Detecting intra-categorical structure through contextual bundling on relevant attributes

Stefan Schneider[1], Rohit Tidke[1,†], Steven Verheyen[2], Armin Schulz[3,†] and Andreas Nürnberger[1]

[1]Data and Knowledge Engineering Group, Faculty of Computer Science, Otto von Guericke University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany

[2]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Post Box 1738, 3000 DR Rotterdam, The Netherlands

[3]Department of Computer Science and Media, Brandenburg University of Applied Science, Magdeburger Strasse 50, 14770 Brandenburg an der Havel, Germany

Abstract

Intra-categorical structuring is an important part of category representation for describing the relationship between categories and their exemplars. Based on contextual bundling [1], the degree of correspondence between the bundles of an approximative model and the original incidence data indicates how typical an exemplar is for a category. It has been shown that contextual bundling can reflect the users' perceived graded category structure. However, this observation is limited to object-attribute matrices that are based on a limited number of attributes generated toward the category. In large object-attribute matrices, irrelevant attributes distort the derived conceptual orders. The question arises whether filtering a subset of the most relevant attributes can help to capture graded category structure particularly if the attribute set size is very large. Our analysis is based on large binary object-attribute matrices and human typicality ratings for the domains of Animals and Artifacts which compromise several normed semantic categories present in the Leuven Concept Data [2]. We reproduce existing results by Ceulemans and Storms [3] and find that the contextual bundling fits well, while the intra-categorical structure can only be reproduced for the domain of Animals. For the first time, we calculated model matrices for attributes generated toward categories' exemplars (instead of the category proper). In a subsequent step, we filtered these large and more concrete object-attribute matrices using various measures of attribute weight, and were able to construct valid models. The results suggest that the attribute filtering methods used do not provide a robust categorization. Although none of the methods examined is effective in capturing the intra-categorical structure of the categories in general, attribute filtering is successful in categories where the unfiltered approach fails, and vice versa.

1. Introduction

What is required to describe a concept sufficiently? It is now commonly accepted that concepts must be represented using an intra-categorical structure - the structure within a category - AND an inter-categorical structure - the structure between categories, which is often presumed to represent a hierarchical order [1].

The HIerarchical CLASs (HICLAS) model [4, 1] is believed to capture the intra- and inter-categorical structuring. For that purpose HICLAS approximates the given incidence in an object-attribute matrix through overlapping rectangles represented within a so called model matrix. A model matrix is fully described by pairs of object and attribute bitsets whereas each of those pairs is interpreted as a factor or bundle of that model matrix. These overlapping bundles capture the inter-categorical structure. The intra-categorical structure of categories is represented in HICLAS using contextual bundling [1], whereby typicality is calculated by measuring the fit between the bundles of a model matrix and the original data matrix, its context. The suitability of contextual bundling has been proven for matrices

CONCEPTS 2026, August 31–September 04, 2026, Montpellier, France

† Undergraduate Institution.

stefan.schneider@sschneider.de (S. Schneider)

0000-0001-6572-5983 (S. Schneider); 0009-0007-4895-3061 (R. Tidke); 0000-0002-6778-6744 (S. Verheyen); 0009-0003-0063-2964 (A. Schulz); 0000-0003-4311-0624 (A. Nürnberger)

covering only a certain basic-level category (e.g. Sports [1]). For matrices consisting of larger sets of categories, the results are not as clear [3]. The typicality of exemplars can be captured quite well for some categories but not for others. One reason for this could be the influence of irrelevant attributes. To improve the ratio between irrelevant and relevant attributes, one can select subsets of attributes based on an additional attribute relevance measure and form a submatrix - a method known as attribute filtering [5].

The following example illustrates the effects of attribute filtering on the structure within a category. Consider a scenario with a binary object-attribute matrix for the category Vehicles, consisting of four objects (i.e., Car, Bicycle, Train, and Boat) and four associated attributes (i.e., Has wheels, Is electric, Has an engine, and Is a watercraft), as well as a non-perfectly grouped incidence structure. We formulate two bundles - Bundle 1: (Car, Bicycle, Train) - (Has wheels, Is electric), Bundle 2: (Train, Boat) - (Has an engine, Is a watercraft) - to approximate the data in a model visualized by two rectangles -labeled City Vehicles and Freight Vehicles. Now we wish to express that some of the attributes are more or less relevant and assign continuous weights w to each attribute. Based on these weights, it becomes possible to filter out attributes below a certain threshold. We set the threshold to $w \geq 0.6$, and the attribute Is Electric (0.1) is filtered out along with the corresponding incidences. To model the reduced data matrix, bundle 1 simply needs to be slightly modified to (Car, Bicycle, Train) - (Has Wheels). On the other hand, the effects on the intra-category structuring are much more drastic, as the first rectangle loses half of all possible incidence relationships. By calculating typicality scores for incidences using the contextual bundling t, the typicality of Bicycle increases from 0.5 to 1.0, while the typicality of Train decreases from 0.5 to 0.33. This means, for example, that by defining Is Electric as an irrelevant attribute, the car is no longer the most typical representative of the category Private Urban Vehicles; in fact, the Bicycle is now just as typical.

Table 1
Example from the Vehicles category - fitting the model (gray) to the data results in changes to the typicality predictions (t) when the "Is electric" attribute has been filtered out as irrelevant.

	Has Wheels	Is Electric	Has Engine	Is Watercraft	t
w	0.9	0.1	0.7	0.8	/
Car	**1**	1	**1**	**0**	$1.0 \rightarrow 1.0$
Bike	**1**	0	**0**	**0**	$0.5 \rightarrow 1.0$
Train	**0**	1	**1**	**0**	$0.5 \rightarrow 0.3$
Boat	**0**	0	**1**	**1**	$1.0 \rightarrow 1.0$

This illustrative example highlights the role of attribute relevance. Even filtering out a single attribute as irrelevant can have a significant impact on the internal structure of a category. The following paper focuses on contextual bundling and examines whether attribute filtering is helpful for capturing intra-category structure. In the following section, we describe some relevant fundamentals of domain order selection and contextual bundling using the terminology of Boolean Matrix Factorization (BMF). In the subsequent analysis, we first examine whether existing results regarding the quality of intra-categorical structuring by Ceulemans and Storms (2010) [3] can be reproduced. These were based on a relatively small object-attribute matrices comprised of attributes generated toward categories. To examine the role of attribute filtering, we also evaluate the quality of intra-categorical structuring on larger matrices comprised of attributes generated toward the category exemplars.

2. Model order selection and contextual bundling

Using a model matrix for inter-categorical structuring was originally proposed by DeBoeck and Rosenberg [4] as part of the HICLAS model. Storms et. al. [1] showed how to extend the notion of HICLAS in terms of intra-categorical structuring.

The following nomenclature has been taken from [6]. Matrices are denoted by upper-case bold letters ($\mathbf{A}$). If $\mathbf{X}$ and $\mathbf{Y}$ are two *n*-by-*m* Boolean matrices, we have the following element-wise matrix operations. The *Boolean sum* $\mathbf{X} \vee \mathbf{Y}$ is the normal matrix sum with addition defined as $1 + 1 = 1$. The *Boolean subtraction* $\mathbf{X} \ominus \mathbf{Y}$ is the normal element-wise substraction with $0 - 1 = 0$. Note that this does not define an inverse of Boolean sum, as $1 + 1 - 1 = 0$. The *Boolean element-wise product* $\mathbf{X} \wedge \mathbf{Y}$ is defined as the normal element-wise matrix product. The *Boolean exclusive or* $\mathbf{X} \oplus \mathbf{Y}$ is the normal matrix sum with addition defined as $1 + 1 = 0$ (i.e., addition is done over the field $\mathbb{Z}_2$). Let $\mathbf{X}$ be *n*-by-*k* and $\mathbf{Y}$ be *k*-by-*m* Boolean matrices. Their *Boolean matrix product*, $\mathbf{X} \circ \mathbf{Y}$, is the Boolean matrix $\mathbf{Z}$ with $\{z_{ij} = \bigvee_{l=1}^{k} x_{il} y_{lj}\}$. Boolean matrix product is the normal matrix product using the Boolean addition.

A formal context $\mathbb{K} := (X, Y, I)$ consists of two sets, X (of objects) and Y (of attributes), and a binary relation I between them, i.e. $I \subseteq X \times Y$. If an object x is in relation I with an attribute y is it written xIy. Attribute weights are an additional component based on a given formal context. They are one way of channeling a user's background knowledge. We follow the basic terminology by Belohlavek and Macko [7] and assume that the weights to be assigned to attributes form a partial ordering $\langle W, \leq \rangle$. (Weights are elements $w \in W$.) A popular case used in various methods of data analysis as well as in our paper is $\langle W, \leq \rangle = \langle \mathbb{R}_0^+, \leq \rangle$ (weights are reals). Given a formal context $\mathbb{K}$ and a partially ordered set $\langle W, \leq \rangle$ of weights, we assume that there is an assignment of weights w from W to attributes y element of Y, i.e. that a function $w : Y \to W$ is specified. The rule of thumb is that the more important the attribute, the larger the weight assigned to it. In this paper, attribute weights are only comparable within a formal context $\mathbb{K}$.

We distinguish between a data matrix, or simply data $\mathbf{D}$, and its formal context $\mathbb{K}_D := (X, Y, D)$ with $\mathbf{D} \subseteq X \times Y$ from its model (or a hypothesis) $\mathbf{H} \in \mathbb{H}$ and its formal context $\mathbb{K}_H := (X, Y, H)$ with $\mathbf{H} \subseteq X \times Y$. Such a model H is called *exact* if $\mathbf{D} = \mathbf{H}$, and *approximate* if $\mathbf{D} \neq \mathbf{H}$. To reconstruct $\mathbf{D}$ from an approximate $\mathbf{H}$, an error matrix $\mathbf{E}$ captures the discrepancy between the data and the model with $\mathbf{E} = \mathbf{D} \oplus \mathbf{H}$, using a unique Boolean matrix of dimensions n by m [6]

An $n \times m$-matrix $\mathbf{H}$ is said to be rectangular [8] if and only if there exist L-sets $\mathbf{B}$ in $1, ..., n$ and $\mathbf{C}$ in $1, ..., m$ such that $H_{ij} = B(i) \otimes C(j)$ holds for $1 \leq i \leq n$ and $1 \leq j \leq m$. For a given rank k, the model $\mathbf{H}$ defined by $\mathbf{B} \circ \mathbf{C}$ defines a $n \times k$-object bundle matrix $\mathbf{B}$ and a $k \times m$ -bundle attribute matrix $\mathbf{C}$ [3]. A bundle of objects and attributes takes on a specific rank value, which is used as an index to obtain a pair of bitsets from $\mathbf{B}$ and $\mathbf{C}$ [1]. Each bundle describes $\mathbf{D}$ in $\mathbf{H}$ to a certain extent. All bundles together enable the factorization of the model $\mathbf{H}$, which is why we speak of a Boolean matrix factorization (BMF). For this purpose, the following association rule is applied [9]: $H_{ij} = \bigoplus_r^R = b_{ir} c_{jr}$.

The step for capturing structural regularities of a given data matrix $\mathbf{D}$ using a method for factorizing a model matrix $\mathbf{H}$ is called *model order selection* [6]. In a pre-step, re-ordering $\mathbf{D}$ by permutating its columns and / or rows simplifies the model order selection process. Alternating least squares takes an expected model rank value r and constructs a model matrix $\mathbf{H}$ minimizing the following loss function [9, 3]: $L = \sum_{i=1}^{X} \sum_{j=1}^{Y} (d_{ij} - h_{ij})^2$.

Model order selection, which was introduced as part of HICLAS [4] in close connection with contextual bundling, originally used the Annealing Least-Squares method [9]. Contextual bundling is about calculating the fit between model $\mathbf{H}$ and data $\mathbf{D}$ for a certain object $x \in X$, attribute $y \in Y$ or over all objects and attributes model-wise. The Jaccard goodness of fit J [10] is a model data fit measure used together with HICLAS [1, 3]. It takes values between 0 and 1 and is calculated as follows:

$$J = \frac{\mathfrak{n}_{D=1,H=1}}{\mathfrak{n}_{D=1,H=1} + \mathfrak{n}_{D=0,H=1} + \mathfrak{n}_{D=1,H=0}}$$

3. Method

The Leuven Concept Data [2] contains both extensive human-created binary object-attribute matrices and human-assigned typicality ratings, feature generation frequencies, and feature importance ratings. The Leuven Concept Data consist of 16 normed semantic categories, with 5 categories (i.e., Birds, Fish, Insects, Mammals, Reptiles) grouped under Animals and 6 categories (i.e., Clothing, Kitchen utensils, Musical instruments, Tools, Vehicles, Weapons) grouped under Artifacts. Respondents were asked to list either attributes for a specific category (so-called category attributes or category-level) or for a specific exemplar (exemplar attributes or exemplar level). These two types of attributes differ from one another. Category attributes identified for the Fish category for example only partially overlap with the union of the attributes identified for the respective exemplars in that category. The former tend to be more abstract (e.g., for Artifacts "is different for men and women", "can be dangerous") than the latter ones (e.g., for Artifacts "works on electricity", "is round"), which are more concrete.

The result are two versions of the applicability matrices represented as binary object-attribute matrices: 'Exemplar by Category Attribute' matrices and 'Exemplar by Exemplar Attribute' matrices. Four participants independently completed each type of applicability matrix for the domains of animals and artifacts (i.e., the participants indicated whether objects possessed certain attributes or not). To obtain a single binary object-attribute matrix from these four matrices, at least two participants had to agree in order to disqualify an item under the majority rule. The typicality ratings range from 1 to 20, as they are calculated as averages of the typicality scores assigned by 112 respondents to the Animal and Artifact exemplars with respect to their respective categories (e.g., Fish, Vehicles) on a 20-point scale with higher values indicating higher typicality.

The relevance of the attributes in the applicability matrices is evaluated based on two measures of user-based attribute weighting: namely, the feature generation frequencies and the feature importance ratings. Feature generation frequencies can be determined when a feature generation task has been conducted with multiple participants. In general, feature generation frequencies are a distribution measure that indicates how often an attribute is generated across study participants [2]. The more participants associate an attribute with a stimulus, the more relevant the attribute is considered to be. Note also the associated description of how to perform the feature generation task. Participants are usually encouraged to name as many different attributes as possible that they consider characteristic. To normalize w on feature generation frequencies we take $w = w_i / \sum_i^n w_i$. Feature ratings are an alternative to feature generation frequencies to measure the relevance of attributes through direct user ratings based on a set of already generated attributes [11, 2]. Feature importance ratings are based on the assumption that different attributes may differ in their relevance to the category. The user is asked to rate how important or relevant they consider the attributes to be for the category on a 7-point Likert scale, ranging from -3 (for a very unimportant feature) to +3 (for a very important feature). We take the average of the feature importance ratings provided by all 12 participants in the Leuven Concept Data.

Feature generation frequencies were available for each attribute of the attribute sets of Animals and Artifacts. Since feature importance ratings are only available for each basic category of Animals and Artifacts, we filtered the attributes for each category separately. The set of attributes, objects, and the incidence density for the aggregated categories Animals and Artifacts in the Leuven Concept Data is described in Table 2:

Table 2
Leuven Concept Data overview of the domains Animals and Artifacts

category	data	objects	attributes	density
Animals	categories	129	225	.32
Animals	exemplar	129	764	.13
Artifacts	categories	166	301	.23
Artifacts	exemplar	166	1295	.09

In the following we call the original matrices present in the Leuven Concept Data for simplicity only

data and their related models matrices for short only model.

4. Outline

Attribute filtering has proven useful for capturing inter-categorical structure [12]. We therefore ask whether it is also beneficial for capturing intra-categorical structure. Our study is based on existing methodology proposed by Ceulemans and Storms [3], who use the least-squares method for domain order selection and a Jaccard goodness-of-fit model measure for contextual bundling (for more details, see also the Method section).

To investigate the advantage of attribute filtering, we select attribute subsets on the exemplar level in order to arrive at a comparable attribute set size as on the category level. The attribute filtering procedure selects a subcontext based on a certain attribute weight measure present for each attribute. We use feature generation frequencies and feature importance ratings as two popular user-based attribute weights as well a random attribute selection as a baseline approach. Attribute filtering based on the same metrics and considerations has been used in our previous analysis of inter-categorical structure [12]. These results show that feature importance ratings and feature generation frequencies indicate different aspects of attribute importance.

Since the processing pipeline involves domain order selection, categorization of exemplars into categories, and intra-categorical structuring, each of those steps must be examined separately. Descriptive measures such as number of constructed factors, relationship between factor count and degree of data approximation [1, 13] and (cumulative) coverage - $1 - \frac{E(I,H)}{||I||}$ [13] shed first light onto the quality of domain order selection. A more complete description of the model quality is provided by considering both the model coverage and the residual error [6]. Calculating the Jaccard goodness of fit between model and data matrix, as proposed by Storms et al. [1], is such a measure of model coverage that incorporates the residual error. Ceulemans and Storms [3] used this Jaccard goodness-of-fit and calculated its value for a model using the least-squares method. The only hyperparameter in this process is the number of bundles. They showed that as the number of bundles increases, the loss value indicating the remaining error and the model-wise Jaccard goodness-of-fit are inversely corresponding to each other. Based on this observation, the optimal model can be selected visually - the one where the slope of the model-fit curve flattens out - as this model offers the best balance between coverage, low error, and low model complexity.

Domain order selection provides a set of bundles which are able to approximate the existing data. Categorization afterwards investigates whether the taken bundles are meaningful and plausible (i.e. correspond to existing semantic categories). Belohlavek and Vychodil [13] checked whether the related attribute subset looks plausible for most general bundles (Behlohlavek and Vychodil refer to them as “factor concepts” rather than “bundles.”). In contrast, based on HICLAS analysis, Storms et. al. [1] described bundles based on most certain object classes and directly related attribute classes by taking the most typical and most a-typical exemplars into account. Furthermore, a correlation analysis of the object-specific Jaccard similarity across the entire set of objects and the corresponding human typicity ratings can reveal the extent to which the intra-category structure is reflected.

5. Analysis

Category results are presented at the category level `cat`, and the exemplar level `exe`, filtered by random values `exe(rnd)`, filtered by the feature generation frequencies `exe(frq)` and filtered by the feature importance ratings `exe(imp)`.

5.1. Model selection

To determine a valid model matrix, a suitable configuration is selected during the model selection process. In the Annealing Least Squares method, the number of bundles to be identified must first be

specified using rank r. Analogous to the analysis by Ceulemans and Storms [3], we consider models tested at ranks of $r = [1, 6]$. The quality of a solution is measured by a trade-off between model coverage - specified by the Jaccard goodness-of-fit (model-wise) - and the error ratio. The error ratio results from the bitlength of the error matrix E compared to the bitlength of the data D.

In the following, the results for Animals (Figure 1) and Artifacts (Figure 2) are examined separately. As in Ceulemans and Storms [3], the error curve decreases steeply at first and more gradually later on. In particular, for Animals at the category level and for exemplar data filtered based on feature importance ratings, a slightly better model fit is achieved than for exemplar data filtered at random or based on feature generation frequencies. For Animals, a valid model with 5 bundles van be observed for all variants.

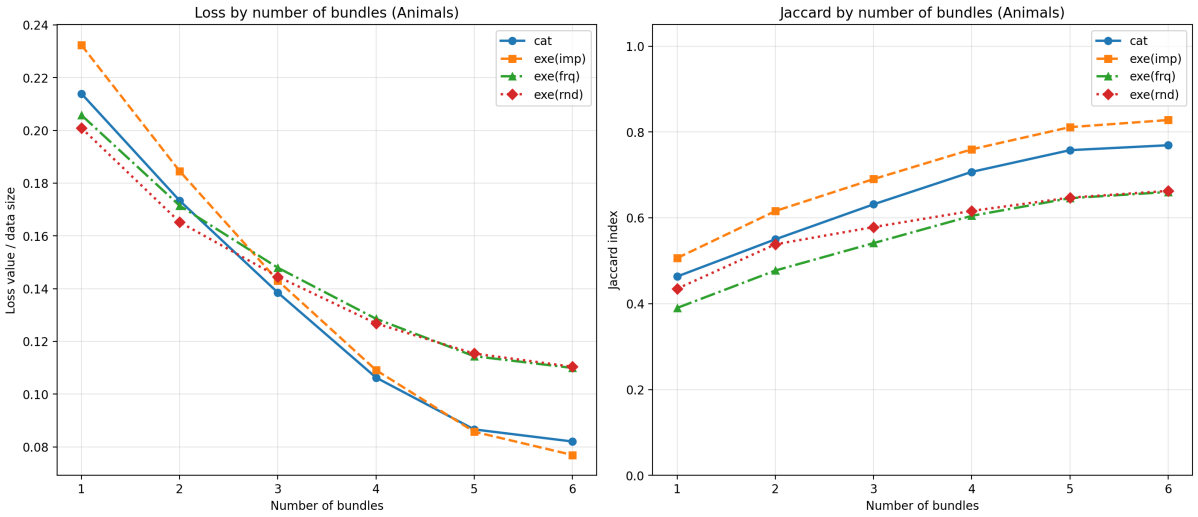

Figure 1: Loss function values (divided by data size) and Jaccard indices of the HICLAS solutions with 1 up to 6 bundles, for the Animals data.

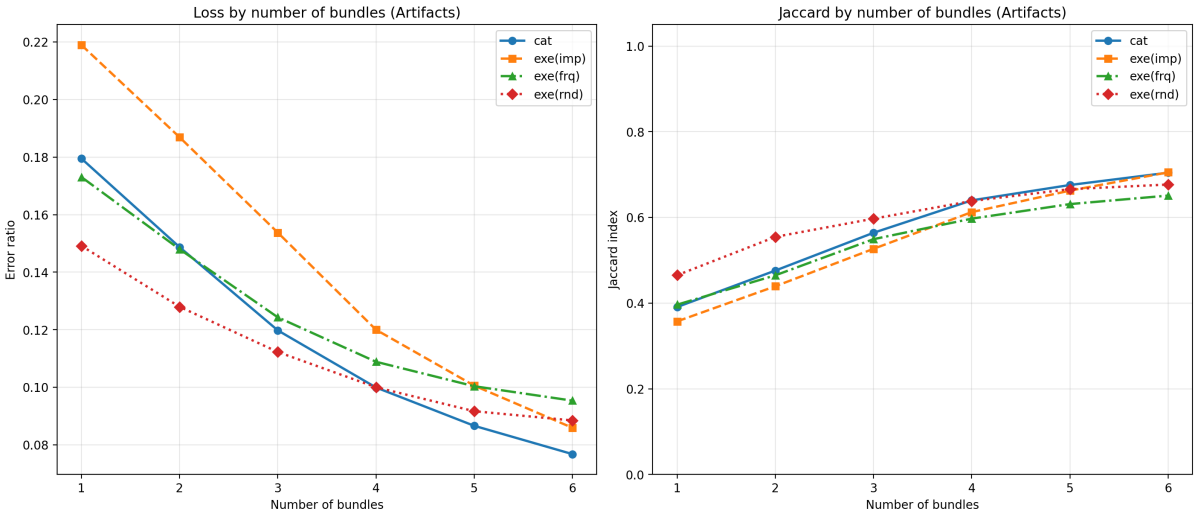

Figure 2: Loss function values (divided by data size) and Jaccard indices of the HICLAS solutions with 1 up to 6 bundles, for the Artifacts data.

If Artifact exemplar data is filtered randomly or based on feature generation frequencies, a valid model (elbow) is present with just 5 bundles. In contrast, for models at the category level or for exemplar data filtered by feature importance ratings, a valid model with 6 bundles appears to be more suitable. In both of these latter cases, no elbow is observed, but the low error rate indicates a valid model at this rank. We follow the recommendations of Ceulemans and Storms [3] and select a model with 5 bundles

for Animals and 6 bundles for Artifacts. Based on our observations as well, this should be the best option, with the exception of Artifacts filtered by random selection or feature generation frequencies.

5.2. Categorization

Based on selected models, we want to investigate first its categorization, specially the question for each bundle which of the exemplars is not matched to the major corresponding pre-defined category. To do this, we assign a target category to each bundle provided that the majority of examplars are assigned to it accordingly. A listing of false-positive exemplars and related categories is provided in Table 3.

Table 3: False-positive exemplars by type and category

Type	Category	False-positive exemplars
cat	Mammals	Dolphin (Fish), Orca (Fish), Whale (Fish)
exe (rnd)	Mammals	Dinosaur (Reptiles), Tortoise (Reptiles)
exe (rnd)	Birds	Squirrel (Mammals), Hamster (Mammals), Cat (Mammals), Rabbit (Mammals), Mouse (Mammals)
exe (rnd)	Insects	Hedgehog (Mammals), Bat (Mammals), Frog (Reptiles), Toad (Reptiles)
exe (rnd)	Fish	Turtle (Reptiles)
exe (frq)	Birds	Hamster (Mammals), Mouse (Mammals), Bat (Mammals)
exe (frq)	Fish	Turtle (Reptiles)
exe (frq)	Mammals	Tortoise (Reptiles)
exe (imp)	Mammals	Canary (Birds), Chicken (Birds), Cockhafer (Insects), Viper (Reptiles)
cat	Clothing	Apron (Kitchen utensils)
cat	Kitchen utensils	Vacuum cleaner (Tools)
cat	Tools	Knife (Kitchen utensils)
cat	Vehicles	Lawn mower (Tools), Tank (Weapons)
cat	Weapons	Go-cart (Vehicles), Cart (Vehicles), Carriage (Vehicles), (Hot air) Balloon (Vehicles), Skateboard (Vehicles), Sled (Vehicles), Kick scooter (Vehicles)
exe (rnd)	Kitchen utensils	Chisel (Tools), Adjustable spanner (Tools), Clamp (Tools), Wrench (Tools), Filling knife (Tools), Screwdriver (Tools), Nail (Tools), Wire brush (Tools), Tongs (Tools), Rope (Tools), Paint brush (Tools), File (Tools), Level (Tools)
exe (rnd)	Kitchen utensils (2)	Drill (Tools), Vacuum cleaner (Tools)
exe (rnd)	Vehicles	Anvil (Tools), Lawn mower (Tools), Wheelbarrow (Tools), Stove (Kitchen utensils), Fridge (Kitchen utensils), Tank (Weapons)
exe (rnd)	Weapons	Axe (tools), Crowbar (Tools), Hammer (Tools), Pickaxe (Tools), Plough (Tools), Plane (Tools), Shovel (Tools), Saw (Tools), Submarine (Vehicles), Knife (Kitchen utensils)
exe (frq)	Clothing	Vacuum cleaner (Tools), Towel (Kitchen utensils), Place mat (Kitchen utensils), Apron (Kitchen utensils)

Continued on next page

Type	Category	False-positive exemplars
exe (frq)	Kitchen utensils	Anvil (Tools), Wheelbarrow (Tools), Wrench (Tools), Plough (Tools), Wire brush (Tools), Sled (Vehicles), Kick scooter (Vehicles), Cymbals (Musical instruments), Triangle (Musical instruments), Knuckle dusters (Weapons), Shield (Weapons)
exe (frq)	Tools	Knife (Kitchen utensils)
exe (frq)	Vehicles	Lawn mower (Tools)
exe (imp)	Kitchen utensils	Anvil (Tools), Vacuum cleaner (Tools), Sled (vehicles)
exe (imp)	Clothing	Towel (Kitchen utensils), Apron (kitchen utensils)
exe (imp)	Tools	Knife (Kitchen utensils), Scissors (Kitchen utensils)

Same as Ceulemans and Storms [3] we observe for Animals a perfect match at the category level expect for Dolphin, Orca, and Whale which were placed in the Fish category in the Leuven Concept data, but here (correctly) categorized with Mammals. For Artifacts, at the category level there are many more outliers. Ceulemans and Storms explain that in terms of the hierarchical object classes. All of the outliers look like exemplars that are semantically very close to the target category. For example, an Apron is usually used as a Kitchen Utensil but may also be considered a piece of clothing, and a Go-Cart is a Vehicle but if people are using it maliciously could also be used as a weapon. The exemplar data filtered based on feature importance ratings contains a similar number of outliers as at the category level. Some of these outliers are plausible, such as classifying a Vacuum cleaner as a Kitchen utensils or Scissors as a Tool, others are not. Among the implausible outliers are an Anvil and a Sled, which were classified as Kitchen utensils and Viper classified with Mammals. Overall, it can be concluded that filtering based on feature importance ratings leads to a valid categorization, even if it appears to be slightly less accurate than categorization based on category data. Exemplar data filtered by randomly or feature generation frequencies leads to many more false-positive exemplars which are not explainable (e.g., an Adjustable Spanner for Kitchen Utensils). In addition, random attribute filtering is not able to map bundles to the expected target categories. The target category Kitchen utensils is represented by two bundles, whereas the Tools category is not covered by an specific bundle, for exemplar.

5.3. Reproducing results by Ceulemans and Storms at the category-level

The graded category structure was evaluated by examining the correspondence between human typicality ratings and Jaccard goodness-of-fit at the object level, using the Pearson correlation coefficient, or $\tilde{\rho}$. The corresponding significance values p are reported using the significance levels $* : p < .05, ** : p < .01, *** : p < .001$. Positive correlation values $\tilde{\rho} \geq 0$ indicate that a natural graded category structure is captured, and if $\tilde{\rho} < 0$, it is interpreted as non-natural. The correlations are calculated across all exemplars of a target category specified in accordance with the Leuven Concept Data.

The original results by Ceulemans and Storms were reported after excluding certain exemplars, namely Orca, Whale, Dolphin (all category Fish category) in the Animals domain, and Scissors, Knife (both Kitchen utensils), Axe, Drill, Hammer, Pickaxe, Crowbar, Shovel, Grinding disc, Knife, Lawnmower, Rope, Saw (all Tools), Go-cart, Cart, Skateboard, Sled, Kickscooter (all Vehicles), Axe, Tank (both Weapons) from the Artifacts domain. According to the original description, exemplars were excluded if they either did not fit into specific bundles or could be assigned to multiple bundles. To compare our results to the originally reported results, we first perform the correlation analysis based on the previously proposed exclusions and then calculate the correlations without excluding exemplars. Results reported by Ceulemans and Storms (2010) [3] are named `old (excl)`, our results with the same objects excluded `new (excl)`, and our results without excluding any objects `new (no-excl)`. The results are summarized in Table 4.

Table 4
Correlation results of goodness-of-fit (object level) compared to human typicality ratings on category level.

Category	old (excl)	new (excl)	new (no-excl)
Birds	.66**	.69**	.69**
Fish	.57**	.61*	.77**
Insects	.43*	.65**	.65**
Mammals	.04	.05	.05
Reptiles	.51*	.45*	.45*
Clothing	.64**	.67***	.67***
Kitchen utensils	–	.17	.16
Musical instruments	–	.22	.23
Tools	.60*	.47*	.33
Vehicles	–	.45*	.60**
Weapons	–	-.40	-.27

When we exclude exemplars of Fish from the Animals domain, we observe significant, moderate positive correlations for Birds ($.66 \rightarrow .69$), Fish ($.57 \rightarrow .61$), Insects ($.43 \rightarrow .65$) and Reptiles ($.51 \rightarrow .45$). For Mammals ($.04 \rightarrow .05$), we do not observe a significant correlation as was the case in Ceulemans and Storms. For Artifacts, correlation results were reported only for Clothing and Tools by Ceulemans and Storms. Taken into account the mentioned exclusions for Artifact exemplars, we observe moderate correlation results for these categories, similar for Clothing ($.64 \rightarrow .67$) and slighty worse for Tools ($.60 \rightarrow .47$). We observe only a weak positive significant correlation for Vehicles ($.45$). The remaining Artifact categories show no significant correlation results - Kitchen utensils ($.17$), Musical instruments ($.22$) - and a negative correlation was observed for Weapons ($-.40$).

If we ignore the exclusions and take the complete set of exemplars for Animals and Artifacts into account, we find mostly comparable results. Only for Tools does the correlation strength decrease and become insignificant. Suprisingly, for Fish ($.61 \rightarrow .77$) and Vehicles ($.45 \rightarrow .60$), we obtain a higher correlation strength. In general, it can be concluded that the exclusions are not necessary; without any exemplar exclusions, the observations are similar. Overall, we can conclude that we have successfully reproduced Ceulemans and Storms' results at the category level. As originally noted, the graded category structure can be observed for Animals with the exception of Mammals. Taking the correlation results for all categories of Artifacts into account, it follows that the graded structure cannot be captured overall.

5.4. Capturing the graded category structuring at the exemplar level

We now examine the correspondence between typicality predictions and human typicality ratings for the categories Animals and Artifacts at the exemplar level. We investigate whether the graded category structure can be captured by exemplar filtered data. Since the Annealing Least-Squares [9] methods becomes too inefficient when dealing with very large binary object attribute matrices, it is not possible to perform computations on the full exemplar data. The results are presented in Table 5. For the category-level, we reuse the results without any exclusion described in the last paragraph.

For the filtered exemplar data, there was no model in which the Jaccard goodness-of-fit (object level) could capture the graded category structure across all categories. For Birds, the results are comparable to the category level. For Fish, the results on the category level slightly outperform the filtered data. For Insects and Reptiles, the results are much worse at the exemplar level than at the category level (no significant correlations). Much like the results at the category level, the exemplar results for Mammals are poor. For Artifacts the pattern is different than for Animals in that predictions tend to be better at the exemplar level than at the category level. Not only are the typicality gradients of Clothing and Vehicles captured to similar Clothing is a bit worse at the exemplar level as they were at the category level, the internal structure of the categories Musical instruments and Weapons is also captured at the

Table 5
Correlation results of Jaccard goodness-of-fit (object level) compared to human typicality ratings.

Category	cat	exe (rnd)	exe (frq)	exe (imp)
Birds	.69**	.69***	**.71***	.36
Fish	**.77****	.63**	.71***	.67***
Insects	**.65****	.38	-.22	.15
Mammals	.05	-.25	-.01	.03
Reptiles	**.45***	.16	.21	.07
Clothing	**.67*****	.40*	.18	.54**
Kitchen utensils	.16	.21	-.15	.08
Musical instruments	.23	.32	.22	**.38***
Tools	.33	.58	**.45***	.27
Vehicles	.60**	.67***	.70***	**.77*****
Weapons	-.27	.65**	**.80*****	.66**

exemplar level (unlike at the category level). The internal structure of Kitchen utensils and Tools is not captured at either level. The general observation (which also applies to Animals) is that random selection does fairly well. This suggests that the attribute weights do not carry particularly much informational value compared to a random selection. Consistent to our previous observations [12], we see that filtering by feature generation frequencies and filtering by feature importance ratings carry different information, in that they do not always lead to comparable correlations.

6. Discussion

In this contribution, we have examined the influence of attribute filtering on the quality of intra-categorical structuring. Following the approach proposed by Storms and colleagues [1], we group incidence data by calculating a model matrix through HICLAS and interpret the intra-categorical structure as the correspondence between the model matrix and the original matrix on object level basis within the constructed bundles. To ensure comparability with the reference work by Ceulemans and Storms (2010) [3] we follow their methodology and use the Leuven Concept Data [2] as dataset. Both for incidence data based on attributes generated toward the categories and toward the exemplars, we were able to identify valid models. Existing results for categorization as well as intra-categorical structure reported by Ceulemans and Storms (2010) [3] were reproduced. Filtering exemplar attributes based on feature importance ratings results in a similar number of outliers (false-positives) as at the category level. However, categorization based on filtered exemplar data in general is much more questionable than at the category level, as it produces many more implausible false-positive category assignments. When it comes to representing the intra-categorical structure, the results are less clear-cut. Attribute filtering was somewhat helpful for the Artifacts domain, but not for Animals. It should be noted that random attribute filtering sometimes achieves results comparable to those of systematic attribute filtering approaches, raising concerns about the informational value of the used attribute weights.

The introduction of approximate binary model matrices enables to construct a considerably smaller number of factors which may account for a large portion of the data [13]. It means also that not only the inter-categorical structure but also the intra-categorical structure is considered. Contextual bundling according to Storms et. al. [1] is a promising approach, as it allows for capturing the graded category structure in binary data while fully accounting for the model information about concepts, which we described as contextual bundling. Attribute filtering - that is, the selection of a submatrix based on attribute weights - is a suitable approach for greatly reducing the attribute space. The role of attribute filtering in binary object-attribute matrices has already been investigated with regard to inter-categorical structure [12]. This earlier work highlighted, an important role of attribute filtering in inter-categorical structuring. Only by filtering the most relevant attributes at the exemplar level was there a match between user-based attribute weights and attribute derivations. Filtering the most

important attributes yielded the best results in this work, while random attribute filtering performed better than considering all attributes. This last observation suggests a negative influence resulting from the fact that many irrelevant attributes distort the space. We did not find a similar result as for inter-category structure in that attribute filtering at the exemplar level did not prove very successful for capturing typicality. Previous studies focused on data representing a single category, whereas the present study examined data with multiple categories in order to investigate not only the intra-categorical structure but also the quality of the categorization. This study shows that a model that is valid in terms of model fit and error rate does not necessarily result in a plausible categorization and does not necessarily capture intra-categorical structure well. The added value of attribute filtering is limited because, compared to category-level categorization, it does not ensure robust categorization that contains only a few explainable outliers. Nevertheless, the basic idea behind attribute filtering - focusing on relevant attributes - has some promise, as it enables the capture of graded structure to some extent for categories where contextual bundling at the category level failed. To enable a more complete comparison, it will be necessary to also consider unfiltered data at the exemplar level. Since model order selection based on the Annealing Least Squares method [9] is too inefficient, other, more efficient alternatives (e.g., ASSO [6] or GreCond [13]) must also be considered.

7. Declaration on Generative AI

The authors used DeepL, Perplexity for grammar correction, Latex code refinement and Latex, MatLab, Python source code bug fixing. All AI-assisted output was checked, revised, and integrated by the authors. The authors remain fully responsible for the accuracy, originality, and integrity of the paper.

References

[1] G. Storms, I. V. Mechelen, P. De Boeck, Structural analysis of the intension and extension of semantic concepts, European Journal of Cognitive Psychology 6 (1994) 43–75.

[2] 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, Behavior Research Methods 40 (2008) 1030–1048.

[3] E. Ceulemans, G. Storms, Detecting intra-and inter-categorical structure in semantic concepts using hiclas, Acta Psychologica 133 (2010) 296–304.

[4] P. De Boeck, S. Rosenberg, Hierarchical classes: Model and data analysis, Psychometrika 53 (1988) 361–381. doi:10.1007/BF02294218.

[5] T. Hanika, M. Koyda, G. Stumme, Relevant attributes in formal contexts, in: Graph-Based Representation and Reasoning: 24th International Conference on Conceptual Structures, ICCS 2019, Marburg, Germany, July 1–4, 2019, Proceedings 24, Springer International Publishing, 2019, pp. 102–116.

[6] P. Miettinen, J. Vreeken, Mdl4bmf: Minimum description length for boolean matrix factorization, ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (2014) 1–31. doi:10.1145/2629680.

[7] R. Belohlavek, J. Macko, Selecting important concepts using weights, in: International Conference on Formal Concept Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 65–80.

[8] R. Belohlavek, Optimal decompositions of matrices with entries from residuated lattices, Journal of Logic and Computation 22 (2012) 1405–1425. doi:10.1093/logcom/exr037.

[9] E. Ceulemans, I. Van Mechelen, I. Leenen, The local minima problem in hierarchical classes analysis: An evaluation of a simulated annealing algorithm and various multistart procedures, Psychometrika 72 (2007) 377–391. doi:10.1007/s11336-005-1496-7.

[10] P. H. Sneath, R. R. Sokal, Numerical Taxonomy: The Principles and Practice of Numerical Classification, W.H. Freeman and Company, San Francisco, 1973.

[11] J. A. Hampton, Inheritance of attributes in natural concept conjunctions, Memory & Cognition 15 (1987) 55–71.

[12] S. Schneider, A. Schulz, S. Verheyen, A. Nürnberger, Limitations of attribute derivations for approximating user-based attribute weighting, in: International Joint Conference on Conceptual Knowledge Structures, Springer Nature Switzerland, Cham, 2025, pp. 67–90.

[13] R. Belohlavek, V. Vychodil, Discovery of optimal factors in binary data via a novel method of matrix decomposition, Journal of Computer and System Sciences 76 (2010) 3–20. doi:10.1016/j.jcss.2009.05.002.