment of typicality reveal several patterns and observations. Most importantly, human similarity results in overall very good predictions

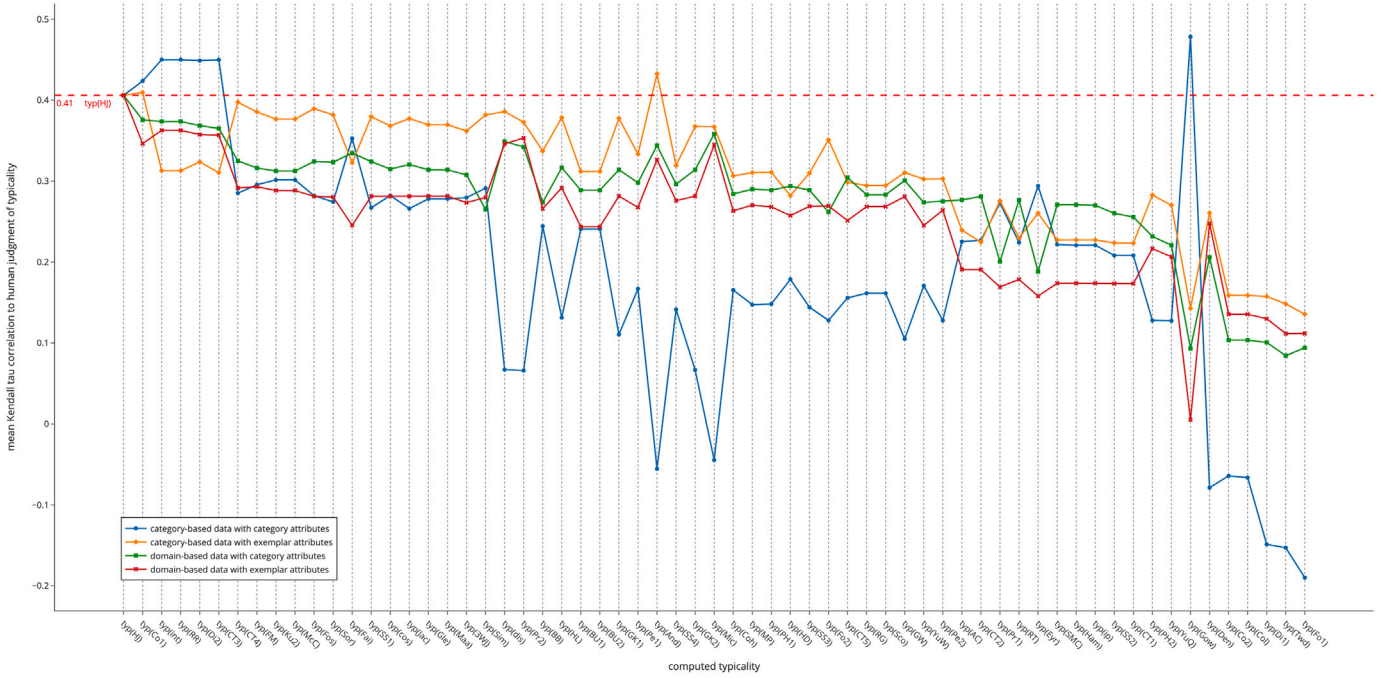


Fig. 9. Mean correlations of computed typicality to human judgment of typicality across categories of the animal domain.

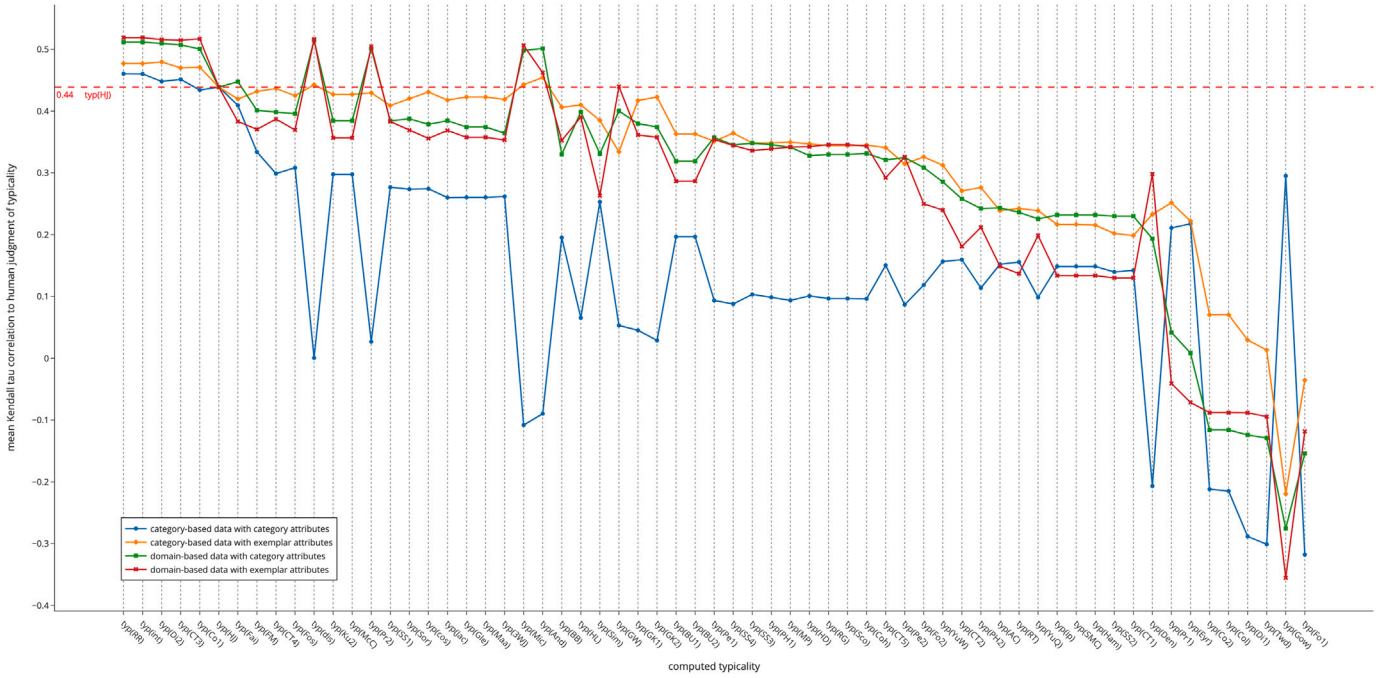


Fig. 10. Mean correlations of computed typicality to human judgment of typicality across categories of the artifact domain.

of typicality. For categories of the animal domain, it provides the best predictions. In this perspective, human similarity has a distinct place among the examined similarities as regards cognitive abilities.

On the other hand, human similarity ranks as the sixth best among all the explored similarities across all typicality predictions involved in our experiments. The experiments reveal a group of similarities, which includes human similarity, whose predictions of similarity are indeed strongly correlated with human assessment of typicality, as well as further observations worth further exploration.

As regards future research, we propose the following topics:

- The present experiments enable to compare similarities as regards their performance in a certain cognitive task (viz. prediction of typicality). A different experiment, however, should also be performed in which the existing similarity measures are compared as regards their ability to predict human judgment of similarity. This may reveal further, possibly different patterns and observations. The Dutch data, used in our experiment, allow for such kind of experiment.
- It became apparent that the quality of attributes which describe the exemplars plays a significant role in prediction of typicality of these exemplars. Since the quality of attributes is generally

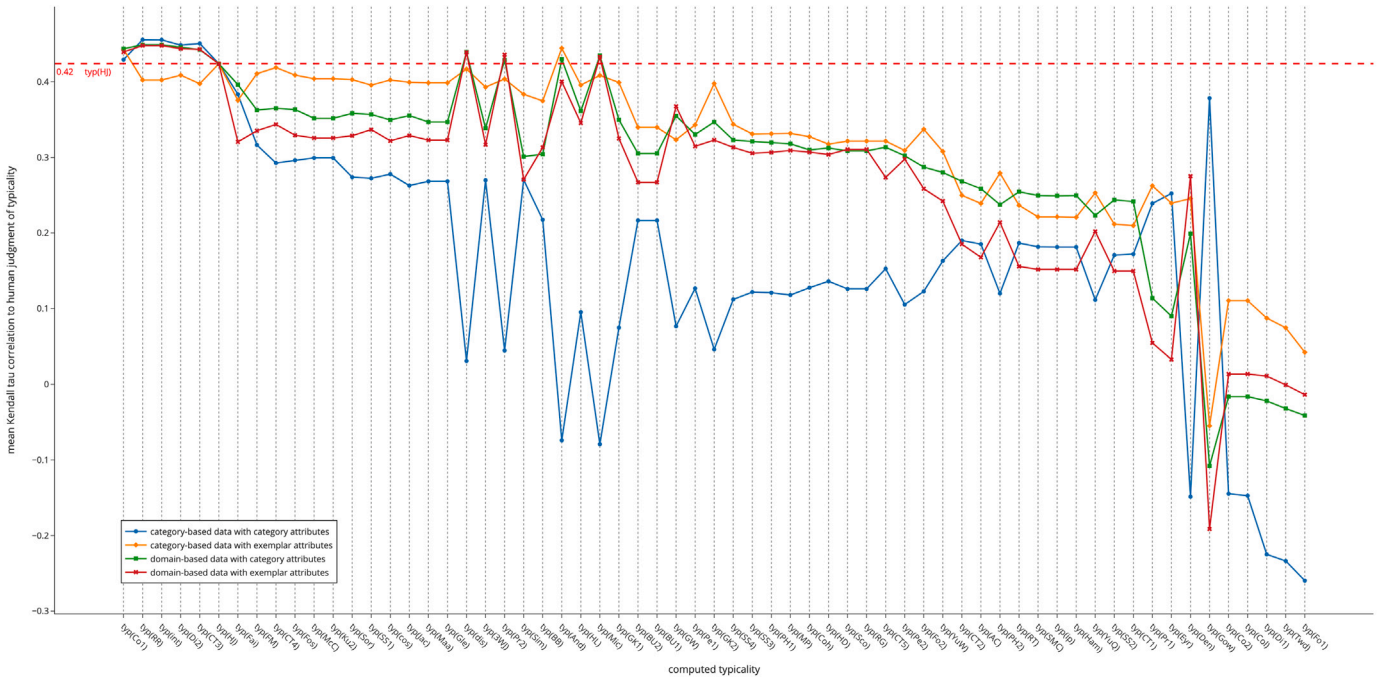


Fig. 11. Mean correlations of computed typicality to human judgment of typicality across all data.

regarded as important in a variety of cognitive tasks in the psychological literature, more focused studies shall be conducted in this direction. This includes possible quantitative measure of quality of a given set of attributes.

- In view of note 4.3, it seems to be of interest to compare the human assessment of similarity in the presence of context with the assessment with no context in the sense of note 4.3, as well as to perform a comparison with similarity degrees computed using similarity measures when binary attributes describing the exemplars are available. Such experiments may improve our understanding of the role of context for human assessment of similarity.
- It is apparent that for some categories (such as "mammal" in Fig. 2), the observed similarity measures differ in their capability to predict typicality to a larger extent compared to other categories (such as "fish" in Fig. 2). It seems of interest to explore in greater detail whether this phenomenon is due to the particular dataset used in our experiments or rather due to some general factor of psychological relevance.

CRediT authorship contribution statement

Radim Belohlavek: Formal analysis, Investigation, Methodology. **Tomas Mikula:** Formal analysis, Investigation, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Reference to the data is presented in the paper.

Acknowledgments

Supported partly by the project IGA 2023, reg. no. IGA_PrF_2023_026, of Palacký University Olomouc, Czech Republic.

Appendix. Similarity measures

The appendix presents 69 similarity measures for binary data we employ in the experiments along with additional information. In the formulas defining the similarity measures we denote for two binary vectors $x, y \in \{0,1\}^n$ by a, b, c, and d the numbers of attributes defined in Section 2.1. Hence, a, b, c, and d denote the number of attributes shared by x and y, possessed by x but not by y, possessed by y but not by x, and possessed neither by x nor by y, respectively. Thus,

$$n = a + b + c + d.$$

The measures are presented in Table 4. For each measure we include its abbreviation, its name (along with alternative names), a formula defining the measure, and a list of significant comparative papers in which this measure appears. The measures are ordered lexicographically by their abbreviations for ease of lookup. In our table, we refer to the following comparative papers, to which refer by the numbers 1–5 in the appendix:

1. Brusco, Cradit, and Steinley [1], which contains 71 similarity measures;
2. Choi, Cha, and Tappert, 2010 [2], which includes 60 similarity (and 16 dissimilarity) measures;
3. Hubálek, 1982 [4], which involves 20 similarity measures (in fact, it lists 43 measures from which 20 are selected after removing certain measures due to their equivalence with other involved measures or due to lack of required properties);
4. Todeschini, Consonni, Xiang, Holliday, Buscema, and Willett, 2012 [5], which employs 44 similarity measures (it includes 51 similarity measures, of which 7 were eliminated due to their equivalence with other measures);
5. Wijaya, Afendi, Batubara, Darusman, Altaf-Ul-Amin, and Kanaya, 2016 [6], which includes 62 similarity (and 17 dissimilarity) measures.

Remark 1. (a) Some similarity measures presented in Table 4 are not defined for certain values of a, b, c, and d, which naturally occur in data. The measures that suffer from this defect on our data are omitted in the graphs presenting results of our experiments in Section 4. In

Table 4
Similarity measures.

Symbol	Name	Formula	Source
AC	Austin-Colwell	$\frac{2}{\pi}\arcsin\sqrt{\frac{a+d}{n}}$	1, 3, 4
And	Anderberg	$\frac{\tau_1-\tau_2}{2n}$ with $\tau_1=\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)$, $\tau_2=\max(a+c,b+d)+\max(a+b,c+d)$	1, 2, 5
BB	Braun-Blanquet	$\frac{a}{\max(a+b,a+c)}$	1, 2, 3, 4, 5
BU1	Baroni-Urbani-Buser 1	$\frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c}$	1, 2, 3, 4, 5
BU2	Baroni-Urbani-Buser 2	$\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}$	1, 2, 3, 4, 5
Coh	Cohen	$\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)}$	1, 4
Col	Cole	$\frac{ad-bc}{(a+b)(b+d)}$ if $ad \geq bc$; $\frac{ad-bc}{(a+b)(a+c)}$ if $ad < bc$ and $d \geq a$; $\frac{ad-bc}{(b+d)(c+d)}$ otherwise	2, 3, 5
Co1	Cole (Cole 1)	$\frac{ad-bc}{(a+c)(c+d)}$	1, 4
Co2	Cole (Cole 2)	$\frac{ad-bc}{(a+b)(b+d)}$	1, 4
cos	cosine (Driver-Kroeber, Ochiai)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	1, 2, 4, 5
CT1	Consonni-Todeschini 1	$\frac{\ln(1+a+d)}{\ln(1+n)}$	1, 4
CT2	Consonni-Todeschini 2	$\frac{\ln(1+n)-\ln(1+b+c)}{\ln(1+n)}$	1, 4
CT3	Consonni-Todeschini 3	$\frac{\ln(1+a)}{\ln(1+n)}$	1, 4, 5
CT4	Consonni-Todeschini 4	$\frac{\ln(1+a)}{\ln(1+a+b+c)}$	1, 4, 5
CT5	Consonni-Todeschini 5	$\frac{\ln(1+ad)-\ln(1+bc)}{\ln(1+n^2/4)}$	1, 4, 5
Den	Dennis	$\frac{ad-bc}{\sqrt{n(a+b)(a+c)}}$	1, 2, 4, 5
dis	dispersion	$\frac{ad-bc}{n^2}$	1, 2, 4, 5
Di1	Dice 1	$\frac{a}{a+b}$	1, 4
Di2	Dice 2	$\frac{a}{a+c}$	1, 4
Eyr	Eyraud	$\frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$	1, 2, 5
Fai	Faith	$\frac{a+0.5d}{n}$	1, 2, 4, 5
FM	Fager-McGowan	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2\sqrt{\max(a+b,a+c)}}$	1, 2, 3, 5
Fos	Fossum	$\frac{n(a-\frac{1}{2})^2}{(a+b)(a+c)}$	1, 2, 4, 5
Fo1	Forbes 1	$\frac{na}{(a+b)(a+c)}$	1, 2, 3, 4, 5
Fo2	Forbes 2	$\frac{na-(a+b)(a+c)}{n\min(a+b,a+c)-(a+b)(a+c)}$	1, 2, 3, 5
Gle	Gleason (Dice, Sørensen, Czekanowski	$\frac{2a}{2a+b+c}$	1, 2, 3, 4, 5
GK1	Goodman-Kruskal 1	$\frac{\tau_1-\tau_2}{2n-\tau_2}$ with $\tau_1=\max(a,b)+\max(c,d)+\max(a,c)+\max(b,d)$, $\tau_2=\max(a+c,b+d)+\max(a+b,c+d)$	1, 2, 5
GK2	Goodman-Kruskal 2	$\frac{2\min(a,d)-b-c}{2\min(a,d)+b+c}$	1, 4
Gow	Gower	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1, 2, 5
GW	Gilbert-Wells	$\ln\frac{a^3}{2a(a+b)(c+d)(a+c)(b+d)}+2\ln\frac{n!a!b!c!d!}{(a+b)!(c+d)!(a+c)!(b+d)!}$	1, 2, 3, 5
Ham	Hamman	$\frac{a+d-b-c}{n}$	1, 2, 3, 4, 5
HD	Hawkins-Dotson	$\frac{1}{2}\left(\frac{a}{a+b+c}+\frac{d}{d+b+c}\right)$	1, 4
HL	Harris-Lahey	$\frac{a(2d+b+c)}{2(a+b+c)}+\frac{d(2a+b+c)}{2(b+c+d)}$	1, 4
int	intersection	a	2, 5
ip	inner product	$a+d$	2, 5
Jac	Jaccard (Jaccard-Tanimoto)	$\frac{a}{a+b+c}$	1, 2, 3, 4, 5
Ku1	Kulczynski 1	$\frac{a}{b+c}$	1, 2, 3, 5
Ku2	Kulczynski 2 (Driver-Kroeber)	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	1, 2, 3, 4, 5
Maa	van der Maarel	$\frac{2a-b-c}{2a+b+c}$	1, 4
McC	McConnaughey	$\frac{a^2-bc}{(a+b)(a+c)}$	1, 2, 3, 4, 5
Mic	Michael	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	1, 2, 3, 4, 5
Mou	Mountford	$\frac{2a}{ab+ac+2bc}$	1, 2, 3, 4, 5
MP	Maxwell-Pilliner	$\frac{2(ad-bc)}{(a+b)(c+d)+(a+c)(b+d)}$	1, 4
Pe1	Pearson 1 (χ^2 statistical significance)	$\frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	1, 2, 3, 5
Pe2	Pearson 2	$\sqrt{\frac{\chi^2}{n+\chi^2}}$ with χ^2 equal to Pe1	1, 2, 3, 5
Pe3	Pearson 3	$\sqrt{\frac{\rho}{n+\rho}}$ with ρ equal to PH1	2, 5
PH1	Pearson-Heron 1 (Phi)	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$	1, 2, 3, 4, 5
PH2	Pearson-Heron 2	$\cos\left(\frac{\pi\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}\right)$	2, 3, 5
Pr1	Peirce 1	$\frac{ad-bc}{(a+b)(c+d)}$	1, 4
Pr2	Peirce 2	$\frac{ad-bc}{(a+c)(b+d)}$	1, 3, 4
Pr3	Peirce 3	$\frac{ad+bc}{ab+2bc+cd}$	1, 2, 3, 5
RG	Rogot-Goldberg	$\frac{a}{2a+b+c}+\frac{d}{2d+b+c}$	1, 4
RR	Russel-Rao	$\frac{a}{n}$	1, 2, 3, 4, 5
RT	Rogers-Tanimoto	$\frac{a+d}{a+2(b+c)+d}$	1, 2, 3, 4, 5
Sco	Scott	$\frac{4ad-(b+c)^2}{(2a+b+c)(2d+b+c)}$	1, 4
Sim	Simpson	$\frac{a}{\min(a+b,a+c)}$	1, 2, 3, 4, 5
SMC	simple matching coefficient (Sokal-Michener)	$\frac{a+d}{n}$	1, 2, 3, 4, 5
Sor	Sorgenfrei	$\frac{a^2}{(a+b)(a+c)}$	1, 2, 3, 4, 5
SS1	Sokal-Sneath 1	$\frac{a}{a+2b+2c}$	1, 2, 3, 4, 5
SS2	Sokal-Sneath 2	$\frac{2a+2d}{2a+b+c+2d}$	1, 2, 3, 4, 5
SS3	Sokal-Sneath 3	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d}\right)$	1, 2, 3, 4, 5
SS4	Sokal-Sneath 4, Ochiai 2	$\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	1, 2, 3, 4, 5
SS5	Sokal-Sneath 5	$\frac{a+d}{b+c}$	1, 2, 3, 5
Sti	Stiles	$\log_{10}\frac{n(\lvert ad-bc\rvert-\frac{n}{2})^2}{bc(a-b)(a-c)}$	1, 2, 5
Tar	Tarantula	$\frac{a(c+d)}{c(a+b)}=\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$	1, 2, 5
Twd	Tarwid	$\frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	1, 2, 3, 5
YuQ	Yule (Yule Q)	$\frac{ad-bc}{ad+bc}$	1, 2, 3, 4, 5
YuW	Yule (Yule W)	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	1, 2, 3, 4, 5
3WJ	3W-Jaccard	$\frac{3a}{3a+b+c}$	1, 2, 4, 5

particular, there are 7 such measures: Ku1, Mou, Pe3, Pr3, SS5, Sti, Tar. For instance, the Mountfond measure given by $\frac{2a}{ab+ac+2bc}$ is not defined if $a = b = 0$ or $a = c = 0$.

(b) Some measures are not defined even for two objects which share the same attributes (i.e., for which $b = c = 0$), which is counterintuitive. One could redefine each such measure so that for $b = c = 0$ it yields its maximal value. We nevertheless refrained from this possible modification to obey the definitions presented in the literature.

Remark 2. We found a number of mistakes in the literature on similarity measures for binary data. In the following list, we include the significant ones pertaining to the measures we employ.

1. AC: 3 lists a slightly different formula for AC, namely $\frac{1}{50\pi}\sqrt{\frac{a+d}{n}}$, i.e., a formula yielding a value 100× smaller than our formula.
2. Col: 2, 3, 5 list a different formula, namely $\frac{\sqrt{2}(ad-bc)}{\sqrt{(ad-bc)^2-(a+b)(a+c)(b+d)(c+d)}}$. This formula also appears in the original paper [18, p. 416] as a so-called mean square contingency, but is not meant as the similarity measure which the authors present in their paper. The Abydos library [19] lists our formula for Col.
3. Eyr: 3 lists a different formula, namely $\frac{a-(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}$.
4. FM: 1, 2, 3, and 5 list a different formula, $\frac{a}{\sqrt{(a+b)(a+c)}} - \frac{\max(a+b,a+c)}{2}$, which is apparently wrong. Namely, the original paper [20] contains the formula we use as FM, and notes that this formula results by a modification of a formula used in [21].
5. Fos: 1 lists a different formula, $\frac{n(a-\frac{1}{2})^2}{\sqrt{(a+b)(a+c)}}$, which is an apparent misprint.
6. GW: 1, 2, 3, and 5 list a different formula, namely $\log a - \log n - \log(\frac{a+b}{n}) - \log(\frac{a+c}{n})$; 1 and 3 refer to [22], 2 does not contain a reference for this measure, and refers to 1 and 3. The original paper [22] includes our formula, as does [19].
   Note also that 1, 2, 3, 4, and 5 list the so-called Johnson measure with a formula $\frac{a}{a+b} + \frac{a}{a+c}$. Clearly, this formula yields the value of $2 \cdot \text{Ku2}$, hence we do not include the Johnson measure [23].
7. SS3: 2 contains a misprint in the formula for SS3.
8. Sti: 1, 2, and 5 list a different formula, namely $\log \frac{n(|ad-bc|-n/2)^2}{(a+b)(a+c)(b+d)(c+d)}$. Our formula comes from the original paper [24] and is also used in the Abydos library [19].

References

[1] M. Brusco, J.D. Cradit, D. Steinley, A comparison of 71 binary similarity coefficients: The effect of base rates, PLoS One 16 (2021) 4.

[2] S.S. Choi, S.H. Cha, C.C. Tappert, A survey of binary similarity and distance measures, J. Syst. Cybern. Inf. 8 (1) (2010) 43–48.

[3] J.C. Gower, P. Legendre, Metric and euclidean properties of dissimilarity coefficients, J. Classification 3 (1986) 5–48.

[4] Z. Hubálek, Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation, Biol. Rev. 57 (4) (1982) 669–689.

[5] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, P. Willett, Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets, J. Chem. Inf. Model. 52 (11) (2012) 2884—2901.

[6] S.H. Wijaya, F.M. Afendi, I. Batubara, L.K. Darusman, M. Altaf-Ul-Amin, S. Kanaya, Finding an appropriate equation to measure similarity between binary vectors: Case studies on Indonesian and Japanese herbal medicines, BMC Bioinformatics 17 (1) (2016) 520.

[7] M.J. Warrens, Similarity Coefficients for Binary Data (Ph.D. thesis), Leiden University, The Netherlands, 2008, https://scholarlypublications.universiteitleiden.nl/access/item%3A2961378/view.

[8] G.L. Murphy, The Big Book of Concepts, MIT Press, Cambridge, Mass, 2002, http://dx.doi.org/10.7551/mitpress/1602.001.0001.

[9] R. Belohlavek, T. Mikula, Typicality: A formal concept analysis account, Internat. J. Approx. Reason. 142 (2022) 349–369.

[10] S. De Deyne, S. Verheyen, E. Ameel, W. Vanpaemel, M.J. Dry, W. Voorspoels, G. Storms, Exemplar by feature applicability matrices and other dutch normative data for semantic concepts, Behav. Res. Methods 40 (4) (2008) 1030–1048, http://dx.doi.org/10.3758/BRM.40.4.1030.

[11] E. Rosch, Principles of categorization, in: E. Rosch, B.B. Lloyd (Eds.), Cognition and Categorization, Erlbaum, Hillsdale, NJ, 1978, pp. 27–48.

[12] E. Rosch, C.B. Mervis, Family resemblances: Studies in the internal structure of categories, Cogn. Psychol. 7 (4) (1975) 573–605, http://dx.doi.org/10.1016/0010-0285(75)90024-9.

[13] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, P. Boyes-Braem, Basic objects in natural categories, Cogn. Psychol. 8 (3) (1976) 382–439, http://dx.doi.org/10.1016/0010-0285(76)90013-X.

[14] D. Dua, C. Graff, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2019, http://archive.ics.uci.edu/ml.

[15] W. Ruts, S. De Deyne, E. Ameel, W. Vanpaemel, T. Verbeemen, G. Storms, Dutch norm data for 13 semantic categories and 338 exemplars, Behav. Res. Methods Instrum. Comput. 36 (2004) 506–515.

[16] R. Belohlavek, T. Mikula, Dutch concepts. GitHub repository, 2023, https://github.com/mikulatomas/dutch-concepts.

[17] SciPy 1.0 Contributors, et al., Scipy 1.0: fundamental algorithms for scientific computing in python, Nature Methods (2020) http://dx.doi.org/10.1038/s41592-019-0686-2.

[18] L.C. Cole, The measurement of interspecific association, Ecology 30 (4) (1949) 411–424.

[19] Abydos: abydos.distance package. https://abydos.readthedocs.io/en/latest/abydos.distance.html.

[20] E.W. Fager, J.A. McGowan, Zooplankton species groups in the north pacific, Science 140 (3566) (1963) 453–460.

[21] E.W. Fager, Determination and analysis of recurrent groups, Ecology 38 (4) (1957) 586—595.

[22] N. Gilbert, T.C.E. Wells, Analysis of quadrat data, J. Ecol. 54 (3) (1966) 675–685.

[23] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (3) (1967) 241–254.

[24] E.H. Stiles, The association factor in information retrieval, J. ACM 8 (2) (1961) 271–279.