Conceptual Data Scaling in Machine Learning

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
(Dr. rer. nat.)

Vorgelegt im
Fachbereich 16
Elektrotechnik/Informatik
der
Universität Kassel

von

Johannes Hirth

Gutachter
Prof. Dr. Gerd Stumme, Universität Kassel
Prof. Dr. Robert Jäschke, Humboldt-Universität zu Berlin

Eingereicht am 15. Dezember 2023
Verteidigt am 26. September 2024

Created using Free Software. In particular LaTeX, TikZ, git, GNU emacs and GNU/Linux.

Scientific Papers that Contribute to this Thesis

In this chapter, we list the publications and pre-prints that are incorporated into this thesis. The works from the first list were envisioned and first authored by the author of this thesis. The writing, reviewing and editing of these them was supervised by Tom Hanika.

[91] T. Hanika and J. Hirth. "On the lattice of conceptual measurements". In: *Information Sciences* 613 (2022), pp. 453–468

[89] T. Hanika and J. Hirth. "Exploring Scale-Measures of Data Sets". In: *Formal Concept Analysis - 16th International Conference, ICFCA 2021, Strasbourg, France, June 29 - July 2, 2021, Proceedings*. Ed. by A. Braud, A. Buzmakov, T. Hanika, and F. L. Ber. Vol. 12733. Lecture Notes in Computer Science. Springer, 2021, pp. 261–269

[92] T. Hanika and J. Hirth. "Quantifying the Conceptual Error in Dimensionality Reduction". In: *Graph-Based Representation and Reasoning - 26th International Conference on Conceptual Structures, ICCS 2021, 2021, Proceedings*. Ed. by T. Braun, M. Gehrke, T. Hanika, and N. Hernandez. Vol. 12879. Lecture Notes in Computer Science. Springer, 2021, pp. 105–118

[87] T. Hanika and J. Hirth. "Conceptual views on tree ensemble classifiers". In: *International Journal of Approximate Reasoning* 159 (2023), p. 108930. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2023.108930

[98] J. Hirth and T. Hanika. *Formal Conceptual Views in Neural Networks*. 2022. arXiv: 2209.13517 [cs.LG]

[99] J. Hirth and T. Hanika. *The Geometric Structure of Topic Models*. 2024. DOI: 10.48550/arxiv.2403.03607. arXiv: 2403.03607 [cs.AI]

The works from the next list were envisioned and first authored by the author of this thesis. The writing, reviewing and editing of them were supervised by Tom Hanika and Bernhard Ganter.

[77] B. Ganter, T. Hanika, and J. Hirth. "Scaling Dimension". In: *Formal Concept Analysis - 17th International Conference, ICFCA 2023, Kassel, Germany, July 17-21, 2023, Proceedings*. Ed. by D. Dürrschnabel and D. López-Rodríguez. Vol. 13934. Lecture Notes in Computer Science. Springer, 2023, pp. 64–77

The works from the next list were envisioned and first authored by the author of this thesis with the exception of Section 5 *Human-Centered Textual Explanations* from Hirth, Horn, Stumme, and Hanika [100]. Said section was first authored by Viktoria Horn. The textual explanations in Section 5 *Human-Centered Textual Explanations* are joint work between the author of this thesis and Viktoria Horn. The writing, reviewing and editing were supervised by Tom Hanika and Gerd Stumme with helpful discussions with Viktoria Horn.

[101] J. Hirth, V. Horn, G. Stumme, and T. Hanika. "Ordinal Motifs in Lattices". In: *Information Sciences* (2023), p. 120009

[100] J. Hirth, V. Horn, G. Stumme, and T. Hanika. "Automatic Textual Explanations of Concept Lattices". In: *Graph-Based Representation and Reasoning*. Ed. by M. Ojeda-Aciego, K. Sauerwald, and R. Jäschke. Cham: Springer Nature Switzerland, 2023, pp. 138–152

Besides these articles the author of this work contributed to the following publications that are not incorporated into this thesis:

[90] T. Hanika and J. Hirth. "Knowledge cores in large formal contexts". In: *Annals of Mathematics and Artificial Intelligence* 90 (2022), pp. 537–567

[88] T. Hanika and J. Hirth. "Conexp-Clj - A Research Tool for FCA.". In: *ICFCA (Supplements)*. Ed. by D. Cristea et al. Vol. 2378. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 70–75

Other research articles that the author of this work contributed to are:

[107] V. Horn et al. "Disclosing Diverse Perspectives of News Articles for Navigating between Online Journalism Content". In: *Nordic Conference on Human-Computer Interaction*. NordiCHI 2024. Uppsala, Sweden: Association for Computing Machinery, 2024. ISBN: 9798400709661. DOI: 10.1145/3679318.3685414

[190] B. Schäfermeier, J. Hirth, and T. Hanika. "Research Topic Flows in Co-Authorship Networks". In: *Scientometrics* (Oct. 2022)

[214] M. Uhlmann, J. Hirth, and V. Horn. *Jenseits der Logik der Empfehlung: Formale Begriffsanalyse als Grundlage für eine neue Variante zur Vermittlung von Nutzenden und journalistischen Inhalten*. Submitted. 2023

To Uhlmann, Hirth, and Horn [214] the author of this work contributed the scaling and encoding of the analyzed data.

Chapter Contributions

The chapters of this work are comprised of the works listed above and contain complete or partially verbatim, or paraphrased quotes.

Chapter 7: Conceptual Data Scaling The introductory part in Section 7.1 is recalled from the literature [74, 80], whereas the additional notions on pre-scaling in Section 7.1.1 are from our work in Ganter, Hanika, and Hirth [77]. Our thoughts on the scaling of categorical values in Section 7.1.2 are based on works from the literature. The separation of conceptual scaling and conceptual data reduction in Section 7.3 is a new addition. Employing scale-measures for conceptual data reduction in Section 7.2 includes parts from Ganter, Hanika, and Hirth [77], Hanika and Hirth [91], and Hirth, Horn, Stumme, and Hanika [101]. Section 7.5 is based on our method presented in Hanika and Hirth [92]. Results on inverse plain scaling in Section 7.4 are recalled from the literature with the exception of the proof of Theorem 5 and Proposition 9. With these additions, we made parts of the inverse scaling procedure more explicit.

Chapter 8: Navigating Conceptual Views Sections 8.2 and 8.3 are based on Hanika and Hirth [91], Section 8.4 on Hanika and Hirth [89] and Section 8.2.3 is based on my contributions to Uhlmann, Hirth, and Horn [214]. All other sections in this chapter, i.e, Sections 8.1 and 8.5 to 8.7 are based on Hanika and Hirth [89, 91].

Chapter 9: Ordinal Motifs in Lattices Chapter 9 is based on Hirth, Horn, Stumme, and Hanika [100, 101]. Proposition 26 and the ordinal and interordinal case in Proposition 27 are new additions. Section 9.4.1 is based on Section 5 from Hirth, Horn, Stumme, and Hanika [100], which was first authored by Viktoria Horn. We summarized and reproduced this section faithfully. The textual templates in Section 9.4.1 are joint work between the author of this thesis and Viktoria Horn.

Chapter 10: The Complexity of Conceptual Views Section 10.1 are from Hanika and Hirth [91]. All other sections, i.e, Sections 10.2 to 10.4 are from Ganter, Hanika, and Hirth [77]. Section 10.2.3 is a new addition. Section 10.3 includes a more comprehensive analysis compared to Ganter, Hanika, and Hirth [77] with the addition of the inverse scaling results.

Chapter 11: Conceptual Scaling Error in Dimensionality Reduction Chapter 11 is based on Hanika and Hirth [92].

Chapter 12: Conceptual Views on Tree Classifiers Chapter 12 is based on Hanika and Hirth [87].

Chapter 13: Conceptual Views on Neural Networks Chapter 13 is based on Hirth and Hanika [98].

Chapter 14: The Geometric Structure of Topic Models Chapter 14 is based on Hirth and Hanika [99].

Preface

Information can be measured and represented on many different scales and encoded in multiple formats. The operations that can be used on the data depend on the type of scale used in measurement. Some scales allow for combining values with numerical operations while others are only equipped with relational information, e.g., (hierarchical) order relations. There are at least two principles on how to deal with data that is given in mixed formats with different scale types. The first is to artificially define operations on the measured scales that are needed for an analytical method, e.g., numeric addition and multiplication on categorical values. While this approach is very successful in the field of machine learning, the artificially introduced operations lack explainability with respect to the data domain. Thus, resulting in black box models. The second principle is to perform an analysis solely based on operations that are common to all scales. These, often more algebraic, operations are consistent to the underlying scales and allow for a rich interpretation of analytical results.

With this work, we focus on the latter principle. We analyze data with ordinal methods based on the Formal Concept Analysis framework. An analytical tool of this framework are *valid implications* within a symbolic data domains. These implications are rules of the form *"if . . . then . . . "* and the *abstract concepts* they generate, i.e., closed sets of properties to which no further implication can be applied. Implications and concepts can not only be used as unit of information but also as tool of reasoning to extend a set of properties A by a set B that logically follows from A. The set of concepts is equipped with an underlying hierarchical (order) structure to which we refer to as *conceptual data (structure)*.

Data that is naturally not measured in terms of symbolic properties can be mapped to such domains. For example, for numeric data one may use thresholds δ in combination with comparisons $\leq$ to derive symbolic properties. This process is known as *conceptual data scaling*. The resulting concept hierarchy and the process of deriving them are the core topics of our work. Of special interest to us is the application of conceptual data scaling to data representations in machine learning.

Throughout all our studies we lay a great emphasis on the explainability of these structures. Explaining here means deriving *models* that predict or describe the inherent conceptual structure of a data set or parts of it. These models are themselves conceptual structures with human interpretable features that are possibly more abstract or simpler. We approach the explainability aspect from many different perspectives: models whose structure have a specific semantic, textual explanations generated with principles from human-computer interaction, logical combinations of symbolic properties, a geometric view on conceptual structures and a decomposition into explainable parts. This extends to our presented applications in the realm of machine learning. We not only demonstrate the applicability of our methods but also interpret what our findings mean for the studied machine learning algorithms.

Acknowledgments

This thesis is the result of many years of personal growth, study and research and would not have been possible without the many people that accompanied me during this time.

I am very grateful to the Knowledge and Data Engineering group at the University of Kassel. From my days as an undergraduate student up until now the group has always been a supportive and welcoming learn and research environment. Over all this time, I cannot remember a single day when I did not enjoy my time with the group. With many former and present members I shared countless fruitful discussions on various topics, fun table soccer matches and coffee sessions.

In particular, I want to thank Maximilian Stubbemann and Dominik Dürrschnabel who made the group a fun place with a lot of – never aging – humor. Tobias Hille for stimulating discussions, the development of efficient conversation protocols and Emacs tinkering. My co-authors Viktoria Horn and Bastian Schäfermeier with whom I had many helpful discussions and who refined my research with new perspectives. All the students that I have worked with as teaching assistant. Bernhard Ganter with whom I had many inspiring conversations during our weekly seminars and paper writing sessions.

Björn Fries for maintaining a very productive technical setup for the group and remarkably fast solutions for technical problems. Monika Vopicka who helped me maneuver safely through the administrative jungle. My supervisor Gerd Stumme who gave me the opportunity to pursue my studies, for providing a research environment in which a free scientific spirit is lived out and for always supporting my endeavors. I am especially grateful to Tom Hanika who has been a mentor for me from my days as undergraduate student onward with countless inspiring discussions, paper writing sessions and support in general.

Last but certainly not least, I thank my family for their unconditional support and my partner Julia Marx for always being there for me and for filling my life with joy and positivity.

Contents

Content Lattice

The linear structure of printed documents require that the highly interrelated pieces of this work are presented in list form, i.e., a *linear order*. The result of this linearization can be seen in the table of contents and contains sixteen chapters and four (*disjoint*) parts.

In addition to the table of contents we want to show the reader our view on how the chapters of this thesis are interconnected. For this we propose the **content lattice**, a hierarchical structure that groups the chapters based on meta-topics. The content lattice for this thesis is displayed on the right page and is represented as a line diagram. The diagram includes all chapters from the main part of this thesis. There are two reading rules that are required to understand the content lattice:

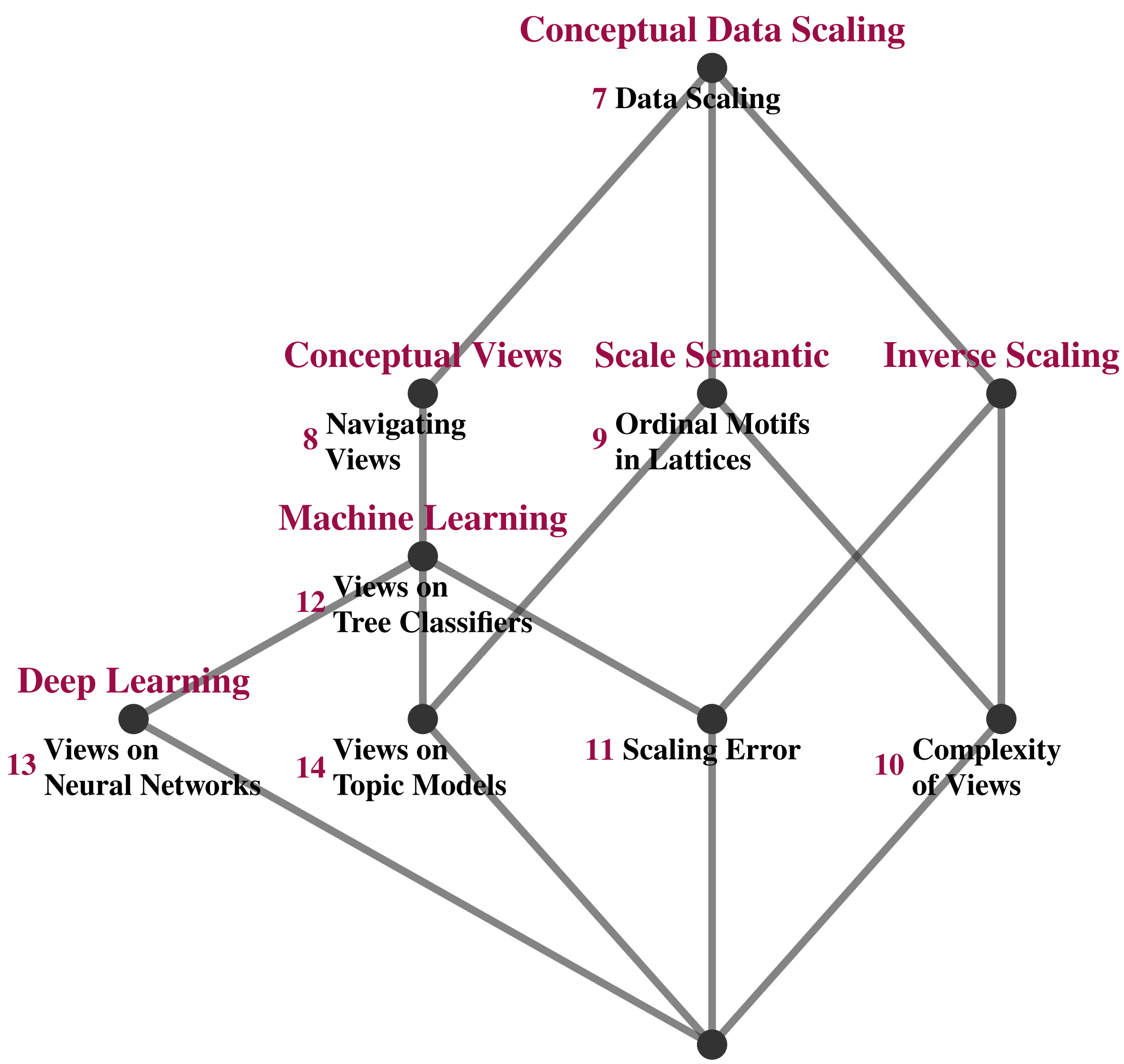
i) A chapter has all topics that can be reached by following upward paths.

ii) A topic is included in all chapters that can be reached by following downward paths.

For example, chapter nine *Ordinal Motifs in Lattices* is about *Scale Semantic* and *Conceptual Views*. The topic of *Conceptual Views* is included in five chapters, i.e., chapter *eight*, *eleven*, *twelve*, *thirteen* and *fourteen*, of which four are also about *Machine Learning*. Readers that open this thesis in PDF format can also use the content lattice as a tool to navigate this thesis by clicking on the chapter titles.

In this thesis, the content lattice not only serves as a means to describe the content, but as a subject of investigation itself. The overall research field of *Conceptual Data Scaling* is composed of three lines of research, the generation of conceptual views via scaling (*Conceptual Views*), the interpretation of the applied scaling (*Scale Semantic*) and the inversion of the applied scaling (*Inverse Scaling*). These three topics are – besides *Conceptual Data Scaling* – the top most in the diagram and build the main dimensions of this works content. Within the field of *Conceptual Views* we study the developed notions in the field of *Machine Learning*, which can be inferred from the lower positions in the diagram. The *view* that is reflected by the table of contents is retrieved from the content lattice as follows: part two (Chapter 7 to 10) deals with the theoretical foundation of this work, which is reflected by the *outer* nodes that form a cube like structure. The third part (Chapter 11 to 14) deals with the application of the introduced methods in the field of machine learning. The latter is reflected by the *inner* nodes below the *Machine Learning* topic.

On the title page of part two and three we present the content lattice restricted to the chapters and topics of the given part. This provides a *local view* on the respective part.



Conceptual Data Scaling
7 Data Scaling
Conceptual Views
8 Navigating Views
Scale Semantic
9 Ordinal Motifs in Lattices
Inverse Scaling
Machine Learning
12 Views on Tree Classifiers
Deep Learning
13 Views on Neural Networks
14 Views on Topic Models
11 Scaling Error
10 Complexity of Views

Abstract

Information that is intended for human interpretation is frequently represented in a structured manner. This allows for a navigation between individual pieces to find, connect or combine information to gain new insights. Within a structure, we derive knowledge from inference of hierarchical or logical relations between data objects. For unstructured data there are numerous methods to define a data schema based on user interpretations. Afterward, data objects can be aggregated to derive (hierarchical) structures based on common properties.

There are four main challenges with respect to the explainability of the derived structures. First, formal procedures are needed to infer knowledge about the data set, or parts of it, from hierarchical structures. Second, what does knowledge inferred from a structure imply for the data set it was derived from? Third, structures may be incomprehensibly large for human interpretation. Methods are needed to reduce structures to smaller representations in a consistent, comprehensible manner that provides control over possibly introduced error. Forth, the original data set does not need to have interpretable features and thus only allow for the inference of structural properties. In order to extract information based on real world properties, we need methods that are able to add such properties.

With the presented work, we address these challenges using and extending the rich tool-set of Formal Concept Analysis. Here, data objects are aggregated to closed sets called formal concepts based on (unary) symbolic attributes that they have in common. The process of deriving symbolic attributes is called conceptual scaling and depends on the interpretation of the data by the analyst. The resulting hierarchical structure of concepts is called concept lattice.

To infer knowledge from the concept lattice structures we introduce new methods based on sub-structures that are of standardized shape, called ordinal motifs. This novel method allows us to explain the structure of a concept lattice based on geometric aspects.

Throughout our work, we focus on data representations from multiple state-of-the-art machine learning algorithms. In all cases, we elaborate extensively on how to interpret these models through derived concept lattices and develop scaling procedures specific to each algorithm. Some of the considered models are black-box models whose internal data representations are numeric with no clear real world semantics. For these, we present a method to link background knowledge to the concept lattice structure.

To reduce the complexity of concept lattices we provide a new theoretical framework that allows us to generate (small) views on a concept lattice. These enable more selective and comprehensibly sized explanations for data parts that are of interest. In addition to that, we introduce methods to combine and subtract views from each other, and to identify missing or incorrect parts.

Zusammenfassung

Informationen werden häufig strukturiert repräsentiert, um für Analyst:innen besser verständlich zu sein. Die Struktur dient hierbei nicht nur zur Navigation zwischen Inhalten, sondern erlaubt es auch Informationen zu verknüpfen und zu kombinieren. Innerhalb einer Struktur lässt sich Wissen von strukturellen/hierarchischen Eigenschaften und logischen Verknüpfungen ableiten. Für Daten ohne klare Struktur existieren zahlreiche Methoden, um diese zu interpretieren und daraus ein Datenschema abzuleiten. Anschließend können diese, basierend auf gemeinsamen Eigenschaften, zu hierarchischen Strukturen aggregiert und zusammengefasst werden.

In diesem Umfeld identifizieren wir vier Probleme, bezogen auf die Erklärbarkeit der entstehenden Strukturen. Zuerst werden formale Methoden benötigt, um Wissen von hierarchischen Strukturen, oder Teilen von ihnen, abzuleiten. Darüber hinaus stellt sich die Frage, was das abgeleitete Wissen über die zugrunde liegenden Daten aussagt. Ein weiteres Problem ist die Größe der entstehenden Strukturen. Diese ist nicht immer in einer Größenordnung, die sich von Analyst:innen überblicken lässt. Hierfür werden Methoden benötigt, um die Größe in einer konsistenten und kontrollierten Art zu reduzieren. Für Daten, deren Attribute keine klare Bedeutung haben, braucht es zusätzliche Methoden, um Zusammenhänge mit Hilfe von interpretierbarem Hintergrundwissen zu erklären.

In dieser Arbeit untersuchen wir diese Probleme im Bereich der Formalen Begriffsanalyse. In dieser werden Datenelemente anhand von (unären) symbolischen Eigenschaften, die sie gemeinsam haben, zu Begriffen gruppiert. Der Vorgang, um symbolischen Eigenschaften aus Daten abzuleiten, nennt sich Begriffliche Skalierung. Die resultierende Hierarchie von Begriffen heißt Begriffsverband.

Um Wissen von Begriffsverbänden abzuleiten, haben wir neue Methoden basierend auf der Erkennung von Teil-Strukturen, die eine bestimmte Form haben, genannt *Ordinal Motifs*, entwickelt. Diese erlauben uns, die Struktur, die zwischen Daten Elementen vorliegt, zu erfassen und zu beschreiben. Zudem erläutern wir, wie wir aus *Ordinal Motifs* geometrische Eigenschaften ableiten und damit Begriffsverbände erklären können.

In unseren Analysen verwenden wir Datensätze, die aus den internen Repräsentationen von State-of-the-Art Modellen des Maschinellen Lernens gewonnen wurden. In unseren Analysen gehen wir besonders darauf ein, wie wir Erkenntnisse über die Modelle basierend auf abgeleiteten Begriffsstrukturen gewinnen können. Manche Modelle benutzen numerische Datenrepräsentationen, dessen Werte keine klare Echt-Welt Bedeutung haben. Hierfür stellen wir Methoden vor, um Erklärungen mit Hintergrundwissen zu finden.

Um die Komplexität von Begriffsverbänden zu reduzieren, stellen wir eine neue Methode vor, um kleinere *Sichten* zu generieren. Diese ermöglichen eine gezielte Betrachtung eines Teils des Begriffsverbands. Zusätzlich stellen wir Operationen vor, um Sichten miteinander zu verrechnen und fehlendes Wissen zu identifizieren.

1
Introduction

An important aspect of formal knowledge representations is that they should be understandable and processable by humans and computers. In order to do so, they often incorporate findings from psychology on mental knowledge representations and processing in their design [192]. This results in the definition of formal structures that allow for an organized and consistent manner of storing and accessing information. The derived structures can get complex very quickly, since they incorporate information in the form of relational, hierarchical and logical representations [167]. Prominent knowledge structures implementing these concepts are knowledge graphs [103] and ontologies [36].

Knowledge representations find application in many areas such as for automatic problem-solving in artificial intelligence [187], the development of computer expert systems [111] or to enrich modern large language models with (factual) background knowledge [225]. For human users they are used for searching and accessing information. In addition to that, knowledge structures are used as a means of navigation from one information unit (entry) to another. This enables one to relate information to each other and put information into context. During the navigation one may find more specific, more general or related information. These actions support human understanding and the inference of new knowledge from data.

Another, well studied, class of knowledge representations is based on the human way of thinking in *abstract concepts*. There are multiple definitions to what concepts are ranging from mental representations to abstract objects [151]. These *abstract concepts* are organized in a hierarchy with more general super-concepts and more specific sub-concepts. One theoretical framework that formalizes such conceptual structures [82] are concept lattices from Formal Concept Analysis [80, 223] (FCA). At first glance, a restriction to conceptual structures may seem limiting compared to the expressiveness of knowledge graphs or ontologies. Yet, their simplicity makes them possibly easier to comprehend by humans. On top of that are most problems in FCA computationally feasible while problems of other knowledge representation may not even be decidable.

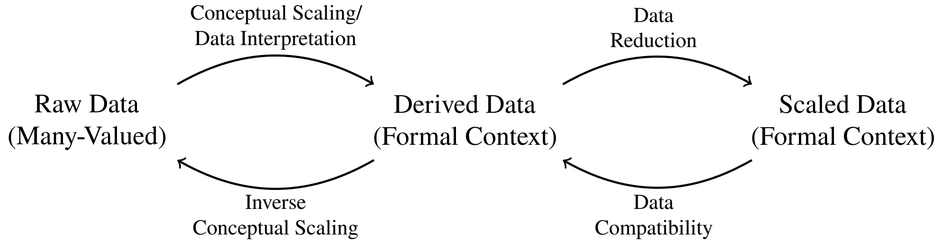


Figure 1.1: An overview of the four main problems in conceptual scaling.

1.1 Conceptual Data Scaling

A particular strength of Formal Concept Analysis is that it can process data from various heterogeneous sources and derive conceptual structures from them. The attributes in a data set can be measured on different scales, whose values allow for different means of comparison. A well-known classification of scales are the *levels of measurements* [201]. These include four levels with increasing capabilities. The first is the *nominal* level which includes all scales whose values can be compared for equality = and inequality ≠. All scales of measurement are included in this category. The second is the *ordinal* level which contains scales that are equipped with an order relation ≤, ≥ on its values. From an order theoretic point of view is the equality relation an order relation as well. Thus, we consider the nominal level as a special case within the ordinal level and therefore, the ordinal level to be the lowest level of measurement. The third level is the *interval* level which allows for a quantification of difference between values. The fourth level is the *ratio* level and extends the interval level by a notion of ratio between values. Decisive for the classification of a scale is not if we are able to come up with a distance measure between values, but if the distance emerges naturally from the scale values and their real-world semantics. For example, the numbers on uniforms in sports are numbers, but their difference has no meaning.

Most machine learning techniques are defined using a vector space model and therefore operate on the ratio level. They often achieve this by mapping scales from lower levels to a numeric scale. *Representational Theory of Measurement* (RTM) [145, 168, 210] provides a formalizing and understanding of this process. RTM relies on homomorphisms from an (empirical) relational structure to a numerical relational structure, where the numerical structure is often chosen to be the real line $\mathbb{R}$ or an n dimensional vector space on it. By performing numeric calculations with the resulting representations, they implicitly define operations on the data that – although they perform well in machine learning – have no real world meaning. This leads to uninterpretable black-box models.

With respect to the levels of measurement, the RTM procedure can be understood as an *up-scaling* of the data. This is in contrast to approaches from Formal Concept Analysis. Here, we interpret data on the ordinal level by a process called **conceptual data scaling**. Attributes that are measured on the interval and ratio level are *down-scaled* to the ordinal level by omitting distance and ratio functions. Analysis of the resulting ordinal data sets can be performed with algebraic methods that are interpretable by design. Thus, no artificial operations are introduced which results in explainable (algebraic) models.

There are four main tasks that we identify in the realm of conceptual scaling for which we provide an overview in the terminology of Formal Concept Analysis (see Figure 1.1).

The first task is the *data interpretation*, in which we define for each data feature a scale that is on the ordinal level. In FCA we distinguish scales within the ordinal level whose order relations have a specific structure and meaning, called *standard scales*. These scales allow for specialized interpretations of the values known as *basic meanings* [80, Figure 1.26]. One of them is the *nominal scale* which induces a partition of the scale values into incomparable and – in terms of the order relation – unrelated parts. Other standard scales are the (linear) *ordinal scale*, *contranominal scale*, *interordinal scale* and the *crown scale*. The composition of the defined scales results in an ordinal data set called the *formal context*.

The by this process derived formal context may be too large and complex for some application or human comprehension. The second task is concerned with *data reduction* methods that reduce the size of the data set. We understand both, the data scaling and reduction processes, as part of the *conceptual data scaling* problem. Essential here is that the data reduction method is consistent to the original data. For this, we extend the notion of *conceptual views* [221] to derive and characterize consistent data reductions. In particular, we allow for any view that allow for continuous maps, which are foundational in many machine learning methods.

The third task is *data compatibility*. Here, we verify the consistency of a data reduction and identify introduced errors. Here we introduce methods for the characterization, identification and quantification of errors.

The processes of data interpretation and reduction are often implicitly done and mixed. However, to have a precise interpretation of the data and its contained patterns we argue that it is important to separate these tasks. In Chapter 7 we elaborate on this is greater detail.

Conceptual views are not only great at capturing knowledge but also for (self-determined) navigation between information using methods from order theory. We can identify more narrow or abstract concepts following the order relation. In addition to that, we can compute greater commons and least multiples of concepts and conceptual views on the data using the meet and join operations. Besides these, we introduce new methods to navigate between multiple conceptual views by adding and removing information, combining views or by identifying missing information. On top of that, we present an exploration algorithm to compute a users view.

The fourth task is concerned with the *inversion of conceptual scaling*. This is done by identifying and extracting scales that are contained in a given formal context. Thereby, we derive a description on how the original raw data set may have looked liked and allows us to derive a possibly cleaner representation with removed noise. In addition to that, we present a new method to describe and explain the concept lattice based on the identified standard scales. We derive local explanations that characterize the interrelationship between objects in the data and automatically generate human-centered textual explanations. On top of that, we introduce new geometric properties that describe the shape and complexity of a concept lattice on a global level.

Lastly, we introduce the theoretical framework to check for the compatibility of conceptual data reductions. The method identifies parts of the conceptual structure of the scaled data set that are not present in the original data set.

1.2 Conceptual Views in Machine Learning

Modern machine learning algorithms achieve astonishing results when solving optimization problems. However, these excellent results are almost always achieved at the price of human explainability. This problem is addressed in research and practice from different standpoints.

There are calls to refrain from non-explainable models for important problems and to rely on explainable methods, even if they give worse results in terms of accuracy [184]. The second major direction is to develop methods for explaining black-box models. Such explanations can be classified into *local explanations*, i.e., why a particular data point was treated in a specific manner [178], and *global* explanations, i.e., approaches for explaining the whole model. The latter can be achieved, e.g., by mapping the non-explainable model to an explainable surrogate. A common approach for locally explaining models is to highlight important inputs [70]. For flat data, e.g., images, this is a viable approach since an essential explanatory component, the human, can be integrated into the process. This is not the case in high-dimensional or complex data. Global approaches are more difficult, in particular for high-dimension, and therefore less frequent. A typical idea is to find an (explainable) surrogate for a model, e.g., symbolic regression [3].

In Part III we contribute to the still growing interest for *global explanations procedures* by scaling internal data representations of machine learning models to conceptual views. We study neural networks, tree ensembles and topic models, and derive scaling methods that are specialized for them. We demonstrate how models can be represented by these views and what implications can be drawn for the model from the conceptual view. We further demonstrate how to compare models based on their views and extract a novel (global) geometric structure for their interpretation. This allows us to assess the structure of the *models view* on the data. Finally, we show how to enrich the conceptual views by interpretable background knowledge in the form of human-comprehensible propositional statements.

Part I

Foundations

2

Data Structures

Data can be given in many different types or structures. These are often accompanied by multiple relations between data (points) and allow for various operations to be used on the data. In this chapter, we introduce (abstract) data structures in a language inspired from **model theory**[1] [102], viewed as a combination of **universal algebra** and **logic**. This provides an algebraic framework to characterize classes of (data) *structures* in a unified notation and lets us introduce *algebraic concepts* that apply to all studied data structures. Another benefit of this approach is that we can focus in the later chapters on specific structures as *analytical tools* to analyze data and special properties of them.

Relation

Before we formally introduce a data structure, as studied within this thesis, we define all components that they can entail. A **relation** R on sets $U_1, U_2, \ldots, U_n$ with $n \in \mathbb{N}_{>0}$ is a subset of $U_1 \times U_2 \times \cdots \times U_n$. The later is called the **domain** $\operatorname{dom}(R)$ of R and the number n the **arity** of R. In case $U = U_i$ for all $1 \leq i \leq n$ we call R an n-ary relation on U and if $n = 2$ we call R a binary relation on U. The **complement** of R, denoted $\bar{R}$, is the set of all $(u_1, \ldots, u_n) \in \operatorname{dom}(R)$ with $(u_1, \ldots, u_n) \notin R$, or short $\operatorname{dom}(R) \setminus R$. For a binary relation R the **inverse** or **dual relation** of R is defined as $R^{-1} \coloneqq \{(b, a) \mid (a, b) \in R\}$. A relation $R \subseteq U_1 \times U_2 \times \cdots \times U_n$ **restricted** to a subset $S \subseteq \operatorname{dom}(R)$ is defined as $R|_S \coloneqq R \cap S$.

Composition and product of relations

The **cartesian product** of two relations $R \subseteq A \times B, S \subseteq C \times D$ is defined as $R \times S \subseteq (A \times C) \times (B \times D)$ with $\big((a, c), (b, d)\big) \in R \times S$ iff $(a, b) \in R$ and $(c, d) \in S$. The **composition** of a relation $R \subseteq A \times B$ and $S \subseteq B \times C$ is defined as $R \circ S \subseteq A \times C$ with $(a, c) \in (R \circ S)$ iff there exists a $b \in B$ with $(a, b) \in R$ and $(b, c) \in S$.

Properties of binary relations

A binary relation R on a set U is **reflexive** iff for all $u \in U$ we have $(u, u) \in R$. Moreover, R is **symmetric** iff for all $u, w \in U$ with $(u, w) \in R$ it follows that $(w, u) \in R$ and R is **transitive** iff for all $u, z, w \in U$ with $(u, z), (z, w) \in R$ it follows that $(u, w) \in R$. Contrary, R is **irreflexive** iff $\bar{R}$ is reflexive, R is **anti-symmetric** iff $\bar{R}$ is symmetric and R is **anti-transitive** iff $(u, v), (v, w) \in R$ implies that $(u, w) \notin R$. For a set U we call the set

[1] While we use the language of model theory to some extent, we do not pursue the study on how to classify classes of models. We only introduce concepts from model theory that we use and extend it to fit the data structures in our setting. For a complete introduction we refer the reader to the literature.

of all reflexive pairs the **diagonal** of U and denote it by $\Delta(U) := \{(u, u) \mid u \in U\}$. The **transitive extension** or **transitive closure** of R is the, with respect to set inclusion, smallest super set $R^* \subseteq \mathrm{dom}(R)$ of R for which $R^* \circ R^* \subseteq R^*$. A relation that is reflexive, symmetric and transitive is called an **equivalence relation**.

Function

A **function** or **map** from U to Y, denoted $f\colon U \to Y$, is a relation $f \subseteq U \times Y$ that is **left-total**, i.e., for all $u \in U$ there exists a $y \in Y$ with $(u, y) \in f$, and **right-unique**, i.e., for all $u \in U$ and $y_1, y_2 \in Y$ does $(u, y_1), (u, y_2) \in f$ imply that $y_1 = y_2$. Throughout the rest of this work, we write $f(u) = y$ instead of $(u, y) \in f$. The set U is called the **domain** $\mathrm{dom}(f)$ and the set Y the **co-domain** $\text{co-dom}(f)$ of f. In some cases, we omit to name a function and denote by $u \mapsto g(u)$ a function that maps inputs u to $g(u)$.

Function composition

The **composition** of two functions $f : A \to B,\ g : C \to D$ with $B \subseteq C$ is defined based on the composition of relations, i.e., $(g \circ f) : A \to D$ with $(g \circ f)(a) := f(g(a))$.

Image and pre-image

The **image** of $u \in U$ in f is the output $f(u)$. The **image** of f is the set of all images, i.e., $\mathrm{img}(f) := \{f(u) \mid u \in U\}$. The **pre-image** of a $y \in Y$ in f is the set $f^{-1}(y) := \{u \in U \mid f(u) = y\}$.

Properties of functions

A function f is **surjective** iff for all $y \in Y$ there exists a $u \in U$ with $f(u) = y$. We call f **injective** iff for all $u_1, u_2 \in U$ with $f(u_1) = f(u_2)$ it follows that $u_1 = u_2$. Moreover, a function is **bijective** iff it is injective and surjective. For a function f, we define the **pre-image map** of f to be $f^{-1} : Y \to \mathcal{P}(U)$ with $f^{-1}(y) := \{u \in U \mid f(u) = y\}$ (possibly empty) where $\mathcal{P}(U)$ denotes the **powerset** of U, i.e., the set of all subsets of U. In case f is bijective, we write $f^{-1}(y) = u$ in short and call f^{-1} the **inverse function** of f. A binary function f on U, i.e., $f : U \times U \to U$, is **associative** iff $f(f(a, b), c) = f(a, f(b, c))$ for all $a, b, c \in U$. Moreover, f is **commutative** iff $f(a, b) = f(b, a)$ for all $a, b \in U$ and f is **idempotent** iff $f(a, a) = a$ for all $a \in U$.

Set systems and undirected binary relations

Some structures include a **set system** $\mathcal{F}$ of a (domain) U, i.e., $\mathcal{F} \subseteq \mathcal{P}(U)$. A set system restricted to a set $V \subseteq U$ is defined as $\mathcal{F}|_V := \{A \cap V \mid A \in \mathcal{F}\}$. Special cases of set systems that we consider are set systems that are closed by intersection, i.e., for $A, B \in \mathcal{F}$ is $A \cap B \in \mathcal{F}$ and **undirected binary relations** $E \subseteq \binom{U}{2}$, where for $n \in \mathbb{N}$

$$\binom{U}{n} := \left\{A \subseteq U \,\middle|\, |A| = n\right\}.$$

In the latter case we define the restriction to be $E|_V := \{A \cap V \mid A \in E\} \cap \binom{U}{2}$. The **complement** of an undirected binary relation is defined as $\bar{E} := \binom{U}{2} \setminus E$.

Function lifts

For a function $f : U \to Y$ we define the lift of f to the powerset as $f : \mathcal{P}(U) \to \mathcal{P}(Y)$ with $f(A) := \{f(a) \mid a \in A\}$. The second lift of f is $f : \mathcal{P}(\mathcal{P}(U)) \to \mathcal{P}(\mathcal{P}(U))$ with $f(\mathcal{A}) := \{f(A) \mid A \in \mathcal{A}\}$. For both lifts of f we use the same symbol but differentiate them by the typographic style of the input or by specifying the employed lift in the text.

Special relations and functions

In case $U \subseteq Y$ we identify by $\iota : U \hookrightarrow Y$ the **inclusion map** $\iota : U \to Y$ with $\iota(u) := u$. The function f restricted to $A \subseteq \mathrm{dom}(f)$ is defined as $f|_A : A \to Y$ with $f|_A(a) := f(a)$. In the later parts of this work we are often interested in structures of standard shape. For easy readability we define them on sets of positive natural numbers up to some value $k \in \mathbb{N}_{>0}$, i.e., $[k] := \{1, \ldots, k\}$. A benefit of this notation is that it lets us directly infer the number of elements in a structure. For relations on $[k]$ we often use the symbols $\neq, =, \leq, \geq$ to denote the relation that includes all pairs numbers that are unequal, equal or comparable by $\leq, \geq$. By abuse of notation we also employ these symbols for the undirected binary relation variant of these relations. The type of relation is to be inferred from the surrounding text and often specified by the employed structure.

2.1 Data Structures

Data structure

A **(data) structure** is a finite tuple composed of a finite sets $U_1, \ldots, U_n$ called the **universe** or **domain sets**, relations $R_1, \ldots, R_k$ on the domains, set families $\mathcal{F}_1, \ldots, \mathcal{F}_m$, functions $f_1, \ldots, f_l$ from a combination of the domains to some set Y and **constants** $c_1, \ldots, c_p$, which are elements of the domains. This definition includes three adaptations from standard model theory. The first is that we allow for multiple, possibly overlapping, universe sets. The second is that we include set systems and the third is that the co-domain of functions may be (partially) composed of domains from external structures.[2] Such configurations can also be found in the definition of **abstract data types** in Ihringer [110].

$$(U_1, \ldots, U_n,\ \mathcal{F}_1, \ldots, \mathcal{F}_m,\ R_1, \ldots, R_k,\ f_1, \ldots, f_l,\ c_1, \ldots, c_p) \qquad \text{(structure)}$$

Signature

The **signature** $\mathcal{L}$ of a structure A is the tuple of all universe symbols $U_1, \ldots, U_n$, set system symbols $\mathcal{F}_1, \ldots, \mathcal{F}_m$, relation symbols $R_1, \ldots, R_k$ including their domain, all function symbols and their domain and co-domain $f_1 : X_1 \to Y_1, \ldots, f_l : X_l \to Y_l$, and all constant symbols $c_1, \ldots, c_p$. The interpretation of the symbols from a signature's definition, i.e., which elements are in relation or how inputs are mapped by a function is dependent on the structure. For a structure A that is of signature $\mathcal{L}$ we often write that A is a $\mathcal{L}$-structure in short. When two structures of the same signature are within scope, we indicate the structure in which to interpret a symbol by an index. For example let K, S be two $\mathcal{K}$-structures and I is a relation symbol in $\mathcal{K}$, then is I_K the relation I in structure K. If not specified otherwise is the size of a structure K, denoted $|K|$, equal to the sum of all domain set, set system, relation and function (in relation form) sizes.

Sub-structure

A structure A is a **sub-structure** of B iff A and B are $\mathcal{L}$ structures and

i) for all domain symbols U in $\mathcal{L}$ we have $U_A \subseteq U_B$,

ii) for all set system symbols $\mathcal{F}$ in $\mathcal{L}$ we have $\mathcal{F}_A \subseteq \mathcal{F}_B|_{\mathrm{dom}(\mathcal{F}_A)}$,

iii) for all relation symbols R in $\mathcal{L}$ we have $R_A \subseteq R_B|_{\mathrm{dom}(R_A)}$,

iv) for all function symbols f in $\mathcal{L}$ we have $f_A(u) = f_B(u)$ for all $u \in \mathrm{dom}(f_A)$,

v) for all constant symbols c in $\mathcal{L}$ we have $c_A = c_B$.

We denote the sub-structure relation between two structures A, B of signature $\mathcal{L}$ by $A \leq B$. A sub-structure A is an **induced sub-structure** of B, denoted $A \trianglelefteq B$, iff $A \leq B$ and for all relations R and set systems $\mathcal{F}$ of $\mathcal{L}$-structures it holds that $R_A = R_B|_{\mathrm{dom}(R_A)}$ and $\mathcal{F}_A = \mathcal{F}_B|_{\mathrm{dom}(\mathcal{F}_A)}$. An induced sub-structure is uniquely identified by its domain sets. Thus, we often write $B[U_{1A}, \ldots, U_{nA}]$ to denote the induced sub-structure given by the domain sets $U_{1A}, \ldots, U_{nA}$. For functional structures, both definitions are equivalent. For those we simply refer to as sub-structures since this is more consistent to the literature (cf. lattices and sub-lattices). A structure A is **coarser** than a structure B, denoted $A \preceq B$, iff $A \leq B$ and A, B have equal domain sets $U_A = U_B$ for all domains U of $\mathcal{L}$-structures. Dually, the structure B is called **finer** than A, i.e., $B \succeq A$.

Finer and coarser

Classes of structures

A **class** or **type** of structures is a signature $\mathcal{L}$ together with a set of axioms Φ. A structure A belongs to a class iff A is of signature $\mathcal{L}$ and it satisfies all axioms in Φ, written A is a **model** of Φ or $A \models \Phi$. More on the logic related notations can be found in Section 2.3.

[2]Unless stated otherwise, we employ for numerical values Y the usual arithmetic operations.

Relational structures

Structures that, besides the domain sets, contain only (undirected binary) relations are also called **relational (data) structures**. The **inverse/dual** and **complement** of a relational structure $A := (U_1, \dots, U_n, R_1, \dots, R_k)$ are defined to be $A^{-1} := (U_1, \dots, U_n, R_1^{-1}, \dots, R_k^{-1})$ and $\bar{A} := (U_1, \dots, U_n, \bar{R}_1, \dots, \bar{R}_k)$ respectively. In case the inverse/dual or complement of a structure belongs to the same class of structures we also call them the inverse/dual structure or complement structure. For some structure classes there are special versions of complement or inverse relations which we introduce when needed.

2.2 Morphisms of (Data) Structures

Morphisms motivation

In this section we introduce several types of maps from a structure A into a structure B of the same signature $\mathcal{L}$ that *preserve* or *reflect* algebraic properties of A. These maps are called **morphisms** and can be used to interpret a structure A via a possibly smaller or easier to comprehend structure B. Let $(U_1, \dots, U_n)$ be the domain symbols of $\mathcal{L}$ then is a morphism from A to B a tuple of maps $(\alpha_1 : U_{1A} \to U_{1B}, \dots, \alpha_n : U_{nA} \to U_{nB})$. In case $\mathcal{L}$ has only one domain symbol U we refer to a morphism by the map $\alpha : U_A \to U_B$ directly without the tuple notation. We say a morphism is injective/surjective iff all maps α_i are injective/surjective. The image of a morphism is the tuple of all image sets of the individual maps α, i.e., $\mathrm{img}(\alpha_1, \dots, \alpha_n) := (\mathrm{img}(\alpha_1), \dots, \mathrm{img}(\alpha_n))$. If the structure B belongs to the same class C as A we also call a morphism from A to B a C-morphism. The same naming convention applies to all morphism variants that we introduce in this section.

Relation preserving and reflecting

Let A, B be two structures of signature $\mathcal{L}$ with domain symbols $U_1, \dots, U_n$ and let R be a relation symbol of $\mathcal{L}$ with domain $U_{i_1} \times \dots \times U_{i_s}$. A morphism from A to B is called **R-preserving** iff

$$(u_1, \dots, u_s) \in R_A \implies (\alpha_{i_1}(u_1), \dots, \alpha_{i_s}(u_s)) \in R_B.$$

A morphism that is preserving for all relations is called **relation preserving**. Analogously, a morphism is R-reflecting iff

$$(u_1, \dots, u_s) \in R_A \impliedby (\alpha_{i_1}(u_1), \dots, \alpha_{i_s}(u_s)) \in R_B.$$

A morphism that is reflecting for all relations is called **relation reflecting**. **Set system reflecting** and **set system preserving** maps are defined analogously.

Function preserving

Let f be a function symbol of $\mathcal{L}$ with $\mathrm{dom}(f) = U_{i_1} \times \dots \times U_{i_s}$ and $\text{co-dom}(f) = Y$, then is a morphism from A to B is called **f-preserving** iff

$$\begin{aligned} f(u_{i_1}, \dots, u_{i_s}) &= f(\alpha_{i_1}(u_{i_1}), \dots, \alpha_{i_s}(u_{i_s})) && \text{if } f \text{ maps to an external structure,} \\ \alpha_j(f(u_{i_1}, \dots, u_{i_s})) &= f(\alpha_{i_1}(u_{i_1}), \dots, \alpha_{i_s}(u_{i_s})) && \text{if } f \text{ maps to a } U_j \text{ of } \mathcal{L}. \end{aligned}$$

A morphism that is f-preserving for all functions is called **function preserving**.

Constant preserving

Let c be a constant symbol of $\mathcal{L}$ with $\mathrm{dom}(c) = U_i$, then is a morphism from A to B c-preserving iff

$$\alpha_i(c_A) = c_B.$$

A morphism that is c-preserving for all constants is called **constant preserving**.

From homomorphism to automorphism

A morphism that is relation preserving, preserving with respect to the set systems, function preserving and constant preserving is called a **homomorphism** from A to B. If there exists a homomorphism from A to B we call A **homomorphic** to B, denoted $A \leqq B$. Furthermore, is an injective homomorphism an **embedding** iff it is relation and set systems reflective. If there exists an embedding from A to B we call A **embeddable** into B, denoted $A \leqq B$. A surjective embedding is called an **isomorphism** and if additionally we have $A = B$ it is called an **automorphism**. We call A and B **isomorphic**, denoted $A \cong B$, iff there exists an isomorphism between A and B. The inverse maps of an isomorphism from A to B form an isomorphism from B to A. By restricting the structure B to the image of an embedding we get an isomorphism.

2.3 Logic

Logical theories

In this section we introduce basic notions and terminology of logic that we use throughout this work. A **theory** is a set of **sentences**, i.e., variable free formulas, from a formal language of logical expressions F. By $F[M,\Omega]$ we identify the formal language of all well-formed (quantor-free) sentences over symbols M and logical operations Ω. For example by $F[M,\{\wedge,\vee,\neg\}]$ we identify the formal language of all well formed logical expressions from **propositional logic**. Another logic that we employ is a fragment of **Horn logic** [106, 148]. Horn logic is the formal language of all logical expressions from $F[M,\{\vee,\neg\}]$ that include at most one positive symbol. We employ the fragment $\mathcal{H}(M)$ that contains all horn sentences that have exactly one positive symbol. In a later chapter we study an equivalent theory to $\mathcal{H}(M)$ which includes all sentences of the form $\varphi \to \psi$ where $\varphi,\psi \in F[M,\{\wedge\}]$.

Formal logics

Model Relation

For any set $N \subseteq M$ we *interpret* $m \to \top$ for $m \in N$ and $m \to \bot$ for $m \in M \setminus N$, where $\top$ denotes *true* and $\bot$ denotes *false*. A set $N \subseteq M$ is called a *model* of $\varphi \in F[M,\{\wedge,\vee,\neg\}]$, denoted by $N \models \varphi$, iff the interpretation of N satisfies φ, i.e., φ evaluates to true given the interpretation above.

Logical basis

A sentence $\varphi \in F$ is **entailed** in a theory $T \subseteq F$ iff it logically follows from T (by Armstrong rules [9]), denoted $T \vdash \varphi$. Analogously, T entails a set $\Phi \subseteq F$ iff all $\varphi \in \Phi$ are entailed in T. The **transitive closure** of T is the set of all $\varphi \in F$ that are entailed in T, i.e., $T^* := \{\varphi \in F \mid T \vdash \varphi\}$. A theory is **closed** in F iff $T = T^*$. A **basis** is an inclusion minimal set $B \subseteq T^*$ such that $B \vdash T^*$. Two sentences or two theories are equivalent $T_1 \cong T_2$ iff $T_1 \vdash T_2$ and $T_2 \vdash T_1$.

Theory of a structure

The **theory** of a structure A with respect to F is the set of all sentences that are true in A:

$$\mathrm{Th}_F(A) := \{\varphi \in F \mid A \models \varphi\}.$$

In case F can be inferred from the surrounding text we may simply write $\mathrm{Th}(A)$.

3 Graphs

The first data structure that are relevant to this work are graphs [55]. They are commonly applied in several research fields to model entities V and relations between them. Early problems based on graph structures are routing problems in street networks. Well-known instances are the *Seven Bridges of Königsberg* problem [67] or the *traveling salesman problem*. More modern applications of graphs can be found in social network analysis and the study of social structures [71, 105] or knowledge graphs [103]. Graph theory is, besides its use for analyzing data, well studied with respect to the computational complexity of problems. The following notions are recalled from the literature.

Computational aspect

3.1 Graph Data Structures

Throughout the remainder of this chapter, we formally introduce all notions from graph theory [55] that are used in this thesis.

Graph

Definition 1 (Graph). *A **graph** structure is a tuple (V, E) where V is the domain called **vertices** or **nodes** and E an undirected binary relation on V. A graph is **complete** iff $E = \binom{V}{2}$.*

An example graph can be seen in Figure 3.1 which encodes the co-authorship relation of the *KDE*[1] research group. This graph has nine authors as vertices which are in relation iff they are co-authors in at least one research article. This results in a total of nineteen edges. The graph is displayed as a *line diagram* which depicts each vertex of the graph by a node in the diagram and connects two nodes by a line iff they are in relation in the graph. Graphs that encode this type of relation are also known as *collaboration network* [163].

Example

The **neighborhood** of a node n in a graph (V, E) is defined as the set of all nodes that are in relation to n, i.e., $\mathrm{N}(n) := \{a \in V \mid \{a, n\} \in E\}$. In the KDE example we find that the two authors with the largest neighborhood are *GS* and *TH* which both have seven authors in

Neighborhood

[1]The Knowledge and Data Engineering Research Group of the University of Kassel.

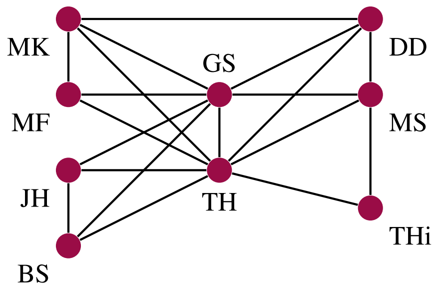


Figure 3.1: A graph encoding the co-authorship network of the KDE research group.

their neighborhood. The other authors have a neighborhood size of three or four. Following on from the concept of neighborhood is the connectivity of nodes in a graph by a series of edges.

Definition 2 (Path). *A **path** in a graph $G := (V, E)$ is a sequence of edges $(e_1, \ldots, e_k)$ such that for $1 \leq i \leq k$ is $e_i \in E$ and for $i \neq 1$ it holds that $e_i \cap e_{i-1} \neq \emptyset$.*

Paths and connections

A commonly investigated class of graph sub-structures are sub-graphs for which there is a path between any two nodes.

Definition 3 (Connectedness in Graphs). *Two nodes u, w of a graph $G := (V, E)$ are **connected** iff there exists a path $(e_1, \ldots, e_k)$ in G with $u \in e_1$ and $w \in e_k$. A **connected component** of G is an induced sub-graph $G[C] \trianglelefteq G$ such that C is an inclusion maximal set for which all pairs of nodes are connected. A graph G is connected iff V is a connected component in G.*

Distance

One of the most well-known investigations of connectedness in graphs is the study of **Erdös numbers**. This number is equal to the shortest path distance between an author and the researcher Paul Erdös, who is a mathematician famous for having many collaborations and research papers. In the KDE example we find that there are no two authors in this example that have a larger path distance than two.

Cycle

A special type of path is a cycle, i.e., a path from a node to itself that visits every node at most once.

Definition 4 (Cycle). *A path $(e_1, \ldots, e_k)$ in a graph (V, E) is called a **cycle** iff $e_1 \cap e_k \neq \emptyset$ and for all $1 \leq i < j \leq k$ it holds that $j \neq i + 1$ and $(i, j) \neq (1, k)$ implies $e_i \cap e_j = \emptyset$. A cycle is a **Hamiltonian cycle** if for all nodes $n \in V$ there is an $1 \leq i \leq k$ with $n \in e_i$. A graph that contains no cycle is called **acyclic**.*

Hamiltonian cycle

The problem of finding cycles is an important problem for in the analysis of many types of graphs. For example, in route planning they encode a round trip. Other examples can be found in artificial intelligence and strategy planning of games. A cycles in a graph of all possible game configurations may be used to determine draws.

Deciding for a graph if there exists a Hamiltonian cycle or a **Hamiltonian path**, i.e., a path that visits every node exactly once, is a computational costly problem:

Problem 1: Hamiltonian Cycle Problem	
Input: A graph $G := (V, E)$	
Output: True iff there is a Hamiltonian cycle in G	
Complexity:	NP-complete

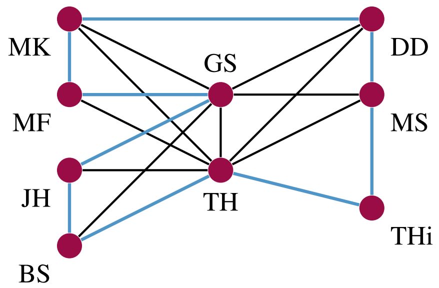


Figure 3.2: The graph given in Figure 3.1 with a Hamiltonian cycle colored in cyan.

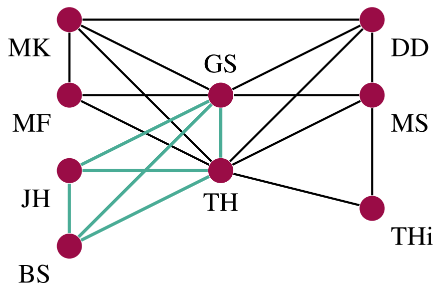


Figure 3.3: The graph given in Figure 3.1 with a maximal clique colored in cyan.

Our example graph in Figure 3.1 does contain multiple Hamiltonian cycles. One of them is highlighted in cyan in Figure 3.2. Further investigations on cycles can be found in Chapters 9 and 14.

Hamiltonian cycle example

Other sub-structures of interest are cliques, i.e., complete sub-graphs.

Clique

Definition 5 (Clique). *A **clique** of a graph G is an induced sub-graph $S \trianglelefteq G$ that is complete. A **maximal clique** of G is a clique S such there is no clique $H \neq S$ of G with $S \trianglelefteq H \trianglelefteq G$.*

In a clique the neighborhoods of all nodes are highly overlapping. Due to this property they are often used to model community structures [71]. Another useful property of cliques is their heredity, i.e., all induced sub-graphs of cliques are cliques. Determining if a graph contains a clique with k elements is known as the *clique problem* [112] and is a computational costly task:

Problem 2: Clique Problem	
Input: A graph $G := (V, E)$ and $k \in \mathbb{N}$	
Output: True iff $([k], \neq) \leqq G$	
Complexity:	NP-complete

The co-authorship graph in Figure 3.1 contains three maximal cliques with four elements and two with three elements. One of the cliques on four elements is highlighted in Figure 3.3. The cliques with four elements overlap in two nodes which are *TH* and *GS*. These authors were the postdoc and professor and *central* to the group, since they often supervised the published papers.

Clique examples

Finding sub-graphs

A more general instance of the sub-structure problem is the identification of arbitrary sub-structures and induced sub-structure in a graphs [48].

Problem 3: Sub-Graph Isomorphism Problem	
Input: Two graphs H and G	
Output: True iff there exists a $S \leq G$ with $H \cong S$	
Complexity:	NP-complete

Problem 4: Induced Sub-Graph Isomorphism Problem	
Input: Two graphs H and G	
Output: True iff there exists a $S \trianglelefteq G$ with $H \cong S$ (i.e., $H \leqq G$)	
Complexity:	NP-complete

The hardness of Problem 3 follows immediately from Problem 3.1, and Problem 4 from Problem 2.

Graph isomorphism

Deciding if the isomorphism [229] applies to the entire graphs, i.e., $H \cong G$, is known as the **graph isomorphism problem**. This problem forms its own complexity class GI which is in NP. It is unknown if GI is entailed in NP-complete or P, or neither.

Problem 5: Graph Isomorphism Problem	
Input: Two graphs H and G	
Output: True iff $H \cong G$	
Complexity:	GI-complete

3.2 Hypergraphs

Motivation

While graphs are great at modeling relational data as seen in Figure 3.1, they are limited to binary relations. Modeling relations of arbitrary cardinality either requires reification, e.g., as used in the *RDF 1.1 Semantics*[2] standard, or relational data structures with relations of higher arity. In this work, we use **hypergraphs** which use set systems to model relations between data points of varying arity.

Definition 6 (Hypergraph). *A* ***hyperpgraph*** *structure is a tuple* $(V, \mathcal{E})$ *with* $\mathcal{E} \subseteq \mathcal{P}(V)$ *where* V *is a set of* ***nodes*** *and* $\mathcal{E}$ *is a set system on* V*. An* $e \in \mathcal{E}$ *is called a* ***hyperedge****.*

Example: Document lift

With hypergraphs we are able to lift the KDE co-authorship example in Figure 3.1 to the document level. Formally, we depict in Figure 3.1 $(V, \mathcal{E})$, where V are the authors from the example in the previous section and an $e \subseteq V$ is a hyperedge in the graph iff there is a paper with authors e. This modeling results in a hypergraph with eighteen hyperedges. We refrain from drawing this hypergraph since we do not think that the resulting diagram would be nicely readable with this many edges.

Example: Hypergraph of maximal cliques

Another example based on the graph in Figure 3.1 can be given using the maximal cliques as hyperedges. This results in a hypergraph on nine vertices with eight hyperedges. There are many possibilities on how to visualize a hypergraph. In this section, we draw hyperedges as colored boxes that surround the nodes they contain. From the resulting diagram (see Figure 3.4) we can infer, again, that the authors *TH* and *GS* are *central* to

[2] `https://www.w3.org/TR/rdf11-mt/`, W3C Recommendation 25 February 2014

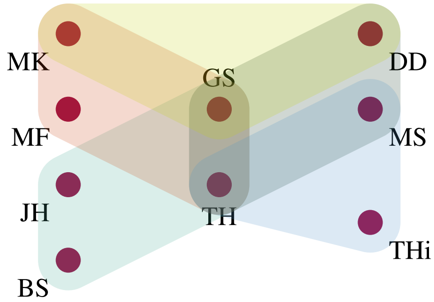


Figure 3.4: A hypergraph modeling the co-authorship network on a document level.

the group, since they occur in almost all cliques. In addition to that, we can see how cliques overlap. The structure of their intersections forms a *hierarchy* which may reveal commonalities of (research) communities. In the next section, we investigate this structure in more detail.

Multirelational hypergraph

In the later parts of this work, we derive a principle on how to model more types of sub-structures, other than cliques, of a graph simultaneously. The resulting structure is a multi-relational hypergraph with one hyperedge relation per sub-structure type.

Definition 7 (Multirelational Hypergraph). *A **multirelational hypergraph** structure is a tuple* $(V, \mathcal{E}_1, \ldots, \mathcal{E}_m)$ *where* $(V, \mathcal{E}_i)$ *is a hypergraph for all* $1 \leq i \leq m$*.*

4

Ordered Sets

A common type of relation in data structures are hierarchical relationships between elements where one element is more, larger or higher than another [207]. Be it food chains, dominance hierarchies or genealogies, hierarchies can be observed in all kinds of relationships. These, however, are not limited to nature example. Everywhere objects are sorted, listed or ranked we induce a successor/predecessor relation forming a hierarchical structure. These type of relations are called **order relations** [23] and are the basis of the in this work employed data analysis paradigm [80, 223]. The following notions are recalled from the literature.

4.1 Ordered Data Sets

Ordered Sets

Definition 8 (Ordered Set). *An **ordered set** is a tuple $(P, \leq)$ where P is a set and $\leq$ is a binary relation on P that is reflexive, anti-symmetric and transitive. Two elements $x, y \in P$ are **comparable** if $(x, y) \in \leq$ or $(y, x) \in \leq$. Otherwise x, y are **incomparable**. For $(x, y) \in \leq$ we say that x is **smaller than or equal to** y or that y is an **upper bound** of x. Analogously, we define **greater or equal** and **lower bound**.*

Naming conventions

Sometimes we refer to an ordered set as **order** or use the relation symbol in infix notation, i.e., $x \leq y$ instead of $(x, y) \in \leq$, for simplicity. Moreover, we use the symbol $\geq$ for the inverse/dual $\leq^{-1}$, $\nleq$ for the complement $\bar{\leq}$ and $<$ for $\leq \setminus \Delta(P)$. For a $A \subseteq P$ we define by $A^{\geq} := \{u \in P \mid \forall a \in A : a \leq u\}$ the set of all upper bounds of A and analogously $A^{\leq} := \{l \in P \mid \forall a \in A : l \leq a\}$ the set of all lower bounds of A.

Duality principle

We often switch between an order and its inverse. This often simplifies formal expressions since many algebraic concepts in order theory are the dual of each other, e.g., smaller and greater.

Theorem 1 (Duality Principle [23, Thm. 2]). *The dual of an ordered set is an ordered set.*

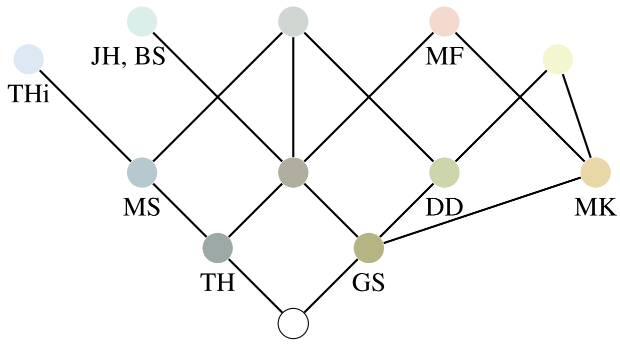


Figure 4.1: The ordered set of all intersections of cliques of Figure 3.4. The node color is chosen in accordance to the colors in Figure 3.4.

The just recalled duality principle is not limited to ordered sets but extends to structures that we analyze in the upcoming sections and chapters.

Cover relation

A diagram that displays all relations of $\leq$ includes a lot of redundant information and would become very complex even for smaller ordered sets. This is due to the transitive property of order relations. This can be avoided by displaying only the *direct neighbors* of an element. The remaining order can be followed using the transitive closure, i.e., paths in the diagram.

Definition 9 (Cover Relation). *For an ordered set $(P, \leq)$, the* ***cover relation*** *is the inclusion minimal subset $\prec \subseteq \leq$ such that $(\prec \cup \Delta(P))^* = \leq$.*

The cover relation $\prec$ of an ordered set $(P, \leq)$ is unique and contains all elements $x \leq y$ such that there exists no z with $x \leq z \leq y$.

Order examples

Besides the examples given in the beginning of this chapter we have already seen a couple of examples throughout this work. The *table of contents*, *content lattice* or the layout of a page are order structures. In Figure 4.1 we provide an example that builds on the graph example from the previous chapter. We encode via an ordered set the hierarchy of the node sets that form maximal cliques and their intersections in the diagram. The order relation is the set inclusion.

Line diagram

We visualize ordered sets by a line diagram of the cover relation. The diagram follows the convention that if $x \leq y$ we depict y above x in the diagram. The converse does not necessarily hold. Thus, for a line diagrams there is an underlying order preserving map from the ordered set to the ordinate of the diagram. In our example order we have the special case that the domain is a set system. For this we employ the **shorthand notation** which we formally introduce in Chapter 5. For now, it is sufficient to know that a node encodes the set of all elements that can be reached by following downward paths.

Order example: hierarchy of cliques

The order diagram (see Figure 4.1) depicts eleven nodes and sixteen edges. The ordered set enables us, in contrast to the hypergraph, see the hierarchy and contained relations more clearly. Elements that are of special interest in an ordered set are *minimal* and *maximal* elements with respect to the order structure.

Pareto optima

Definition 10 (Pareto Optima). *For an ordered set $(P, \leq)$, an element $e \in P$ is a* ***pareto optima*** *or* ***maximal element*** *of $(P, \leq)$ if $(e, d) \in \leq$ implies that $e = d$. The* ***pareto front*** *is the set of all pareto optima. An element e is* ***minimal*** *in $(P, \leq)$ iff it is* ***maximal*** *in $(P, \geq)$.*

The pareto front of our example is equal to the set of all maximal cliques of the underlying graph.

Filter and ideal

Besides these extreme elements we can study the set of their upper and lower bounds. In our order example we find that all non-empty cliques bound from below either by $\{TH\}$ or $\{GS\}$. This observation is consistent with our considerations from the previous section.

Definition 11 (Filter and Ideal). *The* ***(order) filter*** *of an element $e \in P$ in an ordered set $(P, \leq)$ is the set of elements that are bound by e from below:*

$$\uparrow e := \{o \in P \mid e \leq o\}.$$

The (order) filter of a set $A \subseteq P$ is the set of elements that are bound by an $e \in A$ from below:

$$\uparrow A := \{o \in P \mid \exists e \in A : e \leq o\} = \bigcup_{e \in A} \uparrow e$$

Analogously, is the ***(order) ideal*** *of an element $\downarrow e$ or set $\downarrow A$ equal to the filter in the dual order. An* ***interval*** *$[a, b]$ is defined as the set of elements bound from above by b and from below by a, i.e., $[a, b] := \uparrow a \cap \downarrow b = \{c \in P \mid a \leq c \leq b\}$.*

Chain and anti-chain

Similar to cliques in graphs, there are ordered sets that are of standard structure. A **chain** (or **linear order**) is an ordered set $(P, \leq)$ in which every pair of elements is comparable. A **linear extension** of an order relation $\leq$ is a relation $<$ with $\leq \subseteq <$ and $(P, <)$ is a linear order. An **anti-chain** is an ordered set $(P, \leq)$ in which no elements $e, d \in P$ with $e \neq d$ are comparable. More ordered sets of standard scale are introduced and analyzed more thoroughly in the upcoming chapters.

4.2 Dimensions of Ordered Sets

Height and width

There exist several notions of dimensions and quantities to measure structural properties of an ordered set. The first are the **height** and **width** of an ordered set.

Definition 12 (Height and Width). *The* ***height*** *of an ordered set P is the largest number $h \in \mathbb{N}$ such that there exists a chain $C \trianglelefteq P$ with h elements. The* ***width*** *of an ordered set P is the largest number $w \in \mathbb{N}$ such that there exists an anti-chain $A \trianglelefteq P$ with w elements.*

Order dimension

A commonly applied measure of dimension to quantify the complexity of an ordered set is the *order dimension*. The following definition uses the (cross-)product of ordered sets $(P_i, \leq_i)_{i \in N}$ which is defined as

$$\bigtimes_{i \in N} (P_i, \leq_i) := \Big(\bigtimes_{i \in N} P_i, \leq\Big), \text{ with } (x_i)_{i \in N} \leq (y_i)_{i \in N} \iff x_i \leq_i y_i \text{ for all } i \in N.$$

Definition 13 (Order Dimension [80, Definition 82]). *The* ***order dimension*** odim(P) *of an ordered set P is equal to the smallest number $d \in \mathbb{N}$ such that there is an order embedding from P into the product of d chains.*

Another common characterization of the order dimension of $(P, \leq)$ is that it is equal to the smallest number of chains $(P, \leq_1), \ldots, (P, \leq_n)$ such that

$$\bigcap_{1 \leq i \leq n} \leq_i = \leq .$$

Determining the order dimension of an ordered set is computational costly. In the following we present the problem and complexity [226] of deciding upper bounds for the order dimension.

Problem 6: Order Dimension Problem	
Input: An ordered set P and $d \in \mathbb{N}$	
Output: True iff $\mathrm{odim}(P) \leq d$	
Complexity:	NP-complete

4.3 Lattices

Semilattice

A natural extension of the investigation of bounds is to find the *greatest lower bound* and *least upper bound* of a set of elements. However, in arbitrary ordered sets there does not need to exist for a set $A \subseteq P$ an upper bound $t \in P$, i.e., $A \subseteq \downarrow t$. Even if such an element exists there does not need to be a least element with that property.

Definition 14 (Semilattice). *An ordered set $(P, \leq)$ is a **meet-semilattice** iff for all two-element subsets $A \in \binom{P}{2}$ the set of all lower bounds $A^{\leq}$ has a greatest element, called the **meet** $\bigwedge A$ of A. Dually, is $(P, \leq)$ a **join-semilattice** iff for all $A \in \binom{P}{2}$ the set of all upper bounds $A^{\geq}$ has a least element, called the **join** $\bigvee A$ of A. The least element in a meet-semilattice is also called the bottom element $\bot$ and the greatest element in a join-semilattice is called top $\top$.*

For the meet and join of two elements $u, v \in P$ we often use the infix notation, i.e., $u \wedge v$. The meet of a set $A \subseteq P$ is equal to the meet of all its elements and the meet of the empty set is $\top$, if it exists. The dual applies to the join.

Semilattice example

The example given in Figure 4.1 is a meet-semilattice. The meet of the nodes annotated with *DD* and *MK* is the node annotated with *GS*. The meet of the *THi* and *DD* nodes is bottom. However, the example ordered set is not a join-semilattice since the join of the *JH,BS* and *MF* node does not exist.

Binary tree

A join-semilattice is a **binary tree** iff the graph (P, E) with $E := \{\{d, e\} \mid (d, e) \in \prec\}$ is acyclic and for all $e \in P$ is the set of direct lower neighbors $|\{p \in P \mid p \prec e\}| \leq 2$.

Definition 15 (Lattice). *A **lattice** is an ordered set that is a meet- and join-semilattice.*

Lattice example

By extending our example in Figure 4.1 by a *KDE* node, i.e., the set of all authors, as top element we do get a lattice order. As seen in Figure 4.2 is the *KDE* node the join of all subsets for which the join was previously undefined. Also, the product of lattices or chains as used in Definition 13 is lattice ordered. Another example is the content lattice presented in the introductory part of this work. The content lattice carries, besides its order properties, a special semantic on its elements which we discuss in more detail in the next chapter.

Duality principle

The duality principle of ordered sets extends naturally to lattices and semi-lattices.

Remark 1 (Duality Principle – Lattices). *The dual of a meet-semilattice is a join-semilattice and vise versa. The dual of a lattice order is a lattice order.*

Finite lattices

At this point we want to remind the reader that we consider data structures to be finite. For (semi)lattices one usually differentiates between (semi)lattices and **complete (semi)lattices**. In the finite case all (semi)lattices are complete. Therefore, we do not make this differentiation and do not go further into this. Theorems and propositions that we recall from the literature may include the completeness requirement.

Irreducible elements

Elements that are integral to a lattice order are those that can not be represented as the meet or join of other elements.

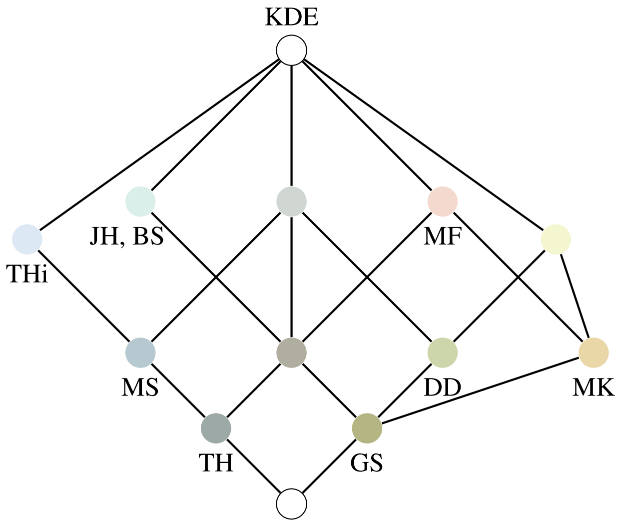


Figure 4.2: The lattice of all intersections of cliques Figure 3.4. The node color is chosen in accordance to the colors in Figure 3.4.

Definition 16 (Irreducible Elements). *An element v of a lattice order $(L, \leq)$ is called **meet-irreducible** iff there does not exist an $A \subseteq L \setminus \{v\}$ with $\bigwedge A = v$. By* $\mathrm{M}(L)$ *we denote the set of all meet-irreducible elements. An element is called **meet-reducible** iff it is not meet-irreducible. The dual is defined for the join operation where the set of all join-irreducible elements is denoted* $\mathrm{J}(L)$. *An element that is meet- and join-irreducible and called **double-irreducible** and **double-reducible** iff it is meet- and join-reducible.*

We may note that $\bigvee\{\} = \bot$ which is therefore not join-irreducible. Analogously, the top element is not meet-irreducible. The meet- and join-irreducible elements can be read directly from an order diagram via the following characterization: An element is meet-irreducible iff it has exactly one upper neighbor in the cover relation and an element is join-irreducible iff it has exactly one lower neighbor in the cover relation.

Atoms and Co-atoms

A special type of lattices are those in which every element can be written as the meet or join of the (besides $\bot$ and $\top$) top and bottom most elements.

Definition 17 (Atomistic and Co-Atomistic). *An element a of a lattice L is called **atom** iff $\bot \prec a$. The set of all atoms is denoted* $\mathrm{At}(L)$. *Moreover, L is called **atomistic** iff* $\mathrm{At}(L) = \mathrm{J}(L)$. *Analogously, is* $\mathrm{coAt}(L)$ *the set of all **co-atoms**, i.e., the elements c of L with $c \prec \top$. A lattice L is **co-atomistic** iff* $\mathrm{coAt}(L) = \mathrm{M}(L)$.

Alternative lattice definition

There is a second (functional) definition for lattice structure using the signature $(L, \vee, \wedge)$. Here, a $(L, \vee)$-structure is a join-semilattice iff $\vee$ is a binary commutative and associative function on L. Analogously, is a $(L, \wedge)$-structure is a meet-semilattice iff $\wedge$ is a binary commutative and associative function on L. A $(L, \vee, \wedge)$-structure is a lattice iff for all $a, b \in L$ the **absorption law** holds:

$$a \vee (a \wedge b) = a,$$
$$a \wedge (a \vee b) = a.$$

Sub-lattice

Although both definitions can be used interchangeably, they do result in different notions for morphisms and sub-structures. In this work, we use the definition based on ordered sets since we find it to be easier to comprehend. However, the notions of morphisms and sub-structures are derived from the $(L, \vee, \wedge)$ signature. Thus, for two lattices L, O an embedding and lattice embedding are with respect to the $(L, \vee, \wedge)$-structures and an order embedding with respect to the $(L, \leq)$-structures. This combination is also common in the literature [80], even though it may not be stated explicitly. The implication is that for a lattice $(L, \leq_L)$ not every induced sub-structure $(S, \leq_S)$ that is lattice ordered is also an induced sub-lattice. For this it is also required that the meets and joins are preserved, i.e., the inclusion $S \hookrightarrow L$ is a meet and join preserving morphism.

Dedekind-MacNeille completion

The join and meet operations are very useful for calculations with elements of an ordered set and often carry a specific semantic with respect to the data domain, e.g., the greater common of community structures as seen in Figure 4.1. For ordered sets it may be useful to extend the structure to a lattice to enable such computations. This can be done by computing the Dedekind-MacNeille completion, which we recall for finite data structures.

Definition 18 (Dedekind-MacNeille completion). *For an ordered set P is the* ***Dedekind-MacNeille completion (DM completion)*** $\mathrm{DM}(P)$ *the smallest lattice L (up to isomorphism) such that $P \leqq L$. The lattice L is isomorphic to the ordered set*

$$(\{(A^{\geq})^{\leq} \mid A \subseteq P\}, \subseteq).$$

In the next chapter we analyze the DM completion more thoroughly by the means of *closure systems* and will see that $A \mapsto (A^{\geq})^{\leq}$ is a *closure operator* on P.

5

Formal Concept Analysis

In this chapter we introduce a formalism for ordinal data analysis. This formalism is called Formal Concept Analysis (FCA) [80, 223] and provides an expressive language to access data objects as well as their properties, inherent ordinal structure and implicational theory. A defining characteristic of FCA is that its notions were designed with human comprehensibility in mind. All notions of this chapter are, unless stated otherwise, recalled form Ganter and Wille [80].

5.1 Contextual Data Structure

Formal context

The main data structure in which we represent data in Formal Concept Analysis are **formal contexts**. They are relational structures that describe objects by attributes that they have. We remind the reader that, in this work, data structures are finite.

Definition 19 (Formal Context). *A **formal context** $\mathbb{K}$ is a tuple (G, M, I) where G is a finite set called **objects**, M is a finite set called **attributes** and $I \subseteq G \times M$ is a binary relation between them. The relation I is called the **incidence relation** of $\mathbb{K}$.*

Formal context semantic

We often write context instead of formal context for simplicity reasons and denote contexts in *blackboard bold* typeface style, e.g, $\mathbb{K}$ or $\mathbb{H}$. The interpretation of a pair $(g, m) \in I$ is that *"object g has attribute m"*. This relation naturally extends to sets of objects that have common attributes or sets of attributes that are shared by some objects.

Definition 20 (Derivation Operators). *For a context $\mathbb{K} := (G, M, I)$ we define two derivation operators:*

$$(\cdot)' : \mathcal{P}(G) \to \mathcal{P}(M) \text{ with } A' := \{m \in M \mid \forall g \in A\colon (g, m) \in I\} \quad \text{(object derivation)}$$

$$(\cdot)' : \mathcal{P}(M) \to \mathcal{P}(G) \text{ with } B' := \{g \in G \mid \forall m \in B\colon (g, m) \in I\} \quad \text{(attribute derivation)}$$

In case there are multiple contexts present we identify the derivation operator of a context by is incidence relation, i.e., $(\cdot)^{I_{\mathbb{K}}}$ is the derivation operator of $\mathbb{K}$.

	Conceptual Data Scaling	Conceptual Views	Machine Learning	Deep Learning	Scale Semantic	Inverse Scaling
7 Data Scaling	×					
8 Navigating Views	×	×				
9 Ordinal Motifs in Lattices	×				×	
10 Complexity of Views	×				×	×
11 Scaling Error	×	×	×			×
12 Views on Tree Classifiers	×	×	×			
13 Views on Neural Networks	×	×	×	×		
14 Views on Topic Models	×	×	×		×	

Figure 5.1: The formal context of all chapters of this thesis and their topics. The chapter titles are shortened to improve readability.

Visualizing contexts

Contexts are commonly visualized as *cross-tables*, where each object is a row and each attribute is a column in the table. A cell at the intersection of row g and column m has a cross $\times$ iff $(g, m) \in I$. This way we can easily derive all attributes of an object or all objects that have an attribute by looking at the corresponding row or column. The derivation of a set of objects can be inferred by the crosses that are common to all rows of the objects.

Context example

Context interpretation

In Figure 5.1 we have depicted an example context visualized as a cross-table. The set of objects are the main chapters of this thesis and the attributes are topics that they are about. In the table we can see that the context is composed of eight chapters and eight topics with twenty-two incidences between them. The *Conceptual View* topic has the most incidences followed by *Conceptual Measurability* with five and *Machine Learning* with four incidences. The topics with the least number of incidences are *Unsupervised Learning* and *Deep Learning*. The number of incidences reflects the number of chapters in which a topic is present and can be used as a measure of relevance. The chapter *Conceptual Views on Neural Networks* includes the largest number of topics. Therefor, we can interpret chapter thirteen to be the most diverse. Chapter seven *Conceptual Data Scaling* has only a single topic and is therefore more specialized.

Duality principle of contexts

Similar to the prior data structures there is a notion of duality for formal contexts. The **dual** of a context $\mathbb{K} := (G, M, I)$ is defined as $\mathbb{K}^d := (M, G, I^{-1})$. For formal contexts we use the term dual instead of inverse since it is more commonly used in the literature and also switches the attribute and object set.

Remark 2 (Duality Principle – Contexts). *The dual of a formal context is a formal context.*

Derivation properties

There are some properties of the derivation operators that we use in the next sections.

Proposition 1 (Derivation Operators). *For a context (G, M, I) and $A, C \subseteq G$ we find the* *following properties:*

$$\begin{aligned} A \subseteq C &\implies C' \subseteq A' && (5.1)\\ A &\subseteq A'' && (5.2)\\ A' &= A''' && (5.3) \end{aligned}$$

The analogue applies to the attribute derivation due to the principle of duality for contexts.

5.2 Conceptual Data Structures

Among all sets of objects and attributes there are some that form a closed sub-structure with respect to the derivation operator. Those are called **formal concepts**. Formal concept

Definition 21 (Formal Concept). *A **formal concept** of a context $\mathbb{K} := (G, M, I)$ is a pair (A, B) with $A \subseteq G$ and $B \subseteq M$ such that*

$$A' = B \quad \text{and} \quad A = B'.$$

*The set A is called the **extent** and B the **intent** of the formal concept (A, B).*

Analogue to contexts, we often write concept instead of formal concept for simplicity. By $\text{Ext}(\mathbb{K})$ we denote the set of all extents and by $\text{Int}(\mathbb{K})$ the set of all intents of $\mathbb{K}$. Given Equation (5.3) from Proposition 1 we can infer that for an object $g \in G$ is the set $\{g\}'$ an intent of $\mathbb{K}$ called the **object intent** of g. The second application of the derivation operator yields that $\{g\}''$ is an extent of $\mathbb{K}$ which we call the **object extent** of g. Combining these two observations we get that for an object g of $\mathbb{K}$ the tuple $(\{g\}'', \{g\}')$ is a concept which is called the **object concept** of g. The analogue is defined for attributes by the duality principle.

The set of extents and intents are set systems for which there is a natural ordering. By $\underline{\text{Ext}}(\mathbb{K}) := (\text{Ext}(\mathbb{K}), \subseteq)$ and $\underline{\text{Int}}(\mathbb{K}) := (\text{Int}(\mathbb{K}), \subseteq)$ we identify the ordered sets of all extents and intents respectively. Of special interest is the set of all formal concepts of $\mathbb{K}$. Concept lattice

Definition 22 (Concept Lattice). *For a context $\mathbb{K}$ we define by $\mathfrak{B}(\mathbb{K})$ the set of all formal concepts of $\mathbb{K}$. A concept $(A, B) \in \mathfrak{B}(\mathbb{K})$ is a **sub-concept** of $(C, D) \in \mathfrak{B}(\mathbb{K})$, denoted $(A, B) \leq (C, D)$, iff $A \subseteq C$. The **concept lattice** $\underline{\mathfrak{B}}(\mathbb{K})$ of $\mathbb{K}$ is the ordered set $(\mathfrak{B}(\mathbb{K}), \leq)$.*

An important result in FCA is *The Basic Theorem on Concept Lattices*. This theorem proofs that the concept lattice of a formal context is a complete lattice. Another result of the same theorem is that any complete lattice is isomorphic to a concept lattice. We recall this theorem adapted to finite data structures in the language of this work. Basic Theorem

Theorem 2 (The Basic Theorem on Concept Lattices [80, Theorem 3]). *The concept lattice $\underline{\mathfrak{B}}(G, M, I)$ is a lattice order in which meet and join are given by:*

$$\bigwedge_{t \in T} (A_t, B_t) = \left(\bigcap_{t \in T} A_t \quad , \left(\bigcup_{t \in T} B_t \right)'' \right),$$

$$\bigvee_{t \in T} (A_t, B_t) = \left(\left(\bigcup_{t \in T} A_t \right)'', \quad \bigcap_{t \in T} B_t \quad \right).$$

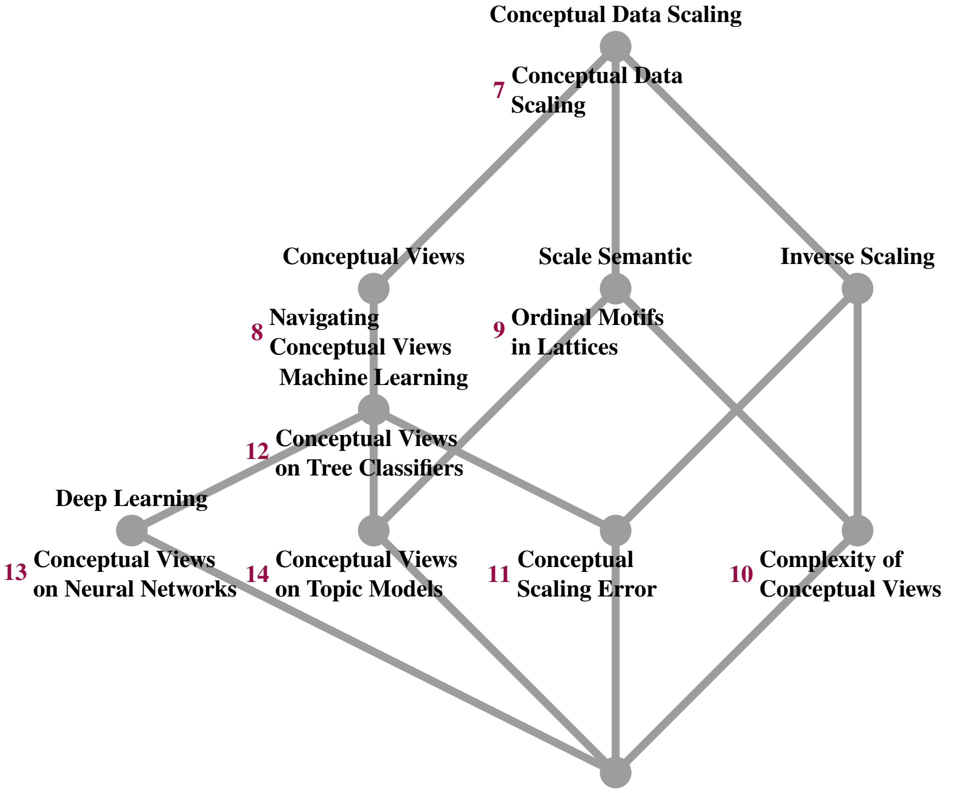


Figure 5.2: The concept lattice of the context given in Figure 5.1.

A lattice L is isomorphic to $\underline{\mathfrak{B}}(G, M, I)$ if and only if there are maps $\alpha : G \to L$ and $\beta : M \to L$ such that $\mathrm{J}(L) \subseteq \alpha(G)$, $\mathrm{M}(L) \subseteq \beta(M)$ and

$$(g, m) \in I \iff \forall g \in G, m \in M : \alpha(g) \leq \beta(m)$$

holds. In particular, $L \cong \underline{\mathfrak{B}}(L, L, \leq_L)$.

Concept lattice diagrams

We visualize concept lattices by order diagrams, as seen in Figure 4.2. However, annotating concepts to nodes would result in a very complex diagram with a lot of redundant information. By double application of Proposition 1 (5.1) we can deduce for a context $\mathbb{K} := (G, M, I)$ and $g \in G$ that the object extent of g is the smallest extent that contains g, i.e., for $A \subseteq G$ with $\{g\} \subseteq A''$ it holds that $\{g\}'' \subseteq A'''' = A''$. Thus, it is sufficient to annotate g to its object extent in the order diagram. The dual applies to the attributes of $\mathbb{K}$. Per convention, we annotate objects below nodes and attributes above. The resulting annotation style is known as **shorthand notation**. The reading rules are as follows:

i) An object is in all extents that can be reached by following upward paths.

ii) An attribute is in all intents that can be reached by following downward paths.

Example concept lattice

In Figure 5.2 we display the concept lattice of the context given in Figure 5.1 which is equal to the content lattice of this work. A description of the content lattice can be found alongside its introduction in the content section.

We frequently use the relation between the set of extents of an induced sub-context $\mathbb{K}[H, M_{\mathbb{K}}]$ and $\mathbb{K}$. We provide this relation by the following proposition which follows directly from i) in Proposition 3.2 [90] and i) in Corollary 3.3 [90].

Extents of induced sub-contexts

Proposition 2 (Extents of induced sub-contexts). *For a context $\mathbb{K} := (G, M, I)$ and $H \subseteq G$, the set of extents of $\mathbb{K}[H, M]$ is given by the following equality*

$$\operatorname{Ext}(\mathbb{K}[H, M]) := \{A \cap H \mid A \in \operatorname{Ext}(\mathbb{K})\}.$$

In the last chapter we studied with the Dedekind-MacNeille completion (cf. Definition 18) for an ordered set P, i.e., the smallest lattice L such that $P \leqq L$. A lattice that satisfies this property can be given by the concept lattice of the **general ordinal scale**, i.e., $(P, P, \leq_P)$.

Dedekind completion

Theorem 3 (Dedekind's Completion Theorem [80, Theorem 4]). *For an order $(P, \leq)$ is the map $x \mapsto (\{x\}^{\geq})^{\leq} = \downarrow x$ an order embedding from $(P, \leq)$ into $\underline{\operatorname{Ext}}(P, P, \leq)$. Moreover, $\operatorname{DM}(P, \leq)$ is isomorphic to $\underline{\mathfrak{B}}(P, P, \leq)$.*

The size of a concept lattice can be exponential [4] in the number of attributes or objects. For example the context $([k], [k], \neq)$ has 2^k concepts. The algorithm with the best worst case complexity has a computational cost of $O\big(|G| \cdot (|G| + |M|)\big)$ per concept [165].

Concept lattice size

Problem 7: Concept Lattice Computation	
Input: A formal context $\mathbb{K} := (G, M, I)$	
Output: $\underline{\mathfrak{B}}(\mathbb{K})$	
Complexity:	$O\big(\lvert\underline{\mathfrak{B}}(\mathbb{K})\rvert \cdot \lvert G\rvert \cdot (\lvert G\rvert + \lvert M\rvert)\big)$

While this algorithm has the best worst case complexity we often use the `next_closure` [76] in our theoretical investigations, since it lets us determine the order in which concepts are computed. This algorithm has a runtime complexity of $O\big(|G|^2 \cdot |M|\big)$ per concept. For an arbitrary linear order $\leq$ on the set of attributes the algorithm enumerates the set of concepts in the **lectic order**, i.e., $A \subseteq B$ iff the smallest m in the **symmetric difference** $A \vartriangle B := (A \setminus B) \cup (B \setminus A)$ is in A. An analogy that is often used for the lectic order is that of enumerating words in $\{0, 1\}^{|M|}$ with the order relation for binary numbers.

The associated counting problem [130] is known to be #P-complete.

Counting concepts

Problem 8: Concept Lattice Size Problem	
Input: A formal context $\mathbb{K}$	
Output: $\lvert\underline{\mathfrak{B}}(\mathbb{K})\rvert$	
Complexity:	#P-complete

Not every object or attribute is integral to the structure of the concept lattice. Some of them can be removed without changing the order relation.

Clarified context

Definition 23 (Clarified Context). *A formal context (G, M, I) is called **object clarified** iff for any two objects $g, h \in G$ the equality $\{g\}' = \{h\}'$ implies that $g = h$. A context is called **attribute clarified** iff $\mathbb{K}^d$ is object clarified. A context that is object and attribute clarified is called **clarified**.*

By clarifying a context we address the problem of computing an induced sub-context that is clarified with isomorphic concept lattice. This can be done polynomial in the size of the context.

Problem 9: Clarify Context Problem	
Input: A formal context $\mathbb{K}$	
Output: An induced sub-context $\mathbb{S}$ of $\mathbb{K}$ that is reduced with $\underline{\mathfrak{B}}(\mathbb{S}) \cong \underline{\mathfrak{B}}(\mathbb{K})$	
Complexity:	$O(\|G\|^2 \cdot \|M\| + \|G\| \cdot \|M\|^2)$

Reduced context

Other objects (attributes) that can be removed are those whose object (attribute) concepts are join-reducible (meet-irreducible) in the concept lattice.

Definition 24 (Reduced Context). *A clarified formal context (G, M, I) is* ***object reduced*** *iff every object concept is join-irreducible in $\underline{\mathfrak{B}}(\mathbb{K})$. A context is* ***attribute reduced*** *iff its dual context is object reduced. A context that is object and attribute reduced is called* ***reduced****. An object or attribute that contradicts the reduced property is called* reducible.

For a formal context there is up to context isomorphism a unique induced sub-context that is reduced. This context can be computed by clarifying a context and then checking for each object and attribute if it satisfies Proposition 13 from Ganter and Wille [80].

Problem 10: Reduce Context Problem	
Input: A formal context $\mathbb{K}$	
Output: An induced sub-context $\mathbb{S}$ of $\mathbb{K}$ that is reduced with $\underline{\mathfrak{B}}(\mathbb{S}) \cong \underline{\mathfrak{B}}(\mathbb{K})$	
Complexity:	$O(\|G\|^2 \cdot \|M\|^2)$

Duality principle for concept lattices

The duality principle of formal contexts extends to the concept lattice.

Remark 3 (Duality Principle – Concept Lattice). *For a formal context $\mathbb{K}$ is*

$$\underline{\mathfrak{B}}(\mathbb{K}^d) \cong \underline{\mathfrak{B}}(\mathbb{K})^{-1} \text{ and } \underline{\mathrm{Ext}}(\mathbb{K}) \cong \underline{\mathrm{Int}}(\mathbb{K})^{-1}.$$

5.2.1 Closure Systems

Connection to closure systems

The double application of the derivation operator, i.e., $A \mapsto A''$, yields a **closure operator** on the object set G and attribute set M respectively. For a set S is a map $\mathrm{cl} : \mathcal{P}(S) \to \mathcal{P}(S)$ a closure operator iff

$$X \subseteq \mathrm{cl}(X) \quad \text{(extensive)}$$
$$X \subseteq Y \implies \mathrm{cl}(X) \subseteq \mathrm{cl}(Y) \quad \text{(increasing)}$$
$$\mathrm{cl}(X) = \mathrm{cl} \circ \mathrm{cl}(X) \quad \text{(idempotent)}$$

By $\mathrm{cl}_{\mathbb{K}}$ we refer to either the object or attribute closure operator of context $\mathbb{K}$. At each occurrence, we declare which of them is used. The set of all **fixed points** or **closed sets**, i.e., $\{\mathrm{cl}(X) \mid X \subseteq S\}$, is the **closure system** of cl.

Definition 25 (Closure System). *A set system $\mathcal{A}$ on a set S is a* ***closure system*** *iff it is closed with respect to intersections $\bigcap \mathcal{B}$ with $\mathcal{B} \subseteq \mathcal{A}$. The finer and coarser, and (induced) sub-closure system definitions are given in accordance to $(S, \mathcal{A})$-structures.*

The intersection of the empty subset $\bigcap\{\}$ is equal to S. Thus, S is in every closure system on S. For finite sets S is the set system $\mathcal{A}$ a closure system iff $S \in \mathcal{A}$ and $\mathcal{A}$ is closed with respect to pairwise intersections. The set of extents $\mathrm{Ext}(\mathbb{K})$ as well as the set of intents $\mathrm{Int}(\mathbb{K})$ are closure systems on G and M respectively.

5.3 Context Products

Combining and joining data is an important aspect when deriving knowledge. In the realm of FCA there are several context operations that we make use of.

Semi-product of contexts

The first operation that we consider is the **semi-product** of two formal contexts $\mathbb{K} := (G, M, I)$ and $\mathbb{S} := (H, N, J)$, where $I \diamond J := \{((g,h),(m,n)) \mid (g,m) \in I \wedge (h,n) \in J\}$:

$$\mathbb{K} \between \mathbb{S} := (G \times H,\ \dot{M} \cup \dot{N},\ I \diamond J) \qquad \text{(Context semi-product)}$$

The concepts of $\underline{\mathfrak{B}}(\mathbb{K} \between \mathbb{S})$ are given by the pairs $(A \times B, C \cup D)$ where $(A, C) \in \underline{\mathfrak{B}}(\mathbb{K})$, $(B, D) \in \underline{\mathfrak{B}}(\mathbb{S})$ and $A = \bot_{\mathbb{K}} \iff B = \bot_{\mathbb{S}}$. Thus, the concept lattice of the context semi-product is isomorphic to the product of the individual concept lattices with a combined bot element.

(Disjoint) union of contexts

The next operation on contexts $\mathbb{K}, \mathbb{S}$ that we recall are the **union** and **disjoint union**:

$$\mathbb{K} \cup \mathbb{S} := (G \cup H, M \cup N, I \cup J) \qquad \text{(Context Union)}$$
$$\mathbb{K} \dot\cup \mathbb{S} := (G \dot\cup H, M \dot\cup N, I \dot\cup J) \qquad \text{(Disjoint Context Union)}$$

The operation $\dot\cup$ is the union of disjoint sets. Our theoretical investigations are not concerned with the naming of objects therefore we often assume without loss of generality that the operands are disjoint. Otherwise, they can be *colored* with some identifier i, i.e., $A_1 \dot\cup A_2 := \{1\} \times A_1 \cup \{2\} \times A_2$.

Apposition of contexts

The last operation that we use is the **apposition** of contexts $\mathbb{K}, \mathbb{S}$ with $G = H$:

$$\mathbb{K} \mid \mathbb{S} := (G, M \dot\cup N, I \dot\cup J) \qquad \text{(Context Apposition)}$$

Extents of appositions

The **subposition** of two contexts $\frac{\mathbb{K}}{\mathbb{S}}$ is equivalently defined by the disjoint union of the object sets. The extents of context apposition $\mathbb{K} \mid \mathbb{S}$ are known to be equal to the set of all $A \cap B$ where $A \in \mathrm{Ext}(\mathbb{K})$ and $B \in \mathrm{Ext}(\mathbb{S})$. We include a small proof for this statement, since this is not explicitly stated in the literature.

Lemma 1 (Context Apposition). *For two contexts $\mathbb{K}$ and $\mathbb{S}$ with $G = G_{\mathbb{K}} = G_{\mathbb{S}}$ is a set $D \subseteq G$ an extent of $\mathbb{K} \mid \mathbb{S}$ iff it can be written as $A \cap B$ for $A \in \mathrm{Ext}(\mathbb{K})$ and $B \in \mathrm{Ext}(\mathbb{S})$.*

Proof. We split the proof in the following to parts:

$\Longrightarrow$ An extent D of $\mathbb{K} \mid \mathbb{S}$ can be written as the intersection of all attribute derivations $\{m\}'$ with $m \in D'$. Let E, F be a bi-partition of D given by the context the attributes are from, i.e., $E = D' \cap M_{\mathbb{K}}$ and $F = D' \cap M_{\mathbb{S}}$. Concluding, D equals $E' \cap F'$ with $E' \in \mathrm{Ext}(\mathbb{K})$ and $F' \in \mathrm{Ext}(\mathbb{S})$.

$\Longleftarrow$ For an $A \in \mathrm{Ext}(\mathbb{K})$ is $A^{I_{\mathbb{K}}} \subseteq M_{\mathbb{K}|\mathbb{S}}$. For a $B \subseteq M_{\mathbb{K}}$ is $B^{I_{\mathbb{K}}} = B^{I_{\mathbb{K}|\mathbb{S}}}$. Following, is $A^{I_{\mathbb{K}} I_{\mathbb{K}|\mathbb{S}}} = A^{I_{\mathbb{K}} I_{\mathbb{K}}} = A$ and thus $A \in \mathrm{Ext}(\mathbb{K} \mid \mathbb{S})$. The same applies to a $B \in \mathrm{Ext}(\mathbb{S})$. The remainder follows from the intersection property of closure systems. □

The semantics of the introduced context operations is discussed more thoroughly in Chapter 7 with respect to data scalings.

5.4 Implicational Theory

Alongside formal concepts we often study the theory of the attribute closure system $\text{Int}(\mathbb{K})$ of a context $\mathbb{K} := (G, M, I)$ with respect to the Horn fragment

$$\mathcal{H}(M) := \{\varphi \to \psi \mid \varphi, \psi \in F[M, \{\wedge\}]\}.$$

Model relation

To define the model relationship for set systems and logical expressions from $\mathcal{H}(M)$ we identify the variable symbols of a sentence $\varphi \in F[M, \{\wedge\}]$, denoted $\text{var}(\varphi)$, with elements from the attribute set M. Moreover, we often write $A \to B$ with $A, B \subseteq M$ in short to identify the implication $\varphi \to \psi$ with $\text{var}(\varphi) = A$ and $\text{var}(\psi) = B$.

Definition 26 (Model Relation – Formal Contexts). *A set $C \subseteq M$ is a* ***model*** *of an implication $A \to B \in \mathcal{H}(M)$ iff $A \nsubseteq C$ or $B \subseteq C$. A set system $\mathcal{A}$ on M is a model of an implication iff all elements of $\mathcal{A}$ are a model of the implication. A set system is a model of an implicational $\mathcal{H}(M)$ theory $\mathcal{T}$ iff all elements of $\mathcal{A}$ are in model relation to all implications of $\mathcal{T}$.*

A context $\mathbb{K} := (G, M, I)$ is a model of $A \to B \in \mathcal{H}(M)$ iff $\text{Int}(\mathbb{K}) \models A \to B$. *An object $g \in G$ is a model of $A \to B \in \mathcal{H}(M)$ iff $\{g\}' \models A \to B$.*

Theory of a context

The model relationship between contexts and implications is often tested using the following characterization

$$(G, M, I) \models A \to B \quad \iff \quad A^I \subseteq B^I.$$

Per default we identify by the theory of a context $\text{Th}(\mathbb{K})$ the theory on its attribute closure system $\text{Th}_M(\mathbb{K})$. The latter is defined as the set of all implications from $\mathcal{H}(M_{\mathbb{K}})$ for which $\mathbb{K}$ is a model. Besides the theory based on the attributes, we also study the theory of object implications $\mathcal{H}(G_{\mathbb{K}})$, i.e., $\text{Th}_G(\mathbb{K}) := \text{Th}_M(\mathbb{K}^d)$.

Implications in the concept lattice diagram

Using the characterization of attribute implications from above, we are able to infer implications $A \to B$ from the sub-concept relation of their generated concepts $(A', A'') \leq (B', B'')$. Thus, we can read attribute/object implications from concept lattice diagrams. Namely, from upward/downward paths between concept annotations in the order diagram. For example, we find that *Machine Learning* implies *Conceptual Views*. This indicates that we study machine learning only with respect to their conceptual views. This implication is true within the context of this thesis but does not need to be true in general.

Implicational theory and closure systems

There is a one-to-one relationship between closed $\mathcal{H}(M)$ theories and closure operators on M. A closure system is uniquely defined by its implicational theory and for a closed theory $T \subseteq \mathcal{H}(M)$ is the mapping $A \mapsto \text{cl}_T(A) := \bigcup\{Y \mid X \to Y \in T \text{ and } A \subseteq X\}$ a closure operator. Moreover, for two closed theories $T, S \subseteq \mathcal{H}(M)$ does the equality $\text{cl}_T = \text{cl}_S$ imply that $T \cong S$. We use this connection to switch between representations.

Implicational basis

The theory of a context is too large and includes a lot of redundant information. Therefor, we study an implicational basis $\mathcal{B} \subseteq \text{Th}(\mathbb{K})$ with $B \vdash \text{Th}(\mathbb{K})$. There are many proposed basis in the literature [220]. The **Duquenne-Guigues base** [85] or **canonical base** is known to be the smallest base in size and is recursively defined using **pseudo-intents**.

Definition 27 (Pseudo-Intent). *A set $A \subseteq M$ of a context (G, M, I) is a* ***pseudo-intent*** *iff 1) $P \notin$* $\text{Int}(\mathbb{K})$ *and 2) $Q'' \subseteq P$ for every pseudo-intent $Q \subseteq P$ with $Q \neq P$.*

The canonical base is given as the set of all implications generated from pseudo-intents.

Definition 28 (Canonical Base). *The **canonical base** of a context $\mathbb{K}$ is the set of all implications*

$$\mathfrak{C}(\mathbb{K}) := \{P \to P'' \mid P \text{ is a pseudo-intent of } \mathbb{K}\}.$$

Pseudo-intents are hard to recognize [13] due to the recursive definition. The associated decision problem is given below. Complexity

Problem 11: Pseudo-Intent Problem	
Input: A formal context $\mathbb{K} := (G, M, I)$ and $A \subseteq M$	
Output: True iff A is a pseudo-intent in $\mathbb{K}$	
Complexity:	co-NP-complete

The number of pseudo-intents and thereby the size of the canonical base can grow exponential in the number of attributes [129]. The complexity of the counting problem [133] is given below.

Table 5.1: Canonical base of the content lattice context.

Implication			support
{}	→	*Conceptual Data Scaling*	1
Conceptual Data Scaling *Machine Learning*	→	*Conceptual Views*	0.5
Conceptual Data Scaling *Conceptual Views* *Inverse Scaling*	→	*Machine Learning*	0.125
Conceptual Data Scaling *Deep Learning*	→	*Machine Learning* *Conceptual Views*	0.125
Conceptual Data Scaling *Scale Semantic* *Conceptual Views*	→	*Unsupervised Learning*	0.125
Conceptual Data Scaling *Machine Learning* *Deep Learning* *Conceptual Views* *Inverse Scaling*	→	*Scale Semantic*	0.0
Conceptual Data Scaling *Conceptual Views* *Inverse Scaling* *Machine Learning* *Scale Semantic*	→	*Deep Learning*	0.0
Conceptual Data Scaling *Conceptual Views* *Machine Learning* *Deep Learning* *Scale Semantic*	→	*Scale Semantic*	0.0

Problem 12: Canonical Base Size Problem	
Input: A formal context $\mathbb{K}$	
Output: $\lvert\mathfrak{C}(\mathbb{K})\rvert$	
Complexity:	# P-complete

Example implications

The list of all implications in the canonical base of the content lattice context is depicted in Table 5.1. The base contains eight implications with varying premise and conclusion size. The previously read implication from the diagram, i.e, *Machine Learning* implies *Conceptual Views* can also be found in this list. In addition to that we find three implications for which there are no chapters that satisfy the premise. These rules are a description for the closure operator of attribute sets whose output is the entire set of attributes.

Support and confidence

A quantity that describes this observation is the **support** of an implication in a context:

$$\mathrm{supp}(A \to B) := \frac{|(A \cup B)'|}{|G|} \qquad \text{(support)}$$

In Table 5.1 we report alongside each implication their support values. From the table we can infer that *Machine Learning* → *Conceptual Views* has the highest support of all implications. This observation is supported by the aim of this thesis to deepen the understanding of conceptual measurement and apply it to machine learning data representations.

Association rule

A more relaxed notion of implications among attribute sets are **association rules** which are only partially in model relation with a context. The *correctness* of an association rule in a context is described by the **confidence**:

$$\mathrm{conf}(A \to B) := \frac{|(A \cup B)'|}{|A'|} \qquad \text{(confidence)}$$

Luxenburg basis

Analogously, to the implicational theory of a context there is a basis for the set of all association rules that have a minimum confidence and the premise as well as conclusion have a minimum support. The **Luxenburg basis** [146, 208] which contains all rules $X \to Y$, such that 1) X and Y satisfy the support and confidence criterion and 2) there are two concepts $(A, B), (C, D) \in \underline{\mathfrak{B}}(\mathbb{K})$ that are in cover relation $(A, B) \prec (C, D)$, and $X = D$ and

Table 5.2: Luxenburger basis of the content lattice context for minimum support of 0.2 and minimum confidence of 0.5.

Association Rules			support	confidence
Conceptual Data Scaling	→	*Conceptual Views*	1.0	0.625
Conceptual Data Scaling Conceptual Views	→	*Machine Learning*	0.625	0.8
Conceptual Data Scaling Inverse Scaling	→	*Machine Learning Conceptual Views*	0.25	0.5
Conceptual Data Scaling Inverse Scaling	→	*Scale Semantic*	0.25	0.5

$X \cup Y = B$. Thus, the rules in the Luxenburg basis can be read from downward lines in the concept lattice diagram. The deduction rules are the usual where for two rules $X \to Y$ and $Y \to Z$ the confidence $\operatorname{conf}(X \to Z)$ is equal to $\operatorname{conf}(X \to Y) \cdot \operatorname{conf}(Y \to Z)$. In Table 5.2 we depict all association rules of the Luxenburg basis of the content lattice context.

5.5 Conceptual Scaling

Many-valued data

Data is not always in a format that is compatible with the formal context structure. For this we have to translate attributes with respect to some formal prescription on how to interpret the data on the ordinal level. The process of defining such prescription and applying them to derive a formal context is referred to as **conceptual scaling** [74]. In this section we recall only the basic notions from conceptual scaling, since we discuss and contribute to this topic more in the upcoming parts. The most common procedure applied to scale data sets is known as **plain scaling** of **many-valued contexts**.

Many-valued context

Definition 29 (Many-Valued Context). *A **many-valued context** is a tuple $\mathbb{D} := (G, M, W, I)$ where G is a set of objects, M a set of **many-valued attributes**, W a set of **attribute values** and $I \subseteq G \times M \times W$ a relation with $(g, m, w), (g, m, v) \in I$ implies $w = v$. A pair $(g, m, w) \in I$ is interpreted as* "object g has value v for attribute m". *A many-valued context is called **complete** iff for all $g \in G, m \in M$ there exists is a $w \in W$ with $(g, m, w) \in I$.*

By $m_{\mathbb{D}}(g)$ we denote the value w that g has for attribute m in $\mathbb{D}$ and the **domain** of m by

$$\operatorname{dom}_{\mathbb{D}}(m) := \{w \in W_{\mathbb{D}} \mid \exists g \in G : (g, m, w) \in I_{\mathbb{D}}\}.$$

The absence of values is allowed. By $m(g) = \bot$ we indicate that a value is missing.

Scale context

In plain scaling we define for each many-valued attribute m in $\mathbb{D}$ a **scale** context.

Definition 30 (Scale Context). *A **scale** for the many-valued attribute m of a many-valued context $\mathbb{D}$ is a context $\mathbb{S}_m := (G_m, M_m, I_m)$ where $\operatorname{dom}_{\mathbb{D}}(m) \subseteq G_m$. The elements of G_m are called the **scale values** and M_m are called the **scale attributes**.*

Defining scales

By defining a scale for an attribute we provide an interpretation of the attribute's domain on the ordinal level with respect to the scale's concept lattice. There are many ways on how to define such a scale depending on the interpretation and perspective of the data analyst. There are some *standard interpretations* of attribute domains via scales that reflect specific semantics from the realm of measurability, e.g., *linear ordered*, *incomparable*, *interval* or *partitioning* values.

Plain scaling

The derived context with respect to plain scaling is defined as follows.

Definition 31 (Plain Scaling). *Let $\mathbb{D} = (G, M, W, I)$ be a many-valued context and $\mathbb{S}_m$ be a scale context for each many-valued attribute. The **derived context** is defined as*

$$(G, \bigcup_{m \in M} \{m\} \times M_m, J) \text{ with } (g, (m, v)) \in J :\Longleftrightarrow (m_{\mathbb{D}}(g), v) \in I_m.$$

6

Metric Data and Similarity Measures

In Part III we focus on applying methods from conceptual measurability to analyze the inherent data representations of machine learning models. Those are often numeric representation on the ratio level of measurement. For those, we recall in this section basic structures and operations from the literature.

The most commonly used structure for numeric data is the k-dimensional $\mathbb{R}$-**vector space**, i.e., a structure associated with the following operations:

Vector space on real numbers

$$+ : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^k \text{ with } (x_1, \ldots, x_k) + (y_1, \ldots, y_k) := (x_1 + y_1, \ldots, x_k + y_k), \qquad \text{(vector addition)}$$

$$\cdot : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^k \text{ with } a \cdot (x_1, \ldots, x_k) := (a \cdot x_1, \ldots, a \cdot x_k), \qquad \text{(scalar multiplication)}$$

$$\cdot : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R} \text{ with } (x_1, \ldots, x_k) \cdot (y_1, \ldots, y_k) := x_1 \cdot y_1 + \cdots + x_k \cdot y_k. \qquad \text{(dot product)}$$

Elements from $\mathbb{R}^k$ are called **vectors** for which we use symbols $\vec{x} := (x_1, \ldots, x_k)$. We use for both the dot product and the scalar multiplication the $\cdot$ symbol. The present operation can be inferred from the input type, the used operand symbols or the surrounding text. For the **absolute** of a vector $\vec{x}$ exist several variants that are characterized as L**-norms**:

L-norms

$$\|\vec{x}\|_p := \Big(\sum_{1 \le i \le k} |x_i|^p \Big)^{p^{-1}} \qquad \text{for } p \in \mathbb{N}_{>0}. \qquad \text{(generalized } L\text{-norm)}$$

For the special case of $p = 1$ we identify the $L1$-norm by $|\vec{x}| := \|\vec{x}\|_p$ and for $p = 2$ the $L2$-norm by $\|\vec{x}\| := \|\vec{x}\|_2$.

The dot product of vectors does naturally extend to products of matrices A of size $m \times n$

Matrix operations

and B of size $n \times p$:

$$\begin{pmatrix} \vec{a}_1 \\ \vdots \\ \vec{a}_m \end{pmatrix} \cdot \begin{pmatrix} \vec{b}_1 & \cdots & \vec{b}_p \end{pmatrix} = \begin{pmatrix} \vec{a}_1\vec{b}_1 & \cdots & \vec{a}_1\vec{b}_p \\ \vdots & \ddots & \vdots \\ \vec{a}_m\vec{b}_1 & \cdots & \vec{a}_m\vec{b}_p \end{pmatrix} \qquad \text{(matrix product)}$$

Other spaces

Other spaces of vector like elements that we encounter explicitly and implicitly throughout this work are the **binary space** using $\{0,1\}^k$ where $+_B, \cdot_B$ are the usual with the exception of $1 +_B 1 = 1$. In addition to that we use the **Boolean space** $\{\perp, \top\}^k$ with the standard logical operations. Although these spaces are isomorphic, we use the Hamming cube for numeric representations and the Boolean space in logical settings. We may note that the later two are no vector spaces. Hence, we refer to them simply as **space** or **coordinate system**.

6.1 Quantitative Comparison

Besides comparing elements using order relations there are special data representations that allow for a quantitative comparison of elements. These measures can not only be used to interpret results but are also used as data representations in many machine learning models.

Metric Space

The first approach measures the **distances** between object representations.

Definition 32 (Pseudo Metric Space). *A **pseudo-metric space** is a tuple (M, d) with $d : M \times M \to \mathbb{R}$ where d satisfies the following properties for $x, y, z \in M$:*

$$d(x,x) = 0$$
$$d(x,y) = d(y,x) \qquad \text{(symmetry)}$$
$$d(x,z) \leq d(x,y) + d(y,z) \qquad \text{(triangle inequality)}$$

In some settings we have multiple objects with the same numeric vector representation. This leads to unequal elements that have zero distance between them. In case only equal elements have a zero distance we get a **metric space**.

Definition 33 (Metric Space). *A **metric space** is a pseudo-metric space (M, d) that satisfies*

$$x \neq y \implies d(x,y) > 0 \qquad \text{(positivity)}$$

*for $x, y \in M$. The function d is called a **metric**.*

In case the function d satisfies the $d(x,x) = 0$, symmetry and positivity property, we call the structure a **distance space** and the function d a **distance function**.

Embedding

Metric spaces use a relaxed notion for embeddings that allows for some difference between the distance functions. A map $\alpha : A \to B$ between two metric spaces is an embedding with **distortion** $C > 0$ and constant $c > 0$ iff

$$c \cdot d_A(x,y) \leq d_B(\alpha(x), \alpha(y)) \leq c \cdot C \cdot d_A(x,y)$$

for all $x, y \in A$. In case this inequality holds for $c, C = 1$ the embedding is called isometric. For finite metric spaces there is for any injective map α a pair of c, C such that α can be considered an embedding. Thus, the term embedding is often simply used for maps that aim to preserve the distances.

For numeric data we employ a distance function based on the L-norms of the vector difference: Metric examples

$$d_p(\vec{x}, \vec{y}) := \|\vec{x} - \vec{y}\|_p$$

The special case of $p = 2$ is called the **Euclidean distance** and **Hamming distance** for $p = 1$. For sets and set systems we use the absolute of the symmetric difference $\triangle$ and for connected graphs G the length of the shortest path d_G.

A second space to compare elements uses **similarity functions**. Similarity Space

Definition 34 (Similarity Space). *A* ***similarity space*** *is a tuple* (M, s) *with* $s : M \times M \to [0, 1]_{\mathbb{R}}$ *where* s *satisfies the following properties for* $x, y \in M$*:*

$$d(x, x) = 1$$
$$d(x, y) = d(y, x) \qquad \text{(symmetry)}$$

The map s *is called a* ***similarity measure****.*

A commonly applied similarity function for numeric data is the **cosine similarity** between two non-zero vectors in the $\mathbb{R}$-vector space: Similarity examples

$$\cos(\vec{x}, \vec{y}) := \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} \qquad \text{(cosine similarity)}$$

For sets or set systems we use the **Jaccard index** between two sets A, B:

$$\mathrm{J}(A, B) := \frac{|A \cap B|}{|A \cup B|} \qquad \text{(Jaccard index)}$$

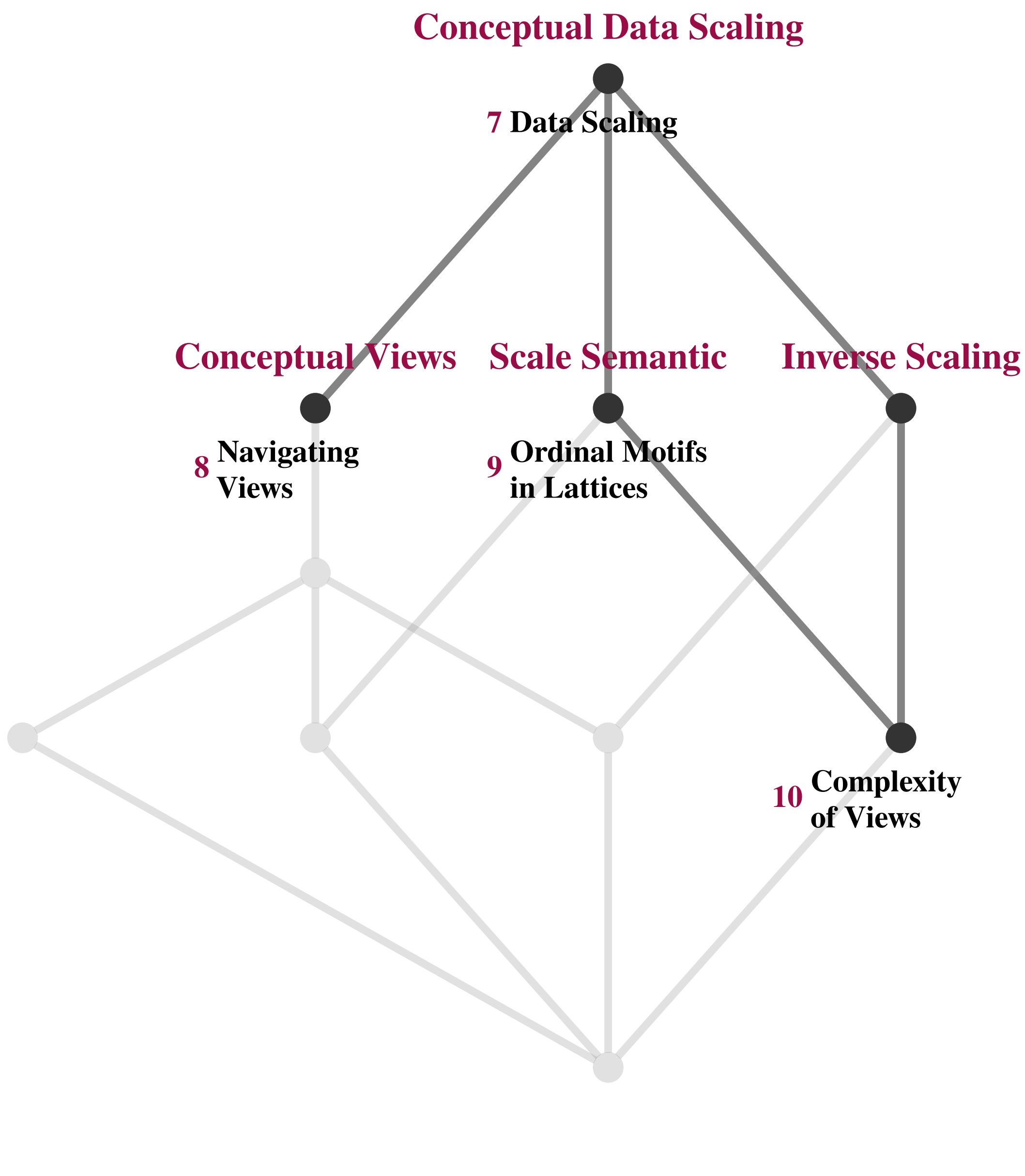
Distance functions d and similarity functions s can be used interchangeably by the following inversion and normalization: From distance to similarity

$$d_s(x, y) := 1 - \quad \text{s(x,y)} \qquad \text{(similarity to distance)}$$
$$s_d(x, y) := 1 - \frac{d(x, y)}{\max\limits_{x,y \in X} d(x, y)} \qquad \text{(distance to similarity)}$$

Other distance and similarity measures will be defined in the following parts when they are needed. In some cases we will define non-symmetric distance and similarity functions.

Part II

Conceptual Measurability



7 Conceptual Data Scaling

Stages in data scaling

In order to analyze heterogeneous data on the ordinal level, e.g., for (closed) pattern mining [114], ontology learning [41] or machine learning [87, 98], conceptual scaling [74] is a tool of choice. There are three stages that we identify in the scaling process (see Figure 7.1).

Stage: raw data

First, there is the original data set whose attributes can be given on any level of measurement [201]. We describe this data set by a relational data structure, namely a many-valued context $\mathbb{D} := (G, M, W, I)$. The next stage is a formal context $\mathbb{K}$ derived from $\mathbb{D}$ by interpreting each attribute $m \in M$ on the ordinal level. This is done by defining scale contexts $\mathbb{S}_m$. There are many ways to do this, but not all methods are meaningful in every situation. Often the data has an implicit structure that guides the scaling. For example, it is natural to encode the order relation of attributes that are on the ordinal level in the defined scales. We propose the notion of *pre-scaling* to encode implicit or background knowledge and make them accessible to scaling.

Stage: ordinal interpretation

Stage: ordinal data reduction

The size of the derived context increases with the number of values $w \in W$ and their scale attributes in the scales $\mathbb{S}_m$. Thus, even relatively small many-valued contexts may result

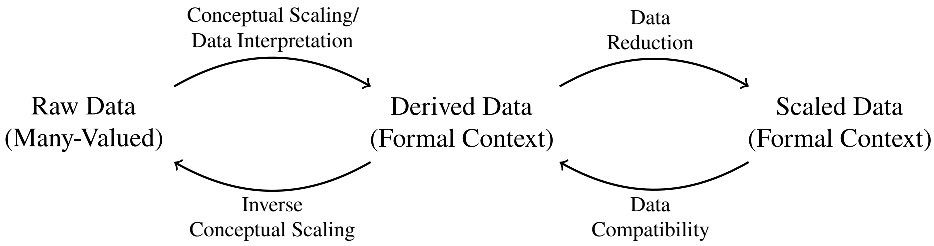


Figure 7.1: An overview of methods of scaling methods for conceptual structures.

in very large derived contexts and concept lattices. In a second step we apply data reduction to compute a *scaled context* $\mathbb{S}$ from $\mathbb{K}$ with reduced complexity. In this work, we propose a framework to characterize consistent conceptual data reductions that is based on continuous maps. Within this framework we understand the scaled context $\mathbb{S}$ as a *view* of $\mathbb{K}$ that reflects parts of interest from the conceptual structure of $\mathbb{K}$. This framework is agnostic with respect to the data reduction method and can therefore be applied to many different methods from machine learning. In the following chapters. we present plenty of contributions in this area that range from the computation, combination and interpretation of views $\mathbb{S}$.

Reductions as views

Separate stages for better interpretation

The process of data reduction is often integrated into the scaling process by defining scales $\mathbb{S}_m$ whose values include some form of aggregation. We argue that for a clear and precise interpretation of the data and resulting analytical findings it is important to separate these tasks. We elaborate on this in greater detail in Section 7.3.

Data compatibility and error

The introduced framework of views can also be used to verify whether the output of a data reduction method is consistent with respect to its conceptual structure. This is the case iff the reduction map is continuous. When using inconsistent data representations it is crucial to identify the parts that cause the inconsistencies. By doing so, we can assess the quality of individual concepts and implications in the scaled context $\mathbb{S}$ with respect to the derived context $\mathbb{K}$. We introduce foundation a notion of error in conceptual scaling in Section 7.5. Methods on how to deal with this error are presented in Chapter 11 based on Binary Matrix Factorization.

Inverse scaling

With inverse conceptual scaling we study from which scales $\mathbb{S}_m$ and data sets $\mathbb{D}$ a context $\mathbb{K}$ can be derived from. We recall in Section 7.4 the state-of-the-art from the literature. In the following chapters we have developed many new applications based on this theory. In Chapter 10 we introduce a new notion of dimension and show how to find a representation of $\mathbb{D}$ with fewer – possibly compressed – features. In Chapter 9 we introduce a new method based on the identification of individual scales that allows us to automatically interpret the concept lattice of $\mathbb{K}$ and generate textual explanations for them. This is, to the best of our knowledge, the first method that is able to do this and opens a promising new line of research to improve the interpretability of Formal Concept Analysis for untrained users.

Applications of inverse scaling

7.1 Conceptual Scaling

Standard scales

We introduced in Section 5.5 the basics on how to scale a many-valued context with plain scaling. In this section we introduce the **standard scales** from Ganter and Wille [74] that we use to interpret data on the ordinal level. For each scale we provide formal definitions, their basic meaning of interpretation [74, 80] and example data.

Nominal scale

The first scale is the **nominal scale** and it is used to describe that attribute values are incomparable and have nothing in common. Hence, the *basic meaning* of *partition* [80, Figure 1.26]. The scale attributes induce a partition on the set of objects based on their values. Let $[n] \coloneqq \{1, \ldots, n\}$ be the set of natural numbers from one to n, the scale is defined to be

$$\mathbb{N}_n \coloneqq ([n], [n], =) \tag{Scale}$$

and its object theory is equivalent to

$$\mathrm{Th}_G(\mathbb{N}_n) \cong \{i, j \Rightarrow 1, \ldots n \mid 1 \leq i < j \leq n\}. \tag{Object Theory}$$

Nominal features

In Figure 7.2 we depict the scale as well as its concept lattice. This scale is commonly used to scale categorical values such as categories, classes, names or clusters. A special variant of nominal scales are **dichotomic scales**. They are isomorphic to the $\mathbb{N}_2$ scale and are defined on the values *True* and *False*.

Nominal scales in ML

In machine learning a nominal interpretation of the data is also understood as *OneHotEncoding*[1], which is often used for all attributes that are not numeric.

$\mathbb{N}_n$	1	2	3	...	$n-1$	n
1	×					
2		×				
3			×			
...						
$n-1$					×	
n						×

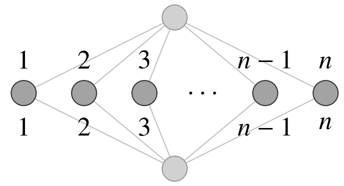


Figure 7.2: This figure shows the formal context of the nominal scale $\mathbb{N}_n$ on n elements (left) and its concept lattice (right).

Ordinal scale

The next two scales are for attributes whose domains are linearly ordered. Examples are physical quantities, scores, rankings or food chains. The first of the two is the **ordinal scale**. Its attributes values induce a linear order on the scale values and its basic meaning is *ranking* [80, Figure 1.26]. The scale is defined to be

$$\mathbb{O}_n := ([n], [n], \geq) \qquad \text{(Scale)}$$

and its object theory is equivalent to

$$\text{Th}_G(\mathbb{O}_n) \cong \{i \Rightarrow i+1 \mid 1 \leq i \leq n-2\} \cup \{\emptyset \Rightarrow n\}. \qquad \text{(Object Theory)}$$

In Figure 7.3 we depict the scale as well as its concept lattice.

$\mathbb{O}_n$	1	2	3	...	$n-1$	n
1	×					
2	×	×				
3	×	×	×			
...	×	×	×	×		
$n-1$	×	×	×	×	×	
n	×	×	×	×	×	×

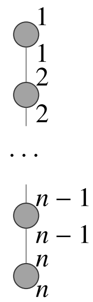


Figure 7.3: This figure shows the formal context of the *ordinal scale* $\mathbb{O}_n$ on n elements (left) and its concept lattice (right).

Sometimes we also consider $([n], [n+1], \geq)$ to be an ordinal scale. This scale has in addition to $\text{Ext}([n], [n], \geq)$ the empty set as extent and its theory is equivalent to $\{i \Rightarrow i-1 \mid 2 \leq i \leq n\}$. The ordinal scale can be used to analyze sets of objects that are bound from below by some value. Upper bounds can be analyzed using the dual scale. To

Interordinal scale

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html`, 06.2023

analyze upper and lower bounds at the same time, i.e., arbitrary intervals within a linear order, we use the **interordinal scale**. The scale is defined to be

$$\mathbb{I}_n := ([n], [n], \leq) \mid ([n], [n], \geq), \qquad \text{(Scale)}$$

its object theory is equivalent to

$$\mathrm{Th}_G(\mathbb{I}_n) \cong \{i, j \Rightarrow i, \ldots, j \mid 1 \leq i < j \leq n\}, \qquad \text{(Object Theory)}$$

and its basic meaning is *betweenness relation* [80, Figure 1.26]. It is used for linear ordered domains in which we are interested in intervals, such as time periods in time data, spectral colors in wavelengths or in tree classifiers. We discuss the later in more detail in Chapter 12.

$\mathbb{I}_n$	$1_\leq$	$2_\leq$	$3_\leq$	$\ldots$	$(n-1)_\leq$	$n_\leq$	$1_\geq$	$2_\geq$	$3_\geq$	$\ldots$	$(n-1)_\geq$	$n_\geq$
1	×	×	×	×	×	×	×					
2		×	×	×	×	×	×	×				
3			×	×	×	×	×	×	×			
...				×	×	×	×	×	×	×		
$n-1$					×	×	×	×	×	×	×	
n						×	×	×	×	×	×	×

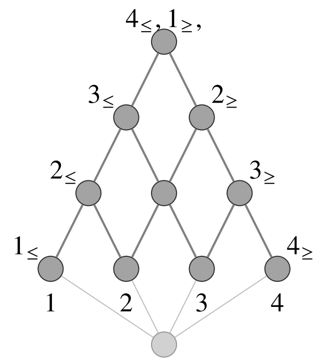


Figure 7.4: This figure shows the formal context of the *interordinal scale* $\mathbb{I}_n$ on n elements (left) and its concept lattice for $n = 4$ (right).

General ordinal scale

For non-linear orders $(P, \leq)$, e.g., hierarchies, the **general ordinal scale** $(P, P, \leq)$ can be used. In our work, we do not consider this scale to be a standard scale since general ordinal scales can rarely be used for attributes other than the ordered set they are defined on.

Contranominal scale

The next scale is the **contranominal scale**, which is defined as

$$\mathbb{B}_n := ([n], [n], \neq) \qquad \text{(Scale)}$$

and its object theory is equivalent to

$$\mathrm{Th}_G(\mathbb{B}_n) \cong \{\}. \qquad \text{(Object Theory)}$$

Its basic meaning is *partition and independence* [80, Figure 1.26]. The concept lattice of this scale is isomorphic to the Boolean scale $(\mathcal{P}([n]), \mathcal{P}([n]), \subseteq)$. Both are used for attributes for which we are interested in any combination of values. Examples are spices planners for recipes or transactions in retail data sets.

Contranominal scales and size

The concept lattice (see Figure 7.5) of the contranominal scale on n elements has 2^n formal concepts (see Figure 7.5). Contexts that include contranominal scales as sub-context are often considered to be very complex [4]. Many works deal with the problem of finding and isolating such scales [63, 125].

Crown scale

The last scale that we consider as standard scale is the **crown scale** [116]. For $n \geq 3$ is this scale defined as

$$\mathbb{C}_n := ([n], [n], J), \text{ where } (a, b) \in J :\Longleftrightarrow a = b \text{ or } (a, b) = (n, 1) \text{ or } b = a + 1 \qquad \text{(Scale)}$$

$\mathbb{B}_n$	1	2	3	...	$n-1$	n
1		×	×	×	×	×
2	×		×	×	×	×
3	×	×		×	×	×
...	×	×	×		×	×
$n-1$	×	×	×	×		×
n	×	×	×	×	×	

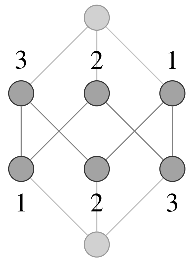


Figure 7.5: This figure shows the formal context of the *contranominal scale* $\mathbb{B}_n$ on n elements (left) and its concept lattice for $n = 3$ elements (right).

and its object theory is equivalent to

$$\mathrm{Th}_G(\mathbb{C}_n) \cong \{i, j \Rightarrow 1, \ldots, n \mid 1 \le i < j \le n \text{ and } 2 \le j - i < n - 1\}. \quad \text{(Object Theory)}$$

This scale is not among the scales considered by Ganter and Wille [80] and does thereby not have a basic meaning. This scale has a strong connection to Hamiltonian cycles and we often find it as sub-structure in concept lattices. We interpret this scale as a round trip between values. In Figure 7.6 we depict its context and concept lattice.

$\mathbb{C}_n$	1	2	3	...	$n-1$	n
1	×	×				
2		×	×			
3			×	×		
...						
$n-1$					×	×
n	×					×

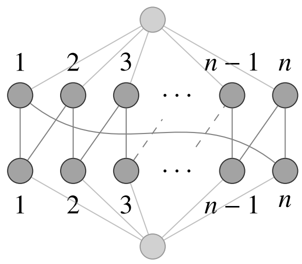


Figure 7.6: This figure shows the formal context of the *crown scale* $\mathbb{C}_n$ on n elements (left) and its concept lattice (right).

Lattices of standard scale

When a concept lattice L, or parts of it, are isomorphic to the concept lattice of a standard scale we say that L is of standard scale. For example a concept lattice that is linearly ordered is of ordinal scale.

7.1.1 Pre-Scalings

Scaling with additional information

Some (many-valued) data sets are equipped with richer information, for which we introduce additional definitions. A **pre-scaling** of a many-valued context $\mathbb{D} := (G, M, W, I)$ is a family $(W(m) \mid m \in M)$ of sets $W(m) \subseteq W$ such that $W = \bigcup_{m \in M} W(m)$ and

$$(g, m, w) \in I \implies w \in W(m)$$

for all $g \in G, m \in M$. We call $W(m)$ the **value domain** of the many-valued attribute m. The value domain is a super-set of the domain $\mathrm{dom}_{\mathbb{D}}(m)$ and also includes values that m accepts but are not supported in the data. A tuple $(v_m \mid m \in M)$ **matches** a pre-scaling iff $v_m \in W(m) \cup \{\perp\}$ holds for all $m \in M$. $(G, M, W(m)_{m\in M}, J)$ is called a **stratified** many-valued context.

Many-valued attributes with structure

It is also allowed that the value domains additionally carry a structure, e.g., are ordered. This also falls under the definition of "pre-scaling". We remain a little vague here, because its seems premature to give a sharp definition. Prediger [171] suggests the notion of a **relational many-valued context**. This may be formalized as a tuple

$$\big(G, M, (W(m), \mathcal{R}_m)_{m\in M}, I\big),$$

where $(G, M, W(m)_{m\in M}, I)$ is a stratified many valued-context as defined above, where on each value domain $W(m)$ a family $\mathcal{R}_m$ of relations is given. Prediger and Stumme [170] then discuss deriving one-valued attributes using expressions in a suitable logical language, such as one of the OWL-variants. They call this *logical scaling*. Moreover, we can envision that an extension to arbitrary data structures (cf. Chapter 2) is beneficial.

Ordinal context

The special case where each R_m is an order relation is also known as **ordinal context** [169, 202]. The kind of *pre-scaling*, as mentioned above, may suggest the scales to use. An ordinal pre-scaling naturally leads to an *interordinal* interpretation of data, using only interordinal scales.

Scaling based on pre-scaling

Given the equivalence in Proposition 4, we find suggestions for scales based on structural properties. Pollandt and Wille [169] suggest the use of contra-ordinal scaling for ordinal contexts, i.e., for a given ordinal pre-scaling $(W(m), \leq_m)$ is the **contra-ordinal scale** defined as $\big(W(m), W(m), \not\geq := \overline{\leq_m^{-1}}\,\big)$. Such a scaling leads to a natural correspondence between order preserving maps of ordinal pre-scalings and scale-measures between the derived contexts. This is a useful connection for theoretic investigations. However, in practice the choice for a scale should be made based on the analyst's interpretation and the relations he or she is interested in.

Context with incomplete knowledge

For data with unknown information Burmeister and Holzer [33] proposed **incomplete contexts** $(G, M, \{\times, ?, -\}, J)$. The value + encodes that an object has an attribute, − that an object does not have an attribute and ? encodes that it is not known if an objects has an attribute. Suggested scalings either project the data only to either value, or use the $- \leq ? \leq \times$ or $? \leq -, \times$ ordinal scaling [68].

7.1.2 Dealing with Categorical Values

Are categorical values nominal?

Categorical attributes are most often scaled nominally by which we do not measure any relation between attribute values. However, this is often an oversimplification of the data. For example consider a clothing retail data set with an attribute whose values are types of clothes like *shoes, sandals, trousers, shorts, sweater* and *hats*. While these values are pairwise unequal and none is greater than the other, there are some that are more similar than others. For example, some of the clothes are footwear and some are shorts. As discussed in Stumme [204], such information can be added to the scale based on background knowledge, e.g., from a taxonomy. The additional relations may be very beneficial for tasks like classification, clustering or to derive more generalized implications using the more abstract scale attributes *footwear*. The concept lattices of the described scales are displayed in Figure 7.7.

More examples

Other examples are fruits that can be measured into botanic classification hierarchies or animals into their evolutionary genealogy.

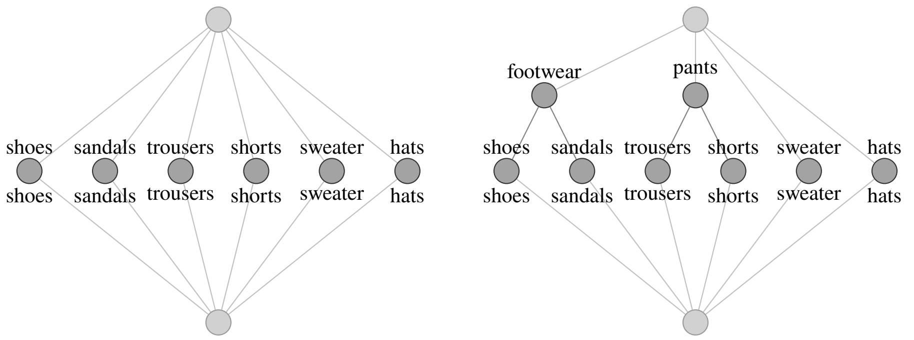


Figure 7.7: Two scales for a categorical attribute. One is a nominal scale, the other is a nominal scale enriched with background knowledge.

7.1.3 Other Scaling Methods

Many extensions to plain scaling can be found in the literature. First, there is **logical scaling** which allows for the definition of scale attributes that are logical combinations of attributes in the many-valued context $\mathbb{D}$. The scale attributes can either be defined using an SQL syntax to generate unary relations [172] or description logics [170]. Logical scaling

Another extension is **local scaling** [206] which conditions the use of a scale for an attribute $m \in M$ based on the values an object has in the scale for a different attribute $n \in M \setminus \{m\}$. The condition is meet for an object $g \in G$ iff the object extent of $n(g)$ in $\mathbb{S}_n$ is within a convex subset $C \subseteq \underline{\mathfrak{B}}(\mathbb{S}_n)$. This procedure is designed for nested representations of concept lattices as used in TOSCANA [123, 218] and allows to only visualize the scale $\mathbb{S}_m$ within a convex subset of $\underline{\mathfrak{B}}(\mathbb{S}_n)$. A variation of local scaling is **nested scaling** [204] which given a concept (A, B) in $\underline{\mathfrak{B}}(\mathbb{S}_n)$ only applies a scale $\mathbb{S}_m$ iff $\mathbb{S}_m$ differentiates objects $g \in G$ with $n(g) \in A$, i.e., $\mathbb{S}_m$ restricted to the scale values $m(A)$ has more than one concept. This applies a scale only if it adds information with respect to nested diagrams. Local scaling Nested scaling

Both scalings are proper extension to the baseline scaling procedure since attribute scales in plain scaling are independent from each other.

The TOSCANA software also allows for the definition of objects in the scale context using SQL expressions [206]. As long as this is only used within the domain of a single attribute, this mainly provides more compact representations of scales. Equivalent scalings can be provided by replacing a SQL expression for each value that satisfies the expression. When multiple attributes are allowed, this procedure extends plain scaling similar to logical scaling. Logical scale values

Furthermore, there are scaling methods to derive other structures from a many-valued context. For many-valued contexts that are not complete there is scaling into incomplete contexts [33]. These are special many-valued contexts with value set $\{\times, ?, -\}$ to encode if an object has an attribute, does not have an attribute or it is unknown. This context does also come with additional derivation operators for each value. For *fuzzy contexts*, i.e., contexts where incidences are quantified within the interval $[0, 1]$, there is **fuzzy scaling** [18] and for *traidic formal contexts* [224] there is **triadic scaling** [115]. **Relational scaling** [65, 173] derives from one many-valued context a family of formal contexts $(\mathbb{K}_j)_{j\in J}$ with $J \subseteq \mathbb{N}$ called a **power context family**. The object set G_j of $\mathbb{K}_j$ is defined as $(G_\mathbb{D})^j$ and each attribute set $N \subseteq M_j$ encodes a j-ary relation on $G_\mathbb{D}$ by $N^{I_{\mathbb{S}_j}} \subseteq (G_\mathbb{D})^j$. These relations are used to define a (directed) hypergraph on $G_\mathbb{D}$ called **concept graph**. Selection methods based on SQL syntax [84] have been proposed to present only parts of the concept graph. Derive other structures

7.2 Conceptual Data Reduction

We explain the notions for conceptual data reduction using the ice cream context (see Figure 7.8). This context has seven ice cream flavors as objects and nine ingredients as attributes. The incidence encodes that an ingredient is contained in an ice cream.

	Brownie (B)	Peanut Butter (PB)	Peanut Ice (PI)	Caramel (Ca)	Caramel Ice (CaI)	Choco Ice (CI)	Choco Pieces (CP)	Dough (D)	Vanilla (V)
Fudge Brownie (FB)	×					×			
Cookie Dough (CD)							×	×	×
Half Baked (HB)	×					×	×	×	×
Caramel Sutra (CS)				×	×	×	×		
Caramel Chew Chew (CCC)				×	×		×		
Peanut Butter Cup (PBC)		×	×				×		
Salted Caramel Brownie (SCB)	×			×			×		×

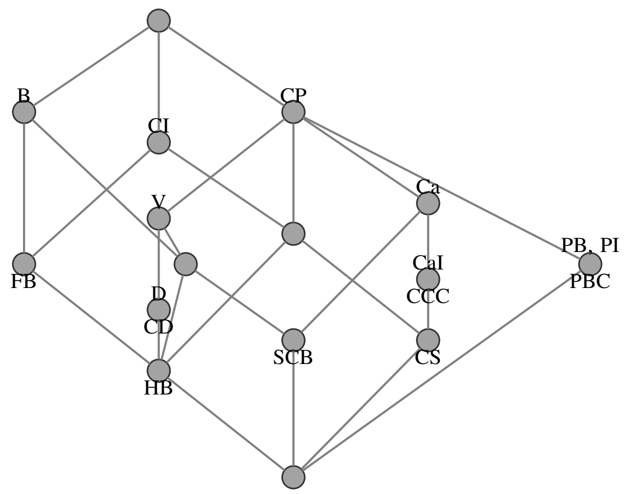


Figure 7.8: This Figure shows the ice cream context and its concept lattice. The incidence describes if an ingredient (attribute) is contained in an ice cream flavor.

Scale-measures for data reduction

With the following formalism we understand conceptual data reduction as a morphism into a conceptually simpler structure.

Definition 35 (Scale-Measure (cf. Definition 91, [80])). *Let $\mathbb{K}$ and $\mathbb{S}$ be two formal contexts. The map $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ is called an* $\mathbb{S}$**-measure** of $\mathbb{K}$ into the scale $\mathbb{S}$ *iff the pre-image $\sigma^{-1}(A) := \{g \in G_{\mathbb{K}} \mid \sigma(g) \in A\}$ of every extent $A \in \mathrm{Ext}(\mathbb{S})$ is an extent of $\mathbb{K}$. An $\mathbb{S}$-measure σ is* ***full*** *if every extent of $\mathbb{K}$ is the pre-image of an extent of $\mathbb{S}$, i.e., $\sigma^{-1}\big(\mathrm{Ext}(\mathbb{S})\big) = \mathrm{Ext}(\mathbb{K})$.*

	Brownie (B)	Peanut (P)	Caramel (Ca)	Choco (Ch)	Dough (D)	Vanilla (V)
Fudge Brownie (FB)	×			×		
Cookie Dough (CD)				×	×	×
Half Baked (HB)	×			×	×	×
Caramel Sutra (CS)			×	×		
Caramel Chew Chew (CCC)			×	×		
Peanut Butter Cup (PBC)		×		×		
Salted Caramel Brownie (SCB)	×		×	×		×

Figure 7.9: A scaled context of the ice cream data set (see Figure 7.8).

This definition corresponds to the notion for *continuity between closure spaces* (G_1, c_1) and (G_2, c_2), i.e., a map $f : G_1 \to G_2$ is continuous iff

Continuous conceptual data reduction

$$\text{for all } A \in \mathcal{P}(G_2) \text{ we have } c_1\big(f^{-1}(A)\big) \subseteq f^{-1}\big(c_2(A)\big). \tag{7.1}$$

This property is equivalent to the requirement in Definition 35 that the pre-image of closed sets is closed, more formally,

$$\text{for all } A \in \mathcal{P}(G_2) \text{ with } c_2(A) = A \text{ we have } f^{-1}(A) = c_1\big(f^{-1}(A)\big). \tag{7.2}$$

Conditions in (7.1) and (7.2) are known to be equivalent, since (7.1) $\Rightarrow$ (7.2) follows from

$$x \in c_1\big(f^{-1}(A)\big) \Rightarrow x \in f^{-1}\big(c_2(A)\big) \xRightarrow{c_2(A)=A} x \in f^{-1}(A).$$

Also, from

$$x \in c_1\big(f^{-1}(A)\big) \Rightarrow x \in c_1\Big(f^{-1}\big(c_2(A)\big)\Big) \xRightarrow{(7.2)} x \in f^{-1}\big(c_2(A)\big)$$

results (7.2) $\Rightarrow$ (7.1).

In the following we may address by $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$ the set of all extents of $\mathbb{K}$ that are *reflected* by the scale context, i.e., $\{\sigma^{-1}(A) \mid A \in \mathrm{Ext}(\mathbb{S})\}$.

Reflected conceptual structure

In case the map σ is not the inclusion map, we understand the map σ as an interpretation of the objects from $\mathbb{K}$ in $\mathbb{S}$. Thus, any object $g \in G$ is expressed in the (potentially limited) language of the scale-context $\mathbb{S}$.

Interpretation of σ

Remark 4 (Surjective Scale-Measures). *It is reasonable to consider only surjective maps when using scale contexts for scale-measures. Since objects that are not contained in the image of the scale-measure σ do not contribute to the set of reflected extents.*

In Figure 7.10 we depict a scale-measure, and the concept lattices of the ice cream context and of the scaled context. This scale-measure uses the same object set as the original

Example scale-measure

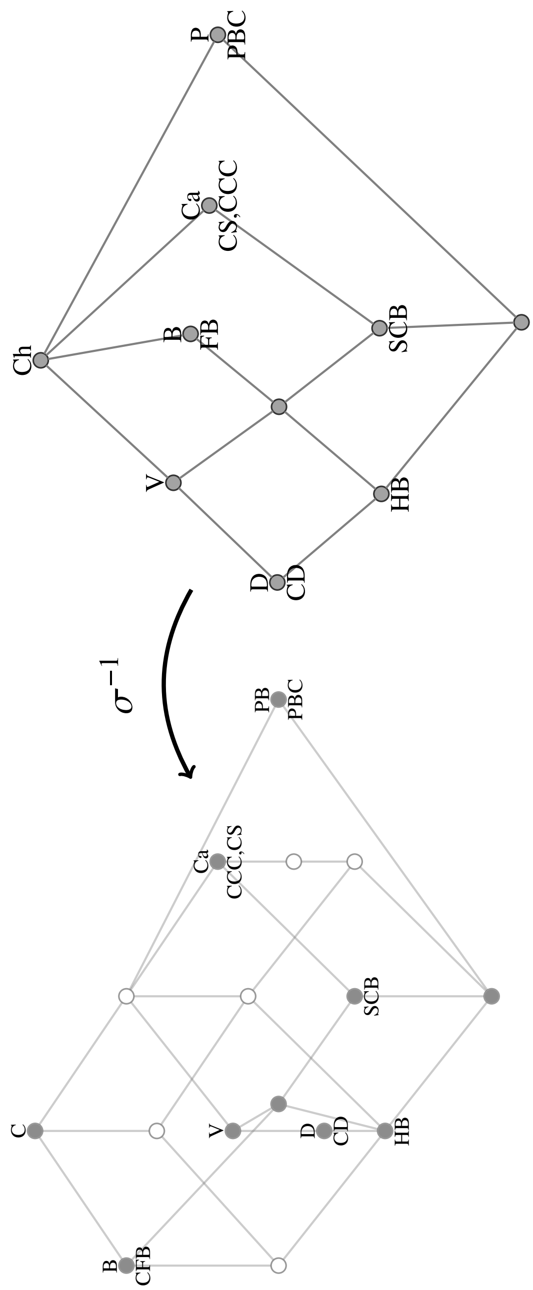


Figure 7.10: This figure depicts the concept lattice (right) of the scaled ice cream context (Figure 7.9). The reflected extents $\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))$ of the ι_G scale-measure are highlighted in the concept lattice diagram of the ice cream context.

context alongside the identity map on G. The attribute set consists of six elements, which may reflect the taste, instead of the original nine attributes that indicated the used ingredients. The specified scale-measure map allows for a human comprehensible interpretation of σ^{-1}, as indicated by the gray colored concepts in Figure 7.10. In this figure we observe that the concept lattice of the scale-measure reflects ten out of the sixteen concepts in $\underline{\mathfrak{B}}(\mathbb{K}_{\text{ICE}})$. The order dimension of $\underline{\mathfrak{B}}(\mathbb{K}_{\text{ICE}})$ is three whereas the order dimension of the concept lattice of the scale-measure given in Figure 7.10 is two. Finding low dimensional scale-measures for large and complex data sets is a natural approach towards comprehensible data analysis. In Chapter 10 we will answer the question: *Is the order dimension of scale-measures bound by the order dimension of* $\underline{\mathfrak{B}}(\mathbb{K})$*?*

Scale-measures and complexity

There is a close connection between scale-measures, embeddings and morphisms between closure systems that we describe by the following propositions.

Scale-measures and other morphisms

Proposition 3 (Scale-Measures and Order Morphisms). *For contexts* $\mathbb{K}$, $\mathbb{S}$ *and a surjective scale-measure* σ *from* $\mathbb{K}$ *to* $\mathbb{S}$ *we find that*

$$\big(\text{Ext}(\mathbb{S}), \subseteq\big) \cong \big(\sigma^{-1}(\text{Ext}(\mathbb{S})), \subseteq\big) \trianglelefteq \big(\text{Ext}(\mathbb{K}), \subseteq\big) \tag{7.3}$$

and in case σ *is full*

$$\big(\text{Ext}(\mathbb{K}), \subseteq\big) \cong \big(\sigma(\text{Ext}(\mathbb{K})), \subseteq\big) \cong \big(\text{Ext}(\mathbb{S}), \subseteq\big) \cong \big(\sigma^{-1}(\text{Ext}(\mathbb{S})), \subseteq\big) \tag{7.4}$$

Proof. Equation (7.3): Since there is no $g \in G_{\mathbb{S}}$ with $\sigma^{-1}(g) = \emptyset$, for every $E, \hat{E} \in \text{Ext}(\mathbb{S})$ with $E \subseteq \hat{E}$ it is true that $\sigma^{-1}(E) \subseteq \sigma^{-1}(\hat{E})$. The second morphism follows directly from Proposition 118 in Ganter and Wille [80], which states that $\big(A, A^{I_{\mathbb{S}}}\big) \mapsto \big(\sigma^{-1}(A), \sigma^{-1}(A)^{I_{\mathbb{K}}}\big)$ defines a meet-preserving embedding of $\underline{\mathfrak{B}}(\mathbb{S})$ to $\underline{\mathfrak{B}}(\mathbb{K})$. Since the order relation of any concept lattice is derived from the inclusion order on its extents, we can infer that the $A \mapsto \sigma^{-1}(A)$ defines a meet-preserving map from $\text{Ext}(\mathbb{S})$ to $\text{Ext}(\mathbb{K})$.

Equation (7.4): From the surjectivity of σ we can deduce via [80, Proposition 118] that the map σ^{-1} exists, which is injective. This also means that every extent $E \in \text{Ext}(\mathbb{S})$ is mapped to a unique extent $\hat{E} \in \text{Ext}(\mathbb{K})$. Moreover, since σ is a full scale-measure, every extent of $\mathbb{K}$ is also a pre-image of an extent of $\mathbb{S}$. From this it follows that σ^{-1} bijectively maps the extents from $\mathbb{S}$ and $\mathbb{K}$.

The last isomorphism follows from Equation (7.3). □

Proposition 4 (CS-morphism and scale-measures). *For two formal contexts* $\mathbb{K}, \mathbb{S}$ *and a map* $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ *TFAE:*

i) σ *is a* $\mathbb{S}$*-measure of* $\mathbb{K}$*,*

ii) for all $A \subseteq G_{\mathbb{K}}$ *and* $x \in G_{\mathbb{K}}$*:* $x \in A^{I_{\mathbb{K}} I_{\mathbb{K}}} \implies \sigma(x) \in \text{cl}_{\mathbb{S}}(\sigma(A))$.

Proof. $\Rightarrow$ Assume there is a $g \in G_{\mathbb{K}}$ and $B \subseteq G_{\mathbb{K}}$ with $g \in B^{I_{\mathbb{K}} I_{\mathbb{K}}}$ but $\sigma(g) \notin \sigma(B)^{I_{\mathbb{S}} I_{\mathbb{S}}}$. Thus, $g \notin \sigma^{-1}(\sigma(B)^{I_{\mathbb{S}} I_{\mathbb{S}}})$, but $g \in \sigma^{-1}(\sigma(B)^{I_{\mathbb{S}} I_{\mathbb{S}}})^{I_{\mathbb{K}} I_{\mathbb{K}}}$ since $B \subseteq \sigma^{-1}(\sigma(B)^{I_{\mathbb{S}} I_{\mathbb{S}}})$ and $g \in B^{I_{\mathbb{K}} I_{\mathbb{K}}}$. Concluding, there is a $D \in \text{Ext}(\mathbb{S})$, i.e., $D = \sigma^{-1}(\sigma(B)^{I_{\mathbb{S}} I_{\mathbb{S}}})$, with $\sigma^{-1}(D) \notin \text{Ext}(\mathbb{K})$ and therefore is σ not a scale-measure.

$\Leftarrow$ Assume there is an extent $A \in \text{Ext}(\mathbb{S})$ with $\sigma^{-1}(A) \notin \text{Ext}(\mathbb{K})$. Thus, there is a $g \in G_{\mathbb{K}}$ with $g \in \sigma^{-1}(A)^{I_{\mathbb{K}} I_{\mathbb{K}}}$, but $g \notin \sigma^{-1}(A)$. From the later we follow that $\sigma(g) \notin \sigma(\sigma^{-1}(A))$. From $\sigma(\sigma^{-1}(A)) = A$ and $A^{I_{\mathbb{S}} I_{\mathbb{S}}} = A$ we result in a contradiction to (ii), i.e., $g \in \sigma^{-1}(A)^{I_{\mathbb{K}} I_{\mathbb{K}}}$ but $\sigma(g) \notin \sigma(\sigma^{-1}(A))^{I_{\mathbb{S}} I_{\mathbb{S}}}$. □

Other morphisms

Besides scale-measures there are other context morphisms in the literature whose relations to each other have been studied [66, 126].

Reflected closure system

Throughout this work, we make use of the fact that $\sigma^{-1}(\text{Ext}(\mathbb{S}))$ constitutes a closure system.

Proposition 5 (Scale-Measure Closure System). *For a context $\mathbb{K}$ and a scale-measure σ of $\mathbb{K}$ into $\mathbb{S}$ is $\sigma^{-1}(\text{Ext}(\mathbb{S}))$ a closure system on $G_{\mathbb{K}}$.*

Proof. From Proposition 3 we can follow that $\sigma^{-1}(A) \wedge \sigma^{-1}(C) = \sigma^{-1}(A \wedge C)$ for $A, C \in \text{Ext}(\mathbb{S})$. From the Basic Theorem of FCA [80, Theorem 3] we know that the meet of any set of extents is exactly their intersection. Thus, for any two extents $A, C \in \text{Ext}(\mathbb{S})$ we find that $\sigma^{-1}(A) \cap \sigma^{-1}(C) = \sigma^{-1}(A \cap C)$. Finally, $\sigma^{-1}(A), \sigma^{-1}(C) \in \sigma^{-1}(\text{Ext}(\mathbb{S}))$ implies that $\sigma^{-1}(A) \cap \sigma^{-1}(C) \in \sigma^{-1}(\text{Ext}(\mathbb{S}))$, and therefore $\sigma^{-1}(\text{Ext}(\mathbb{S}))$ is a closure system. □

Combining Propositions 3 and 5 we find that a scale-measure reflects a coarser closure system of $\text{Ext}(\mathbb{K})$.

Chaining conceptual data reductions

Scale-measures can also be chained to describe relations across multiple data reductions. This relation can be deduced from the continuity property above and will be used frequently throughout our work.

Corollary 1 (Composition Scale-Measures). *Let $\mathbb{K}$ be a formal context, σ an $\mathbb{S}$-measure of $\mathbb{K}$ and ψ a $\mathbb{T}$-measure of $\mathbb{S}$. Then $\sigma \circ \psi$ is a $\mathbb{T}$-measure of $\mathbb{K}$.*

Conceptual view

We want to further nourish the understanding of scale-measures as consistent measurements of the objects in some scale context. The following definition can be understood as an extension of *views* in Wille [221].

Definition 36 (View). *A **view** of a formal context (G, M, I) is a formal context (G, N, J), where for each $n \in N$ there is a set $A_n \subseteq M$ such that for all $g \in G$,*

$$g\, J\, n :\Longleftrightarrow A_n \subseteq \{g\}^I.$$

*A **contextual view** of a many-valued context $\mathbb{D}$ is a view of a context $\mathbb{K}$ derived from $\mathbb{D}$. The concept lattice of such a contextual view is a **conceptual view** of $\mathbb{D}$.*

Views and scale-measures

Based on the following proposition we understand the application of an $\mathbb{S}$-measure σ on $\mathbb{K}$ as view

$$\big(G, M_{\mathbb{S}}, J\big) \text{ with } A_m := \sigma^{-1}(\{m\}^{I_{\mathbb{S}}}) \text{ and } (g, m) \in J :\Longleftrightarrow A_m \subseteq \{g\}^{I_{\mathbb{K}}}$$

on the data.

Proposition 6 (Scale-Measures and Views). *A formal context $\mathbb{K}_1 := (G, N, J)$ is a view of $\mathbb{K} := (G, M, I)$ if and only if the identity map is a $\mathbb{K}_1$-measure of $\mathbb{K}$.*

Proof. When (G, N, J) is a view of (G, M, I), then every extent E of (G, N, J) is of the form $E = B^J$ for some $B \subseteq N$. Then E also is an extent of (G, M, I), since $E = (\bigcup_{n \in B} A_n)^I$. Conversely, if the identity map is a (G, N, J)-measure of (G, M, I), then for each $n \in N$ the pre-image of its attribute extent n^J (which, of course, is equal to n^J) must be an extent of (G, M, I) and therefore be of the form A_n^I for some set $A_n \subseteq M$. □

7.2.1 Deciding Scale-Measures

A computationally important aspect of conceptual data reduction is to decide if a map is a scale-measure. For this we first show that it is sufficient to check the scale-measure condition for all attribute derivations. This equivalence has also been observed in an earlier work by Erné in Lemma 3.1 [66].

Is a conceptual data reduction continuous?

Proposition 7 (Attribute Scale-Measures). *Let $\mathbb{K} = (G, M, I)$ and $\mathbb{S} = (G_{\mathbb{S}}, M_{\mathbb{S}}, I_{\mathbb{S}})$ be two formal contexts and $\sigma : G \to G_{\mathbb{S}}$, then TFAE:*

i) σ is an $\mathbb{S}$-measure of $\mathbb{K}$

ii) σ is a $\big(G_{\mathbb{S}}, \{n\}, I_{\mathbb{S}} \cap (G_{\mathbb{S}} \times \{n\})\big)$-measure of $\mathbb{K}$ for all $n \in M_{\mathbb{S}}$

Proof. $(i) \Rightarrow (ii)$: Assume $\hat{n} \in M_{\mathbb{S}}$ s.t. σ is not a $\big(G_{\mathbb{S}}, \{\hat{n}\}, \overbrace{I_{\mathbb{S}} \cap (G_{\mathbb{S}} \times \{\hat{n}\})}^{=:J}\big)$-measure of $\mathbb{K}$. Then the only non-trivial extent $\{\hat{n}\}^J$ has a pre-image $\sigma^{-1}(\{\hat{n}\}^J) \notin \text{Ext}(\mathbb{K})$. Since $\{\hat{n}\}^J \in \text{Ext}(\mathbb{S})$ we can conclude that σ is not an $\mathbb{S}$-measure of $\mathbb{K}$.

$(ii) \Rightarrow (i)$: From (ii) we can follow that for all $n \in M_{\mathbb{S}}$ is $\sigma^{-1}(\{n\}^{I_{\mathbb{S}}}) \in \text{Ext}(\mathbb{K})$. From this and the fact that the set of meet-irreducible extents $\text{M}(\mathbb{S})$ of $\mathbb{S}$ is entailed in $\{\{n\}^{I_{\mathbb{S}}} \mid n \in M_{\mathbb{S}}\}$ we can follow that for all $A \in \text{Ext}(\mathbb{S})$ it holds that $\sigma^{-1}(A) \in \text{Ext}(\mathbb{K})$ (cf. Proposition 5). Thus is σ a $\mathbb{S}-$measure of $\mathbb{K}$. □

Thus, we do not have to compute the entire conceptual structure of $\mathbb{K}$ which enables us to decide scale-measure in the size of the contexts.

Complexity of deciding scale-measures

Corollary 2 (Deciding the Scale-measure Problem). *Given a formal context (G, M, I), a scale-context $\mathbb{S} := (G_{\mathbb{S}}, M_{\mathbb{S}}, I_{\mathbb{S}})$ and a map $\sigma : G \to G_{\mathbb{S}}$, deciding if $(\sigma, \mathbb{S})$ is a scale-measure of $\mathbb{K}$ is in P. More specifically, to answer this question does require $O\big(|G| \cdot |M| \cdot |M_{\mathbb{S}}| \cdot |G_{\mathbb{S}}|\big)$ time.*

The cost of applying the map σ and σ^{-1} can be neglected since σ and $\mathbb{S}$ can be substituted by the identity map on $G_{\mathbb{K}}$ and a context $\hat{\mathbb{S}} := (G_{\mathbb{K}}, M_{\mathbb{S}}, \sigma \circ I_{\mathbb{S}})$ in $O\big(|G| \cdot |G_{\mathbb{S}}| \cdot |M_{\mathbb{S}}|\big)$.

Problem 13: Deciding Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$ and a map $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$	
Output: True iff σ is a scale-measure of $\mathbb{K}$.	
Complexity:	$O(\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{S}\rvert)$

For full scale-measures we have to check for each meet-irreducible extent A of $\mathbb{K}$ that $\sigma(A) \in \text{Ext}(\mathbb{S})$ and $\sigma^{-1}(\sigma(A)) = A$. However, this problem is dual to the original scale-measure decision problem. Thus, verifying full scale-measures can be done in time $O(|\mathbb{K}| \cdot |\mathbb{S}|)$.

Complexity of deciding full scale-measures

Problem 14: Deciding Full Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$ and a map $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$	
Output: True iff σ is a full scale-measure of $\mathbb{K}$.	
Complexity:	$O(\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{S}\rvert)$

Complexity and the number of attributes

We may note that this result is favorable since the naive solution would be to compute $\text{Ext}(\mathbb{S})$, which is potentially exponential in the size of $\mathbb{S}$, and checking all its elements in $\mathbb{K}$ for their closure, which consumes $O(|G| \cdot |M|)$ for all $A \in \text{Ext}(\mathbb{S})$. In the special case where both the formal context $\mathbb{K}$ as well as $G_{\mathbb{S}}$ and σ are fixed, the computational cost for deciding the scale-measure problem grows linearly in $|M_{\mathbb{S}}|$.

Towards ordinal motifs

In Chapter 9 we investigate the identification of coarser closure systems in greater detail. There we extend this problem to sub-contexts and the analysis of specific scales. In addition to that, we provide a verification procedure that is based on their implicational theories.

7.3 Interpretation of Conceptual Scalings

Separate scaling and data reduction

In this section we discuss that (within plain scaling) a clear and precise interpretation of data requires to separate the scaling and data reduction procedure. We call this notion *separation principle for conceptual scaling*. This general idea is well-known in other branches of computer science. For example, in the `SOLID`[2] principles of software development the *single-responsibility principle*[3] was created in the same spirit. With the following explanations we claim that the separation principle should be applied to data analysis. We justify this with the observation that there is a correspondence between defining data structures and their function in formal language (cf. Chapter 2) with the implementation as data objects and software functions. Furthermore, in the following text we will support our claim with a real-world example. Applying the separation principle results in one formal structure for scaling, i.e., the employed scales and the derived context, as well as a structure for the data reduction, i.e., the map σ and view $\mathbb{S}$.

A data set with two scalings

We support our claim based on the example provided in Figure 7.12. The example depicts the *commercial laser and their wavelengths*[4] data set $\mathbb{D}$. Alongside the data set, we present two different scalings that we distinguish by index one for the first scaling and index two for the second scaling. The first scaling (red) uses a nominal scale $\mathbb{W}_1$ (wavelength scale 1) to interpret the wavelengths by their perceived color.[5] The second approach (green) interprets the values to be on an ordinal scale and defines an interordinal scale $\mathbb{W}_2$. After that, the second scaling computes a view $\mathbb{S}_2$ of the with $\mathbb{W}_2$ derived context $\mathbb{K}_2$. This view reflects the perceived colors based on intervals in the wavelength scale $\mathbb{W}_2$ (see Figure 7.11).

Different scaling - different interpretation

While the end result of both scalings are equivalent contexts, i.e., $\mathbb{K}_1$ and $\mathbb{S}_2$ are equal, they have different interpretations with respect to plain scaling. The interpretation of the attribute *ultraviolet* in the first scaling $\mathbb{K}_1$ is the set $\{157.0\text{nm}, 337.1\text{nm}\}$, i.e.,

$$\{ultraviolet\}^{I_{\mathbb{W}_1}} = \{157.0\text{nm}, 337.1\text{nm}\},$$

while its interpretation in the view $\mathbb{S}_2$ is the interval $[157.0\text{nm}, 337.1\text{nm}]$, i.e.,

$$\left(\sigma^{-1}(\{ultraviolet\}^{I_{\mathbb{S}_2}})\right)^{I_{\mathbb{K}_2}} = \{\geq 157.0\text{nm}, \leq 337.1\text{nm}, \ldots, \leq 694.3\text{nm}\}.$$

This seemingly subtle difference allows for reasoning and comparisons of objects in $\mathbb{S}_2$ that are not possible in $\mathbb{K}_1$. For example, consider the lasers *Nitrogen* and *He-Ne*. Both have no common incidences in $\mathbb{K}_1$ or $\mathbb{S}_2$. However, there are key differences when comparing the interpretations of their attributes in the first and second scaling. In the first scaling, the values

[2]`https://en.wikipedia.org/wiki/SOLID`
[3]`https://en.wikipedia.org/wiki/Single-responsibility_principle`
[4]`https://en.wikipedia.org/wiki/Laser#/media/File:Commercial_laser_lines.svg`, 06.2023
[5]`https://en.wikipedia.org/wiki/Visible_spectrum#Spectral_colors`, 06.2023

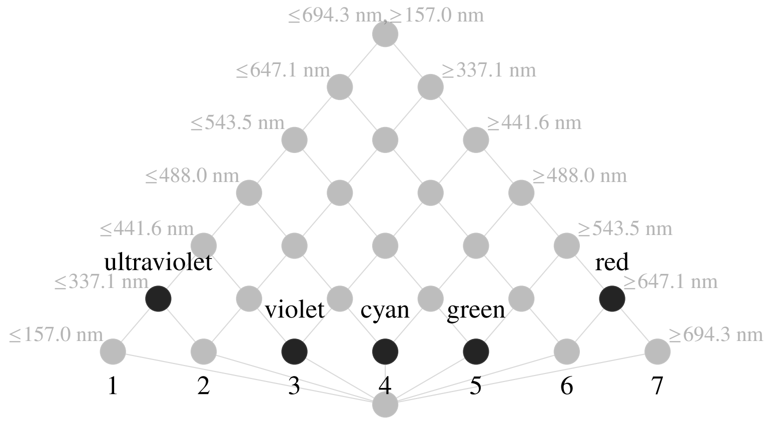


Figure 7.11: This figure depicts the combined scaling and data reduction of the example provided in Figure 7.12.

$\{ultraviolet\}^{I_{\mathbb{W}_1}}$ and $\{green\}^{I_{\mathbb{W}_1}}$ have no common scale attributes. In the second scaling, we find that they share the scale attributes $\{\leq 543.5\text{nm}, \ldots, \leq 694.3\text{nm}, \geq 157.0\text{nm}, \geq 337.1\text{nm}\}$. In this example, the reader may be able to deduce the ordinal property of the wavelength domain, however, this might not be as easy in other domains. For example when the attribute values have a (non-linear) ordinal pre-scaling or are entailed in complex taxonomies.

Note on background knowledge

Besides that, there are other advantages of the second scaling. With the interordinal scale we are able to classify new values that are within the present intervals. The first scaling requires for such a classification that other lasers must have exactly the same emitted wavelength 157.0nm or 337.1nm. One may argue that there are ways around this by defining a scale $\mathbb{W} := (\mathbb{R}, M_{\text{Colors}}, J)$ with all possible scale values. This does yield with $\{ultraviolet\}^{I_{\mathbb{W}}}$ the set of all elements in the interval $[157.0\text{nm}, 337.1\text{nm}]$. However, such a scale is very large and complex. On top of that, determining the incidences for all pairs $\mathbb{R} \times M_{\text{Colors}}$ can be costly.

Capabilities of appropriate scaling

Another advantage of the second scaling is that the derived context $\mathbb{K}_2$ has more views. For example, one might combine the violet and ultraviolet attributes. Such an extension of intervals in the ordinal interpretation does again yield a view on the data. The context $\mathbb{K}_1$ derived from the first scaling allows only for trivial views which can be computed by omitting attributes. This is especially advantages for non domain experts that are not able to define new scales $\mathbb{W}$.

New views

7.4 Inverse Conceptual Data Scaling

In this section, we recall the inverse scaling procedure from the literature [74, 80]. We expand on this in Chapters 9 and 10.

Inverse scaling deals with the task of deciding if a given formal context $\mathbb{K}$ is derived from plain scaling (up to isomorphism). More precisely, one would like to decide whether $\mathbb{K}$ can be derived using a given set scales $\mathcal{S}$, e.g., from interordinal scaling. A key result here is Proposition 122 of the FCA book [80]. This proposition provides a connection between plain scaling and full scale-measure.

Where does my data come from?

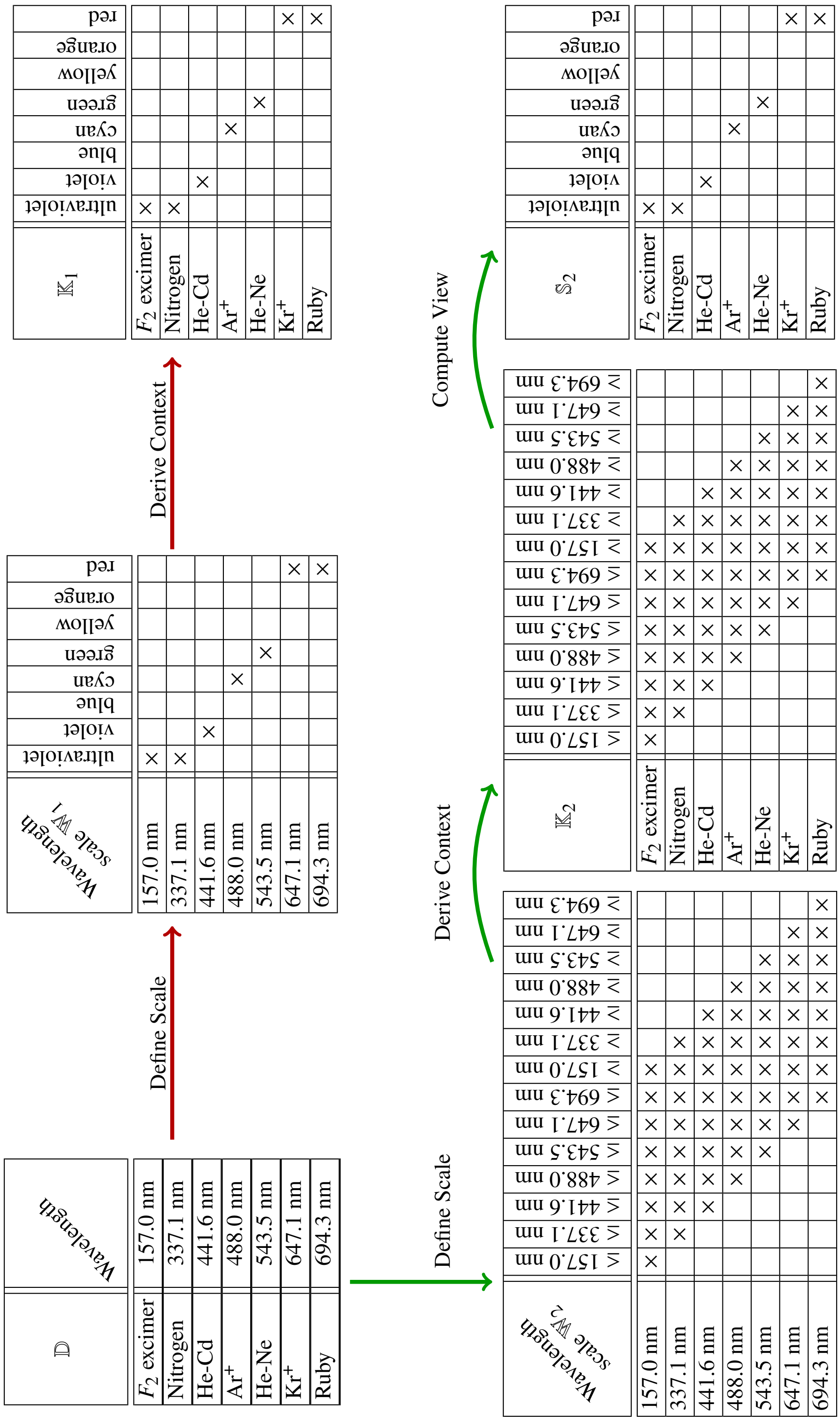


Figure 7.12: Context of *commercial laser and their wavelengths* with two scalings.

Proposition 8 (Inverse Scaling given $\mathbb{D}$, Proposition 122 [80]). *Let $\mathbb{D} := (G, M, W, I)$ be a complete many-valued context and let $\mathbb{S}_m$, $m \in M$, be scales for the attributes of M. Furthermore, let $\mathbb{K}$ be the derived context with respect to plain scaling. Then, for every many-valued attribute $m \in M$ is the map $\sigma_m : G \to G_{\mathbb{S}_m}$ with $\sigma_m(g) := m(g)$ a $\mathbb{S}_m$-measure of $\mathbb{K}$, and $\mathbb{K}$ is isomorphic to the induced sub-context $\mathbb{S}[\sigma(G_{\mathbb{D}}), M_{\mathbb{S}}]$ of the semi-product of the scales $\mathbb{S}_m$ with $\sigma(g) := (\sigma_m(g))_{m\in M}$.*

A slight reformulation of this proposition can be used to answer the following question: Given a formal context $\mathbb{K}$ and a family of scales $\mathcal{S}$, does there exist some complete many-valued context $\mathbb{D}$ such that $\mathbb{K}$ is equal to (up to isomorphism) the context derived from plain scaling of $\mathbb{D}$ using only scales from $\mathcal{S}$? The inverse plain scaling procedure based on the previous result is given by the following proposition.

Inverse plain scaling procedure

Proposition 9 (Inverse Plain Scaling). *Let $\mathbb{K}$ be a formal context, $(\mathbb{S}_j)_{j\in[n]}$ be a family of scales and σ be a full scale-measure from $\mathbb{K}$ into the semi-product $\mathbb{S} := \Join_{j\in[n]}\mathbb{S}_j$ of all scales with $\sigma(g) = (\sigma_1(g), \dots, \sigma_n(g))$. Furthermore, let $\hat{\mathbb{K}}$ be the context derived from $\mathbb{D} := \big(G, [n], \big(G_{\mathbb{S}_j}\big)_{j\in[n]}, I\big)$ via plain scaling and scales $(\mathbb{S}_j)_{j\in[n]}$ where*

$$(g, j, v) \in I :\iff \sigma_j(g) = v.$$

The attribute reduced contexts of $\mathbb{K}$ and $\hat{\mathbb{K}}$ are isomorphic.

Proof. The attribute set of $\hat{\mathbb{K}}$ is equal to $\bigcup_{j\in[n]} M_{\mathbb{S}_j} = M_{\mathbb{S}}$ per definition. For an attribute $m \in M_{\mathbb{S}_i}$ is $(g, m) \in I_{\hat{\mathbb{K}}}$ iff $(\sigma(g), m) \in I_{\mathbb{S}_i}$. Thus, the object g is in incidence with an attribute m in $\hat{\mathbb{K}}$ iff $\sigma(g) = (\sigma_1(g), \dots, \sigma_n(g))$ is in incidence with m in $\mathbb{S}$. Therefore is the pair (σ, ι) a context isomorphism of $\hat{\mathbb{K}}$ into and $\mathbb{S}[\sigma(G), M_{\mathbb{S}}]$.

Since σ is a full scale-measure of $\mathbb{K}$ into $\mathbb{S}$ we can follow that $\sigma^{-1}\big(\operatorname{Ext}(\mathbb{S}[\sigma(G), M_{\mathbb{S}}])\big)$ is equal to $\operatorname{Ext}(\mathbb{K})$. Thus, the attribute reduced contexts of $\mathbb{K}$ and $\hat{\mathbb{K}}$ are isomorphic. □

Theorem 55 of the FCA book also gives some simple characterizations for measurability of lattices. This theorem uses the notion of *pseudocomplements* in a lattice L. An element $x \in L$ has a **meet-pseudocomplement** iff there exists a greatest element $y \in L$ with $x \wedge y = \bot_L$. We recall this theorem in the language of this work:

Lattices derived from standard scales

Theorem 4 (Conceptual Measurability Theorem of Lattices, Theorem 55 [80]). *A lattice L is isomorphic to the concept lattice of a scaled many-valued context $\mathbb{D}$ and a family of scales by the following conditions:*

(Ordinal)	*Every lattice is isomorphic to the concept lattice of an ordinally scaled many-valued context.*
(Nominal)	*A lattice is isomorphic to the concept lattice of a nominally scaled many-valued context iff L is atomistic.*
(Interordinal/ Contranominal)	*A lattice is isomorphic to the concept lattice of an interordinally/contranominally scaled context iff L is atomistic and for each meet-irreducible element there exists a meet-pseudocomplement in L.*

In the following, we provide a similar theorem based on contexts and results from Proposition 9. While the property for ordinal and nominal scales can also be found in the literature [74], our proof provides a construction for the employed full scale-measures. An employed property of said theorem is that of *atomistic* contexts, i.e., a context is **atomistic** iff $\{g\}' \subseteq \{h\}' \implies \{g\}' = \{h\}'$ for all $g, h \in G$.

Contexts derived from standard scales

Theorem 5 (Conceptual Measurability Theorem of Contexts). *An object clarified context is derivable from plain scaling from a complete many-valued context (up to attribute reduction) and a family of scales by the following conditions:*

(Ordinal)	*Every context is derivable from ordinal scaling.*
(Nominal)	*A context is derivable from nominal scaling iff it is atomistic.*
(Interordinal/ Contranominal)	*A context is derivable from interordinal/contranominal scaling iff it is atomistic and the complement of every attribute extent is an extent.*

Proof. WLoG has $\mathbb{K}$ more at least two objects.

(Ordinal) For each extent A of $\mathbb{K}$ we define a scale $\mathbb{S}_A := (G, \{A\}, \in)$. The scale $\mathbb{S}_A$ has extents A and G. Based on Proposition 120 [80] is $\sigma : G \to \bigtimes_{A \in \text{Ext}(\mathbb{K})} G$ with $\sigma(g) := (g, g, \ldots, g)$ a scale-measure that reflects all extents that can be written as the intersection of the extents of the individual scales. Due to the fact that for each $A \in Ext(\mathbb{K})$ there is a scale $\mathbb{S}_A$ that reflects A we can follow that σ is full. For each scale $\mathbb{S}_A$ is the map $\psi_A : G \to [2]$ with $\psi_m(g) = 2$ if $g \in A$ else $\psi(g) = 1$ a full scale-measure into $\mathbb{O}_2$.

The composition $\sigma \circ \psi$ with $\psi : G \to \bigtimes_{m \in M_{\mathbb{K}}}$ and $\psi(g) := (\psi_m)_{m \in M_{\mathbb{K}}}$ yields a full scale-measure (Corollary 1) of $\mathbb{K}$ into the semi-product $\mathbb{O}$ of ordinal scales. The remainder follows directly from Proposition 9. Thus, $\mathbb{K}$ is (up to attribute reduction) derivable from a complete many-valued context and ordinal plain scaling.

(Nominal) [$\Rightarrow$] For a context $\mathbb{K}$ that is derived from nominal scaling and a complete many-valued context, there exists a full scale-measure into the semi-product $\mathbb{N}$ of nominal scales and $\mathbb{K}$ is isomorphic to a sub-scale $\mathbb{N}[\sigma(G), M_{\mathbb{N}}]$ [80, Proposition 122]. Since nominal scales are atomistic and the semi-product of atomistic scales is again atomistic we can follow that $\mathbb{K}$ is atomistic.

[$\Leftarrow$] For each extent A of $\mathbb{K}$ we define a scale $\mathbb{S}_A := (G, \{A\} \cup \{\{g\} \mid g \in G \setminus A\}, \in)$. The extents of $\mathbb{S}_A$ are $A, G, \emptyset$ and the sets $\{g\}$ with $g \in G \setminus A$. Since $\mathbb{K}$ is atomistic we can follow that the sets $\{g\}$ as well as the empty set are closed in $\mathbb{K}$. The remainder follows from an analogous construction as provided for the ordinal case.

(Interordinal/Contranominal) [$\Rightarrow$] A context $\mathbb{K}$ that is derived from interordinal scaling is fully measurable into the semi-product $\mathbb{I}$ of interordinal scales. An interordinal scale has the property that it is atomistic and the complement of every attribute extent is an extent. Both properties are preserved by the semi-product of scales having the same properties. For the case that the maps σ_i from $\mathbb{K}$ into the scales are not surjective we may note that both properties also hold for sub-contexts that only restrict objects. Objects are either removed from an extent A or its complement $G \setminus A$ which results in new complemented extents (cf. Proposition 2).

Given Proposition 122 [80] is $\mathbb{K}$ isomorphic to such an induced sub-context of $\mathbb{I}$ and is therefor atomistic and the complement of every attribute extent is an extent. The analogue applies to contranominal scales.

[$\Leftarrow$] For each attribute extent A of $\mathbb{K}$ we define a scale $\mathbb{S}_A := (G, \{A, G \setminus A\}, \in)$. The set of extents of each scale is equal to $G, A, G \setminus A$ and $\emptyset$, all of which are closed in $\mathbb{K}$ by definition.

The rest follows from an analogous construction as in the ordinal case with the same construction of ψ but into the scale $\mathbb{I}_2$. This also applies to contranominal scales since $\mathbb{I}_2 \cong \mathbb{B}_2$. □

With these theorems we are able to decide the inverse scaling of contexts and standard scales. On top of that we find an order of expressiveness among the scales $\mathbb{O} > \mathbb{N} > \mathbb{I}, \mathbb{B}$ where every context is fully ordinally measurable and every interordinally measurable context is also nominally measurable. Contranominal and interordinal scales are equally expressive.

Expressiveness of standard scales

7.5 Conceptual Data Compatibility

In this section we briefly introduce the basics on conceptual data compatibility to have a broad overview of all scaling tasks in this chapter. In Chapter 11 we analyze this field in greater detail. The basis of conceptual data reduction are extent continuous maps, i.e., scale-measures σ between contexts. The condition here is that the pre-image of every extent in the scaled context $\mathbb{S}$ is an extent in the context $\mathbb{K}$. In Chapter 11 we introduce the set of extents that contradict this property, i.e.,

Towards conceptual scaling error

$$\text{the set of all } A \in \text{Ext}(\mathbb{S}) \text{ with } \sigma^{-1}(A) \notin \text{Ext}(\mathbb{K}),$$

as the **conceptual scaling error**.

7.6 Related Work

Measurement is an important field of study in many (scientific) disciplines that involve the collection, interpretation and analysis of data. According to Stevens [201] there are four feature categories that can be measured, i.e., *nominal*, *ordinal*, *interval* and *ratio* features. There are multiple extensions and re-categorizations of the original four categories, e.g., most recently Chrisman introduced additional levels [39] beyond the interval level.

Level of measurement

Most machine learning techniques are defined using a vector space model on real numbers and thus operate on the ratio level. They often achieve this by mapping scales from lower levels to a numeric scale. *Representational Theory of Measurement* (RTM) [145, 168, 210] provides a formalizing and understanding of this process based on homomorphisms. By performing numeric calculations on the resulting numeric representations, they implicitly define operations on the data that – although they perform well in machine learning – have no real world meaning. This leads to uninterpretable black-box models. In our work, we interpret data on the ordinal level through the use of ordinal scales. With this approach, we derive a unified representation of heterogeneous data and do not artificially introduce metrics or other properties. This results in explainable (algebraic) models. For the interpretation on the ordinal level we use conceptual scaling [74, 79] for the ordinal interpretation.

Working with heterogeneous data

To cope with large data sets, a multitude of methods were introduced to reduce the dimensionality. Popular methods from machine learning are factorization methods like *Latent Semantic Analysis* [46, 62], *principle component analysis* or *binary matrix factorization* [22, 155, 230], or other embedding techniques like *multidimensional scaling* [152, 157]. There also is a large number of data reduction methods that originate from within FCA like the selection of relevant attributes [93], *pq*-cores [90], TITANIC [208] or the use of generalized attributes [134]. The framework of scale-measures allows us to characterize and compare these data reductions based on the views of the data they provide. Relations of scale-measures to other morphisms can be found in the literature [66, 126].

Data reduction in ML

Data reduction in FCA

Characterize consistent conceptual data reductions

The basics on inverse scaling and measurability of contexts and lattices are recalled from Ganter and Wille [74, 80].

Other scaling methods

The relation to the plain scaling procedure to other scaling methods are discussed in Section 7.1.3. While these methods allow for more complex scalings, it is so far not known how these methods relate to the other scaling problems, i.e., inverse scaling, data reduction data compatibility, or how these problems can be formalized for them in a meaningful way. In Chapter 8 we define a hierarchical structure to browse between conceptual data reductions (cf. Figure 7.1). This is related to the *hierarchy of scales* [204] or the implicitly used hierarchy of conceptual scalings in *conceptual data systems* [193] like `TOSCANA` [123, 218].

8

Navigating Conceptual Views

Motivation

In the last chapter we gave an overview on conceptual data scaling and its methods. One of these dealt with continuous conceptual data reductions and how to identify them (cf. data reduction Figure 7.1). In this chapter we broaden the understanding on meaningful conceptual data reductions and how to derive them. For this, we first study the relation of conceptual data reductions and the views they create on the data. We do so, by introducing a refinement order relation on conceptual views. The resulting hierarchy allows us to compare results of various data reductions and navigate between them to refine results. On top of that, we are able to assess differences and the complexity of data reductions.

Ordinal Calculus for views

For more complex data operations to navigate the hierarchy of views, we propose ordinal methods to combine, slice and aggregate views that respect their conceptual structures. These operations are capable of combining the results of heterogeneous data reduction methods since they operate on their generated conceptual views. Moreover, the introduced (algebraic) methods do allow for rich interpretations of the results with respect to the input views. We explain in greater detail on how they can be used to compute the greater common, differences or missing information.

Interpretable representations

An important aspect in data analysis in general is the interpretation of the results. Unfortunately, the features generated in data reductions often elude from human comprehension, e.g., in case they are non-linear combinations of the original attributes. We present in Section 8.3 an approach to compute equivalent views using logical expressions and the original data attributes. Thereby, we preserve the structure of the data reduction while substituting all non-interpretable attributes by human comprehensible logical expressions.

Towards recommendation of views

Not every analyst initially knows what view they want to compute or which data reduction algorithm to use. By adapting the well known exploration [75] algorithm, we developed a semi-automatic procedure that is able to recommend views by querying user preferences. On top of that we investigate several structural properties and importance measures, and show how they can be applied for future automatic recommendations of views.

We demonstrate the applicability of all presented methods on three real world data sets.

8.1 The lattice of Conceptual Views

Comparing scale-measures

A notion for comparing scale-measures is provided by a natural order relation among scales [80, Definition 92]). We may present in the following a more general definition within the scope of scale-measures.

Definition 37 (Scale-Measure Refinement). *Let the set of all scale-measures of a context be denoted by* $\mathfrak{S}(\mathbb{K}) := \{(\sigma, \mathbb{S}) \mid \sigma$ *is an* $\mathbb{S}$*-measure of* $\mathbb{K}\}$. *For* $(\sigma, \mathbb{S}), (\psi, \mathbb{T}) \in \mathfrak{S}(\mathbb{K})$ *we say* $(\sigma, \mathbb{S})$ *is a* ***coarser*** *scale-measure of* $\mathbb{K}$ *than* $(\psi, \mathbb{T})$, *iff* $\sigma^{-1}(\text{Ext}(\mathbb{S}))\subseteq\psi^{-1}(\text{Ext}(\mathbb{T}))$. *Analogously we say* $(\psi, \mathbb{T})$ *is* ***finer*** *than* $(\sigma, \mathbb{S})$. *If* $(\sigma, \mathbb{S})$ *is finer and coarser than* $(\psi, \mathbb{T})$ *we call them* ***equivalent scale-measures*** *and denote this by* $(\sigma, \mathbb{S}) \sim (\psi, \mathbb{T})$.

Preorder

We remark that both the finer relation and the coarser relation constitute a reflexive and transitive relation. The transitivity follows from the continuity of the composition of scale maps (Corollary 1). Binary relations with these properties are also known as **preorder**. For the refinement preorder on $\mathfrak{S}(\mathbb{K})$ we us the symbol $\lesssim$. The symmetric instances of this preorder are given by the $\sim$ relation.

Generate scale-measure through navigation

By computing scale-measures having coarser scale contexts with respect to the *refinement preorder* we can provide a more general or abstract conceptual view on a data set. This kind of analytical approach, is exemplified by the ice cream flavors example in Figure 7.10.

Scale-measure generation methods are needed

Moreover, the set of all scale-measures for some formal context provides an abstract analytical structure to navigate and explore the conceptual structure of a data set. Yet, despite the supposed usefulness of the scale-measures, there are up until now no existing methods, to the best of our knowledge, for the generation and evaluation of scale-measures, in particular with respect to data science applications. To lay the foundation for the navigation methods we start with analyzing the structure of all scale-measures.

Lemma 2 (Scale-Measure Equivalence). *The scale-measure equivalence is an equivalence relation on the set of scale-measures.*

Proof. Let $(\sigma, \mathbb{S}), (\psi, \mathbb{T}), (\omega, \mathbb{O}) \in \mathfrak{S}(\mathbb{K})$ be scale-measures of context $\mathbb{K}$. Using Definition 37 we know from $(\sigma, \mathbb{S}) \sim (\psi, \mathbb{T})$ that $\sigma^{-1}(\text{Ext}(\mathbb{S})) = \psi^{-1}(\text{Ext}(\mathbb{T}))$, from which the reflexivity and the symmetry of $\sim$ can be inferred. Analogously we can infer from $(\sigma, \mathbb{S}) \sim (\psi, \mathbb{T})$ and $(\psi, \mathbb{T}) \sim (\omega, \mathbb{O})$ that $(\sigma, \mathbb{S}) \sim (\omega, \mathbb{O})$. □

Equivalent scale-measures and morphisms

Note that for two given equivalent scale-measures does their scale-measure equivalence not imply the existence of a bijective scale-measure between them. Yet, a minor requirement to the scale-measure map leads to a useful link.

Lemma 3 (Equivalent Scale-Measures and Embeddings). *Let* $(\sigma, \mathbb{S}), (\psi, \mathbb{T}) \in \mathfrak{S}(\mathbb{K})$ *with* $(\sigma, \mathbb{S}) \sim (\psi, \mathbb{T})$ *and* σ, ψ *are surjective maps. Then is* $\sigma^{-1} \circ \psi$ *an order isomorphism from* $\underline{\text{Ext}}(\mathbb{S}) := (\text{Ext}(\mathbb{S}), \subseteq)$ *to* $\underline{\text{Ext}}(\mathbb{T}) := (\text{Ext}(\mathbb{T}), \subseteq)$.

Proof. From [80, Proposition 118] we have that σ^{-1} is a injective $\wedge$-preserving order embedding of $\underline{\text{Ext}}(\mathbb{S})$ into $\underline{\text{Ext}}(\mathbb{K})$ and thereby a bijective $\wedge$-preserving order embedding into $(\sigma^{-1}(\text{Ext}(\mathbb{S})), \subseteq)$. The analogue holds for ψ^{-1} from $\text{Ext}(\mathbb{T})$ into $\psi^{-1}(\text{Ext}(\mathbb{T}))$. Due to $(\sigma, \mathbb{S}) \sim (\psi, \mathbb{T})$ we know that $\sigma^{-1}(\text{Ext}(\mathbb{S})) = \psi^{-1}(\text{Ext}(\mathbb{T}))$, which results in σ^{-1} being a bijective $\wedge$-preserving order embedding into $\psi^{-1}(\text{Ext}(\mathbb{T}))$. Hence, when restricting $\sigma^{-1} \circ \psi : \mathcal{P}(G_{\mathbb{S}}) \to \mathcal{P}(G_{\mathbb{T}})$ to the respective extent set we obtain a bijective map. The fact that all formal contexts are finite (throughout this work) and the monotonicity of the lifts of σ^{-1} and ψ to their respective power sets imply the required order preserving property follow. □

Note on surjectivity

We may stress that the required surjectivity is not constraining the application of scale-measures, since any object g of a scale-context having an empty pre-image may just be removed from the scale-context without consequences to the analysis (cf. Remark 4).

Hierarchy of scale-measures

The just discussed equivalence relation together with the refinement order allow to cope with the set of all scale-measures $\mathfrak{S}(\mathbb{K})$ in a meaningful way.

Definition 38 (Scale-Hierarchy). *For a given formal context $\mathbb{K}$, its set of all scale-measures $\mathfrak{S}(\mathbb{K})$ and $(\sigma,\mathbb{S}) \in \mathfrak{S}(\mathbb{K})$ let $[(\sigma,\mathbb{S})]_\sim := \{(\psi,\mathbb{T}) \mid (\sigma,\mathbb{S}) \sim (\psi,\mathbb{T})\}$ be the equivalence class of $(\sigma,\mathbb{S})$ and ${}^{\mathfrak{S}(\mathbb{K})}/_\sim := \{[(\sigma,\mathbb{S})]_\sim \mid (\sigma,\mathbb{S}) \in \mathfrak{S}(\mathbb{K})\}$ be the set of all equivalence classes of $\sim$. We call $\underline{\mathfrak{S}}(\mathbb{K}) := ({}^{\mathfrak{S}(\mathbb{K})}/_\sim, \leq)$ with $[(\sigma,\mathbb{S})]_\sim \leq [(\psi,\mathbb{T})]_\sim$ iff $(\sigma,\mathbb{S}) \lesssim (\psi,\mathbb{T})$ the **scale-hierarchy** of $\mathbb{K}$.*

The order relation $\leq$ on $\mathfrak{S}(\mathbb{K})/_\sim$ results naturally from the refinement order $\lesssim$, cf. Definition 37 and the paragraph thereafter. The order structure thus given represents all possible consistent data reductions on a contextual data set. Yet, it seems hardly comprehensible, or even applicable, in that form. For this, we present a characterization of said order structure in terms of closure systems.

Towards characterizing scale-measures by closure systems

Lemma 4 (Context with Closure System $\mathcal{A}$). *Let G be a set and $\mathcal{A} \subseteq \mathcal{P}(G)$ be a closure system. Furthermore, let $\mathbb{K}_\mathcal{A} := (G,\mathcal{A},\in)$ be a formal context using the element relation as incidence. The set of extents $\operatorname{Ext}(\mathbb{K}_\mathcal{A})$ is equal to the closure system $\mathcal{A}$.*

Proof. For any set $D \subseteq G$ and $A \in \mathcal{A}$ we find $(*)$ $D \subseteq A \implies A \in D'$. Since $\mathcal{A}$ is a closure system and $D'' = \bigcap D'$ we see that $D'' \in \mathcal{A}$, hence, $\operatorname{Ext}(\mathbb{K}_\mathcal{A}) \subseteq \mathcal{A}$. Conversely, for $A \in \mathcal{A}$ we can draw from $(*)$ that $A'' = A$, thus $A \in \operatorname{Ext}(\mathbb{K}_\mathcal{A})$. □

Characterizing scale-measures by closure systems

We want to further motivate the constructed formal context $\mathbb{K}_\mathcal{A}$ and its particular utility with respect to scale-measures. Since both contexts $\mathbb{K}$ and $\mathbb{K}_\mathcal{A}$ have the same set of objects, we may study the utility of the identity map $\iota : G \to G, g \mapsto g$ as scale-measure map.

Lemma 5 (Canonical Construction). *For a context $\mathbb{K}$ and any $\mathbb{S}$-measure σ the identity map ι is a $\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))}$-measure of $\mathbb{K}$, i.e., $(\iota,\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))}) \in \mathfrak{S}(\mathbb{K})$.*

Proof. Lemma 4 gives that $\operatorname{Ext}(\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))})$ is equal to $\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))$. Since $(\sigma,\mathbb{S}) \in \mathfrak{S}(\mathbb{K})$, i.e., $(\sigma,\mathbb{S})$ is a scale-measure of $\mathbb{K}$, we see that the pre-image $\sigma^{-1}(\operatorname{Ext}(\mathbb{S})) \subseteq \operatorname{Ext}(\mathbb{K})$, and thus $\iota^{-1}\big(\operatorname{Ext}(\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))})\big) \subseteq \operatorname{Ext}(\mathbb{K})$. □

Canonical representation of scale-measures

Using this canonical construction we can facilitate the understanding of the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$.

Proposition 10 (Canonical Representation). *Let $\mathbb{K} = (G,M,I)$ be a formal context with scale-measure $(\sigma,\mathbb{S}) \in \mathfrak{S}(\mathbb{K})$, then $(\sigma,\mathbb{S}) \sim (\iota,\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))})$.*

Proof. Lemma 5 states that ι is a $\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))}$-measure of $\mathbb{K}$. Furthermore, from Lemma 4 we know that the extent set of $\mathbb{K}_{\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))}$ is $\sigma^{-1}(\operatorname{Ext}(\mathbb{S}))$, as required by Definition 37. □

Scale-Hierarchy and coarser closure systems

Equipped with this proposition we are now able to compare sets of scale-measures for a given formal context $\mathbb{K}$ solely based on their respective attribute sets in canonical representation. Furthermore, since these representation sets are coarser closure systems of $\operatorname{Ext}(\mathbb{K})$, following from Definition 35, we may reformulate the problem for navigating scale-measures using coarser closure systems and their relations. For this we want to nourish the understanding of the correspondence of scale-measures and coarser closure systems. For this let $\operatorname{CS}(G)$ denote the ordered set of all closure systems on a set G and $\operatorname{CS}(\mathbb{K}) \subseteq \operatorname{CS}(G)$ the induced sub-order of all coarser closure systems of $\operatorname{Ext}(\mathbb{K})$.

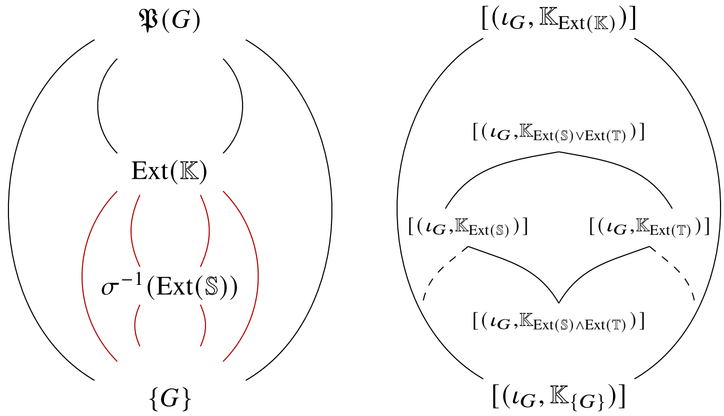


Figure 8.1: Scale-Hierarchy of $\mathbb{K}$ (right) and embedded in Boolean $\mathbb{B}_G$

Proposition 11 (Scale-Hierarchy and Lattice of Closure Systems). *For a formal context $\mathbb{K}$ and the ordered set of all coarser closure systems $\text{CS}(\mathbb{K}) \subseteq \text{CS}(G)$, the following map is an order isomorphism:*

$$i : \text{CS}(\mathbb{K}) \to \underline{\mathfrak{S}}(\mathbb{K}),\ \mathcal{A} \mapsto i(\mathcal{A}) := [(\iota, \mathbb{K}_{\mathcal{A}})]_{\sim}$$

Proof. Let $\mathcal{A}, \mathcal{B} \subseteq \text{Ext}(\mathbb{K})$ be two different closure systems on G. Then the images of $\mathcal{A}$ respectively $\mathcal{B}$ under i are a scale-measures of $\mathbb{K}$, according to Lemma 5, with extents $\mathcal{A}$ and $\mathcal{B}$, respectively. Since $\mathcal{A} \neq \mathcal{B} \iff \text{Ext}(\mathbb{K}_{\mathcal{A}}) \neq \text{Ext}(\mathbb{K}_{\mathcal{B}})$ we can follow that $(\iota, \mathbb{K}_{\mathcal{B}}) \not\sim (\iota, \mathbb{K}_{\mathcal{B}})$, and that i is an injective map. For the surjectivity of i let $[(\sigma, \mathbb{S})]_{\sim} \in \underline{\mathfrak{S}}(\mathbb{K})$, then $(\iota, \mathbb{K}_{\sigma^{-1}(\text{Ext}(\mathbb{S}))}) \sim (\sigma, \mathbb{S})$, i.e., an equivalent representation having extents $\sigma^{-1}(\text{Ext}(\mathbb{S})) \subseteq \text{Ext}(\mathbb{K})$ and $i\big(\sigma^{-1}(\text{Ext}(\mathbb{S}))\big) = [(\iota, \mathbb{K}_{\sigma^{-1}(\text{Ext}(\mathbb{S}))})]_{\sim}$. Finally, for $\mathcal{A} \subseteq \mathcal{B}$ we find that $i(\mathcal{A}) \subseteq i(\mathcal{B})$, since $\text{Ext}(\mathbb{K}_{\mathcal{A}}) \subseteq \text{Ext}(\mathbb{K}_{\mathcal{B}})$, as required. □

In the following we identify equivalence classes of the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$ with any of the respective representatives. This choice of notation leads to a more comprehensible presentation of the upcoming statements and considerations.

Number of coarser closure systems

The order isomorphism i allows us to analyze the structure of the scale-hierarchy by studying the related closure systems. For instance, the problem of computing $|\underline{\mathfrak{S}}(\mathbb{K})|$, i.e., the size of the scale-hierarchy. In the case of the context $\mathbb{K}_{\mathcal{P}(G)}$ this problem is equivalent to calculating the number of closure systems on G, sometimes referred to as Moore families. This number grows tremendously in $|G|$ and is known up to $|G| = 7$, for which it is known [47, 86, 97] to be 14 087 648 235 707 352 472. In the general case the size of the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$ is equal to the size of the order ideal $\downarrow \text{Ext}(\mathbb{K}) = \text{CS}(\mathbb{K})$ in $\text{CS}(G)$. Hence, the scale-hierarchy in its entirety may be computational intractable and therefore demands for efficient navigation algorithms. In order to derive these, we must derive useful theoretical properties that can be applied in scale-measures.

Lattice of conceptual data reductions

The fact that the set of all closure systems on G is again a closure system [34], which is lattice ordered by set inclusion, and the isomorphism in Proposition 11 allow for the following statement.

Proposition 12 (Scale-Hierarchy Order). *For a formal context $\mathbb{K}$ is the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$ lattice ordered.*

We depicted this lattice order relation in the form of abstract visualizations in Figure 8.1. In the bottom (right) we see the most simple scale which has only one attribute, G. The top (right) element in this figure is then the scale which has all extents of $\mathbb{K}$. On the left we see the lattice ordered set of all closure systems on a set G, in which we find the embedding of the hierarchy of scale-measures.

Properties of lattices

Equipped with this structure we have to recall a few notions and definitions for a (complete) lattice $(L, \leq)$. We want to remind the reader that by $\prec$ we denote the cover relation of $\leq$. Furthermore, we say L is 1) **lower semi-modular** if and only if $\forall x, y \in L : x \prec x \vee y \implies x \wedge y \prec y$, 2) **join-semidistributive** iff $\forall x, y, z \in L : x \vee y = x \vee z \implies x \vee y = x \vee (y \wedge z)$, 3) **meet-distributive** (**lower locally distributive**, cf [34]) iff L is join-semidistributive and lower semi-modular, 4) **join-pseudocomplemented** iff for any $x \in L$ the set $\{y \in L \mid y \vee x = \top\}$ has a least, 5) **ranked** iff there is a function $\rho : L \mapsto \mathbb{N}$ with $x \prec y \implies \rho(x) + 1 = \rho(y)$, 6) **atomistic** iff every $x \in L$ can be written as the join of atoms in L. In addition to the just introduced lattice properties, there are properties for elements in L that we consider. An element $x \in L$ is 1) **neutral** iff every triple $\{x, y, z\} \subseteq L$ generates a distributive sub-lattice of L, 2) **distributive** iff the equalities $x \vee (y \wedge z) = (x \vee y) \wedge (x \vee z)$ and $x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z)$ for every $y, z \in L$ hold.

Properties of lattice elements

Characterizing covering scale-measures

We can derive from literature [34, Proposition 19] the following statement.

Corollary 3 (Scale-Hierarchy Cover Relation). *For a context* $\mathbb{K} = (G, M, I)$ *and* $\mathcal{R}, \mathcal{R}' \in \mathrm{CS}(\mathbb{K})$ *TFAE:*

i) $\mathcal{R}' \prec \mathcal{R}$*,*

ii) $\mathcal{R}' \cup \{A\} = \mathcal{R}$ *with* A *is meet-irreducible in* $\mathcal{R}$*.*

Join-irreducible scale-measures

Of special interest in lattices are the (meet-) join-irreducibles, since every element of a lattice can be represented as a (meet) join of these elements.

Proposition 13 (Scale-Hierarchy Join-Irreducibles). *For a context* $\mathbb{K}$ *and* $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$*:* $\mathcal{R}$ *is join-irreducible in* $\mathrm{CS}(\mathbb{K}) \iff \exists A \in \mathrm{Ext}(\mathbb{K}) \setminus \{G\} : \mathcal{R} = \{G, A\}$

Proof. $\Leftarrow$: For $A \in \mathrm{Ext}(\mathbb{K}) \setminus \{G\}$ is $\{A, G\}$ a closure system on G and thereby in $\mathrm{CS}(\mathbb{K})$. Further, the set $\{A, G\}$ is of cardinality two and thereby an atom of $\mathrm{CS}(\mathbb{K})$ and thus join-irreducible. $\Rightarrow$: By contradiction assume that $\nexists A \in \mathrm{Ext}(\mathbb{K}) \setminus \{G\} : \mathcal{R} = \{G, A\}$, then for every $D \in \mathcal{R} \setminus \{G\}$ is $\{D, G\}$ an atom of $\mathrm{CS}(\mathbb{K})$, hence, $\mathcal{R} = \bigvee_{D \in \mathcal{R} \setminus \{G\}} \{D, G\}$, i.e., not join-irreducible. □

Towards meet-irreducible scale-measures

Next, we investigate the meet-irreducibles of $\mathrm{CS}(\mathbb{K})$ using a similar approach as done for $\mathrm{CS}(G)$ [34] (denoted $\mathcal{K}$ in Caspard and Monjardet [34]) based on propositional logic. We recall, that an (object) implication for some context $\mathbb{K}$ is a pair $(A, B) \in \mathcal{P}(G) \times \mathcal{P}(G)$, shortly denoted by $A \to B$. We say $A \to B$ is valid in $\mathbb{K}$ iff $A' \subseteq B'$ (cf. Section 5.4).

Coarsoning closure systems by implications

The set $\mathcal{F}_{A,B} := \{D \subseteq G : A \not\subseteq D \vee B \subseteq D\}$ is the set of all *models* of $A \to B$. Additionally, $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} := \mathcal{F}_{A,B} \cap \mathrm{Ext}(\mathbb{K})$ is the set of all extents $D \in \mathrm{Ext}(\mathbb{K})$ that are models of $A \to B$. The set $\mathcal{F}_{A,B}$ is a closure system [34] and therefor $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$, too. Furthermore, we can deduce that $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \in \mathrm{CS}(\mathbb{K})$.

Lemma 6 (Scale-Measures by Implications). *For a context* $\mathbb{K}$ *and* $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ *with closure operator* $\mathrm{cl}_{\mathcal{R}}$ *we find* $\mathcal{R} = \bigcap\{\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \mid A, B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\}$.

Proof. We know that $\mathcal{R} = \bigcap\{\mathcal{F}_{A,B} \mid A, B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\}$ [34, Proposition 22]. Since $\mathcal{R} \subseteq \mathrm{Ext}(\mathbb{K})$ it holds that $\mathcal{R} = \bigcap\{\mathcal{F}_{A,B} \mid A, B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\} \cap \mathrm{Ext}(\mathbb{K})$ and thus equal to $\bigcap \left\{ \mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \middle| A, B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\right\}$. □

Meet-irreducible scale-measures

Note that for any $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ the set $\left\{ \mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \,\middle|\, A,B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\right\}$ contains only closure systems in $\mathrm{CS}(\mathbb{K})$ and thus possibly meet-irreducible elements of $\mathrm{CS}(\mathbb{K})$.

Proposition 14 (Scale-Hierarchy Meet-Irreducibles). *For a context $\mathbb{K}$ and $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ TFAE: 1. $\mathcal{R}$ is meet-irreducible in $\mathrm{CS}(\mathbb{K})$ 2. $\exists A \in \mathrm{Ext}(\mathbb{K}), i \in G$ with $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} (A \cup \{i\})''$ such that $\mathcal{R} = \mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})}$*

Proof. [1. $\Rightarrow$ 2.] Due to Lemma 6 we can represent $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ by the equation $\mathcal{R} = \bigcap\{\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \mid A,B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\}$. Moreover, since $\mathcal{R}$ is meet-irreducible in $\mathrm{CS}(\mathbb{K})$, we can infer that $\mathcal{R} \in \{\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \mid A,B \subseteq G \wedge B \subseteq \mathrm{cl}_{\mathcal{R}}(A)\}$. In particular there exist $A,B \subseteq G$ with $B \subseteq \mathrm{cl}_{\mathcal{R}}(A)$ such that $\mathcal{R} = \mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$, and thus $\mathcal{R} = \mathcal{F}_{A'',B}\big|_{\mathrm{Ext}(\mathbb{K})}$.Using the fact that $\mathcal{F}_{A,\{i\}} \cap \mathcal{F}_{A,\{j\}} = \mathcal{F}_{A,\{i,j\}}$ we can infer that $\mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})} \cap \mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})} = \mathcal{F}_{A,\{i,j\}}\big|_{\mathrm{Ext}(\mathbb{K})}$. Therefore, there must exist $A,\{i\} \subseteq G$ with $\mathcal{R} = \mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})}$ $(*)$.

In the case that $A = (A \cup \{i\})''$ the set $\mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})} = \mathrm{Ext}(\mathbb{K})$ and $\mathcal{R}$ is thereby not meet-irreducible. Assume that $A \not\prec_{\underline{\mathrm{Ext}(\mathbb{K})}} (A \cup \{i\})''$, then there is a $D \in \mathrm{Ext}(\mathbb{K})$ with $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} D \subseteq (A \cup \{i\})''$ and $i \notin D$. Hence $A, D \not\models A \to \{i\}$ (see $*$) and thus $A, D \notin \mathcal{R}$. Using this, we construct two sets $\mathcal{R} \cup \{A\}$ and $\mathcal{R} \cup \{D\}$. The set $\mathcal{R} \cup \{D\}$ is closed by intersection, since an intersection of D with an element in $\mathcal{R}$ is a model of $A \to i$, thus $\mathcal{R} \cup \{D\} \in \mathrm{CS}(\mathbb{K})$. The same holds for $\mathcal{R} \cup \{A\}$ respectively. The intersection of $\mathcal{R} \cup \{A\}$ and $\mathcal{R} \cup \{D\}$ is equal to $\mathcal{R}$ which is thereby not meet-irreducible, a contradiction.

[1. $\Leftarrow$ 2.] Consider a closure system $\hat{\mathcal{F}} \in \mathrm{CS}(\mathbb{K})$ with $\hat{\mathcal{F}}$ covers $\mathcal{R}$ in $\mathrm{CS}(\mathbb{K})$. By Corollary 3, we can represent $\hat{\mathcal{F}} = \mathcal{R} \cup \{D\}$ $(*)$ with $D \notin \mathcal{R}$ and D is meet-irreducible in $\hat{\mathcal{F}}$ (and therefore $D \in \mathrm{Ext}(\mathbb{K})$). Due to $\mathcal{R} \subseteq \hat{\mathcal{F}}$ the set $(A \cup \{i\})''$ is an element of $\hat{\mathcal{F}}$ and thereby the intersection $(A \cup \{i\})'' \cap D \in \hat{\mathcal{F}}$. Since $D \notin \mathcal{R}$, we can deduce that $D \not\models A \to i$ and therefor $A \subseteq D$ and $i \notin D$. From $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} (A \cup \{i\})''$ we know that $(A \cup \{i\})'' \cap D = A$. Finally, $D \in \hat{\mathcal{F}} \implies A \in \hat{\mathcal{F}}$, and using $(*)$, we can infer that $D = A$. Hence, $\mathcal{R} \cup \{A\}$ is the sole upper neighbor of $\mathcal{R}$ in $\mathrm{CS}(\mathbb{K})$ and thereby $\mathcal{R}$ is meet-irreducible. □

Maximum meet-irreducible scale-measures

Propositions 13 and 14 provide a characterization of irreducible elements in $\mathrm{CS}(\mathbb{K})$ and thereby in the scale-hierarchy of $\mathbb{K}$. Those may be of particular interest, since any element of $\mathrm{CS}(\mathbb{K})$ is representable by irreducible elements.

Proposition 15 (Scale-Hierarchy Maximum Meet-Irreducible Element). *For a context $\mathbb{K} = (G,M,I)$, $A,B \in \mathrm{Ext}(\mathbb{K})$ with $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} B$ and A is meet-irreducible in $\mathrm{Ext}(\mathbb{K})$, then $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$ is a maximum meet-irreducible element in $\mathrm{CS}(\mathbb{K}) \subseteq \mathrm{CS}(G)$.*

Proof. For $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} B$, A is the only extent that is not a model of implication $A \to B$, since every other superset of A in $\mathrm{Ext}(\mathbb{K})$ is also a superset of B. Hence $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$ is equal to $\mathrm{Ext}(\mathbb{K}) \setminus \{A\}$. The only superset in $\mathrm{CS}(\mathbb{K})$ is $\mathrm{Ext}(\mathbb{K})$, which is not meet-irreducible. □

Counting meet-irreducible scale-measures

Equipped with this characterization we look into counting the meet-irreducibles.

Proposition 16 (Number of Meet-Irreducibles in $\underline{\mathfrak{S}}(\mathbb{K})$). *For context $\mathbb{K}$, the number of meet-irreducible elements in the lattice $\mathrm{CS}(\mathbb{K})$ is equal to $|\prec_{\mathrm{Ext}\,\mathbb{K}}|$.*

Proof. According to Proposition 14, an element $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ is meet-irreducible iff it can be represented as $\mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})}$ for some $A \in \mathrm{Ext}(\mathbb{K})$ and some $i \in G$ with $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} (A \cup \{i\})''$. Hence, the number of meet-irreducible elements is bound from above by the number of covering pairs $A \prec_{\underline{\mathrm{Ext}(\mathbb{K})}} B$ in $\mathrm{Ext}(\mathbb{K})$. It remains to be shown that for $\mathcal{R}$ there is only one pair

$(A,B) \in \prec_{\underline{\mathrm{Ext}}(\mathbb{K})}$ with $B = (A \cup \{i\})''$ for some $i \in B \setminus A$ such that $\mathcal{R} = \mathcal{F}_{A,\{i\}}\big|_{\mathrm{Ext}(\mathbb{K})}$. Assume there are $(A,B),(C,D) \in \prec_{\underline{\mathrm{Ext}}(\mathbb{K})}$ with $(A,B) \neq (C,D)$ and $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} = \mathcal{F}_{C,D}\big|_{\mathrm{Ext}(\mathbb{K})}$. First, consider the case $A \neq C$. Without loss of generality let $A \not\subseteq C$, then we have $C \models A \to B$, but $C \not\models C \to D$. Therefore $C \in \mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$ but $C \notin \mathcal{F}_{C,D}\big|_{\mathrm{Ext}(\mathbb{K})}$. In the second case, $A = C$, we have $B \neq D$ and thus $B \not\models C \to D$, but $B \models A \to B$. This implies that $B \in \mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})}$ but $B \notin \mathcal{F}_{C,D}\big|_{\mathrm{Ext}(\mathbb{K})}$. Thus, $\mathcal{F}_{A,B}\big|_{\mathrm{Ext}(\mathbb{K})} \neq \mathcal{F}_{C,D}\big|_{\mathrm{Ext}(\mathbb{K})}$. □

Next, we turn ourselves to other lattice properties of $\mathrm{CS}(\mathbb{K})$ and its elements.

Lemma 7 (Join Pseudocomplement). *For a context $\mathbb{K}$ and $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$, the set $\hat{\mathcal{R}} = \bigvee_{A \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \mathrm{M}(\mathcal{R})} \{A, G\}$ is the inclusion minimum closure system with $\mathcal{R} \vee \hat{\mathcal{R}} = \mathrm{Ext}(\mathbb{K})$.*

Proof. A set $\mathcal{A} \subseteq \mathrm{Ext}(\mathbb{K})$ is a generator of $\mathrm{Ext}(\mathbb{K})$ iff all meet-irreducible elements of $\mathrm{Ext}(\mathbb{K})$ are in $\mathcal{A}$. Hence, for every $\mathcal{D} \in \mathrm{CS}(\mathbb{K})$ with $\mathcal{R} \vee \mathcal{D} = \mathrm{Ext}(\mathbb{K})$, we have D is a superset of $\mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \mathrm{M}(\mathcal{R})$ and thus of $\hat{R}$, since $\hat{R}$ it is the closure of $\mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \mathrm{M}(\mathcal{R})$ in $\mathrm{CS}(\mathbb{K})$. □

All the above results in the following statement about $\mathrm{CS}(\mathbb{K})$:

Lattice properties of the scale-hierarchy

Proposition 17 (Lattice Properties of $\underline{\mathfrak{S}}(\mathbb{K})$). *For any context $\mathbb{K}$, the lattice $\mathrm{CS}(\mathbb{K})$ is:*

i) join-semidistributive

ii) lower semi-modular

iii) meet-distributive

iv) join-pseudocomplemented

v) ranked

vi) atomistic

Proof. i) According to [34, Corollary 30] $\mathrm{CS}(G)$ is join-semidistributive and therefor $\mathrm{CS}(\mathbb{K})$ too, since the meet and join operations of $\mathrm{CS}(G)$ are closed in $\mathrm{CS}(\mathbb{K})$. ii) Analogue to i). iii) Follows from i) and ii) (cf. Definition 15 (5) [34]). iv) The join-pseudocomplement of any $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ is given by $\hat{\mathcal{R}}$ according to Lemma 7. v) The lattice $\mathrm{CS}(G)$ is ranked by the cardinality function [34, Corollary 30]. Since $\mathrm{CS}(\mathbb{K})$ is an order ideal in $\mathrm{CS}(G)$, it is ranked by the same function. vi) Follows directly from the characterization of join-irreducibles in Proposition 13. □

Decomposing the scale-hierarchy

This result can be employed for the recommendation of scale-measures, in particular with respect to Libkins decomposition theorem [141, Theorem 1]. This would allow for a divide-and-conquer procedure within the scale-hierarchy, based on the fact: for context $\mathbb{K}$ the lattice $\mathrm{CS}(\mathbb{K}) \subseteq \mathrm{CS}(G)$ is decomposable into the direct product of two lattices $\mathrm{CS}(\mathbb{K}) \cong L_1 \times L_2$ iff $L_1 = (n], L_2 = (\overline{n}]$ and n is neutral in $\mathrm{CS}(\mathbb{K})$. Here $\overline{n}$ indicates the complement of n with respect to $\mathrm{CS}(\mathbb{K})$, which can be computed using Lemma 7. That this approach is reasonable can be drawn from the fact that $\mathrm{CS}(\mathbb{K})$ fullfils all requirements of Lemma 2 and Theorem 1 from Libkin's work [140, 141] by considering Proposition 17.

On identifing decomposition

In the rest of this section we investigate distributive and neutral elements in $\mathrm{CS}(\mathbb{K})$ more deeply. For this, let $\mathrm{cl}_1, \mathrm{cl}_2 \in \mathrm{CL}(L)$, i.e., the set of all closure operators on lattice L. We say that $\mathrm{cl}_2 \leq_{\mathrm{cl}} \mathrm{cl}_1$ iff for all $x \in L : \mathrm{cl}_2(x) \leq_{\mathrm{cl}} \mathrm{cl}_1(x)$.

Lemma 8 (Coarser Closure Systems and Coarser Closure Operators). *For any context $\mathbb{K}$, we find that $i : \mathrm{CS}(\mathbb{K}) \mapsto \mathrm{CL}(\mathrm{Ext}(\mathbb{K}))$ with $i(\mathcal{A}) \to \mathrm{cl}_{\mathcal{A}}|_{\mathrm{Ext}(\mathbb{K})}$ is a dual-isomorphism.*

Proof. For $\mathcal{A}, \mathcal{D} \in \mathrm{CS}(\mathbb{K})$ with $A \in \mathcal{A}, A \notin \mathcal{D}$ is $i(\mathcal{A})(A) = A$ but $i(\mathcal{D})(A) \neq A$. Thus $i(\mathcal{A}) \neq i(\mathcal{D})$ and i injective. For $\mathrm{cl} \in \mathrm{CL}(\mathrm{Ext}(\mathbb{K}))$ is $\mathrm{cl}(\mathrm{Ext}(\mathbb{K})) \subseteq \mathrm{Ext}(\mathbb{K})$ a closure system with $G \in \mathrm{cl}(\mathrm{Ext}(\mathbb{K}))$ an therefor $\mathrm{cl}(\mathrm{Ext}(\mathbb{K})) \in \mathrm{CS}(\mathbb{K})$ with $i(\mathrm{cl}(\mathrm{Ext}(\mathbb{K}))) = \mathrm{cl}$. Hence, i is bijective. For $\mathcal{A}, \mathcal{D} \in \mathrm{CS}(\mathbb{K})$ with $\mathcal{A} \prec_{\mathrm{CS}(\mathbb{K})} \mathcal{D}$ is $\mathcal{A} \cup \{D\} = \mathcal{D}$ for D meet-irreducible in $\mathcal{D}$ (Corollary 3). Thus for all $A \in \mathrm{Ext}(\mathbb{K})$ is $i(\mathcal{A})(A) = i(\mathcal{D})(A)$ except for the pre-images of D, i.e., $i(\mathcal{D})^{-1}(D)$. For $A \in i(\mathcal{D})^{-1}(D)$ is $i(\mathcal{D})(A) = D \subseteq i(\mathcal{A})(A)$ and thus $i(\mathcal{D}) \leq_{\mathrm{cl}} i(\mathcal{A})$, as required. □

Proposition 18 (Neutral and Distributive Elements in $\underline{\mathfrak{S}}(\mathbb{K})$). *For any formal context $\mathbb{K}$ and $\mathcal{R} \in \mathrm{CS}(\mathbb{K})$ TFAE: i) $\mathcal{R}$ is distributive in $\underline{\mathfrak{S}}(\mathbb{K})$ ii) $\mathcal{R}$ is neutral in $\underline{\mathfrak{S}}(\mathbb{K})$ iii) For $A, B, C \in \mathrm{Ext}(\mathbb{K})$ with $C = A \cap B$ and A, B incomparable in $\mathrm{Ext}(\mathbb{K})$, we have $A \in \mathcal{R}$ or $B \in \mathcal{R}$ or $C \in \mathcal{R}$ implies $A, B, C \in \mathcal{R}$.*

Proof. Using Lemma 8, i)⇔ii) due to Thm. 2 [161] and i)⇔iii) due to Thm. 1 [161]. □

On computing decompositions

An additional accompanying property is that the set of elements that are distributive in $\mathrm{CS}(\mathbb{K})$ is a sub-lattice of $\mathrm{CS}(\mathbb{K})$ (Theorem 2 [160]). Thus, the iterative procedure that results from Proposition 18, iii) yields a closure operator on $\mathrm{CS}(\mathbb{K})$ to compute the neutral elements. To nourish our understanding of the neutral elements take the following example: in the lattice $\mathrm{CS}(G)$ only the top and bottom elements are neutral [34, Proposition 33 (5)]. In contrast, for a chain $C \subseteq \mathcal{P}(G)$ with $G \in C$ is $\mathrm{CS}(\mathbb{K}_C) \subseteq \mathrm{CS}(G)$ a distributive lattice and thus every element a neutral element.

8.2 Navigation Methods for Conceptual Views

Browsing between views

Based on the just introduced scale-hierarchy, we provide in this section the means for efficiently browsing this structure. Given a data set, the presented methods are able to compute and combine arbitrary views using structural operations. In Section 8.3 we introduce logical operations that can not only be used for the computation of views but also for their explanation in a humanly comprehensible manner. Together, they ultimately resemble a *navigation* through conceptual measurements.

8.2.1 Lattice based Navigation

Lattice based browsing

The scale-hierarchy constitutes an ordered set which allows us to navigate between scale-measures using the finer and coarser relation. This yields us conceptual views that are more detailed (finer) and complex with more closed object sets or conceptual views that are more general (coarser) with a less complex structure. Each step in the finer order relation adds one element to the structure that becomes meet-irreducible (cf. Corollary 3) allowing for a gradual increase in complexity. The dual applies to the coarser relation. Representations for the resulting scale-measures are given by the canonical construction (Lemma 5).

Meet and join of views

From the isomorphism between the scale-hierarchy and an order ideal in the lattice of all closure systems (cf. Proposition 11) we can infer that the scale-hierarchy is lattice ordered (Proposition 12). With the following proposition we give constructions for the meet and join of scale-measures in the scale-hierarchy.

Proposition 19 (Scale-Hierarchy Meet and Join Operations). *Let $\wedge, \vee$ be the natural lattice operations in $\underline{\mathfrak{S}}(\mathbb{K})$ and let $[(\sigma, \mathbb{S})]_\sim, [(\psi, \mathbb{T})]_\sim \in \underline{\mathfrak{S}}(\mathbb{K})$. We then find that:*

$$Meet : [(\sigma, \mathbb{S})]_\sim \wedge [(\psi, \mathbb{T})]_\sim = [(\iota, \mathbb{K}_{\sigma^{-1}(\mathrm{Ext}(\mathbb{S})) \cap \psi^{-1}(\mathrm{Ext}(\mathbb{T}))})]_\sim,$$

$$Join\ :\ [(\sigma,\mathbb{S})]_{\sim} \vee [(\psi,\mathbb{T})]_{\sim} = [(\iota, \mathbb{K}_{\{A\cap B \mid A\in\sigma^{-1}(\mathrm{Ext}(\mathbb{S})), B\in\psi^{-1}(\mathrm{Ext}(\mathbb{T}))\}})]_{\sim}.$$

Proof. In the following, we write $(\sigma,\mathbb{S})$ instead of $[(\sigma,\mathbb{S})]_{\sim}$ for better readability. 1. For the pre-images $i^{-1}(\sigma,\mathbb{S})$, $i^{-1}(\psi,\mathbb{T})$ (Proposition 11) we can compute their meet [34], which yields

$$i^{-1}(\sigma,\mathbb{S}) \wedge i^{-1}(\psi,\mathbb{T}) = \sigma^{-1}(\mathrm{Ext}(\mathbb{S})) \cap \psi^{-1}(\mathrm{Ext}(\mathbb{T})).$$

2. The join [34] of the scale-measure pre-images under i (Proposition 11) is equal to $\{A\cap B \mid A\in\sigma^{-1}(\mathrm{Ext}(\mathbb{S})), B\in\psi^{-1}(\mathrm{Ext}(\mathbb{T}))\}$, which results in the required expression by applying the order isomorphism i. □

The join and meet of scale-measures can also be used for browsing the scale-hierarchy. With the meet of two scale-measures we compute the *greater common* of their conceptual views. Their join yields the smallest conceptual view such that both of them are entailed in it. Conceptual views and their meet/join carry in some cases a special semantic and allow for additional interpretation. We show this based on an example in Section 8.2.3.

View of missing information

Another useful property of the scale-hierarchy is that it is join-pseudocomplement (see Lemma 7). This allows us to compute for a conceptual view the smallest view needed to reconstruct the context $\mathbb{K}$. We interpret the join-pseudocomplement as a scale-measure that carries all missing information.

8.2.2 Combining Conceptual Views

Ordinal calculus for views

Other methods to navigate in the scale-hierarchy combine or join views. The first operation we transfer to the realm of scale-measures is the context apposition (cf. Lemma 1 and surrounding text).

Apposition of views

Definition 39 (Apposition of Scale-Measures). *Let* $(\sigma,\mathbb{S}), (\psi,\mathbb{T})$ *be scale-measures of* $\mathbb{K}$. *Then the* apposition of scale-measures $(\sigma,\mathbb{S}) \mid (\psi,\mathbb{T})$ *is:*

$$(\sigma,\mathbb{S}) \mid (\psi,\mathbb{T}) := \begin{cases} (\sigma,\mathbb{S} \mid \mathbb{T}) & \text{if } G_{\mathbb{S}} = G_{\mathbb{T}}, \sigma = \psi \\ (\sigma,\mathbb{S}) \vee (\psi,\mathbb{T}) & \text{else} \end{cases}$$

In case both scale-measures are defined on the same set of objects and use the same map, i.e., $G_{\mathbb{S}} = G_{\mathbb{T}}$ and $\sigma = \psi$, we define the scale-measure based on the apposition of their scale contexts. The resulting scale-measure reflects the same extents as their join in the scale-hierarchy. This can be derived from the following proposition and Proposition 19.

Proposition 20 (Apposition Scale-Measure). *Let* $(\sigma,\mathbb{S}), (\psi,\mathbb{T})$ *be two scale-measures of* $\mathbb{K}$. *Then is* $(\sigma,\mathbb{S}) \mid (\psi,\mathbb{T}) \in \mathfrak{S}(\mathbb{K})$.

Proof. 1. In the first case we know that set of extents $\mathrm{Ext}(\mathbb{S}\mid\mathbb{T})$ contains all intersections $A\cap B$ for $A\in\mathrm{Ext}(\mathbb{S})$ and $B\in\mathrm{Ext}(\mathbb{T})$ [80] (cf. Lemma 1). Furthermore, we know that we can represent $\sigma^{-1}(A\cap B) = \sigma^{-1}(A)\cap\sigma^{-1}(B) = \sigma^{-1}(A)\cap\psi^{-1}(B)$. Since $\sigma^{-1}(\mathrm{Ext}(\mathbb{S})), \psi^{-1}(\mathrm{Ext}(\mathbb{T})) \subseteq \mathrm{Ext}(\mathbb{K})$, we can infer that the intersection $\sigma^{-1}(A)\cap\psi^{-1}(B) \in \mathrm{Ext}(\mathbb{K})$. 2. The second case follows from Proposition 19. □

The apposition operator combines two scale-measures of a data context to a new single scale-measure. We may note that the special case of $(\sigma,\mathbb{S}) = (\iota_G,\mathbb{K})$ was already discussed by Ganter and Wille [80].

Attribute selection

Other than using the coarser relation in the scale-hierarchy we present the following operation to reduce the complexity of scale-measures.

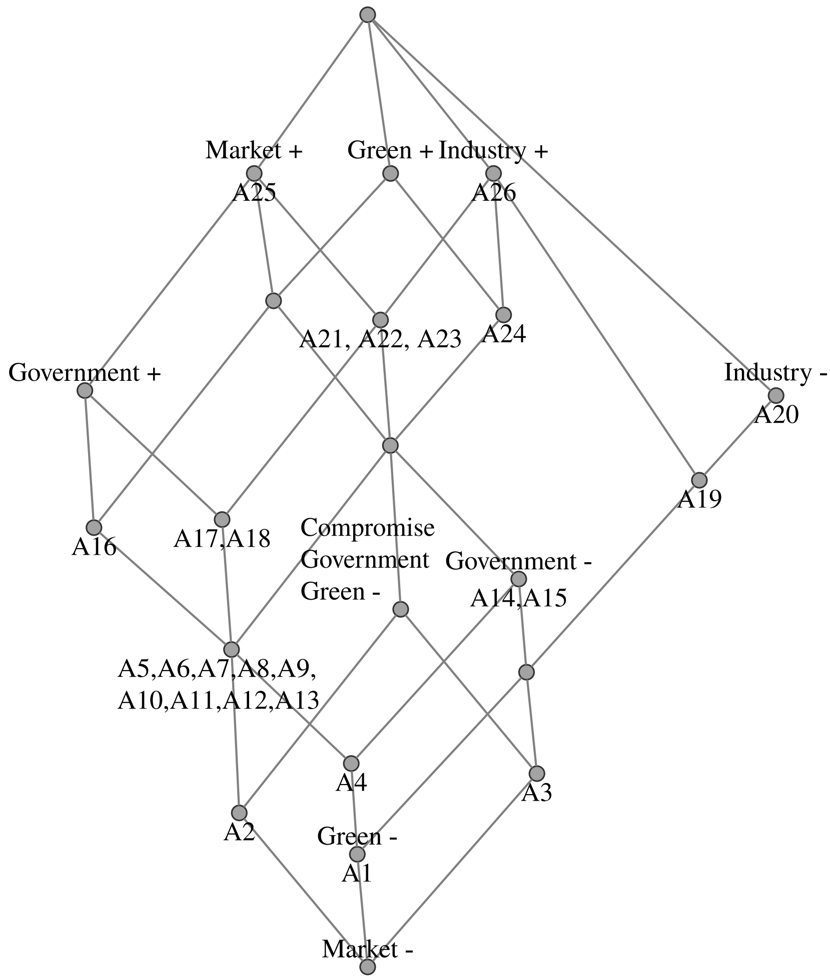


Figure 8.2: Concept Lattice of journalistic articles and their conventions.

Corollary 4 (Attribute Projection). *Let* $\mathbb{K} = (G, M, I)$ *be a formal context,* $M_{\mathbb{S}} \subseteq M$*, and* $I_{\mathbb{S}} := I \cap (G \times M_{\mathbb{S}})$*, then is* ι_G *a* $(G, M_{\mathbb{S}}, I_{\mathbb{S}})$*-measure of* $\mathbb{K}$*.*

Proof. The map id_G is a $\mathbb{K}$-measure of $\mathbb{K}$, hence id_G is a $(G, \{n\}, I \cap (G \times \{n\}))$-measure of $\mathbb{K}$ for every $n \in M$, and in particular $n \in M_{\mathbb{S}}$, by Proposition 7, leading to $\big(\iota_G, (G, M_{\mathbb{S}}, I_{\mathbb{S}})\big)$ being a scale-measure of $\mathbb{K}$, cf. Proposition 20. □

8.2.3 Contexts with special Semantics

Contexts with special semantics

The meet and join operation of views carry in some cases special semantics. An example is the context of journalistic articles and their conventions [214] (Figure 8.2). The attributes in this context are classifications based on the sociological theory *economics of conventions* [54, 199]. The theory states that within a conflict arguments can be traced back to a small set of dimensions of justification, i.e., *conventions*. The context includes four of these conventions, i.e., *Governmental, Green, Market* and *Industry*, including two signs +/- that indicate if an article relates positively or negatively to a convention. The data set contains articles on the conflict of electric mobility as objects and the incidence encodes how the arguments of an article relate to a convention.

Meet as negotiation of compromises

The meet in this concept lattice is of special interest since articles in the intersection of conventions indicates that a *compromise* between perspectives may be formed. One such compromise is encoded by *compromise Government/Green* in the concept lattice. Based on the interpretation of the domain experts does the newly formed compromise become a new dimension and does not need to coincide with its individual parts, i.e., *Government* and *Green*.

8.3 Explaining Conceptual Data Reductions

The need of interpretable views

Although the canonical representation of scale-measures is complete up to equivalence (Proposition 10), this representation eludes human explanation to some degree. The use of the extentional structure of $\mathbb{K}$ as attributes provides insights to the scale-hierarchy itself, but it does not do so for the data, i.e., the objects, attributes, and their relation. A formulation of scales using attributes from $\mathbb{K}$, and their combinations, seems more natural and more comprehensible. In order to facilitate such a formulation, we employ an approach similar to logical scaling [170] where logical expressions are used to introduce new attributes. The thereby newly introduced attributes have a real-world semantic in terms of the attributes and the logic, which is an advantage with respect to (human-)comprehensibility.

Use logic to explain views

Interpretable navigation problem

In this work we use propositional logic on the attributes $F[M, \{\vee, \wedge, \neg\}]$. We identify the variables of the logic using attributes of $\mathbb{K} := (G, M, I)$. For example, *Choco Ice* $\vee$ *Choco Pieces* is an expression of $F[L, \{\wedge, \vee, \neg\}]$ for the ice-cream context and $\{$*Choco Pieces*, *Caramel*$\} \models$ *Choco Ice* $\vee$ *Choco Pieces*. This leads to the following problem description.

Problem 15.1 (Navigation Problem). *For a formal context $\mathbb{K}$, a scale-measure $(\sigma, \mathbb{S}) \in \mathfrak{S}(\mathbb{K})$, compute an equivalent scale-measure $(\psi, \mathbb{T}) \in \mathfrak{S}(\mathbb{K})$ such that $G_{\mathbb{T}} = G, \psi = \iota_G$ and $M_{\mathbb{T}} \subseteq F[M, \{\vee, \wedge, \neg\}]$ where $(g, \varphi) \in I_{\mathbb{T}}$ iff $\{g\}^I \models \varphi$.*

Example logic base explanations

With this problem we are interested in computing for an $\mathbb{S}$-measure σ, which may be in canonical representation, an equivalent scale-measure with interpretable attributes that are formed as logical expressions on M. For example, we can express the *Choco* taste attribute of the example scale-measure in Figure 7.10 as the disjunction of the ingredients *Choco Ice* and *Choco Pieces*, i.e. *Choco* := *Choco Ice* $\vee$ *Choco Pieces*. For any scale-measure $(\sigma, \mathbb{S})$, such an equivalent scale-measure, as searched for in Problem 15.1, is not necessarily unique, and the problem statement does not favor any of the possible solutions. The decision which logical expressions are meaningful depends on the contexts or the analyst [205].

Interpreting logical attributes

To understand the semantics of the logical operations in terms of a context $\mathbb{K}$, we first investigate their contextual derivations.

Lemma 9 (Logical Derivations). *Let $\mathbb{K} = (G, M, I)$ be a context, $\varphi_\wedge \in F[M, \{\wedge\}]$, $\varphi_\vee \in F[M, \{\vee,\}]$, $\varphi_\neg \in F[M, \{\neg\}]$, with scale contexts $(G, \{\varphi\}, I_\varphi)$ having the incidence $(g, \varphi) \in I_\varphi \iff \{g\}^I \models \varphi$ for $\varphi \in \{\varphi_\vee, \varphi_\wedge, \varphi_\neg\}$. Then we find*

i) $\{\varphi_\wedge\}^{I_{\varphi_\wedge}} = \mathrm{var}(\varphi_\wedge)^I$,

ii) $\{\varphi_\vee\}^{I_{\varphi_\vee}} = \bigcup_{m \in \mathrm{var}(\varphi_\vee)} \{m\}^I$,

iii) $\{\varphi_\neg\}^{I_{\varphi_\neg}} = G \setminus \{n\}^I$ *with* $\varphi_\neg = \neg n$ *for* $n \in M$.

$\mathbb{K}$	A	B	C
1	×		
2		×	
3			×

	$A \vee B$
1	×
2	×
3	

	$\neg A$
1	
2	×
3	×

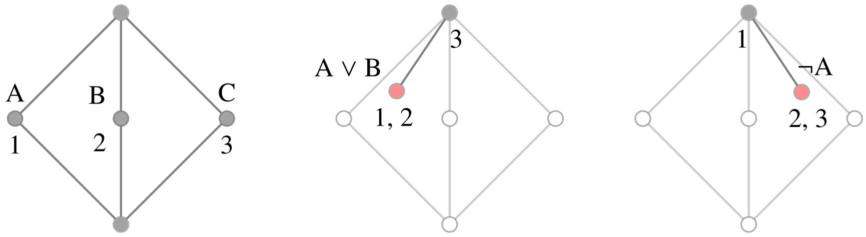


Figure 8.3: Counter examples for which ι_G is not a $(G, \{\varphi_\vee\}, I_{\varphi_\vee})$- or $(G, \{\varphi_\neg\}, I_{\varphi_\neg})$-measure of a $\mathbb{K}$. The conflicting extents are marked in red.

Proof. i) For $g \in G$ if $(g, \varphi_\wedge) \in I_{\varphi_\wedge}$, then $\{g\}^I \models \varphi_\wedge$ and thereby $\text{var}(\varphi_\wedge) \subseteq \{g\}^I$. Hence $g \in \text{var}(\varphi_\wedge)^I$. In case $(g, \varphi_\wedge) \notin I_{\varphi_\wedge}$ it holds that $\text{var}(\varphi_\wedge) \nsubseteq \{g\}^I$ and thereby $g \notin \text{var}(\varphi_\wedge)^I$. ii) For $g \in G$ if $(g, \varphi_\vee) \in I_{\varphi_\vee}$ we have $\{g\}^I \models \varphi_\vee$. Hence, $\exists m \in \text{var}(\varphi_\vee)$ with $g \in m^I$ and therefore g is in the union. If $(g, \varphi_\vee) \notin I_{\varphi_\vee}$ there does not exists such a $m \in \text{var}(\varphi_\wedge)$ and $g \notin \bigcup_{m \in \text{var}(\varphi)} m^I$. iii) For any $n \in M$ we have $\varphi_\neg = \neg n$. Hence, for $g \in G$ if $(g, \varphi_\neg) \in I_{\varphi_\neg}$ we find $g \notin \{n\}^I$. Conversely, if $(g, \varphi_\neg) \notin I_{\varphi_\neg}$ it follows that $g \in \{n\}^I$. □

Naturally, the results from the lemma above generalizes to scale contexts with more than one logical expression in the set of attributes. How this is done is demonstrated in Section 8.2.2. Moreover, more complex formulas, i.e., $\varphi \in F[M, \{\vee, \wedge, \neg\}]$, can be recursively deconstructed and then treated with Lemma 9.

Which logical sentences can be used

To decide if a logical expression yields a scale-measure we can use the result in Proposition 7.

Proposition 21 (Logical Scale-Measure). *Let $\mathbb{K}$ be a context and let $\varphi \in F[M, \{\vee, \wedge, \neg\}]$, then ι_G is a $(G, \{\varphi\}, I_\varphi)$-measure of $\mathbb{K}$ iff $\{\varphi\}^{I_\varphi} \in \text{Ext}(\mathbb{K})$.*

Proof. Since $|\{\varphi\}| = 1$ we find that $(G, \{\varphi\}, I_\varphi)$ has at least one and at most two possible extents, $\{\{\varphi\}^{I_\varphi}, G\}$. If the map ι_G is a scale-measure of $\mathbb{K}$, then $\iota_G^{-1}(\{\varphi\}^{I_\varphi}) = \{\varphi\}^{I_\varphi} \in \text{Ext}(\mathbb{K})$. Conversely, if $\{\varphi\}^{I_\varphi} \in \text{Ext}(\mathbb{K})$ so is $\iota_G^{-1}(\{\varphi\}^{I_\varphi})$, hence, ι_G is a $(G, \{\varphi\}, I_\varphi)$-measure of $\mathbb{K}$. □

Counter examples

This result raises the question for which formulas φ is ι_G a $(G, \{\varphi\}, I_\varphi)$-measure of $\mathbb{K}$ and for which it is not. Counter examples for which ι_G is not a $(G, \{\varphi_\vee\}, I_{\varphi_\vee})$- or $(G, \{\varphi_\neg\}, I_{\varphi_\neg})$-measure of a $\mathbb{K}$ are depicted in Figure 8.3.

Conjunctive navigation

Corollary 5 (Conjunctive Logical Scale-Measures). *Let $\mathbb{K} = (G, M, I)$ be a formal context and $\varphi_\wedge \in F[M, \{\wedge\}]$, then $(\iota_G, (G, \{\varphi_\wedge\}, I_{\varphi_\wedge})) \in \mathfrak{S}(\mathbb{K})$.*

Proof. According to Lemma 9 $(\varphi_\wedge)^{I_{\varphi_\wedge}} = \text{var}(\varphi)^I$, hence, by Proposition 21 we know that $\big(\iota_G, (G, \{\varphi_\wedge\}, I_{\varphi_\wedge})\big) \in \mathfrak{S}(\mathbb{K})$. □

Combining our results on the scale-measure apposition (Proposition 20) with the logical attributes (Proposition 21) we now tackle the navigation problem as stated in Problem 15.1.

Logical navigation without solutions

When we look at this problem again, we find that in its generality it does not always permit a solution. For example, consider the contranominal formal context $\mathbb{B}_n := ([n], [n], \neq)$. This context allows a scale-measure into the nominal scale $\mathbb{N}_n := ([n], [n], =)$, namely the map $\iota_{[n]}$. Restricted to any disjunctive combination of attributes, i.e., $M_{\mathbb{T}} \subseteq F[M, \{\vee\}]$, and $n \geq 3$ does the afore mentioned scale-measure not have an equivalent logical scale-measure. This is due to the fact that (1) in nominal contexts there exists for every object g an attribute m such that $\{m\}' = \{g\}$ and $|\{m\}'| = 1$, (2) all attribute derivations in a contranominal context $\mathbb{B}_n$ are of cardinality $n-1$, (3) the derivation of a disjunctive formula (over $[n]$) is the union of the elemental attribute derivations (Lemma 9). Hence, the derivation of a disjunctive formula is at least of cardinality $n-1$ in $\mathbb{T}$ and therefore there must not exist an $m \in M_{\mathbb{T}}$ such that $|\{m\}^{I_{\mathbb{T}}}| = 1$, and therefore $\text{Ext}(\mathbb{N}) \neq \text{Ext}(\mathbb{T})$.

Conjunctive explanation of scale-measures

Despite this result, we may also report positive answers for particular instances of Problem 15.1 that use conjunctive formulas for $M_{\mathbb{T}}$.

Proposition 22 (Conjunctive Normalform of Scale-Measures). *Let* $\mathbb{K}$ *be a context and* $(\sigma, \mathbb{S}) \in \mathfrak{S}(\mathbb{K})$. *Then the scale-measure* $(\psi, \mathbb{T}) \in \mathfrak{S}(\mathbb{K})$ *given by*

$$\psi = \iota_G \quad \textit{and} \quad \mathbb{T} = |_{A \in \sigma^{-1}(\text{Ext}(\mathbb{S}))} (G, \{\varphi = \wedge A^I\}, I_\varphi)$$

is equivalent to $(\sigma, \mathbb{S})$ *and is called* ***conjunctive normalform*** *of* $(\sigma, \mathbb{S})$.

Proof. We know that every formal context $(G, \{\varphi = \wedge A^I\}, I_\varphi)$ together with ι_G is a scale-measure (Corollary 5). Moreover, every apposition of scale-measures (for some formal context $\mathbb{K}$) is again a scale-measure (Proposition 20). Hence, the resulting pair $(\psi, \mathbb{T})$ is a scale-measure of $\mathbb{K}$.

It remains to be shown that $\sigma^{-1}(\text{Ext}(\mathbb{S})) = \iota_G(\text{Ext}(\mathbb{T}))$. Scale-measure equivalence holds if $(\psi, \mathbb{T})$ reflects the same set of extents in $\text{Ext}(\mathbb{K})$ as $(\sigma, \mathbb{S})$, thus if each context $(G, \{\varphi = \wedge A^I\}, I_\varphi)$ has the extent set $\{G, (\wedge A^I)^{I_\varphi}\}$. In this set we find that $(\wedge A^I)^{I_\varphi} = A$ by Lemma 9. Due to the apposition property the resulting context has the intersections of all subsets of $\sigma^{-1}(\text{Ext}(\mathbb{S}))$ as extents. This set is closed under intersection. Therefor, $\sigma^{-1}(\text{Ext}(\mathbb{S})) = \iota_G(\text{Ext}(\mathbb{T}))$. □

Conjunctive explanation of views

We want to point out that the construction of the conjunctive normalform is very similar to the construction of views (cf. Definition 50) in which we implicitly used attribute sets N such that each $n \in N$ encodes a conjunction of the attributes A_n.

The conjunctive normalform $(\psi, \mathbb{T})$ of a scale-measure $(\sigma, \mathbb{S})$ may constitute a more human-accessible representation of a conceptual data reduction output. We demonstrate this using a practical example in Section 8.5.

8.4 Recommending Conceptual Scale-Measures

On structure based recommendations

Our theoretical findings unveil several possibilities to recommend scale-measures. First, there are meet- and join-irreducible elements of the scale-hierarchy (Propositions 13 and 14). These elements are a minimum representation from which every other scale-measure can be retrieved. However, the number of meet- and join-irreducible elements is in the size of the concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$ (Proposition 13) and thereby potentially exponential large. Hence, it is necessary to narrow down the set of join-irreducible scale-measures, for example, by

Algorithm 1: Scale-measure Exploration: A modified Exploration with Background Knowledge

```
Input   : Context 𝕂 = (G, M, I)
Output : (ι_G, 𝕊) ∈ 𝔖(𝕂) and optionally ℒ ≅ Th_G(𝕊)
Init Scale 𝕊 = (G, ∅, ∈)
Init A = ∅, ℒ = ℭ(𝕂) = CanonicalBase(𝕂) (or ℒ = {} for larger contexts)
while A ≠ G do
    while A ≠ A^{I_𝕊 I_𝕊} (or A^{I_𝕂 I_𝕂} = A^{I_𝕊 I_𝕊}) do
        if Further differentiate objects having A^{I_𝕊 I_𝕊 I_𝕂} by attributes in A^{I_𝕂} \ A^{I_𝕊 I_𝕊 I_𝕂} ?
         then
            ℒ = ℒ ∪ {A → A^{I_𝕊 I_𝕊}}
            Exit While
        else
            Enter B ⊆ A^{I_𝕂} \ A^{I_𝕊 I_𝕊 I_𝕂} that should be considered
            Add (A^{I_𝕊 I_𝕊 I_𝕂} ∪ B)^{I_𝕂} to M_𝕊
    (or ℒ = ℒ ∪ {A → A^{I_𝕂 I_𝕂}})
    A = Next_Closure(A, G, ℒ)
Th_G(𝕊) ≅ ℒ
return : (ι_G, 𝕊) and optionally ℒ
```

constraining the selection to irreducible elements in $\underline{\mathfrak{B}}(\mathbb{K})$ or by applying some conceptual importance measure.

Other scale-measures of interest can be depicted based on their structural placement in the scale-hierarchy, i.e., element-wise modularity, distributivity, or neutrality. A further advantage of the latter two selection criteria is that they allow a decomposition of the scale-hierarchy using divide-and-conquer strategies. The existence of such neutral elements, however, cannot be guarantied.

8.4.1 Exploration based Recommendation

Towards semi-automatic navigation

For the task of efficiently determining a scale-measure, based on human preferences, we propose the following approach. Motivated by the representation of meet-irreducible elements in the scale-hierarchy through object implications of the context (see Proposition 14), we employ the dual of the *attribute exploration* algorithm [75] by Ganter. We modified said algorithm toward exploring scale-measures and present its pseudo-code in Algorithm 1. In this depiction we highlighted our modifications with respect to the original exploration algorithm [78, Algorithm 19] with darker print. This algorithm semi-automatically computes a scale context $\mathbb{S}$ and an implicational base. In each iteration of the inner loop of our exploring algorithm the query that is stated to the *scaling expert* is if an object implication $A \implies B$ is true in the closure system of user preferences. If the implication holds, it is added to the implicational base of $\mathbb{S}$ and the algorithm continues with the next implication query. Otherwise a counter example in the form of a closed set $C \in \text{Ext}(\mathbb{K})$ with $A \subseteq C$ but $B \nsubseteq C$ has to be constructed. This closed set is then added as attribute to the scale context $\mathbb{S}$ with the incidence given by $\in$. If $C \notin \text{Ext}(\mathbb{K})$ the scale $\mathbb{S}$ would contradict the scale-measure property (Proposition 7).

(Object) exploration based scaling

Less queries

The object implicational theory $\text{Th}_G(\mathbb{S})$ is initialized to the object canonical base of $\mathbb{K}$, which is an instance of the object exploration with background knowledge [75]. This initialization can be neglected for larger contexts, however it may reduce the number of queries. The algorithm terminates when the implication premise of the query is equal to G. The returned scale-measure is in canonical form, i.e., the canonical representation $(\iota_G, (G, \text{Ext}(\mathbb{S}), \in))$ (cf. Proposition 10). The motivation behind attribute exploration queries is to determine if an implication holds in the unknown representational context of the learning domain. In contrast, the exploration of scale-measures determines if a given $\text{Ext}(\mathbb{K})$ can be coarsened by implications $A \implies B$, resulting in a smaller and thus more human comprehensible concept lattice $\underline{\mathfrak{B}}(\mathbb{S})$, adjusted to the preferences (or view) of the scaling expert.

On the used queries

Querying object implications may be less intuitive compared to attribute implications, hence, we suggest to rather not test for $A \implies A^{I_\mathbb{S} I_\mathbb{S}}$ for $A \subseteq G$ but to test if the difference of the intents $A^{I_\mathbb{K}}$ and $(A^{I_\mathbb{S} I_\mathbb{S}})^{I_\mathbb{K}}$ in $\mathbb{K}$, is of relevance to the scaling expert. In doing so, only extents of $\mathbb{K}$, i.e., $C = (B \cup A^{I_\mathbb{S} I_\mathbb{S} I_\mathbb{K}})^{I_\mathbb{K}} \in \text{Ext}(\mathbb{K})$ are inserted preserves thereby the scale-measure property. Finally, as a post-processing, one may apply the *conjunctive normalform* (cf. Proposition 22) of scale-measures to further increase the human-comprehension.

More interpretable queries

The implications in $\mathcal{L}_\mathbb{K}$ can be used to explain and reproduce the data reduction by applying $\mathcal{L}_\mathbb{K}$ to the closure system $\text{Ext}(\mathbb{K})$.

Implications as scaling witnesses

8.4.2 Concept Importances based Recommendation

In addition to the semi-automatic exploration of the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$, we outline an automatic procedure using important concepts $(A, B) \in \underline{\mathfrak{B}}(\mathbb{K})$. In Formal Concept Analysis, there have been numerous importance measures introduced [132] that can be used to select a set of concepts $\mathcal{A}$. Some of the more well known measures are the *support*, *concept probability* [122], *concept separation-index* [122], *concept robustness* [213] and *concepts stability* [131]. For the selection $\mathcal{A}$ the canonical construction (Lemma 5) enables us to compute the smallest scale-measure that reflects these concepts. Combined with the conjunctive normalform (Proposition 22) we result in a human interpretable representation.

On importance based recommendation

The scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$ is lattice ordered which allows us to apply the above mentioned importance measures to identify important scale-measures using the construction in Theorem 2. Which importance measure is the most preferable by users and the problem of efficiently computing the importance in $\underline{\mathfrak{S}}(\mathbb{K})$ are open problems.

View of important concepts

8.5 Small Case Study

Accompanying our theoretical findings and to demonstrate the expressiveness of the scale-measure navigation, we provide an example analysis using the *Spices Planner* [147] (see also [90]) data set. This data set is comprised of 56 dishes, here understood as objects, and 37 spices, which are considered to be attributes for the dishes. The incidence $I_{\mathbb{K}_{\text{Spices}}}$ indicates that a spice m is necessary to cook a dish g. The resulting context, in the following denoted by $\mathbb{K}_{\text{Spices}}$, is comprised of 357 incidences, which corresponds to a density of 0.228. Additionally, all dishes in the data set bear exactly one of nine categories, such as vegetable, meat, or fish. The resulting concept lattice of $\mathbb{K}_{\text{Spices}}$ has 532 concepts and is therefore too large for a meaningful diagrammatic representation, and thus human comprehension. Consequently, a data reduction that correctly reflects the data is necessary. Using scale-measures, we are able to generate small-scaled views of readable size. In order to discover *"interesting"* scale-measures, we need to find a starting point for navigating the lattice of scale-measures. A natural approach for this is selecting a subset of extents

Spices data set

Spices data set is to large for human interpretation

Spices for vegetable dishes

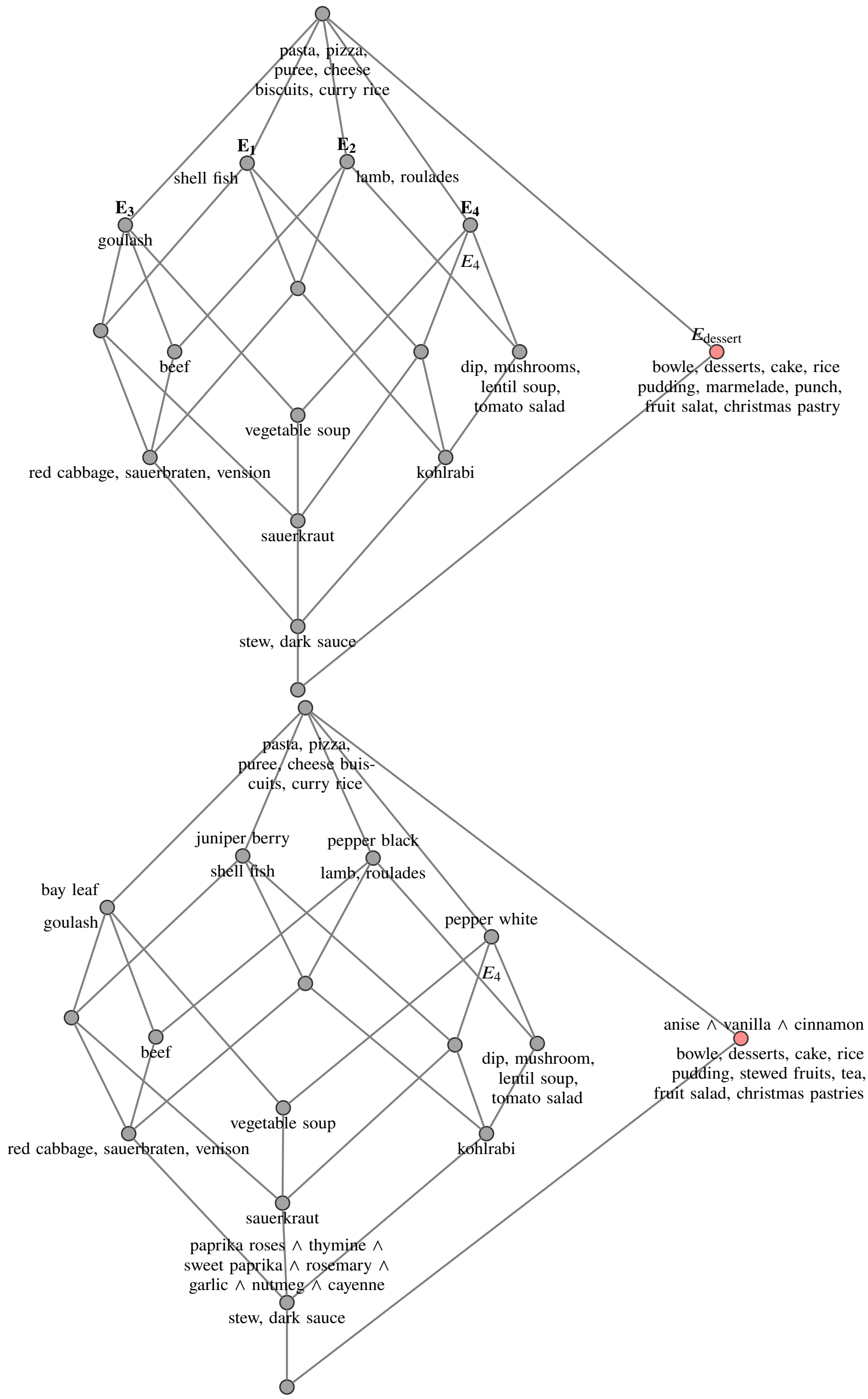


Figure 8.4: The lattices diagrams are views of the spices data set as computed in Section 8.5. The concept lattices reflect a focused view on vegetable dishes that use similar spices (gray) and one selected dessert dish (red). The top view is in canonical representation whereas the bottom view is in conjunctive normalform and thus has speaking attributes.

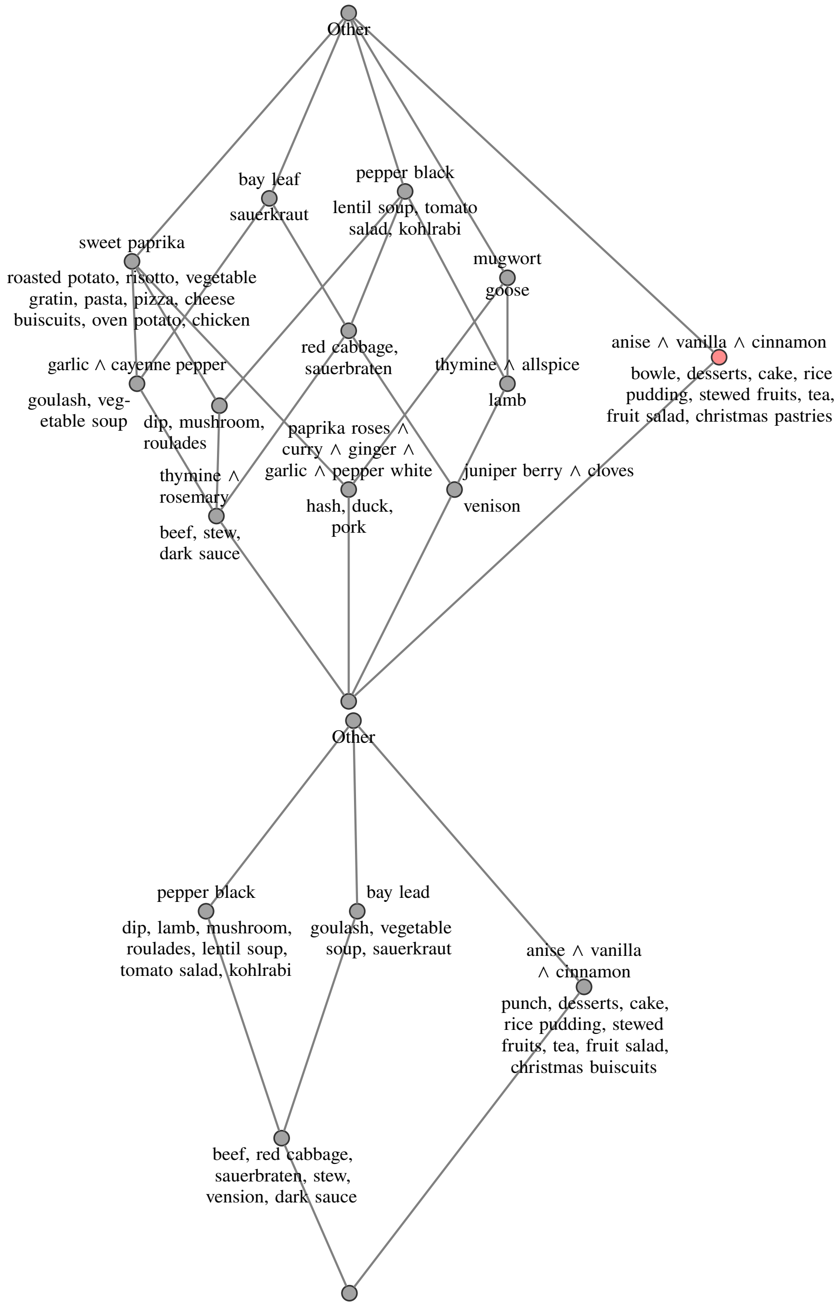


Figure 8.5: The two presented concept lattices are views of the spices data set. The top view is analogously built to the views in Figure 8.4 and reflect similar spiced meat dishes (gray) and a selected desserts (red). The view at the bottom is the greater common of the meat view (top in this figure) and the vegetable view (bot in Figure 8.4) using the meet (Proposition 19) operator of the scale-hierarchy.

of $\mathbb{K}_{\text{Spices}}$ that are considered relevant by the data analyst. In our first example setting, we construct a scale-measure on vegetable dishes:

$$G_V = \{\text{carrots, red cabbage, green salad, spinach, vegetable gratin, broccoli, cauliflower, lentil soup, vegetable soup, cucumber salad, stew, beans, sauerkraut, tomato salad, kohlrabi}\}.$$

Concepts on vegetable dishes

Of course it suffices to restrict the corresponding extents by the set of meet-irreducible concepts of $\mathbb{K}_{\text{Spices}}$, i.e., meet-irreducibles in $\underline{\mathfrak{B}}(G_{\text{V}}, M_{\text{Spices}}, I_{\text{Spices}} \cap (G_{\text{V}} \times M_{\text{Spices}}))$. To fulfill the condition of Proposition 10 we compute their object-closures in $\mathbb{K}_{\text{Spices}}$ which yields the set $\mathcal{E}_{\text{V}} = \{E_1, E_2, E_3, E_4\}$ with:

$$E_1 = \{\text{red cabbage, sauerbraten, shellfish, stew, venison, dark souce, sauerkraut, kohlrabi}\},$$

$$E_2 = \{\text{dip, beef, red cabbage, lamb, sauerbraten, mushrooms, roulades, lentil soup, stew, vension, dark souce, tomato salad, kohlrabi}\},$$

$$E_3 = \{\text{beef, goulash, red cabbage, sauerbraten, vegetable soup, stew, vensio, dark souce, sauerkraut}\},$$

$$E_4 = \{\text{steamed fish, dip, roast potato, carrots, roasted fish, risotto, green salad, omelette, white sauce, potato gratin, spinach, vegetable gratin, hash, broccoli, grilled fish, mushrooms, potato soup, asian rise, baked fish, oven potato, cauliflower, duck, lentil soup, vegetable soup, cucumber, salad, veal, goose, stew, beans, pork, dark souce, sauerkraut, chicken, tomato salad, kohlrabi}\}.$$

View on vegetable dishes

View applications

Interpretable vegetable view

By applying Proposition 10 we can compute a scale-measure in canonical form reflecting $\mathcal{E}_{\text{V}}$, i.e., $(G_{\text{Spices}}, \mathcal{E}_{\text{V}}, \in)$. The concept lattice of this view consists of 15 concepts and is depicted on in Figure 8.4 (top) and is indicated by the gray colored concepts. This lattice is more readable due to a fewer number of concepts. This conceptual view could, for example, be useful to compose, curate or refine instructions for a vegetable kitchen. The features of this view are the chosen extents, which are, due to the constructive nature of this process, in canonical representation. For more meaningful attributes, with respect to a human reader of the diagram, we can employ the conjunctive normalform Corollary 5. The thus altered, but equivalent, view is depicted on the bottom in Figure 8.4.

Add desserts

Starting with this data reduction, we can further navigate in the scale-hierarchy (Definition 38) to a finer or coarser view (Definition 37). In an envisioned setting of restaurants it might be suitable to add a dessert extent. We selected the dessert extent using the meet-irreducibles in $\mathbb{K}_{\text{Spices}}$ restricted to dessert objects.

$$\mathcal{E}_{\text{dessert}} = \{\text{punch,desserts,cake,rice pudding,marmelade,tea,fruit salad ,christmas buiscuits}\}$$

We can add $\mathcal{E}_{\text{dessert}}$ using the context apposition (Proposition 20), i.e., the join operation in the scale-hierarchy (Proposition 12). The result is highlighted in red in the respective diagrams in Figure 8.4.

Second view on meat dishes

In the next step, we demonstrate how to browse within the scale-hierarchy using a second conceptual view. For this, we analogously build a scale-measure on meat dishes. The concept lattice diagram of this view is depicted Figure 8.5 (top). Using the meet operator

(Proposition 12) of the scale-hierarchy we can compute the greater common of conceptual views concerning meat and vegetables. The conjunctive normalform of their meet is depicted on the bottom in Figure 8.5. Such a scale-measure might be useful to find meat dishes and vegetable dishes that go well together or to build a menu that reflects a compromise between the meat and vegetable menu, with respect to spices.

Commonalities of meat and vegetable dishes

Given our theoretical findings, all operations and transformations used above are in fact scale-measures and are thus consistent data reductions of $\mathbb{K}_{\text{Spices}}$. Of course, views of interest could be derived using different criteria, apart from meet-irreducibles. Nonetheless, as long as the data reduction process is based on selecting extents, the resulting views will be scale-measures of the original data set.

Notes on consistency

With Figure 8.6, which is a scale-measure of the well-known *Zoo*[1] data set by R. S. Forsyth, obtained from the UCI repository [60], we provide a second example for our scaling theory. We proceeded analogously to the spices example. Initially, we compute a view based on five logical animal taxons T (red in Figure 8.6) to classify the animals of the zoo data set with respect to (G, T, I_T).

Zoo data set

View based on animal classifications

Their join (Proposition 12) can be computed using the apposition operator (Proposition 20) within the scale-hierarchy. In addition to that, we added in our depiction six attributes that were used in the taxons T, such that the hierarchical connections among the taxons are reflected. Hence, this view on the data enables the human reader to grasp how the selected animal types can be classified and to identify similarities.

8.5.1 Explaining Existing Data Scalings

One ultimate application that we envision for the introduced scale-measures techniques in the realm of data science is the possibility to extract, and therefore interpret, data reductions that were derived through other methods, such as LSA or Boolean matrix factorization. This can be done by computing for the consistent part of every data scaling an equivalent logical scale-measure (see Proposition 21). Moreover, we may identify, and quantify, parts of a data scaling that are inconsistent with respect to the original data. We elaborate on this in greater detail in Chapter 11. To give the reader a hint for this, we revisit the example views of $\mathbb{K}_{\text{ICE}}$ depicted in Figure 7.10. The conjunctive normalform leads to the following interpretations of the view attributes:

Explaining data reductions

Small example

$$\begin{aligned} M_{\mathbb{T}} = \{&\textbf{Choco} = \texttt{Choco Ice} \vee \texttt{Choco Pieces}, \textbf{Brownie} := \texttt{Brownie},\\ &\textbf{Caramel} = \texttt{Caramel Ice} \vee \texttt{Caramel}, \textbf{Dough} := \texttt{Dough},\\ &\textbf{Peanut} = \texttt{Peanut Ice} \vee \texttt{Peanut Butter}, \textbf{Vanilla} := \texttt{Vanilla}\} \end{aligned}$$

8.5.2 (Semi-)Automatic Large Data Set Scaling

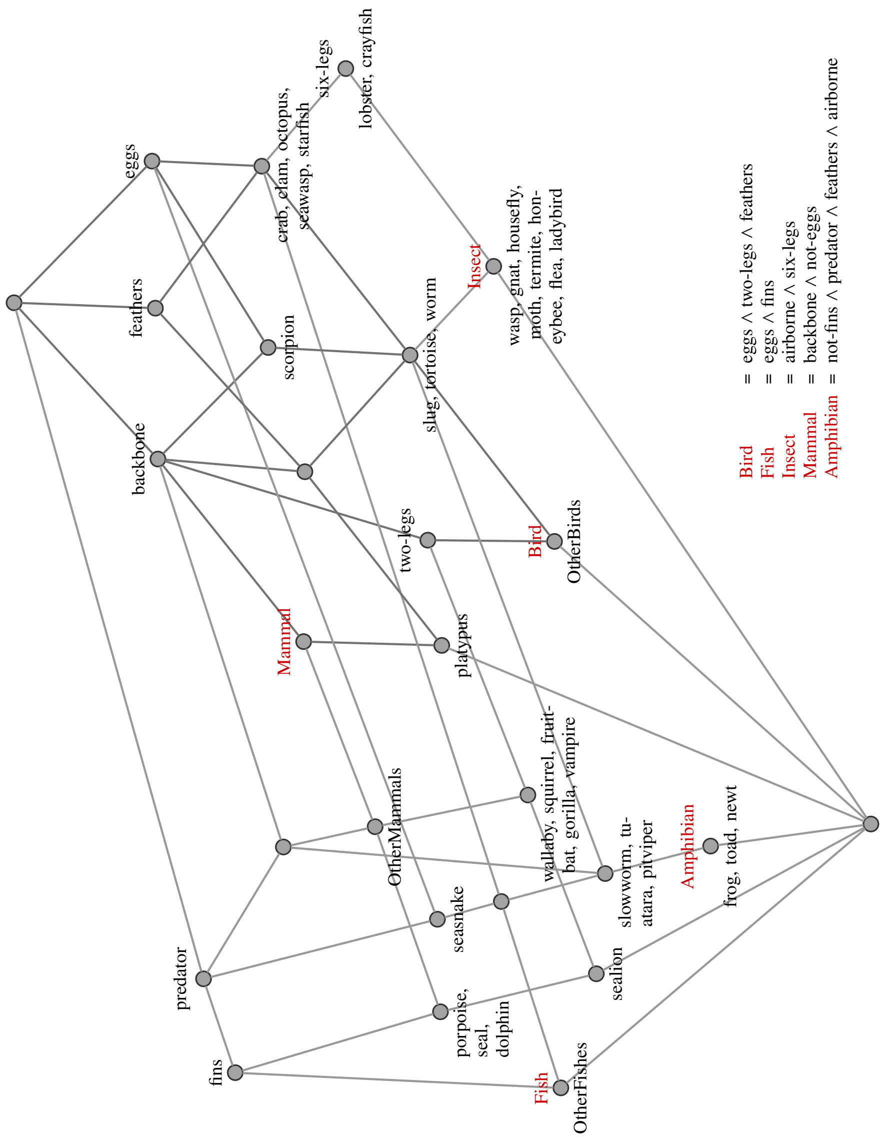
To demonstrate the applicability of the presented exploring algorithm, we have implemented it in the `conexp-clj` [88] software for Formal Concept Analysis. We apply the scale-measure exploration (Algorithm 1) on the *Living Beings and Water* [80] context $\mathbb{K}_W$ (see Figure 8.7). In Figure 8.8 (left) we depicted the evaluation steps of the algorithm. The first two columns represent the object implication that is queried, the third column contains the query translated in terms of attributes. For example, in row two the implication $\{\} \implies \{\text{D}, \text{FL}, \text{Br}, \text{F}\}$ is true in the so far generated contextual view $\mathbb{S}$ and is queried if it should hold. All objects of the implication do have at least the attributes *can move* and *needs water to live*, as

Semi-automatic scaling of the water context

Exploration steps

Scaling and *exploration counter examples*

[1]The objects *girl,frogB* were omitted.



OtherFishes={seahorse, sole, herring, piranha, pike, chub, haddock, stingray, carp, bass, dogfish, catfish, tuna}

OtherMammals={reindeer, aardvark, polecat, wolf, mole, vole, hare, boar, cavy, antelope, goat, puma, mongoose, pony, bear, pussycat, lynx, elephant, calf, mink, opossum, leopard, buffalo, lion, giraffe, cheetah, oryx, deer, hamster, raccoon}

OtherBirds={gull, parakeet, crow, skua, swan, hawk, sparrow, lark, wren, dove, vulture, penguin, duck, flamingo, pheasant, rhea, ostrich, skimmer, chicken, kiwi}

Figure 8.6: Conceptual view on the zoo context with 27 of the original 4579 concepts.

	has limbs (L)	breast feeds (BF)	needs chlorophyll (Ch)	needs water to live (W)	lives on land (LL)	lives in water (LW)	can move (M)	monocotyledon (MC)	dicotyledon (DC)
dog	×	×		×	×		×		
fish leech				×		×	×		
corn			×	×	×			×	
bream	×			×		×	×		
water weeds			×	×		×		×	
bean			×	×	×				×
frog	×			×	×	×	×		
reed			×	×	×	×		×	

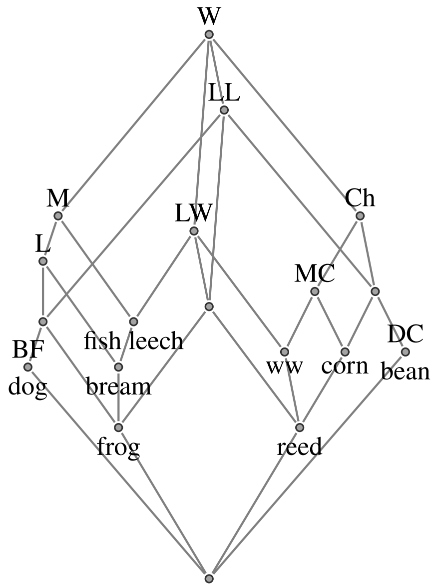


Figure 8.7: This figure shows the *Living Beings and Water* [80] context in the top. Its concept lattice is displayed at the bottom and contains nineteen concepts.

indicated in the third column (left). In the same column (right) we find attributes from $\{1\}^{I_{\mathbb{K}}} \setminus \{2\}^{I_{\mathbb{K}}} \subseteq M_W$ that can be considered by the scaling expert to narrow the object implication, i.e., to shrinken the size of the conclusion. The answer of the scaling expert envisioned by us is given in column four, the attribute *lives on land*. Thus, the object counter example is the attribute-derivation the union $\{M, W, LL\}^{I_W} = \{D, F\}$. In our example of the scale-measure exploration the algorithm terminates after the scaling expert provided nine counter examples and four accepts. The output is a contextual view in canonical representation with twelve concepts as depicted in Figure 8.8 (right).

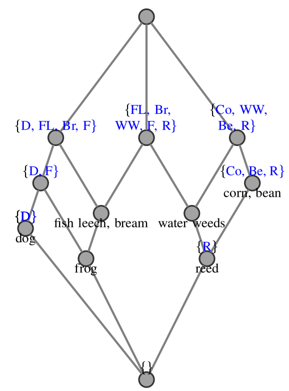
Exploration results

The just demonstrated application of the scale-measure exploration can be supported in every step by conceptual importance measures [132]. Furthermore, these measures can also be used to automate the exploration algorithm by randomly selecting the counterexample from the top-k of the list of outstanding concepts with respect to one or more of said conceptual measures. We illustrate this idea for the spices planer data set $\mathbb{K}_{\text{Spices}}$ and depict

Further automatizations of the exploration

Automatic exploration setup

(Object) Premise (1)	(Object) Conclusion (2)	Attribute Question (2)′		(1)′ \ (2)′	CE?	Edit
{}	{D, FL, Co, Br, WW, Be, F, R}	{W}:	add more detail using	{L, BF, Ch, LL, LW, M, MC, DC}?	{M}	Add {D, FL, Br, F} to $M_{\mathbb{S}}$
	{D, FL, Br, F}	{W, M} :	add more detail using	{L, BF, Ch, LL, LW, MC, DC}?	{LL}	Add {D, F} to $M_{\mathbb{S}}$
	{D, F}	{L, W, LL, M}:	add more detail using	{BF, Ch, LW, MC, DC}?	{BF}	Add {D} to $M_{\mathbb{S}}$
	{D}	{L, BF, W, LL, M}:	add more detail using	{Ch, LW, MC, DC}?	{LW}	Add {} to $M_{\mathbb{S}}$
	{}					
{R}	{D, FL, C, Br, WW, Be, F, R}	{W}:	add more detail using	{Ch, LL, LW, MC}?	{Ch}	Add {C, WW, Be, R} to $M_{\mathbb{S}}$
	{C, WW, Be, R}	{W, Ch}:	add more detail using	{LL, LW, MC}?	{LL}	Add {C, Be, R} to $M_{\mathbb{S}}$
	{C, Be, R}	{W, Ch, LL}:	add more detail using	{LW, MC}?	{LW}	Add {R} to $M_{\mathbb{S}}$
	{R}					
{F}	{F, D}	{L, W, LL, M}:	add more detail using	{LW}?	{LW}	Add {F} to $M_{\mathbb{S}}$
	{F}					
{F, R}	{D, FL, Co, Br, WW, Be, F, R}	{W}:	add more detail using	{LW, LL}?	{LW}	Add {FL, Br, WW, F, R} to $M_{\mathbb{S}}$
	{FL, Br, WW, F, R}	{W}:	add more detail using	{LW, LL}?	no	Add (1) $\Longrightarrow$ (2) to $\mathcal{L}$
{F, Br}	{FL, Br, F}	{W, LW, M}:	add more detail using	{L}?	no	Add (1) $\Longrightarrow$ (2) to $\mathcal{L}$
{FL, Br, F}	{FL, Br, F}					
{WW, R}	{WW, R}					
{FL, Br, WW, F, R}	{FL, Br, WW, F, R}					
{D}	{D}					
{D, F}	{D, F}					
{D, FL, Br, F}	{D, FL, Br, F}					
{Co, R}	{Co, Be, R}	{W, Ch, LL}:	add more detail using	{Mc}?	no	Add (1) $\Longrightarrow$ (2) to $\mathcal{L}$
{Be}	{Co, Be, R}	{W, Ch, LL}:	add more detail using	{Dc}?	no	Add (1) $\Longrightarrow$ (2) to $\mathcal{L}$
{Co, Be, R}	{Co, Be, R}					
{Co, WW, Be, R}	{Co, WW, Be, R}					DONE



	{D, FL, Br, F}	{}	{Co, WW, Be, R}	{Co, Be, R}	{FL, Br, WW, F, R}	{D, F}	{D}	{F}	{R}
dog (D)	×					×	×		
fish leech (FL)	×				×				
corn (Co)			×	×					
bream (Br)	×				×				
water weeds(WW)			×		×				
bean (Be)			×	×					
frog (F)	×				×	×		×	
reed (R)			×	×	×				×

{Be} $\Longrightarrow$ {Co, Be, R},
{F,Br} $\Longrightarrow$ {FL, Br, F},
{Co, R} $\Longrightarrow$ {Co, Be,R},
{F, R} $\Longrightarrow$ {FL, Br, WW, F, R}

Figure 8.8: Scale-measure exploration results (left) for the *Living Beings and Water* context, the resulting context (bottom right) and its concept lattice (top right).
The employed object order is: Be > Co > D > WW > FL > Br > F > R

the resulting scale-measure in Figure 8.9.

Example automatic importance based exploration

For this example of automatic scale-measure exploration, we considered the importance measure *separation index* [122, 132] on the set of objects. We consider the maximum number of concepts that are human readable to be thirty and therefore we restricted the number of counter examples to be computed accordingly. We depicted the concept lattice of the resulting scale-measure in Figure 8.9 using the conjunctive normalform. To improve the readability, we only annotated meet-irreducible attribute concepts in the lattice diagram and omitted redundant attribute conjunctions, e.g., for Anis∧Vanilla∧Cinnamon∧Pastry we annotate …∧Pastry, since Anis∧Vanilla∧Cinnamon is already given by an upper neighbor. The so given scale-measure concept lattice seems empirically more human readable and displays extensive information with respect to the original data set $\mathbb{K}_{\text{Spices}}$ and the employed importance measure. The exploring algorithm outputs 44 implications that can be used to explain the scaling.

8.6 Navigation Methods in Conceptual Scaling

OLAP and logic based views

The navigation methods presented in this chapter combine, slice and aggregate views on the data and are thereby related to *OLAP* [44] but on the conceptual level. The methods themselves are inspired from *relational algebra* [45]. The computation of equivalent logical terms are related to the field of *symbolic regression* [3, 42, 219]. To what extent this field can contribute to solving Problem 15.1 is unexplored.

Navigate scalings

For conceptual scaling (see Figure 7.1, top left) with plain scaling and its extensions (cf. Section 7.1.3) the `TOSCANA` framework [123, 218] provides an extensive tool-set. The `TOSCANA` system provides several ways to define scales, add logical attribute combinations [170, 172] or extend them to encode relations of higher arity [65, 173]. Given a many-valued context $\mathbb{D}$ and a set of defined scales $\mathbb{S}_m$ for each $m \in M_{\mathbb{D}}$, the analyst explores the derived context $\mathbb{K}$ of $\mathbb{D}$ on subsets of the many-valued attributes $N \subseteq M_{\mathbb{D}}$. The concept lattice of the sub-context $\mathbb{K}[G_{\mathbb{D}}, \bigcup_{m\in N} M_{\mathbb{S}_m}]$ is investigated through nested drawing. To navigate between different scalings, the analyst can define new scaling, logical attributes, or constrain the use of a scale based on other attribute values of other attributes [206]. The navigation procedure requires both, an analyst and a domain expert (cf. user and preparator [193]).

Navigate conceptual data reductions

With this chapter, we present, a navigation paradigm for conceptual data reduction (see Figure 7.1, top right) which takes place after the context $\mathbb{K}$ is derived from $\mathbb{D}$. In contrast to conceptual scaling, the reduction process does not require additional knowledge of a domain expert. The notion of scale-measures enables us to verify if a reduction is consistent to the interpretation of the data by the expert. Although, the analyst is able to study $\mathbb{K}$ and its views alone, the domain expert needs to be consulted to assess the meaningfulness of the reduction.

Navigation Structure

The results of selecting scales in `TOSCANA`, if defined with respect to plain scaling, can be considered as views and navigation between views as given by Corollary 4. However, this allows only for a very limited exploration of the scale-hierarchy. Moreover, does the characterization of the hierarchy of views enable us to identify outstanding data reductions (cf. Section 8.4), identify missing information in data reductions (cf. Lemma 7), the introduction of conceptual scaling error (cf. Chapter 11), as well as, a (semi-)automatic conceptual data reduction procedure (cf. Section 8.5.2).

On top of that, the navigation in the scale-hierarchy is independent from the scales used to derive the context $\mathbb{K}$. This allows us, for example, to express that the visual color spectrum is in view relation to an (inter)ordinal interpretation of the wavelength feature (see Figure 7.12).

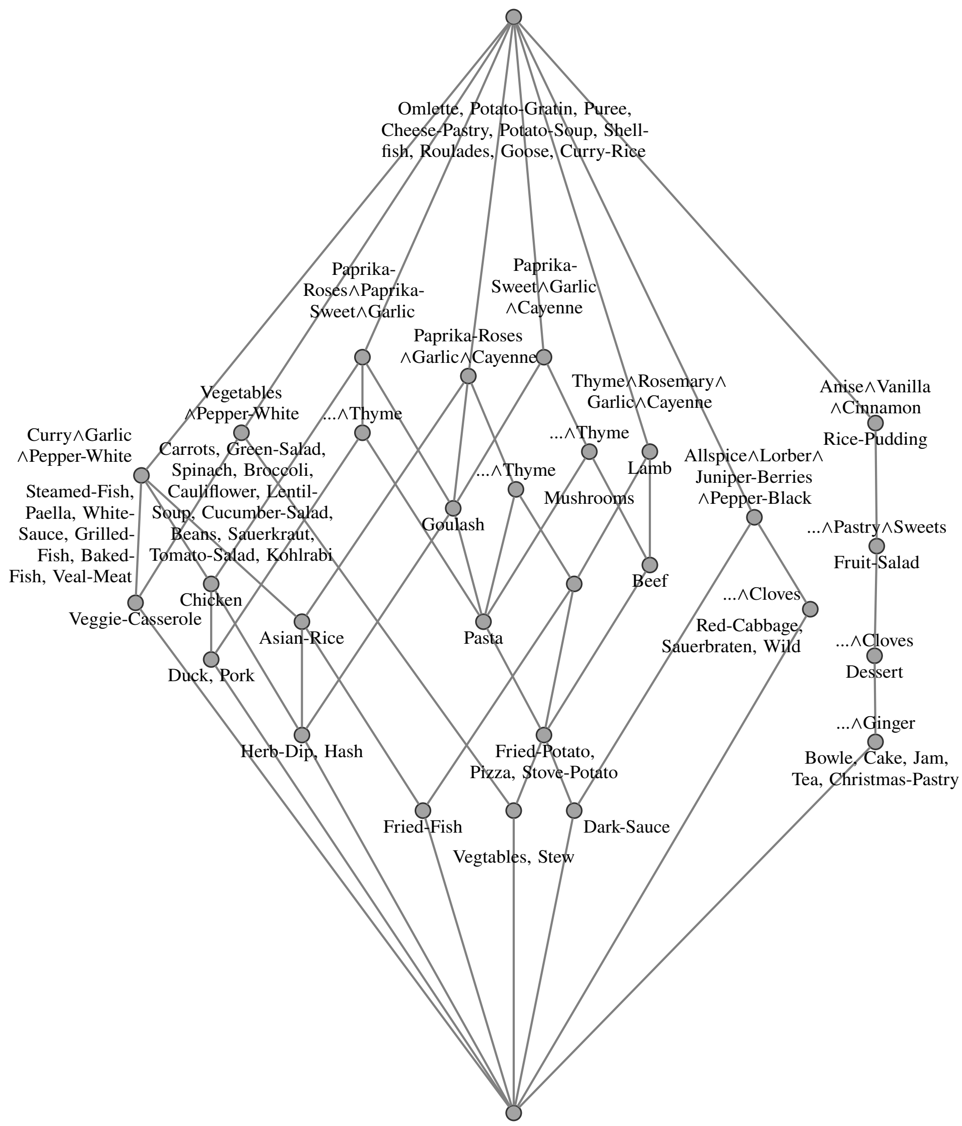


Figure 8.9: Automatically generated scale-measure of the spices context using the most outstanding concepts by the separation index importance measure. The scale consists of 30 of the original 523 concepts and is in conjunctive normalform.

Nonetheless, we believe that a combination of TOSCANA and a navigation of the scale-hierarchy can be beneficial. An example application scenario could be that a domain and scaling expert defines scales to interpret the data and provides the derived context $\mathbb{K}$. In a second step, users that are interested in $\mathbb{K}$ can use the scale-hierarchy to explore different abstractions in a self-determined manner. The methods presented in this chapter provide the guaranty that users remain consistent to $\mathbb{K}$ with respect to scale-measures. This combination enables an end-to-end scaling and reduction framework for conceptual data scaling (Figure 7.1, upper part). Combination

Another hierarchy proposed in the realm of conceptual scaling is the *hierarchy of scales* [204]. For a given conceptual scale $\mathbb{S}_m$ the hierarchy includes all scales that generate finer closure systems with respect to scaling. This hierarchy is used to add information to conceptual scales for plain scaling, e.g., from taxonomy or other background knowledge. This process is opposite to computing coarser closure systems in conceptual data reduction. Hierarchy of scales

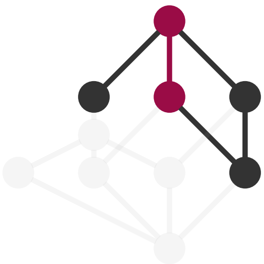
8.7 Discussion

The methods that were presented in this chapter are very useful to navigate between conceptual views, to refine data reductions and to develop recommendation systems. For the later we have shown how user preferences can be used for (semi-)automatic recommendations. For automatic recommendations without prior knowledge on user preferences, further studies are needed. A thorough investigation on this, including a user study, is outside the scope of this work and deemed future work. Limitations on recommendation

Following on from this, computations in the scale-hierarchy are costly due to the incomprehensible size. It is not clear if the importance measures from FCA can be efficiently computed for scale-measures in the scale-hierarchy. In addition to that we can envision that methods like TITANIC [208] or the outlined decomposition at the end of Section 8.1 can improve the efficiency of navigating conceptual views. Further research is needed on this topic to increase the capabilities of recommending views. Computational aspects of the scale-hierarchy

With the computation of the conjunction normalform we provided a first solution to Problem 15.1 using comprehensible features. This approach is very expressive, since it is agnostic to the data reduction method which makes it applicable to many methods in the realm of machine learning. In Chapter 11 we demonstrate this by explaining results from Boolean matrix factorization. This method can be improved in future work by deriving shorter or more comprehensible solutions. A promising line of research that may help here is the field of *symbolic regression* [3, 42, 219]. Other logical representations

Beyond the scope of navigation between conceptual views, we can envision an application of our methods to identify structural similarities within a data set [57, 58]. Consider a formal context $\mathbb{K} := (G, M, I)$ with $A, H \subseteq G$ and a full scale-measure σ from $\mathbb{K}[H, M]$ to $\mathbb{K}[A, M]$. We may interpret for $\sigma(h) = a$ that h relates to H like a to A. Structural similarities



9

Ordinal Motifs in Lattices

Identify sub-structures

A fundamental principle of the formal analysis of data is the identification of unique and meaningful sub-structures and properties. The realm of ordinal, lattice and conceptual structures is no exemption to that. With this chapter, we provide theoretical foundations for analyzing concept lattices by means of ordinal sub-structures. The resulting framework is based on methods from conceptual data reduction and inverse data scaling (cf. Figure 7.1). These allow us to study ordinal sub-structures of concept lattices independently from the elements that generate them, i.e., attributes and incidences in a context $\mathbb{K}$. On top of that, the approach generalizes to any type of ordinal sub-structure due to the capabilities of scale-measures.

Ordinal motifs

We call this approach, in analogy to the notion established in network science [104, 105, 156], *ordinal motifs*. Another analogy can be drawn from feature engineering for graph data from *geometric deep learning* [32]. They represent a graph based on the number of homomorphisms into smaller graphs [28]. The resulting sequence is called *homomorphism numbers* and can be used in deep learning [164]. However, in contrast to network science and deep learning on graphs, where motifs are recurrent and statistically significant sub-graphs (or patterns), we understand motifs as a user-defined set $\mathcal{O}$ of ordered sets, usually represented as formal contexts [80]. The elements of this set can be of different sizes and (ordinal) complexities. They shall allow us to analyze any lattice or conceptual structure, by means of frequency and sizes of ordinal patterns. Thus, the set $\mathcal{O}$ can be considered as an *ordinal tool-set*. In addition to the standard scales mentioned in Section 7.1, any pattern deemed relevant by a user lends itself to be in $\mathcal{O}$. However, we show in our work that already for standard scales the recognition of these motifs is a computationally difficult problem.

Theoretical contributions

In terms of theoretical results, we show the computational complexity of several decision problems for recognizing and finding scale-measures. In particular, we show that for finding a scale-measures for a given ordinal motif we have to solve an NP-complete problem. Moreover, we show that motifs which have the special property of belonging to a hereditary class of scales offer many advantages in computation.

Textual explanations

An advantage of employing sets of standard scales is their well-known structural semantic (cf. *basic meaning* [80, Figure v1.26]). Based on these, we constructed textual templates for every standard scale based on principles from human computer interaction. In detail, we applied the five goodness criteria [149] for explainability in machine learning to ensure that the textual templates are human comprehensible.

Automatic textual explanations

Combining the recognition of ordinal motifs and the textual explanations of them, we yield an automatic procedure to generate textual explanations of concept lattices. To the best of our knowledge, this is the first method that is able to do this. While this approach is very expressive there may be exponentially many ordinal motifs. Therefore, we introduce an importance measure of ordinal motifs based on the proportion of the conceptual structure that they reflect. Based on this, our method can identify few ordinal motifs that cover most of the concept lattice.

Application

To demonstrate the applicability of the ordinal motif method, we demonstrate our findings based on standard scales in a medium-sized data set, the spice planner data set [147].

9.1 Ordinal Motifs

Goal

The overall goal for ordinal motifs is to identify frequent recurring ordinal patterns of user defined shape that allow for analyzing large and complex ordinal structures. The use of scale-measures $(\sigma, \mathbb{S})$ is not limited to contexts and concept lattices, but can be extended to ordered sets $(P, \leq)$. This is done through the general ordinal scale $(P, P, \leq)$ whose concept lattice is isomorphic to the Dedekind-MacNeille completion of $(P, \leq)$, i.e., the smallest lattice in which $(P, \leq)$ can be order-embedded (cf. Section 5.2).

Choosing motifs

The formal context $\mathbb{S}$ in scale-measures is not restriction on what can be used as a scale context. Any scale and scale-measure in the scale-hierarchy $\underline{\mathfrak{S}}(\mathbb{K})$ can be applied here. An important factor for the choice of $\mathbb{S}$ is its interpretation whith respect to its structure. As long as $\mathbb{S}$ carries structural information that we are interested in, we can use $\mathbb{S}$ as ordinal motif and analyze $\underline{\mathfrak{B}}(\mathbb{K})$ through the lens of scale-measures $(\sigma, \mathbb{S})$.

Ordinal motif evaluation

In doing so, we want to consider the following aspects: **scope** and **coverage**. We will first give an informal explanation of the two properties and then derive the mathematical tools and a precise problem definition.

Scope

Starting from a given context $\mathbb{K} := (G, M, I)$ and an ordinal motif $\mathbb{S}$, the scope of the ordinal motif is

- **global**, if it covers the entire data, i.e., all objects G, or
- **local**, if it covers only parts of G.

Coverage

The coverage of an ordinal motif concerns the portion of the ordinal structure that is captured by the motif. We say an ordinal motif

- has **full coverage**, if every element of the concept lattice of $\mathbb{K}$ has a correspondence in the ordinal structure of the motif, or
- has **partial coverage**, otherwise.

For example, the latter case exists if there are extents of $\mathbb{K}$ that are not the pre-image of an extent of $\mathbb{S}$.

Coverage interpretation

In case there is a full scale-measure from a context $\mathbb{K}$ to a context $\mathbb{S}$, we can infer that the closure system of $\mathbb{K}$ on G is, except for relabeling, identical to that of $\mathbb{S}$ (see Proposition 3). A scale-measure from $\mathbb{K}$ to $\mathbb{S}$, on the other hand, only guarantees that the closure system of $\mathbb{K}$ on G has at least all closed sets that the context $\mathbb{S}$ has, up to relabeling.

Definition 40 (Local Scale-Measures). *Let $\mathbb{K}$ and $\mathbb{S}$ be two formal contexts. The map $\sigma : H \to G_{\mathbb{S}}$ is a* ***local scale-measure*** *of $\mathbb{K}$, iff*

1. *$H \subseteq G_{\mathbb{K}}$ and*
2. *σ is a scale-measure from $\mathbb{K}[H, M_{\mathbb{K}}]$ to $\mathbb{S}$.*

We say a local scale-measure is ***full****, iff σ is a full scale-measure from $\mathbb{K}[H, M_{\mathbb{K}}]$ to $\mathbb{S}$.*

local scale-measures and morphisms

For local and full scale-measures the relation between the respective concept lattices is captured by the following proposition.

Proposition 23 (local and full scale-measure). *For contexts $\mathbb{K}$, $\mathbb{S}$, the closure operator $\mathrm{cl}_{\mathbb{K}}$ on $\mathrm{Ext}(\mathbb{K})$ and a surjective local scale-measure $\sigma : H \to G_{\mathbb{K}}$, we find that*

$$\left(\mathrm{Ext}(\mathbb{S}), \subseteq\right) \leqq \left(\mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}]), \subseteq\right) \cong \left(\mathrm{cl}_{\mathbb{K}}\left(\mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}])\right), \subseteq\right).$$

Proof. The first morphism follows directly from Proposition 3 and the fact that σ is a scale-measure from $\mathbb{K}[H, M_{\mathbb{K}}]$ to $\mathbb{S}$. For the final isomorphism we can note that for $A \in \mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}])$ the difference $\mathrm{cl}_{\mathbb{K}}(A) \setminus A$ is in $G \setminus H$. This means, the closure of A in $\mathrm{Ext}(\mathbb{K})$ adds only elements from $G \setminus H$. Thus, since $\mathrm{cl}_{\mathbb{K}}$ is a closure operator we find that for $A, C \in \mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}])$ with $A \subset C$ we have $\mathrm{cl}_{\mathbb{K}}(A) \subset \mathrm{cl}_{\mathbb{K}}(C)$. Hence, $\mathrm{cl}_{\mathbb{K}} : \mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}]) \to \mathrm{Ext}(\mathbb{K})$ is an injective map and by restricting the co-domain we find a bijective map $\hat{\mathrm{cl}}_{\mathbb{K}} : \mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}]) \to \{\mathrm{cl}_{\mathbb{K}}(E) \mid E \in \mathrm{Ext}(\mathbb{K}[H, M_{\mathbb{K}}])\}$. □

Remark on isomorphisms

Propositions 3 and 23 reveal the relations between a context $\mathbb{K}$ and an ordinal motif $\mathbb{S}$. Analyzing $\mathbb{K}$ via ordinal motifs in the full scale-measure setting would mean to simply speak about $\mathbb{K}$ with different labels. For the local full case we find that scale-measures reflect a sub-closure system, i.e., $\mathrm{Ext}(\mathbb{K})$ restricted to a subset $H \subseteq G_{\mathbb{K}}$.

Ordinal motif problem

The following problem summarizes the technical observations so far and (finally) states all notions for ordinal motif. For the surjective property we refer to Remark 4.

Problem 15.2 (Finding Ordinal Motifs). *Given a formal context $\mathbb{K}$ and an ordinal motif $\mathbb{S}$ find a surjective map from $\mathbb{K}$ into $\mathbb{S}$ that is a:*

	global	local
partial	*scale-measure*	*local scale-measure*
full	*full scale-measure*	*local full scale-measure*

Connection to inverse scaling

The **global** and **full** case of this problem can be seen as a special instance of the inverse scaling of a formal context $\mathbb{K}$ (Proposition 9). With inverse scaling of $\mathbb{K}$ we derive full scale-measures into the semi-product of scale context. For both, the inverse scaling and ordinal motifs we employ standard scales due to their explainability. We discuss this case in greater detail in Chapter 10 together with a new notion of complexity.

9.1.1 Ordinal Motifs in Concept Lattices

Example ordinal motifs

In Figure 9.1 we present local full ordinal motifs for our example in Figure 8.7. The ordinal motifs are highlighted in color and represent three common structures analyzed in Formal Concept Analysis and order theory in general. The top diagram highlights a contranominal scale context, i.e, an ordinal motif encoding a Boolean lattice. The diagram in the middle highlights an ordinal scale context, i.e., a chain order. The bottom diagram highlights a

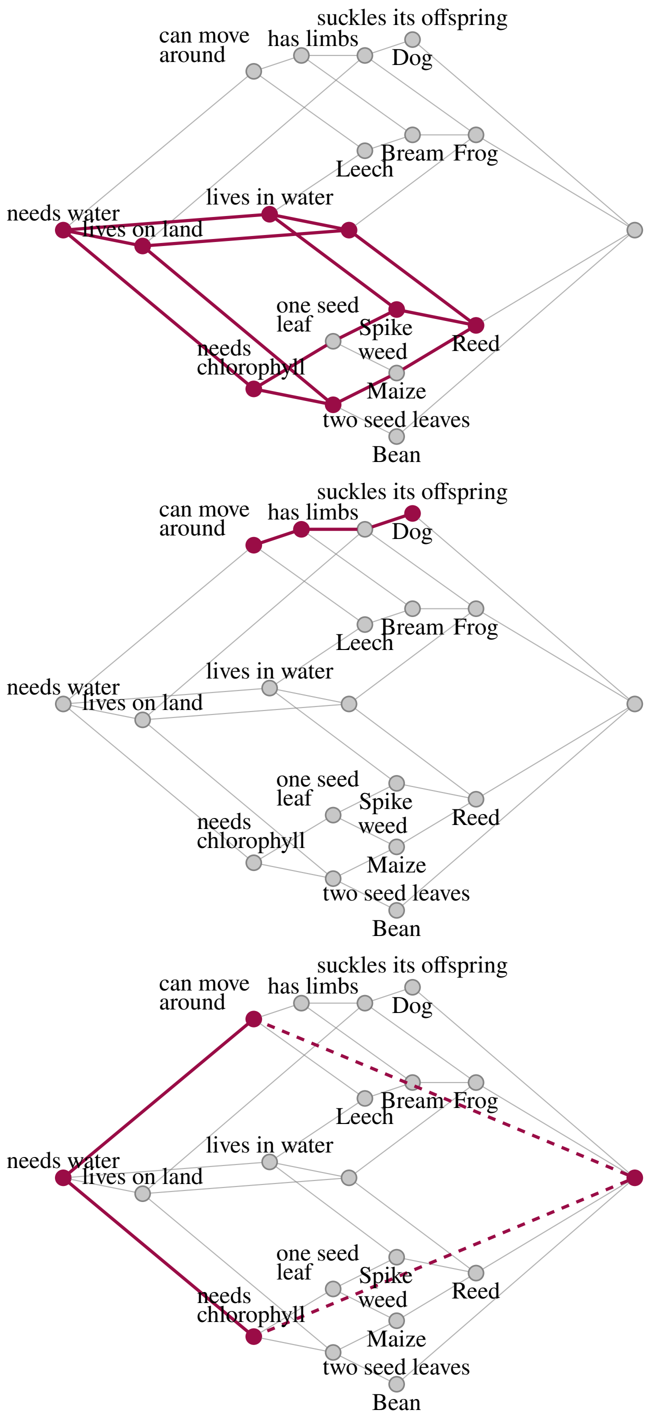


Figure 9.1: The concept lattice of the water context (see Figure 8.7) and highlighted nominal, ordinal and contranominal ordinal motifs.

nominal scale context, i.e., an anti-chain order. Given Problem 15.2, we are able to identify these sub-structures by local full scale-measures into the respective scale contexts, i.e., into $\mathbb{B}_3$, $\mathbb{O}_3$ and $\mathbb{N}_2$. Applied to the set of attributes, the contranominal scale motif encodes that the attributes *lives in water*, *lives on land* and *needs chlorophyll* are independent and any combination of these attributes is closed. The ordinal scale motif encodes that there is a ranking or increase in capability measured by the attributes *can move around*, *has limbs* and *suckles its offspring*. From the nominal scale motif, we can infer that the attributes *can move around* and *needs chlorophyll* are independent of each other, and, moreover, since their meet is empty, they are mutually exclusive. We may note that these findings are true in the present data, but do not need to be true in general.

9.2 Recognizing Ordinal Motifs

Recognize local scale-measures

To recognize ordinal motifs, we have to decide if a given map is a local or local full scale-measure. This problem can be dealt with analogously to Problems 7.2.1 and 14 with the additional check that $H \subseteq G_{\mathbb{K}}$ and a restriction of $\mathbb{K}$ to $\mathbb{K}[H, M_{\mathbb{K}}]$. The resulting check for local (full) scale-measures has the same complexity.

Problem 15: Deciding Local Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$ and a map $\sigma : H \to G_{\mathbb{S}}$	
Output: True iff σ is a local scale-measure of $\mathbb{K}$.	
Complexity:	$O(\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{S}\rvert)$

Problem 16: Deciding Local Full Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$ and a map $\sigma : H \to G_{\mathbb{S}}$	
Output: True iff σ is a local full scale-measure of $\mathbb{K}$.	
Complexity:	$O(\lvert\mathbb{K}\rvert \cdot \lvert\mathbb{S}\rvert)$

Scale-Measures and Implicational Theories

Decide scale-measures with implications

Before we now turn to finding ordinal motifs in data, i.e., finding scale-measures, we want to point out one more practical relevant observation with following proposition. In practice, context like data sets are large, however, mostly only in one dimension. The usual case is that the number of objects in a formal context is many times larger than the number of attributes. The reverse case, of course, also occurs. The most expensive computation for context and scales is the derivation, in particular in the direction of the larger dimension, i.e., objects or attributes. We therefore want to present an alternative representation using implications in contexts.

Reflected implications

To syntactically link implications with scale-measures, we use the short-hand notation $\sigma^{-1}(A \to B) := \sigma^{-1}(A) \to \sigma^{-1}(B)$. For the theory $\mathrm{Th}(\mathbb{K})$ we define $\sigma^{-1}(\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S})) := \{\sigma^{-1}(A \to B) \mid A \to B \in \mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S})\}$.

Proposition 24 (Recognizing (full) Scale-Measures using Implications). *For a context $\mathbb{K}$ a scale $\mathbb{S}$ and a map $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ we find that*

i) *σ is a scale-measure* $\iff \sigma^{-1}(\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S})) \vdash \mathrm{Th}_{G_{\mathbb{K}}}(\mathbb{K})$

ii) *σ is a full scale-measure* $\iff \mathrm{Th}_{G_{\mathbb{K}}}(\mathbb{K}) \cong \sigma^{-1}(\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S}))$.

Proof. First, we note that for two closed implicational theories $\mathrm{Th}_1, \mathrm{Th}_2$, i.e., transitive closures of implication sets, it holds that $\mathrm{Th}_1 \subseteq \mathrm{Th}_2 \iff \mathrm{Th}_2 \vdash \mathrm{Th}_1$. Secondly, we note that there is a Galois connection between the lattice of all implicational theories and the lattice of all closure systems [34, Theorem 57] to which the hierarchy of scale-measures is isomorph (Proposition 10).

Next, we elaborate on the relation between $\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S})$ and $\sigma^{-1}(\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S}))$. (1) The closure system $\mathrm{Ext}(\mathbb{S})$ is a model of $A \to B$ iff for all $E \in \mathrm{Ext}(\mathbb{S})$ with $A \subseteq E$ we have $B \subseteq E$. The map σ^{-1} is monotone and for $g_1, g_2 \in G_{\mathbb{S}}$ it holds that $\sigma^{-1}(g_1) \cap \sigma^{-1}(g_2) = \emptyset$. Therefor we find that: (2) $A \subseteq E$ iff $\sigma^{-1}(A) \subseteq \sigma^{-1}(E)$. Combining (1) and (2) we find that $\mathrm{Ext}(\mathbb{S}) \models A \to B$ iff $\sigma^{-1}(\mathrm{Ext}(\mathbb{S})) \models \sigma^{-1}(A \to B)$.

i) The map σ is a scale-measure iff the closure system $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$ is a coarser closure system of $\mathrm{Ext}(\mathbb{K})$ on $G_{\mathbb{K}}$. Given our preliminary considerations this is the case if and only if the theory of $\mathrm{Th}_{G_{\mathbb{K}}}(\mathbb{K})$ is a model of $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$, i.e., $\sigma^{-1}(\mathrm{Th}_{G_{\mathbb{S}}}(\mathbb{S})) \vdash \mathrm{Th}_{G_{\mathbb{K}}}(\mathbb{K})$.

ii) The map σ is per definition a full scale-measure iff the closure system $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$ is equal to $\mathrm{Ext}(\mathbb{K})$. Given our preliminary considerations this is the case if and only if their object theories are equal. □

With the help of Proposition 24 one may use already existent logical inference checkers for the verification of (local) (full) scale-measures. Such a procedure is especially efficient in case $|G| \ll |H|$ and if we are looking for many ordinal motifs of $\mathbb{K}$.

9.2.1 Ordinal Motif Problems

Problem Definitions

Starting from Problem 15.2, we now want to formulate a decision problem to investigate the complexity of Problem 15.2. In the following we refer by *RsSM* to the decision problem, if for two formal contexts $\mathbb{K}$ and $\mathbb{S}$ there exists a surjective scale-measure from $\mathbb{K}$ to $\mathbb{S}$, i.e., the *Recognizing Surjective Scale-Measures* problem. Analogously, we refer by *RfSM* to the decision problem, if for two formal contexts there exists a full scale-measure.

Theorem 6 (Ordinal Motif Problems). *For two formal contexts $\mathbb{K}$ and $\mathbb{S}$, RsSM and RfSM are NP-complete.*

Proof. To avoid any peculiarities, we consider in the following reductions graphs of size at least three.

a) **hardness:** To show NP-hardness of the RsSM problem, we reduce the sub-graph isomorphism (SI) problem to RsSM. For two Graphs G, H consider the formal context $\mathbb{G} = (V_G \cup \{\bot\}, E_G \cup \{\{v\} \mid v \in V_G\} \cup \{\emptyset\}, \in)$ and analogously constructed formal context $\mathbb{H}$. The set of extents of $\mathbb{G}$ is equal to $\{\{v\} \mid v \in V_G\} \cup E_G \cup \{\emptyset, V_G \cup \{\bot\}\}$. This reduction is polynomial in the size of G, H.

   ⇒ Let σ be a surjective scale-measure of $\mathbb{G}$ into $\mathbb{H}$. Then $\sigma^{-1}(\mathrm{Ext}(\mathbb{H})) \subseteq \sigma^{-1}(\mathrm{Ext}(\mathbb{G}))$. In particular for every $e \in E_H$ we have $\sigma^{-1}(e) \in \mathrm{Ext}(\mathbb{G})$. Since σ is surjective, we can infer that $2 \leq |\sigma^{-1}(e)| < |V_G|$. The only extents with a cardinality in that interval are the edge extents of $\mathbb{G}$. Thus $\sigma^{-1}(e) \in E_G$ and all nodes of e have a unique pre-image. Since $E_H \subseteq \mathrm{Ext}(\mathbb{H})$, all nodes with at least one edge have a unique pre-image. WLOG we assume that the pre-image of all $v \in V_H$ have a unique pre-image, otherwise change the map σ for all but one node to $\bot$. Hence the map $\sigma^{-1} : V_H \to V_G$ is edge preserving and an isomorphism of (H, E_H) into a sub-graph of G, i.e., into the sub-graph given by $(\text{co-dom}(\sigma^{-1}) \setminus \sigma^{-1}(\bot), \{e \in E_G \mid \exists l \in E_H : \sigma^{-1}(l) = e\})$.

$\Leftarrow$ Let σ be an isomorphism of $\mathbb{H}$ into a sub-graph of $\mathbb{G}$, i.e., an edge preserving map from H into G. Based on this consider the map $\theta : V_G \cup \{\bot\} \to V_H \cup \{\bot\}$ where $\theta(v) = \sigma^{-1}(v)$ and $\bot$ otherwise. The map θ is surjective by definition. For the node extents, the empty extent and the top extent $V_H \cup \{\bot\}$ of $\mathbb{H}$ we have that their pre-images are in $\mathrm{Ext}(\mathbb{G})$. For an extent e in E_H we have that $\theta^{-1}(e) = \sigma(e) \in E_G$, since σ is edge preserving. Thus θ is a surjective scale-measure from $\mathbb{G}$ into $\mathbb{H}$.

completeness: An algorithm for identifying if there is a surjective scale-measure for two context $\mathbb{O}, \mathbb{K}$ can be constructed by guessing non-deterministically a mapping σ. The check for a surjective scale-measure can be done deterministically in polynomial time in the size of both contexts.

b) **hardness:** To show NP-hardness of the RfSM problem, we reduce the induced sub-graph isomorphism (ISI) problem to the RfSM problem. For two graphs G, H consider the contexts $\mathbb{G} = (V_G, E_G \cup \{\{v\} \mid v \in V_G\} \cup \{\emptyset\}, \in)$ and $\mathbb{H}$ analogously. The set of extents of $\mathbb{G}$ is equal to $\{\{v\} \mid v \in V_G\} \cup E_G \cup \{\emptyset, V_G\}$. This reduction is polynomial in the size of G, H.

$\Rightarrow$ Let σ be a full scale-measure of $\mathbb{H}$ into $\mathbb{G}$. Then for every $v \in V_H$ the extent $\{v\} \in \mathrm{Ext}(\mathbb{H})$ is the pre-image of an extent A of $\mathrm{Ext}(\mathbb{G})$. Since $v \in \sigma^{-1}(A)$ we have $\sigma(v) \in A$ and from $\{v\} = \sigma^{-1}(A)$ we can infer that there exists no other $w \in V_H$ with $w \neq v$ and $\sigma(w) = \sigma(v)$. Thus σ is injective.
For an edge $e \in E_G$ where $e \subseteq \text{co-dom}(\sigma)$ we have $\sigma^{-1}(e) \in \mathrm{Ext}(\mathbb{H})$ and since σ is injective we can infer $|\sigma^{-1}(e)| = 2$ and thus $\sigma^{-1}(e) \in E_H$. For an edge $e \in E_H$ there must be an $A \in \mathrm{Ext}(\mathbb{G})$ with $\sigma^{-1}(A) = e$. Thus $\sigma(e) \subseteq A$. Since the only extents of $\mathbb{G}$ for which this applies are V_G extents of cardinality two, i.e., the edges of G. Thus, $\sigma(e) \in \mathrm{Ext}(\mathbb{G})$ and further $\sigma(e) \in E_G$. Concluding, σ is an isomorphism between H and $\sigma(H)$.

$\Leftarrow$ Let σ be an isomorphism between H and an induced sub-graph of G. Then for every $v \in V_G$ is $\sigma^{-1}(\{v\})$ either in V_H or empty since σ is injective. For edges $e \in E_G$ where $e \subseteq \text{co-dom}(\sigma)$ we have that $\sigma^{-1}(e) \in E_H \subseteq \mathrm{Ext}(\mathbb{H})$ since σ is an isomorphism restricted to co-dom(σ). In case $e \subseteq \text{co-dom}(\sigma)$ does not hold, the pre-image is equal to a node or the empty set. Thus $\sigma^{-1}(E_G) \subseteq \mathrm{Ext}(\mathbb{H})$. Furthermore, $\sigma^{-1}(\emptyset) = \emptyset \in \mathrm{Ext}(\mathbb{H})$ and $\sigma^{-1}(V_G) = V_H \in \mathrm{Ext}(\mathbb{H})$. Thus σ is a scale-measure of $\mathbb{H}$ into $\mathbb{G}$. For an edge $e \in E_H$ we have that $\sigma^{-1}(\{\sigma(v) \mid v \in e\}) = e$ and $\{\sigma(v) \mid v \in e\} \in E_G \subseteq \mathrm{Ext}(\mathbb{G})$ since σ is isomorphism restricted to co-dom. Thus σ is a full scale-measure.

completeness: An algorithm for identifying if there is a full scale-measure for two context $\mathbb{O}, \mathbb{K}$ can be constructed by guessing non-deterministically a mapping σ. The check for a full scale-measure can be done deterministically in polynomial time in the size of both contexts.

With this theorem we have shown the complexities of the ordinal motif problem (cf. Problem 15.2) in general. Further algorithmic improvements are needed to deal with these problems on large data sets. In the next subsection, we analyze these problems with respect to specific classes of ordinal motifs. Complexities

Problem 17: Recognizing Surjective Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$	
Output: True iff there exists a surjective scale-measure σ from $\mathbb{K}$ to $\mathbb{S}$	
Complexity:	NP-complete

Problem 18: Recognizing Full Scale-Measures Problem	
Input: Formal contexts $\mathbb{K}, \mathbb{S}$	
Output: True iff there exists a full scale-measure σ from $\mathbb{K}$ to $\mathbb{S}$	
Complexity:	NP-complete

9.2.2 Recognizing Standard Scales

Recognize standard scales

In practice we consider families of standard scales for ordinal motifs due to their interpretability. In the following, we demonstrate the complexities for the problems above with respect to families of standard scale contexts $\mathcal{S}$.

Proposition 25 (Recognizing Full Standard Scales). *Let $\mathbb{K}$ be a formal context. Deciding whether there is a surjective full scale-measure into $\mathbb{N}_n$ with $|G_{\mathbb{K}}| = n$ is in* P *with respect to the size of $\mathbb{K}$. The analogue is true for $\mathbb{O}_n$, $\mathbb{I}_n$, $\mathbb{C}_n$ and $\mathbb{B}_n$.*

Proof. WLoG we assume that $\mathbb{K}$ is clarified. We first show the claim for $\mathbb{B}_n$.

For a contranominal scale $\mathbb{B}_n := ([n], [n], \neq)$ every pair of bijective maps $(\alpha : [n] \to [n],\ \beta : [n] \to [n])$ is a context automorphism of $\mathbb{B}_n$. Thus, we can select an arbitrary bijective mapping from G into $[n]$ and check if it is a full scale-measure from $\mathbb{K}$ into the contranominal scale $\mathbb{B}_n$. The verification of full scale-measures is in P [101]. The same reasoning can be applied for $\mathbb{N}_n$.

For ordinal scales we need to verify that for each pair of objects their object concepts are comparable in the concept lattice. Hence, the recognition for ordinal scales is in P. For an interordinal scale $\mathbb{I}_n := ([n], [n], \leq)|([n], [n], \geq)$ we can infer from the extents of $\mathbb{K}$ of cardinality 2 two candidate mappings $\sigma_\leq, \sigma_\geq$ in the following way. For interordinal scales the extents of cardinality two overlap on one object each and form a chain. From said chain we can infer an order relations of the objects G given by position in which they occur in the chain. From the total order on G we define the mapping $\sigma_\leq : G \to [n]$ where the objects are mapped according to their position. The map $\sigma_\geq$ is defined based on the dual order. All maps other than $\sigma_\leq$ and $\sigma_\geq$ would violate the extent structure of the chain. For $\sigma_\leq$ and $\sigma_\geq$ we can verify in P if either is a full scale-measure. Moreover, the extents of cardinality two can be computed in polynomial time using `TITANIC` or `next_closure`. Hence, the recognition for interordinal scales is in P.

For crown scales $\mathbb{C}_n := ([n], [n], J)$, where $(a, b) \in J \iff a = b$ or $(a, b) = (n, 1)$ or $b = a + 1$, we can select an arbitrary object $g \in [n]$ and draw repeatedly without putting back a different $h \in [n]$ with $\{g\}' \cap \{h\}' \neq \{\}$. Starting from g there is a (up to duality) unique drawing order, i.e., $1, 2, \ldots, n$ and the dual $1, n, n-1, \ldots, 2$. In order to find a full scale-measure we have to find an isomorphic drawing order for the elements of G in the same manner. From this we can derive a map $G \to [n]$ with respect to the drawing order and verify if it is a full scale-measure. The computational cost of the drawing procedure as well as the verification is in P. □

Problem 19: Recognizing Full Standard Scales	
Input: A formal context $\mathbb{K}$	
Output: True iff there exists a full scale-measure from $\mathbb{K}$ into either $\mathbb{N}_n$, $\mathbb{O}_n$, $\mathbb{I}_n$, $\mathbb{C}_n$ or $\mathbb{B}_n$ with $\lvert G_{\mathbb{K}}\rvert = n$	
Complexity:	P

The polynomial recognition of full scale-measures into standard scales allows for an efficient generation of full explanations. For partial explanations we provide the following propositions.

Full explanations with standard scales

Proposition 26 (Recognizing Standard Scales I). *Let $\mathbb{K}$ be a formal context. Deciding whether there is a surjective scale-measure into $\mathbb{N}_n$ with $|G_{\mathbb{K}}| = n$ is in* P *with respect to the size of $\mathbb{K}$. The analogue is true for $\mathbb{B}_n$.*

Proof. Since $|G_{\mathbb{K}}| = n$, we can follow that a surjective scale-measure from $\mathbb{K}$ to $\mathbb{N}_n$ is also injective. For a nominal scale is any pair of bijective maps a context automorphism. Thus, either no bijective map from $G_{\mathbb{K}}$ to $[n]$ is a surjective scale-measure of $\mathbb{K}$ into $\mathbb{N}_n$, or all are. Concluding, it is sufficient to check one arbitrary bijective map for the scale-measure property which can be done in polynomial time (see Problem 7.2.1). The analogue reasoning can be applied to $\mathbb{B}_n$. □

Problem 20: Recognizing Standard Scales (P)	
Input:	A formal context $\mathbb{K}$
Output:	True iff there exists a surjective scale-measure from $\mathbb{K}$ into either $\mathbb{N}_n$ or $\mathbb{B}_n$ with $|G_{\mathbb{K}}| = n$
Complexity:	P

Proposition 27 (Recognize Standard Scales II). *Let $\mathbb{K}$ be a formal context. Deciding whether there is a surjective scale-measure into $\mathbb{O}_n$ with $|G_{\mathbb{K}}| = n$ is* NP*-complete. The analogue is true for $\mathbb{I}_n$ and $\mathbb{C}_n$.*

Proof. [**Case** $\mathbb{O}_n$] **hardness:** To show that this problem is NP-hard we reduce the Hamiltonian path problem for **directed graphs** (di-graph) to it. A directed graph is a pair (V, E) where $E \subseteq V \times V$. The notion of paths in directed graphs translates naturally from graphs.

For a di-graph $G := (V, E)$ let $\mathbb{G} := (V, V, E^+)$ be a formal context with $E^+ := E^* \setminus \Delta(V)$, where E^* is the transitive extension of E and $\Delta(V)$ the diagonal of V. Since $|G_{\mathbb{G}}| = |V| = n$, we can follow that a surjective map into $[n]$ is also injective. This reduction is polynomial in the size of the input. WLoG is $n \geq 3$.

- $\Rightarrow$ Let $(v_1, \ldots, v_n)$ be a Hamiltonian path in G. For $i < j$ is $(v_i, v_j) \in E^+$ per definition. For $\{v_1, \ldots, v_i\}$ with $i \leq n$, we find that the object derivation $\{v_1, \ldots, v_i\}^{E^+} \supseteq \{v_{i+1}, \ldots v_n\}$. For all $g \in \{v_1, \ldots, v_i\}$ we find that $g \notin \{v_1, \ldots, v_i\}^{E^+}$, since E^+ excludes $\Delta(V)$ per definition. Thus, the object derivation $\{v_1, \ldots, v_i\}^{E^+}$ is equal to $\{v_{i+1}, \ldots v_n\}$. Applying the analogue reasoning for the attribute derivation yields $\{v_1, \ldots, v_i\}^{E^+E^+} = \{v_1, \ldots, v_i\}$. Thus, for an interval $[1, i]$ is $\{v_1, \ldots, v_i\} \in \text{Ext}(\mathbb{K})$. For the ordinal scale $\mathbb{O}_n$ we find that the set of extents is comprised of the intervals $[j, n]$. Hence, the map $\sigma : V \to [n]$ with $\sigma(v_i) = n - (i - 1)$ is a scale-measure from $\mathbb{G}$ into $\mathbb{O}_n$.
- $\Leftarrow$ For a surjective scale-measure σ from $\mathbb{G}$ into $\mathbb{O}_n$ let $\sigma^{-1}\big(n - (i - 1)\big) = v_i$, cf. σ is also injective. For an extent $[n - (i - 1), n]$ of $\mathbb{O}_n$ is $\sigma^{-1}([n - (i - 1), n]) = \{v_1, \ldots, v_i\}$ in $\text{Ext}(\mathbb{G})$ and therefor $\{v_1, \ldots, v_i\}^{E^+E^+} = \{v_1, \ldots, v_i\}$. Thus, for all intervals $[1, i]$ with $i \leq n$ is $\{v_1, \ldots, v_i\}$ an extent of $\mathbb{G}$. On top of that is $\{\}$ an extent of $\mathbb{G}$. This follows directly from the definition. Hence, (1) the interval extents together with the emptyset form a chain order of $n + 1$ extents.

For an extent $\{v_1, \ldots, v_i\}$ we can deduce that (2) $\{v_1, \ldots, v_i\}^{E^+} \subseteq \{v_{i+1}, \ldots, v_n\}$, since E^+ excludes the diagonal per definition. Due to the antitone property of the derivation operator (cf. 5.1 Proposition 1) we find that $\sigma^{-1}([i,n])^{E^+} \subseteq \sigma^{-1}([i+1,n])^{E^+}$. Moreover, since $\sigma^{-1}([i,n]) \neq \sigma^{-1}([i+1,n])$, and the pre-images $\sigma^{-1}([i,n]), \sigma^{-1}([i+1,n])$ are in Ext($\mathbb{G}$), we can deduce that (3) $\sigma^{-1}([i,n])^{E^+} \subseteq \sigma^{-1}([i+1,n])^{E^+}$ and $\sigma^{-1}([i,n])^{E^+} \neq \sigma^{-1}([i+1,n])^{E^+}$. Combining (1), (2) and (3) we find that $\{v_1, \ldots, v_i\}^{E^+} = \{v_{i+1}, \ldots, v_n\}$. Hence, for $i < j$ we find that $(v_i, v_j) \in E^+$. From the covering relation of E^+ we find that for $i, i+1$ is $v_i, v_{i+1} \in E$. Concluding, $(v_1, \ldots, v_n)$ is a Hamiltonian path.

completeness: This problem can be decided by non-deterministically guessing a map σ and check for the scale-measure and surjective property. The checks can be done in polynomial time (see Problem 7.2.1).

[**Case** $\mathbb{I}_n$] **hardness:** For the interordinal problem we present an analogue reduction to the ordinal case where for a di-graph $G := (V, E)$ we define a formal context $\mathbb{G} := (V, V, E^+) \mid (V, V, E^+)^d$ and $(V, V, E^+)^d := (V, \bar{V} := \{\bar{v} \mid v \in V\}, \{(v_i, \bar{v_j}) \mid (v_j, v_i) \in E^+\})$. In other words, $\mathbb{G}$ is the apposition of the context from the ordinal case and its dual. This reduction is polynomial in the size of the input. WLoG is $n \geq 3$.

- ⇒ For a Hamiltonian path $(v_1, \ldots, v_n)$ in G let $\sigma : V \to [n]$ be a map with $\sigma(v_i) = n - (i-1)$. Based on the ordinal case we can deduce that σ is a surjective scale-measure into $\mathbb{O}_n$. Due to the apposition of the dual context $(V, V, E^+)^d$ we can deduce that σ is a surjective scale-measure into $\mathbb{O}_n^d$. From the apposition of scale-measures (Proposition 20), we find that σ is a surjective scale-measure into $\mathbb{O}_n \mid \mathbb{O}_n^d$, which is equal to $\mathbb{I}_n$.
- ⇐ For a surjective scale-measure σ from $\mathbb{G}$ into $\mathbb{I}_n$ let $\sigma^{-1}\big(n-(i-1)\big) = v_i$. Since $\mathbb{I}_n$ is equal to the apposition $\mathbb{O}_n \mid \mathbb{O}_n^d$ and since the scale-measure property is preserved when removing attributes (cf. Corollary 4), we can deduce that σ is a surjective scale-measure into $\mathbb{O}_n$. From the ordinal case we can deduce that $(v_1, \ldots, v_n)$ is a Hamiltonian path.

completeness: This problem can be decided by non-deterministically guess a map σ and check for the scale-measure and surjective property. The checks can be done in polynomial time (see Problem 7.2.1).

[**Case** $\mathbb{C}_n$] **hardness:** To show the NP-hardness of this problem we reduce the Hamiltonian cycle (HC) problem for undirected graphs to it, i.e., decide for a graph G if there is a circle visiting every node of G exactly ones. This problem is known to be NP-complete.

For the reduction, we map the graph $G := (V, E)$ (WLoG $|V| \geq 3$) to a formal context $\mathbb{G} := (V, \hat{V} \cup E, \in)$ where $\hat{V} := \{\{v\} \mid v \in V\}$. This map is polynomial in the size of the input. The set of extents of $\mathbb{G}$ is equal to $\hat{V} \cup E \cup \{V, \{\}\}$. The context $\mathbb{G}$ accepts a surjective scale-measure into the crown scale of size $|G|$ iff there is a sequence of extents $A_1, \ldots, A_n \in \text{Ext}(\mathbb{K})$ with $|A_i| = 2$ such that $(V, \{A_1, \ldots, A_n\}) \leq G$ is a cycle visiting each object $v \in V$ exactly ones. This is the case iff G has a Hamilton cycle.

completeness: This problem can be decided by non-deterministically guess a map σ and check for the scale-measure and surjective property. The checks can be done in polynomial time (see Problem 7.2.1). □

Problem 21: Recognizing Standard Scales (NP)	
Input:	A formal context $\mathbb{K}$
Output:	True iff there exists a surjective scale-measure from $\mathbb{K}$ into either $\mathbb{O}_n$, $\mathbb{I}_n$ or $\mathbb{C}_n$ with $\lvert G_{\mathbb{K}}\rvert = n$
Complexity:	NP-complete

Given these complexity results there are some partial explanations based on standard scales that are more difficult to generate than others. Despite these results, more research is needed to derive efficient algorithms.

Multiple explanations

9.2.3 Ordinal Motifs and Implications

Within the set of ordinal motifs that a context accepts, there is a subset that is closely related to implications in formal contexts.

Derive implications from ordinal motifs

Proposition 28 (Ordinal Motifs and Implications). *For a* $\mathbb{K} := (G, M, I)$ *and* $g, h \in G$ *with* $\{g\}' \neq \{h\}'$ *TFAE:*

i) $\sigma : \{g, h\} \to \{1, 2\}$ *with* $\sigma(g) = 1$ *and* $\sigma(h) = 2$ *is a l.f. scale-measure from* $\mathbb{K}$ *to* $\mathbb{O}_2$*,*

ii) $g \to h$ *is valid in* $\mathbb{K}$*.*

Proof. The implication $g \to h$ is valid in $\mathbb{K}$ iff $\{g\}' \subseteq \{h\}'$. Given the condition that $\{g\}' \neq \{h\}'$ the context $\mathbb{K}[\{g, h\}, M]$ has two extents, i.e., $\{h\}$ and $\{g, h\}$. Thus, σ is a full scale-measure from $\mathbb{K}[\{g, h\}, M]$ into $\mathbb{O}_2$. In the reverse case we can follow from the full scale-measure that $\{h\}$ and $\{g, h\}$ are the extents of $\mathbb{K}[\{g, h\}, M]$. Thus, $\{g\}' \subseteq \{h\}'$. □

This result opens the question on what notions and methods from rule mining can be transferred to ordinal motifs. In particular a notion of confidence may allow the study of ordinal motifs that are contained in $\mathbb{K}$ up to some degree. Further exploration of this connection is deemed future work.

9.3 Heredity of Ordinal Motifs

Now that we understand the computational complexities for Problem 15.2, we want to present an interesting property of scales that may help to reduce the computational efforts.

The families of standard scales have a special property, called **heredity** [80, Proposition 123], i.e., for every scale $\mathbb{S}$ of a family of scales $\mathcal{S}$ it holds that every sub-scale $\mathbb{S}[H, M_{\mathbb{S}}]$ is equivalent (up to attribute reduction) to a scale in $\mathcal{S}$. In this section we will demonstrate how the notion for heredity of scales impacts scale-measures.

Heredity

Lemma 10 (Heredity of Scale-Measures). *Let* $\mathbb{K}$ *be a formal context,* $\mathbb{S}$ *a scale from a heredity scale family and* $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ *a surjective (full) scale-measure. For any* $H \subseteq G_{\mathbb{K}}$ *is the map* $\sigma\mid_H$ *a surjective (full) scale-measure from* $\mathbb{K}[H, M_{\mathbb{K}}]$ *into* $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$*.*

Proof. First we show that $\sigma\mid_H$ is a scale-measure from $\mathbb{K}[H, M_{\mathbb{K}}]$ into $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$. Since $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$ is an induced sub-context of $\mathbb{S}$ with equal attribute set, we can write every extent $A \in \operatorname{Ext}(\mathbb{S}[\sigma(H), M_{\mathbb{S}}])$ as the intersection $\check{A} \cap \sigma(H)$ for some $\check{A} \in \operatorname{Ext}(\mathbb{S})$. The pre-image $(\sigma\mid_H)^{-1}(\check{A} \cap \sigma(H))$ is equal to $(\sigma\mid_H)^{-1}(\check{A}) \cap (\sigma\mid_H)^{-1}(\sigma(H))$. Since $\check{A}$ and $\sigma(H)$ are entailed in the image of σ on H we can follow that $(\sigma\mid_H)^{-1}(\check{A}) = \sigma^{-1}(\check{A})$ and

$(\sigma\ |_H)^{-1}(\sigma(H)) = H$. Moreover, since σ is a scale-measure we can follow that $\sigma^{-1}(\check{A})$ is an extent of $\mathbb{K}$. Summarizing, the pre-image $(\sigma\ |_H)^{-1}(A)$ is equal to the intersection of an extent of $\mathbb{K}$ and H. Hence, $(\sigma\ |_H)^{-1}(A)$ is an extent of $\mathbb{K}[H, M_{\mathbb{K}}]$ and $\sigma\ |_H$ a scale-measure of $\mathbb{K}[H, M_{\mathbb{K}}]$ into $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$.

In case σ is a full scale-measure it remains to be shown that for every $D \in \text{Ext}(\mathbb{K}[H, M])$ there exists a $C \in \mathbb{S}[\sigma(H), M_{\mathbb{S}}]$ with $(\sigma\ |_H)^{-1}(C) = D$. We can write the extent D as the intersection $\check{D} \cap H$ where $\check{D} \in \text{Ext}(\mathbb{K})$. Since σ is a full scale-measure we can follow for $\check{D}$ that there is a $\check{C} \in \text{Ext}(\mathbb{S})$ with $\sigma^{-1}(\check{C}) = \check{D}$. Since $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$ is an induced sub-context of $\mathbb{S}$ with equal attribute set we find that $\check{C} \cap \sigma(H)$ is an extent of $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$. Thus, for $C := \check{C} \cap \sigma(H)$ we find that $(\sigma\ |_H)^{-1}(\check{C} \cap \sigma(H)) = (\sigma\ |_H)^{-1}(\check{C}) \cap (\sigma\ |_H)^{-1}(\sigma(H))$ and furthermore that $(\sigma\ |_H)(\check{C}) \cap (\sigma\ |_H)^{-1}(\sigma(H)) = \check{D} \cap H = D$. Hence, $\sigma\ |_H$ is a full scale-measure.

The map $\sigma\ |_H$ is surjective, since the object set of $\mathbb{S}[\sigma(H), M_{\mathbb{S}}]$ is equal to the co-domain of $\sigma\ |_H$. □

This property transfers naturally to ordinal motif of contexts.

Proposition 29 (Heredity of Ordinal Motifs). *Let $\mathbb{K}$ be a formal context, $\mathbb{S}$ a scale from a heredity scale family and $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ a surjective (full) scale-measure. Then for any $H \subseteq G_{\mathbb{K}}$ is the map $\sigma\ |_H$ a surjective (full) scale-measure from $\mathbb{K}[H, M_{\mathbb{K}}]$ into an ordinal motif of the same family as $\mathbb{S}$.*

Proof. This proposition follows directly from Lemma 10 and the definition of heredity scales. □

This proposition is essential when applying ordinal motifs for the analysis of ordinal data sets using heredity scales. When computing all candidates for (full) scale-measures this statement allows us to discard a large proportion. Fortunately, many families of scales, such as the nominal scales, ordinal scales, interordinal scales, contranominal scales, etc, have the heredity property [80, Proposition 123]. Crown scales do not have this property.

9.4 Automatic Textual Explanations

Template explanations

The basis for the automatic generation of textual explanation for concept lattices are **local full** explanations. For a given formal context we identify ordinal motifs through local (full) scale-measures and template explanations. A template explanation for an ordinal motif on n objects may look like this:

The elements $g_1, \ldots, g_{n-1}$ and g_n are in … related to each other.

Text generation

Given a context $\mathbb{K}$ and a local (full) scale-measure σ of $\mathbb{K}$ into $\mathbb{S}$ we can replace every instance of an object $g_i \in G_{\mathbb{S}}$ in a textual explanation template of the ordinal motif $\mathbb{S}$ by its pre-image $\sigma^{-1}(g) \subseteq G$. This yields a textual explanation of $\mathbb{K}$ with respect to σ. An example explanation for interordinal scales and the color spectrum in Figure 7.11 can be seen in the following:

*The elements **violet, blue** and **red** are in an interordinal relation to each other.*

This explanation, however, is very technical and focuses on theoretical properties. To generate textual explanations that are understandable to untrained users, we employ human-centered explanations developed by Viktoria Horn [100, Section 5]. For this, we summarize and reproduce section 5 of Hirth, Horn, Stumme, and Hanika [100] in the following subsection faithfully.

9.4.1 Human-Centered Textual Explanations

HCI explainability principles

To derive explanations of a concept lattice that are understandable to untrained users, a *human-centered* approach is recommended. The explanation templates that we recall in this subsection were developed with state-of-the-art principles from *human-centered explanations* [37, 149] in mind, namely the five goodness criteria Mamun et al. [149] for Explainable AI (XAI) [196]. These are the *accuracy*, *scope*, *explanation form*, *simplicity* and *knowledge base criterion*. For a detailed discussion on these criteria and on how all of them are met by the following explanation templates we refer the reader to Section 5 in Hirth, Horn, Stumme, and Hanika [100].

Nominal Scale: *"The elements $n_1, \dots, n_{k-1}$ and n_k are incomparable, i.e., all elements have at least one property that the other elements do not have."*

Ordinal Scale: *"There is a ranking of elements $n_1, \dots, n_{k-1}$ and n_k such that an element has all the properties its successors has."*

Interordinal Scale: *"The elements $n_1, \dots, n_{k-1}$ and n_k are ordered in such a way that each interval of elements has a unique set of properties they have in common."*

Contranominal Scale: *"Each combination of the elements $n_1, \dots, n_{k-1}$ and n_k has a unique set of properties they have in common."*

Crown Scale: *"The elements $n_1, \dots, n_{k-1}$ and n_k are incomparable. Furthermore, there is a closed cycle from n_1, over $n_2, \dots n_{k-1}$ and n_k back to n_1 by pairwise shared properties."*

Domain independence

The proposed textual explanations are designed to be domain independent. This allows us to apply our method in a general setting. Though, some data domains may come with specific terminology and requirements. In these cases it is advisable to develop dedicated explanation templates. Moreover, the presented templates are meant to provide the theoretical foundation for automatically generating human-centered explanations of concept lattices. A test of their capability to be understandably by untrained users is deemed future work.

User study outlook

9.5 Ordinal Motif Covering

Which motifs are best?

Due to the large number of ordinal motifs that a formal context accepts we present in the following a method to select a small but meaningful subset of them. Our goal is to derive **partial** explanations by covering large proportions of a concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$ using a small set of scale-measures $\mathcal{S}$ into a given set of ordinal motifs. We say a concept $(A, B) \in \underline{\mathfrak{B}}(\mathbb{K})$ is covered by $(\sigma, \mathbb{S}) \in \mathcal{S}$ iff it is reflected by $(\sigma, \mathbb{S})$, i.e., there exists an extent $D \in \mathrm{Ext}(\mathbb{S})$ with $\sigma^{-1}(D) = A$ or $\sigma^{-1}(D)^{I_\mathbb{K} I_\mathbb{K}} = A$ in the local case. This leads to the formulation of the general *ordinal motif covering* problem.

Problem 22.1 (Ordinal Motif Covering Problem). *For a context $\mathbb{K}$, a family of ordinal motifs $\mathcal{O}$ and $k \in \mathbb{N}$, what is the largest number $c \in \mathbb{N}$ such that there are surjective local full scale-measures $(\sigma_1, \mathbb{O}_1), \dots, (\sigma_k, \mathbb{O}_k)$ of $\mathbb{K}$ with $\mathbb{O}_1, \dots, \mathbb{O}_k \in \mathcal{O}$ and*

$$\left| \bigcup_{1 \le i \le k} (\mathrm{cl}_\mathbb{K} \circ \sigma_i^{-1})(\mathrm{Ext}(\mathbb{O}_i)) \right| = c$$

where $\mathrm{cl}_\mathbb{K}$ denotes the object closure operator of $\mathbb{K}$. If $\mathbb{K}$ does not allow for any scale-measure into an ordinal motif from $\mathcal{O}$ the value of c is 0.

We remind the reader that the maps $\mathrm{cl}_{\mathbb{K}}, \sigma_i$ are lifted to a family of sets (cf. Chapter 2). We call the set $\{(\sigma_1, \mathbb{O}_1), \ldots, (\sigma_k, \mathbb{O}_k)\}$ an **ordinal motif covering** of $\mathbb{K}$.

From covering to replacing

If one is able to find an ordinal motif covering that reflects all formal concepts of $\mathbb{K}$ we can construct a formal context $\mathbb{O}$ which accepts a scale-measure $(\sigma, \mathbb{S})$ if and only if $(\sigma, \mathbb{S})$ is a scale-measure of $\mathbb{K}$.

Proposition 30 (Ordinal Motif Basis of $\mathbb{K}$). *Let $\mathbb{K}$ be a formal context with object closure operator $\mathrm{cl}_{\mathbb{K}}$ and ordinal motif covering $\{(\sigma_1, \mathbb{O}_1), \ldots, (\sigma_k, \mathbb{O}_k)\}$ that covers all concepts of $\mathbb{K}$, i.e., $c = |\underline{\mathfrak{B}}(\mathbb{K})|$. Let*

$$\mathbb{O} :=|_{1\leq i\leq k}\ (G, M_{\mathbb{O}_i}, I_{\mathbb{O}_i,\mathrm{cl}_{\mathbb{K}}}),\ \textit{with}\ (g,m)\in I_{\mathbb{O}_i,\mathrm{cl}_{\mathbb{K}}} \iff g\in \mathrm{cl}_{\mathbb{K}}\big(\sigma_i^{-1}(\{m\}^{I_{\mathbb{O}_i}})\big)$$

where $|$ is the context apposition. Then a pair $(\sigma, \mathbb{S})$ is a local full scale-measure from $\mathbb{K}[H, M]$ to $\mathbb{S}$ iff σ is a local full scale-measure from $\mathbb{O}[H, M_{\mathbb{O}}]$ to $\mathbb{S}$. In this case we call $\mathbb{O}$ an ***ordinal motif basis*** *of $\mathbb{K}$.*

Proof. We have to show that the identity map is a full scale-measure from $\mathbb{K}$ to $\mathbb{O}$. Hence, we need to prove that all attribute extents of $\mathbb{O}$ are extents in $\mathbb{K}$ Proposition 7 and each extent of $\mathbb{K}$ is an extent of $\mathbb{O}$. For an attribute $m \in M_{\mathbb{S}_i}$ is $\mathrm{cl}_{\mathbb{K}}(\{m\}^{I_{\mathbb{S}_i}}) \in \mathrm{Ext}(\mathbb{K})$ per definition. The second requirement follows from the fact that $c = |\underline{\mathfrak{B}}(\mathbb{K})|$. □

The just introduced basis is a useful tool when investigating scale-measures of a context $\mathbb{K}$ given a set of ordinal motifs O. One can perceive O as a set of analytical tools and the existence of $\mathbb{O}$ implies that a found ordinal motif covering $\{(\sigma_1, \mathbb{O}_1), \ldots, (\sigma_k, \mathbb{O}_k)\}$ is complete with respect to scale-measures of $\mathbb{K}$.

9.5.1 Ordinal Motif Covering with Standard Scales

Which found standard scales are best?

The ordinal motif covering problem is a combinatorial problem which is computationally costly, even for standard scales. Thus, we propose in the following a greedy approach which has two essential steps. First, we compute all local full scale-measures $\mathcal{S}$ for standard scales. This step is computationally tame due to the heredity property of local full scale-measures for standard scales, as discussed in Section 9.3. Our goal is now to identify, in a greedy manner, elements of $\mathcal{S}$ that increase c the most. Thus, we select k full scale-measures where at each selection step i with $1 \leq i \leq k$ we select a scale-measure $(\sigma, \mathbb{O}) \in \mathcal{S}$ that maximizes Equation (9.1).

Covering criterion

$$\left|(\mathrm{cl}_{\mathbb{K}} \circ \sigma^{-1})\big(\mathrm{Ext}(\mathbb{O})\big) \setminus \bigcup_{1\leq j<i} (\mathrm{cl}_{\mathbb{K}} \circ \sigma_j^{-1})\big(\mathrm{Ext}(\mathbb{O}_j)\big)\right| \tag{9.1}$$

In the above equation $(\sigma_j, \mathbb{O}_j)$ denotes the scale-measure that was selected at step $j \leq i$. The union is the covering number c of the ordinal motif covering $(\sigma_1, \mathbb{O}_1), \ldots, (\sigma_{i-1}, \mathbb{O}_{i-1})$. Overall, the computed cardinality is equal to the number of concepts reflected by $(\sigma, \mathbb{O})$ that are not already reflected by $(\sigma_1, \mathbb{O}_1), \ldots, (\sigma_{i-1}, \mathbb{O}_{i-1})$.

Novelty criterion

For obvious reasons this approach results in the selection of scale-measures that have the largest number of (so far) uncovered concepts. A downside of this heuristic is that it favors ordinal motifs that have in general more concepts, e.g., contranominal scales over ordinal scales. To compensate for this we propose to normalize the heuristic by the number of concepts of the ordinal motif, i.e., $\left|\sigma^{-1}\big(\mathrm{Ext}(\mathbb{O})\big)\right|$.

In the first step, the normalized heuristic does not account for the total size of the ordinal motif. The first selected scale-measure covers at least the top extent, i.e., G, and thus the scores for all following ordinal motifs are at most ${}^{|\mathrm{Ext}(\mathbb{S})|-1}/_{|\mathrm{Ext}(\mathbb{S})|}$.

Table 9.2: Results for ordinal motifs of the spices planner context. Every column represents ordinal motifs of a particular standard scale family. Maximal lf-sm is the number of local full scale-measures for which there is no lf-sm with a larger domain. Largest lf-sm refers to the largest domain size that occurs in the set of local full scale-measures.

	nominal	ordinal	interordinal	contranominal	crown
local full sm	2342	37	4643	2910	2145
maximal lf-sm	527	37	2550	1498	2145
largest lf-sm	9	1	5	5	6

9.6 Applying Ordinal Motifs to Data Sets

Experiment settings

We demonstrate the applicability of ordinal motifs on real-world data using a medium sized formal context: the *spices planner* data set $\mathbb{K}_{\text{Spices}}$. We conduct our experiment on the dual context, i.e., $\mathbb{K}^d := (M, G, I^{-1})$, to derive ordinal motifs within the spices and food categories.

For our application we employ the standard scales from Section 7.1, as they are the most commonly used. For the family of ordinal scales we include the additional $([n], [n+1], \geq)$ scales whose concept lattice include the empty concept. For the rest of this chapter, we focus on **local** explanations and discuss **global** explanations in Chapter 10 due to the close connection to inverse conceptual scaling.

9.6.1 Ordinal Motifs

Number of ordinal motifs

The number of identified local full scale-measure of the spices data set per standard scale can be found in Table 9.2. In this table we distinguish between local and maximal local (with respect to the heredity). We observe that the spices data set entails a large number of ordinal motifs. The interordinal scale motifs are the most frequent in both cases, i.e., local and maximal local. For crown scales both values are equally 2145, since crown scales do not have the heredity property. All found ordinal scale motifs are trivial, i.e., all 37 found motifs are of size 1. In the last row of Table 9.2 we printed the size of the largest ordinal motif of the respective kind. Thus, the motif on the most objects is nominal and of size nine. The largest crown is of size six. We depicted all largest motifs in Figures 9.5 to 9.8.

Largest ordinal motifs

9.6.2 Basic Meanings

Basic interpretation

The discovered ordinal motifs allow us to interpret parts of the spices data set in terms of their *basic meaning* of standard scales [74, 80]. In the following we provide basic meanings of the largest local full scale-measure with respect to the found motifs.

Nominal: The food categories *miscellaneous (group), fish (group), potato (group), vegetables (group), meat (group), sauce (group), poultry (group), rice (group)* and *pastries (group)* **form a partition**.

Ordinal: There are no non trivial local full ordinal scale-measures. If this motif would exist in the spices data set, it would **form a rank order**.

Interordinal: The spices and food categories *ginger, mugwort, meat (group), black pepper* and *juniper berries* **form a linear betweenness relation**.

Contranominal: The spices *Thyme, Sweet Paprika, Oregano, Caraway* and *Black Pepper* **form a partition and are independent**.

Crown: The literature, precisely Ganter and Wille [74] and Ganter and Wille [80], does not provide a basic meaning for crowns.

Automatic unsupervised extraction

With ordinal motifs we are able to automatically identify these relations in an unsupervised setting. While some of the found ordinal motifs are not surprising, e.g., that the food categories form a partition, no user had to define this scale or provide the background information that within the set of attributes there are some that cluster the set of dishes. This relation was solely extracted based on structural properties.

Explain basic meanings

The explanations derived from the basic meanings are very limited in their given form and very technical. For example the *linear betweenness relation* encodes that the spices and food categories *ginger **(1)**, mugwort **(2)**, meat (group) **(3)**, black pepper **(4)*** and *juniper berries **(5)*** are ordered in such a way that every interval in this order relation is a closed set, i.e., for the indices $1, \ldots, 5$ every set given by the interval $[i, j]$ with $1 \leq i \leq j \leq 5$ is closed. Such a relation is equivalent to interordinal scales or interval orders from conceptual scaling [74].

Link to measurement science

As demonstrated in Section 9.1.1, do ordinal motifs allow for a far more complex and meaningful explanation of the found sub-structures. With our method, we linked order theory into the realm of measurement science [201]. In measurement theory a domain expert defines scales that encode how data values are structured, how they are interpreted and what operations can used within a scale. With ordinal motifs we identify these scales in an underlying data set in an unsupervised manner. In case scales are of a standardized structure, we can automatically apply their interpretation to the objects that are entailed in an ordinal motif. In future work, we use this link to automatically derive textual explanations for concept lattices.

9.6.3 Textual Explanations

Textual explanations

Due to the large number of ordinal motifs, we first compute a small but meaningful selection using the introduced greedy strategy. In Figure 9.3 we report the extent sizes of selected ordinal motifs. In the left diagram we depict in the abscissa the steps of the greedy selection and in the ordinate the number of newly covered concepts. We report the results for the standard scales individually and combined. For the latter we also experimented with the normalized heuristic. In the right diagram we depict the accumulated values, i.e., the value c.

Initial observations

First we observe that the normalized heuristic does not decrease monotonously in contrast to all other results. From the right diagram we can infer that the crown, interordinal and nominal motifs are unable to cover all extents. The contranominal and the combined scale family took the fewest selection steps to achieve complete extent coverage. These are followed by the normalized heuristic on the combined scale family which took about thirty percent more steps. Out of the other scale families the crown scales achieved the highest coverage followed by the interordinal and nominal scales.

Choice of heuristic

With Figure 9.4 we investigate the influence of the normalization on the greedy selection process. For this we depict the relative proportion of selected scale types up to a step i (abscissa). The left diagram shows the proportions for the standard heuristic and the right reports the proportions for the normalized heuristic. We count ordinal motifs that belong to multiple standard scale families relatively. For example we count the contranominal scale of size three half for the crown family. We see in the first diagram that a majority of the selected ordinal motifs are of contranominal scales. This is not surprising since they have the most

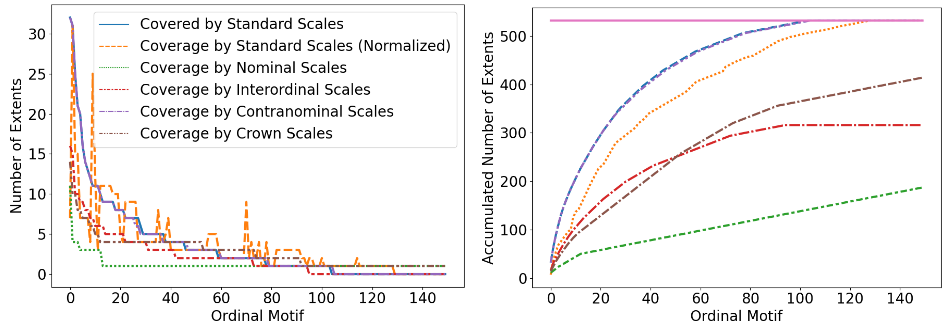


Figure 9.3: The extent coverage (left) for the ordinal motif covering computation for all and each standard scale family individually. The right diagram displays the accumulated coverage at each step in the ordinal motif covering computation. The legend of the left diagram does also apply to the right diagram with the addition of the total number of extents (pink) in the context.

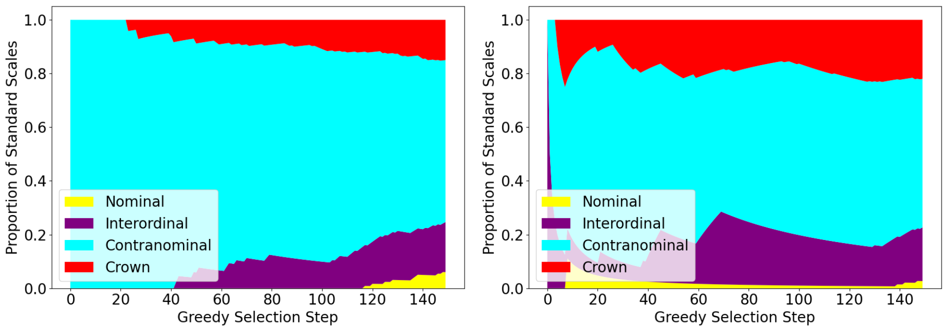


Figure 9.4: The ratio of each standard scale family in the ordinal motif covering computation for the standard (left) and normalized heuristic.

concepts among all standard scales. The interordinal and crown scales are almost equally represented and the nominal motifs are the least frequent. In contrast to this, the normalized heuristic selects crown and interordinal motifs are more frequent (right diagram).

Heuristic trade-offs

Overall, we argue that while the normalized heuristic produces slightly worse coverage scores it provides a more diverse selection in terms of the standard scales. Therefore, the normalized heuristic may result in potentially more insightful explanations.

Results

We conclude by providing automatically generated textual explanations for the spices planner context. For this we report the top ten selections for the standard and normalized heuristic. First we depict the explanations for the standard heuristic which consist solely of contranominal motifs. Thereafter we will turn to the normalized heuristic results.

1. Each combination of the elements *Thyme*, *Sweet Paprika*, *Oregano*, *Caraway* and *Black Pepper* has a unique set of properties they have in common.

2. Each combination of the elements *Curry*, *Garlic*, *White Pepper*, *Curcuma* and *Cayenne Pepper* has a unique set of properties they have in common.

3. Each combination of the elements *Paprika Roses*, *Thyme*, *Sweet Paprika*, *White Pepper* and *Cayenne Pepper* has a unique set of properties they have in common.

4. Each combination of the elements *Paprika Roses*, *Thyme*, *Allspice*, *Curry* and *Curcuma* has a unique set of properties they have in common.

5. Each combination of the elements *Thyme*, *Basil*, *Garlic*, *White Pepper* and *Cayenne Pepper* has a unique set of properties they have in common.

6. Each combination of the elements *Tarragon*, *Thyme*, *Oregano*, *Curry*, and *Basil* has a unique set of properties they have in common.

7. Each combination of the elements *Vegetables*, *Caraway*, *Bay Leef* and *Juniper Berries* has a unique set of properties they have in common.

8. Each combination of the elements *Meat*, *Garlic*, *Mugwort* and *Cloves* has a unique set of properties they have in common.

9. Each combination of the elements *Oregano*, *Caraway*, *Rosemary*, *White Pepper* and *Black Pepper* has a unique set of properties they have in common.

10. Each combination of the elements *Curry*, *Ginger*, *Nutmeg* and *Garlic* has a unique set of properties they have in common.

Observations

These explanations cover a total of 195 concepts out of 532. An interesting observation is that explanation number eight has only four objects compared to the five objects of explanation number nine. Yet, explanation eight was selected first. The reason for this is that number eight has more non-redundant concepts with respect to the previous selections.

Normalization observations

The results for the normalized heuristic are very different compared to the standard heuristic. The ten selected motifs cover a total of 125 concepts. They consist of one interordinal motif, four contranominal, one nominal and four motifs that are crown and contranominal at the same time. For the ordinal motifs that are of crown and contranominal scale we report explanations for both.

1. The elements *Thyme*, *Caraway* and *Poultry* are ordered in such a way that each interval of elements has a unique set of properties they have in common.

2. Each combination of the elements *Curry*, *Garlic*, *White Pepper*, *Curcuma* and *Cayenne Pepper* has a unique set of properties they have in common.

3. Each combination of the elements *Allspice*, *Ginger*, *Mugwort* and *Cloves* has a unique set of properties they have in common.

4. Each combination of the elements *Sweet Paprika*, *Oregano*, *Rosemary* and *Black Pepper* has a unique set of properties they have in common.

5. Each combination of the elements *Sauces*, *Basil* and *Mugwort* has a unique set of properties they have in common.
   The elements *Basil*, *Sauces* and *Mugwort* are incomparable. Furthermore, there is a closed cycle from *Basil* over *Sauces* and *Mugwort* back to *Basil* by pairwise shared properties.

6. Each combination of the elements *Paprika Roses*, *Meat* and *Bay Leef* has a unique set of properties they have in common.
   The elements *Paprika Roses*, *Meat* and *Bay Leef* are incomparable. Furthermore, there is a closed cycle from *Paprika Roses* over *Meat* and *Bay Leef* back to *Paprika Roses* by pairwise shared properties.

7. Each combination of the elements *Saffron*, *Anisey* and *Rice* has a unique set of properties they have in common.
   The elements *Saffron*, *Anisey* and *Rice* are incomparable. Furthermore, there is a closed cycle from *Saffron* over *Anisey* and *Rice* back to *Saffron* by pairwise shared properties.

8. Each combination of the elements *Vegetables*, *Savory* and *Cilantro* has a unique set of properties they have in common.
   The elements *Savory*, *Cilantro* and *Vegetables* are incomparable. Furthermore, there is a closed cycle from *Savory* over *Cilantro* and *Vegetables* back to *Savory* by pairwise shared properties.

9. The elements *Tarragon*, *Potatos* and *Majoram* are incomparable, i.e., all elements have at least one property that the other elements do not have.

10. Each combination of the elements *Paprika Roses*, *Thyme*, *Sweet Paprika*, *White Pepper* and *Cayenne Pepper* has a unique set of properties they have in common.

9.7 Related Work

Ordinal sub-structures

The foundation for ordinal motifs is the identification of unique and meaningful sub-structures and properties. There are many approaches and notions for sub-structures within the realm of FCA. These are either defined on the formal context [63, 90, 127] or within the resulting concept lattice structure [15, 21, 132] and define specific types of sub-structures.

Generalization using scale-measures

In contrast to those, our approach is based on methods from the realm of conceptual data scaling. These allow us to study ordinal sub-structures of concept lattices independently of the elements that generate them, i.e., attributes and incidences in the context. On top of that, the approach generalizes to various types of ordinal sub-structure due to the capabilities of scale-measures. A characterization of specific types of sub-structures with respect to scale-measures would allow for comparing different types of sub-structures in a unified language and may reveal new relations between them. Such an investigation is deemed future work.

Network motifs

In general, the study of motifs in order structures is related to motifs in network science [104, 105, 156] and homomorphism numbers for graph based geometric deep learning [28, 32, 164]. However, in contrast to those, where motifs are recurrent and statistically significant sub-graphs (or patterns), we understand motifs as user-defined set O which allow for special interpretation of the objects, usually represented as formal contexts.

9.8 Discussion

Motifs in lattices

With ordinal motifs we have developed a new approach for the analysis and interpretation of ordinal data. The introduced notions provide a useful extension of notions from the realm of conceptual data scaling and will find applications in Formal Concept Analysis and beyond, independent of ordinal motifs.

Textual explanations

On top of that, we presented, to the best of our knowledge the first approach for the automatic generation of textual explanations of concept lattices. It is a first step towards making Formal Concept Analysis accessible to users without prior training in mathematics. Our contribution comprises the theoretical foundations as well as the preparation of human-centered textual explanations for ordinal motifs of standard scale.

User study

As a next logical step, we envision a participatory user study. This will lead to improved textual explanations for ordinal motifs that are easier to comprehend by humans. Moreover, the development of domain specific textual explanations, or textual explanations that include attributes, may increase the number of applications for our proposed methods.

Algorithmic aspects

While our approach is capable to extract preset frequent recurring ordinal patterns in order structures, there is room for improvement. First, apart from our theoretical considerations on the computational complexity, we did not address the development of specific algorithms. On the one hand, it is certainly possible to find better algorithms than the naive implementations we used in our experiments.

Other motifs

A new line of research would be an extension of the notion of ordinal motifs towards other context-based patterns, such as *clones* [58], *p-clones* [57] or *complements* [180]. Fourth, the new ability to identify standard scales may help a common conceptual data reduction method which is based on nested representations of concept lattices [166].

Improved drawing algorithms

Wille [222] discusses how full scale-measures of a $\mathbb{K}$ into $\mathbb{R}^2$ and $\mathbb{R}^3$ can be used to generate diagrams for concept lattices. We can envision that ordinal motifs can used to compute partial solutions for contexts that do not allow for such scale-measures.

Implications and motifs

In Section 9.2.3 we have shown that there is a relation between the ordinal motif $\mathbb{O}_2$ and implications. Further investigations on this relation can provide useful insights. In particular, an introduction of bases for the set of all ordinal motifs as well as a degree of confidence by which a set of objects satisfies an ordinal motif can provide further applications and insights.

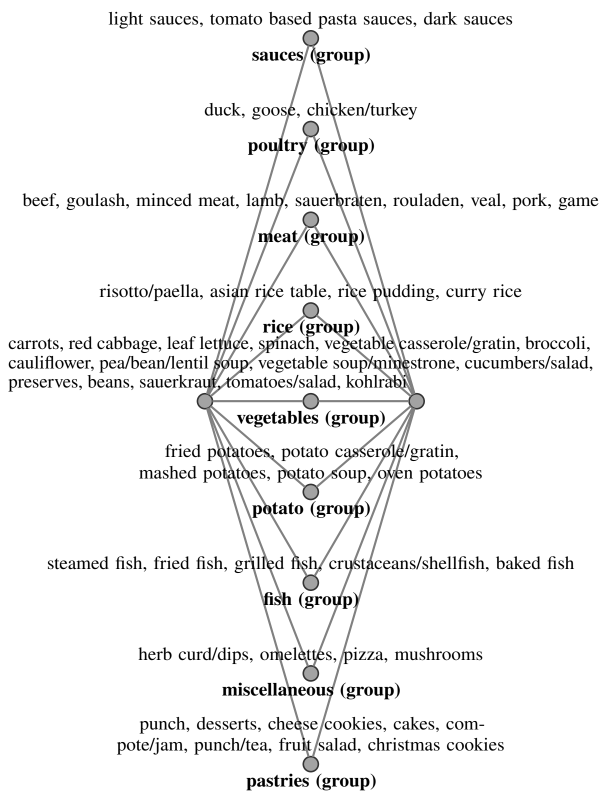


Figure 9.5: The largest local full nominal scale-measure of the spices data sets. We employed the dual context to get conceptual explanations of the attributes (spices). The attributes that induce the local full scale-measure are highlighted with bold font. The diagram was rotated by 90 degrees counter clockwise to improve readability, i.e., the top concept is on the left.

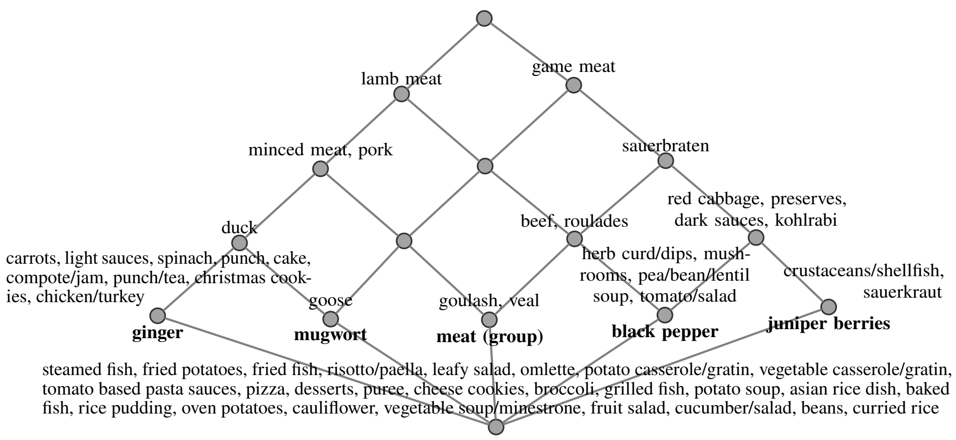


Figure 9.6: The largest local full interordinal scale-measure of the spices data set. We employed the dual context to get conceptual explanations of the attributes (spices). The attributes that induce the local full scale-measure are highlighted with bold font.

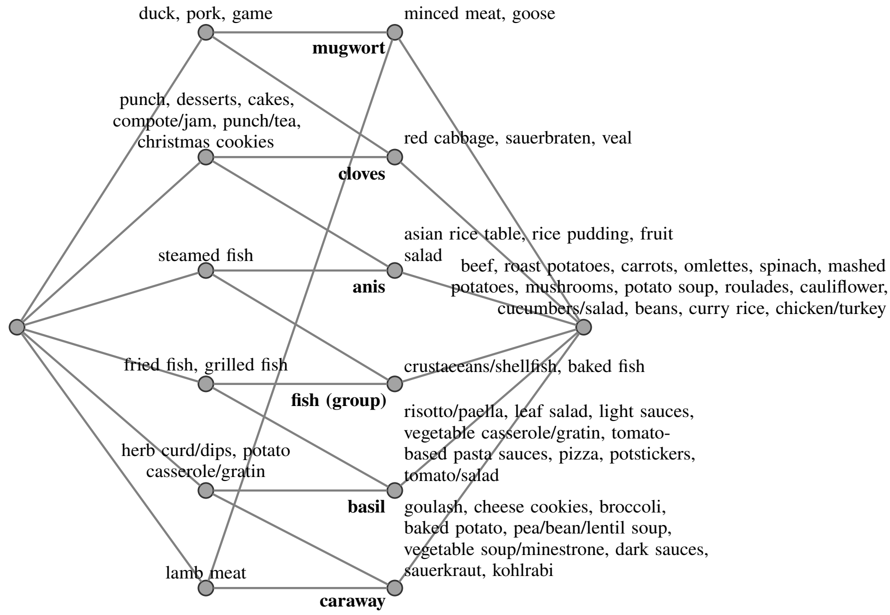


Figure 9.7: The largest local full crown scale-measure of the spices data set. We employed the dual context to get conceptual explanations of the attributes (spices). The attributes that induce the local full scale-measure are highlighted with bold font. The diagram was rotated by 90 degrees counter clockwise to improve readability, i.e., the top concept is on the left.

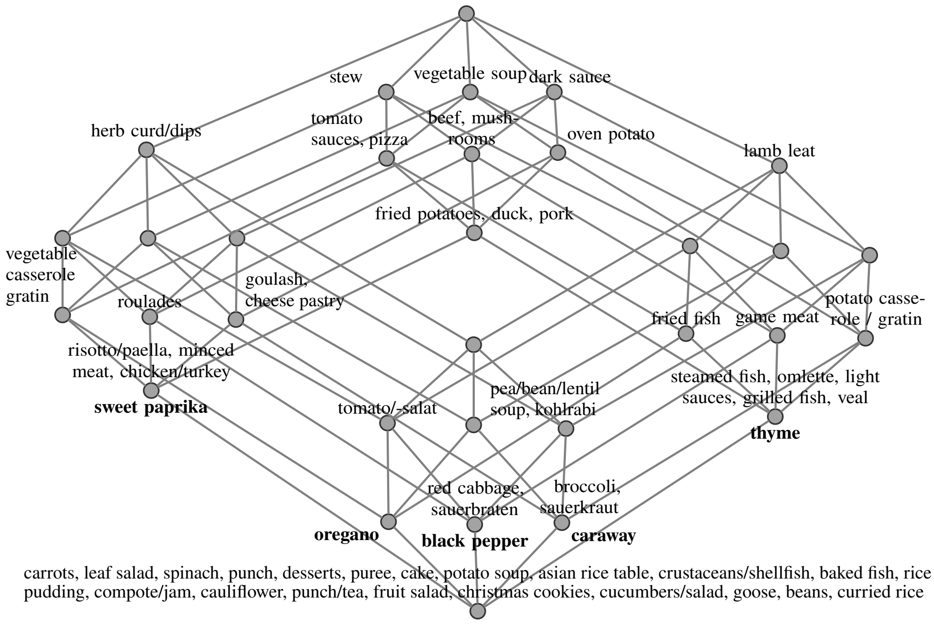


Figure 9.8: The largest local full contranominal scale-measure of the spices data set. We employed the dual context to get conceptual explanations of the attributes (spices). The attributes that induce the local full scale-measure are highlighted with bold font.

10

The Complexity of Conceptual Views

With this chapter, we provide measures to assess the complexity of conceptual data scalings. A commonly used measure of dimensionality in FCA is the *order dimension* of the concept lattice. We show in Section 10.1 that the order dimension of conceptual views is bound from above by the order dimension of the derived context $\mathbb{K}$.

Order dimension of views

A second task in the theory of conceptual scaling is to decide if a given formal context $\mathbb{K}$ is derived (up to isomorphism) from plain scaling. More precisely, one would like to decide whether $\mathbb{K}$ can be derived using a set of given scales, e.g., from interordinal scaling.

Dimension of inverse scaling

An important aspect of this procedure is the complexity of the derived many-valued context $\mathbb{D}$ from inverse scaling. To assess this complexity, we introduce the *scaling dimension* of a context $\mathbb{K}$ which is the smallest number of scales needed to derive $\mathbb{K}$. We study this notion more thoroughly for the special cases of ordinal and interordinal scaling and provide characterizations for them. In addition to our theoretical findings, we demonstrate the applicability of the scaling dimension based on the drive concepts data set.

Scaling dimension

This notion is not only relevant for inverse scaling but also in the study of ordinal motifs (cf. Section 9.1). By a full scale-measure into the semi-product of standard scales, we derive global full explanations of concept lattices.

Complexity of lattice explanations

10.1 Order dimension of conceptual views

The order dimension is a typical (objective) measure for the complexity of lattices. More precisely, the order dimension of the related ordered structure $(L, \leq)$. This quantity is defined using *chain* relations, i.e., subsets of P which are totally ordered. In the following we recall its definition from Section 4.2.

Order dimension

Definition (Order Dimension [80, Definition 82]). *The **order dimension*** $\operatorname{odim}(P)$ *of an ordered set P is equal to the smallest number $d \in \mathbb{N}$ such that there is an order embedding from P into the product of d chains.*

Empirical motivation

In our empirical study in Section 7.2 we observed a decrease in order dimension between the context and one of its conceptual views (see Figure 7.10). These findings lead to question: *Is the order dimension of scale-measures bound by the order dimension of* $\underline{\mathfrak{B}}(\mathbb{K})$*?*.

Order dimension and Ferrers dimension

To formally substantiate our experimental finding we investigate the correspondence between order dimension and scale-hierarchies. For this we employ the Ferrers dimension of a context $\mathbb{K}$, which is equal to the order dimension of $\underline{\mathfrak{B}}(\mathbb{K})$ [80, Theorem 46]. A **Ferrers relation** is a binary relation $F \subseteq G \times M$ such that for $(g,m),(h,n) \in F$ it holds that $(g,n) \notin F \Rightarrow (h,m) \in F$. The **Ferrers dimension** of the formal context $\mathbb{K}$ is equal to the minimum number of ferrers relations $F_t \subseteq G \times M, t \in T$ such that $I = \bigcap_{t\in T} F_t$.

Proposition 31 (Order Dimension in the Scale-Hierarchy). *For a context* $\mathbb{K}$ *and scale-measures* $(\sigma,\mathbb{S}),(\psi,\mathbb{T}) \in \underline{\mathfrak{S}}(\mathbb{K})$ *with* $(\sigma,\mathbb{S}) \lesssim (\psi,\mathbb{T})$*, where* σ *and* ψ *are surjective, it holds that* $\operatorname{odim}(\underline{\mathfrak{B}}(\mathbb{S})) \leq \operatorname{odim}(\underline{\mathfrak{B}}(\mathbb{T}))$.

Proof. We know that $(\sigma,\mathbb{S})$ has the canonical representation $(\iota_G, \mathbb{K}_{\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))})$, cf. Proposition 10, and the same is true for $(\psi,\mathbb{T})$. Since $(\sigma,\mathbb{S}) \lesssim (\psi,\mathbb{T})$ it holds that $\sigma^{-1}(\mathrm{Ext}(\mathbb{S})) \subseteq \psi^{-1}(\mathrm{Ext}(\mathbb{T}))$ and the scale $\mathbb{K}_{\psi^{-1}(\mathrm{Ext}(\mathbb{T}))}$ restricted to the set $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$ as attributes is equal to $\mathbb{K}_{\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))}$. Hence, a Ferrers set F_T such that $\bigcap_{t\in T} F_t$ is equal to the incidence of $\mathbb{K}_{\psi^{-1}(\mathrm{Ext}(\mathbb{T}))}$, can be restricted to the attribute set $\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))$ and is then equal to the incidence of $\mathbb{K}_{\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))}$. Thus, as required, $\operatorname{odim}\left(\underline{\mathfrak{B}}(\mathbb{K}_{\sigma^{-1}(\mathrm{Ext}(\mathbb{S}))})\right) \leq \operatorname{odim}\left(\underline{\mathfrak{B}}(\mathbb{K}_{\psi^{-1}(\mathrm{Ext}(\mathbb{T}))})\right)$. □

Apposition and order dimension

Building on this result we can provide an upper bound for the dimension of the apposition of scale-measures for some formal context $\mathbb{K}$. This also leads to a bound of the order dimension of the join of two conceptual views in the scale-hierarchy.

Proposition 32 (Order Dimension of Scale-Measure Appositions). *For a context* $\mathbb{K}$ *and scale-measures* $(\sigma,\mathbb{S}),(\psi,\mathbb{T}) \in \underline{\mathfrak{S}}(\mathbb{K})$ *with* $(\sigma,\mathbb{S}) \mid (\psi,\mathbb{T}) = (\omega,\mathbb{O})$*. Then is the order dimension of* $\underline{\mathfrak{B}}(\mathbb{O})$ *bound from above by* $\operatorname{odim}\left(\underline{\mathfrak{B}}(\mathbb{O})\right) \leq \operatorname{odim}\left(\underline{\mathfrak{B}}(\mathbb{S})\right) + \operatorname{odim}\left(\underline{\mathfrak{B}}(\mathbb{T})\right)$.

Proof. Without loss of generality we consider for all scale-measures their canonical representation, only. Let F_T be a Ferrers set of the formal context $\mathbb{T}$ such that $\bigcap_{t\in T} F_t = I_{\mathbb{T}}$ and similarly $\bigcap_{s\in S} F_s = I_{\mathbb{S}}$. For any Ferrers relation F of $\mathbb{S}$ it follows that $F \cup (G \times M_{\mathbb{T}})$ is a Ferrers relation of $\mathbb{S} \mid \mathbb{T}$. Hence, the intersection of $\bigcap_{s\in S} F_s \cup (G \times M_{\mathbb{T}})$ and $\bigcap_{t\in T} F_t \cup (G \times M_{\mathbb{S}})$ is a Ferrers set and is equal to $I_{\mathbb{S}\mid\mathbb{T}}$. Since this construction does neither change the cardinality of index set T nor the index set S, the required inequality follows. □

10.2 Scaling Dimension

Recall inverse scaling

In the following, we recall the three results needed for the inverse scaling of contexts from Section 7.4. The first relevant property shows the connection between full scale-measures and plain scaling.

Proposition (Inverse Scaling given $\mathbb{D}$, Proposition 122 [80]). *Let* $\mathbb{D} := (G,M,W,I)$ *be a complete many-valued context and let* $\mathbb{S}_m$*,* $m \in M$*, be scales for the attributes of* M*. Furthermore, let* $\mathbb{K}$ *be the derived context with respect to plain scaling. Then, for every many-valued attribute* $m \in M$ *is the map* $\sigma_m : G \to G_{\mathbb{S}_m}$ *with* $\sigma_m(g) := m(g)$ *a* $\mathbb{S}_m$*-measure of* $\mathbb{K}$*, and* $\mathbb{K}$ *is isomorphic to the induced sub-context* $\mathbb{S}[\sigma(G_{\mathbb{D}}), M_{\mathbb{S}}]$ *of the semi-product of the scales* $\mathbb{S}_m$ *with* $\sigma(g) := (\sigma_m(g))_{m\in M}$.

The second result is a characterization on when a context can be derived from families of standard scales as given by Theorem 5. This result is based on Theorem 55 of the FCA book [80].

Recall measurability theorem

Theorem (Conceptual Measurability Theorem of Contexts). *A context is derivable from plain scaling from a complete many-valued context (up to attribute reduction) and a family of scales by the following conditions:*

(Ordinal)	*Every context is derivable from ordinal scaling.*
(Nominal)	*A context is derivable from nominal scaling iff it is atomistic.*
(interordinal contranominal)	*A context is derivable from interordinal/contranominal scaling iff it is atomistic and the complement of every attribute extent is an extent.*

Third, one needs to determine the many-valued context $\mathbb{D}$ based on the full scale-measure. This result is given in Proposition 9.

Recall inverse plain scaling

Proposition (Inverse Plain Scaling). *Let $\mathbb{K}$ be a formal context, $(\mathbb{S}_j)_{j\in[n]}$ be a family of scales and σ is a full scale-measure from $\mathbb{K}$ into the semi-product $\mathbb{S}$ of all scales with $\sigma(g) = (\sigma_1(g), \ldots, \sigma_n(g))$. Furthermore, let $\hat{\mathbb{K}}$ be the context derived from $\mathbb{D} := \big(G, [n], \big(G_{\mathbb{S}_j}\big)_{j\in[n]}, I\big)$ via plain scaling and scales $(\mathbb{S}_j)_{j\in[n]}$ where*

$$(g, j, v) \in I :\iff \sigma_j(g) = v.$$

The attribute reduced contexts of $\mathbb{K}$ and $\hat{\mathbb{K}}$ are isomorphic.

Equipped with these notions we introduce the *scaling dimension*.

Scaling dimension

Definition 41 (Scaling Dimension). *Let $\mathbb{K} := (G, M, I)$ be a formal context and let $\mathcal{S}$ be a family of scales. The **scaling dimension** of $\mathbb{K}$ with respect to $\mathcal{S}$ is the smallest number d such that there exists a complete many-valued context $\mathbb{D} := (G, M_{\mathbb{D}}, W_{\mathbb{D}}, I_{\mathbb{D}})$ with $|M_{\mathbb{D}}| = d$ and $\mathbb{K}$ has the same extents as the context derived from $\mathbb{D}$ when only scales from $\mathcal{S}$ are used. If no such scaling exists, the dimension remains undefined.*

For a full scale-measure every meet-irreducible element $A \in \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ has to be reflected by at least one of the scale contexts. Thus, we can give an upper bound for the scaling dimension, if it exists, by the number of meet-irreducible elements $|\mathrm{M}(\mathrm{Ext}(\mathbb{K}))| \leq |M_{\mathbb{K}}|$.

The so-defined dimension is related to the feature compression problem. Even when it is known that $\mathbb{K}$ can be derived from a particular many-valued context, it may be the case that there is another, much simpler many-valued context from which one can also derive $\mathbb{K}$ (cf. Figure 10.1 or Figure 10.5 for an example).

Feature compression

10.2.1 Deciding the Scaling Dimension

Determining the scaling dimension is a combinatorial problem whose related decision problem is NP-complete, as can be seen in the following.

Deciding upper bounds

Theorem 7 (Scaling Dimension Complexity). *Deciding for a context $\mathbb{K}$ and a family of scales $\mathcal{S}$ if the scaling dimension is at most $d \in \mathbb{N}$ is NP-complete.*

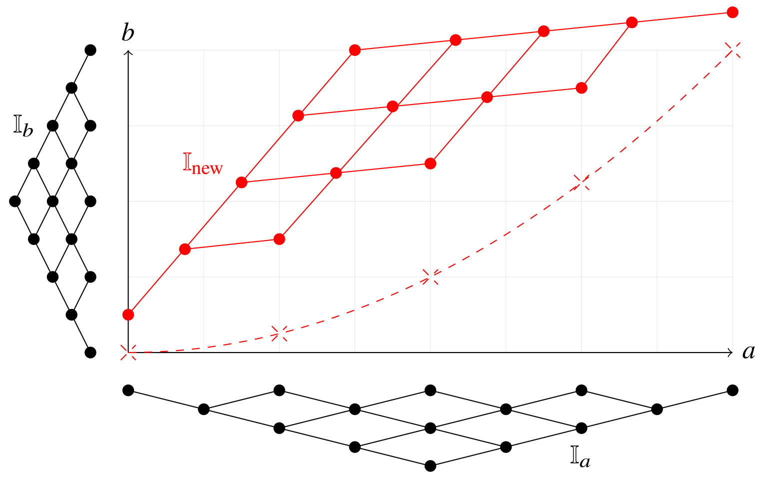


Figure 10.1: This figure displays the a and b feature of five data points and their respective interordinal scales $\mathbb{I}_a$ and $\mathbb{I}_b$ (black). The interordinal scaling dimension of this data set is one and the respective reduced interordinal scale $\mathbb{I}_{\text{new}}$ is depicted in red. The reduction would then remove the a and b many-valued attribute and substitute it for a new attribute given by $\mathbb{I}_{\text{new}}$.

Proof. To show NP-hardness we reduce the recognizing full scale-measure problem (RfSM) Theorem 6 to it. For two input contexts $\hat{\mathbb{K}}$ and $\hat{\mathbb{S}}$ of the RfSM let context $\mathbb{K} := \hat{\mathbb{K}}$. We map $\hat{\mathbb{K}}$ to $\mathbb{K}$ and $\hat{\mathbb{S}}$ to the set of ordinal motifs $\mathcal{S} := \{\hat{\mathbb{S}}\}$ and set $d = 1$. This map is polynomial in the size of the input.

If there is a full scale-measure from $\hat{\mathbb{K}}$ into $\hat{\mathbb{S}}$ we can deduce that there is a full scale-measure of $\mathbb{K}$ into the semi-product that has only one operand and is thus just one element of $\mathcal{S}$. Hence, this element is $\hat{\mathbb{S}}$ and therefore the scaling dimension is at most one. The inverse can be followed analogously.

The scaling dimension is bound from above by the number of attributes. Thus, the scaling dimension problem can be decided in the following way: The scaling dimension is bound from above by d in case $d > |M_{\mathbb{K}}|$. Otherwise, we can non-deterministically guess d scale contexts $\mathbb{S}_1, \ldots, \mathbb{S}_d \in \mathcal{S}$ and d mappings from $\sigma_i = G_{\mathbb{K}} \to G_{\mathbb{S}_i}$. These are polynomial in the size of the input. The verification for full scale-measures in P Problem 14. □

10.2.2 Ordinal Scaling Dimension

Ordinal scaling dimension

A first result on the scaling dimension is easily obtained for the case of ordinal scaling. It was already mentioned that every formal context is fully ordinally measurable, which means that every context is (up to isomorphism) derivable from a many-valued context $\mathbb{D}$ through plain ordinal scaling. With the next proposition we address the problem of how many attributes $\mathbb{D}$ needs to have, i.e., determine the ordinal scaling dimension OSD($\mathbb{K}$) of a context $\mathbb{K}$. An equal result to this special case has been observed by the *grid dimension* [74] or *linear ordinal dimension* [202].

Proposition 33 (Ordinal Scaling Dimension). *The ordinal scaling dimension of a formal context $\mathbb{K}$ equals the width of the ordered set of meet-irreducible concepts.*

Proof. The width is equal to the smallest number of chains C_i covering the ⊆-ordered set of irreducible attribute extents. From these chains we can construct a complete many-valued context $\mathbb{D}$ with one many-valued attribute m_i per chain C_i. The values of m_i are the elements of the chain C_i, and the order of that chain is understood as an ordinal pre-scaling. The derived context by means of ordinal scaling has exactly the set of all intersections of chain extents as extents (Proposition 120 [80]), i.e., the set of all $\bigcap \mathcal{A}$ where $\mathcal{A} \subseteq C_1 \times \cdots \times C_w$. Those are exactly the extents of $\mathbb{K}$. This implies that the scaling dimension is less or equal to the width.

But the converse inequality holds as well. Suppose $\mathbb{K}$ has ordinal scaling dimension w. Then by Proposition 8 every extent of $\mathbb{K}$ is the intersection of pre-images of extents of the individual scales. For ∩-irreducible extents this means that they must each be a pre-image of an extent from one of the scales. Incomparable extents cannot come from the same (ordinal) scale, and thus the scaling must use at least w many ordinal scales. □

Ordinal scaling dimension and order dimension

As a proposition we obtain that the ordinal scaling dimension must be at least as large as the order dimension:

Proposition 34 (Ordinal Scaling Dimension and Order Dimension). *For a context $\mathbb{K}$ is the order dimension of the concept lattice $\underline{\mathfrak{B}}(\mathbb{K})$ a lower bound for the ordinal scaling dimension of $\mathbb{K}$.*

Proof. It is well known that the order dimension of $\underline{\mathfrak{B}}(\mathbb{K})$ equals the Ferrers dimension of $\mathbb{K}$ [80, Theorem 46], which remains the same when $\mathbb{K}$ is the standard context. The Ferrers relation is the smallest number of staircase-shaped relations to fill the complement of the incidence relation of $\mathbb{K}$.

For a context with ordinal scaling dimension equal to w we can conclude that the (irreducible) attributes can be partitioned into w parts, one for each chain, such that for each part the incidence is staircase-shaped, and so are the non-incidences. Thus we can derive w Ferrers relations to fill all non-incidences. □

Different dimensions

A simple example that order dimension and ordinal scaling dimension are not necessarily equal is provided by the $\mathbb{N}_3$ context. Its Ferrers dimension is two, but there are three pairwise incomparable irreducible attributes, which forces its ordinal scaling dimension to be three.

Contra-ordinal scales

Similar investigations based on contra-ordinal scaling with scales $(P, \ngeq)$ and $P = G_{\mathbb{D}}$ can be found in Strahringer and Wille [202] and are related to a coloring problem in hypergraphs.

10.2.3 Nominal Scaling Dimension

Partition covering

The next scaling dimension variant in the order of measurability (see Theorem 5) is the nominal scaling dimension. This dimension is closely related to coverings of set systems by partitions, i.e., sets of disjoint sets.

Proposition 35 (Nominal Scaling Dimension). *The nominal scaling dimension of $\mathbb{K}$, if it exists, is equal to the minimum number of partitions $\mathcal{N}_1, \ldots, \mathcal{N}_d \subseteq \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ the union of which is equal to the set of meet-irreducible extents $\mathrm{M}(\mathrm{Ext}(\mathbb{K}))$.*

Proof. For a context $\mathbb{K}$ with defined nominal scaling dimension, we can infer the following: $\mathbb{K}$ is atomistic (Theorem 5) and there exists a full scale-measure into the semi-product of nominal scales $\mathbb{S}_1, \ldots, \mathbb{S}_d$ (Proposition 9 and Proposition 8). WLoG, $\mathbb{K}$ and the scales $\mathbb{S}_i$ have more than two objects.

Let N_i be the set of extents reflected by $\mathbb{S}_i$ without the top and bottom elements, i.e., without $G, \emptyset$. For all N_i it holds that N_i is a partition of G. Since σ is a full scale-measure, there is for every $A \in \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ a N_i with $A \in N_i$. Thus, there are d subsets $N_i \subseteq \mathrm{Ext}(\mathbb{K})$, each forming a partition on G and $\bigcup_{1\leq i\leq d} N_i \supseteq \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$. The equality can be achieved by omitting elements that are not in $\mathrm{M}(\mathrm{Ext}(\mathbb{K}))$.

Conversely, suppose there are d sets of extents $\mathcal{N}_1, \ldots, \mathcal{N}_d$ of $\mathbb{K}$ with $\bigcup_{1\leq i\leq d} N_i = \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ and each N_i is composed of pairwise disjoint sets. Such a covering exists for any context, e.g., by the sets $N_i = \{A_i\}$ for each $A_i \in \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$. For each $N_i := \{A_1, \ldots, A_k\}$ let B_i be the set of objects not in N_i, i.e., $B_i := G \setminus \bigcup_{1\leq j\leq k} A_k$. By $\hat{N}_i$ we define the set $N_i \cup \{\{g\} \mid g \in B_i\}$. Since the elements of $\hat{N}_i = \{A_1, \ldots, A_{|\hat{N}_i|}\}$ are disjoint is the following a map: $\sigma_i : G \to [|\hat{N}_i|]$ with $\sigma(g) = j$ iff $g \in A_j$. Moreover, since $\mathbb{K}$ is atomistic we can follow that $\hat{N}_i \subseteq \mathrm{Ext}(\mathbb{K})$. Thus, σ_i is a scale-measure of $\mathbb{K}$ into $\mathbb{N}_{|\hat{N}_i|}$ (cf. Lemma 5). Due to $\bigcup_{1\leq i\leq d} N_i = \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ we can follow that $g \mapsto (\sigma_1, \ldots, \sigma_d)$ is a full scale-measure of $\mathbb{K}$ into the semi-product of d nominal scales. □

By the following proposition, we provide an upper and lower bound for the interordinal scaling dimension.

Proposition 36 (Nominal Scaling Dimension Bounds). *For a context $\mathbb{K} := (G, M, I)$ is the nominal scaling dimension, if it exists, bounded from below by*

$$\max_{g\in G}\left|\{A \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid g \in A\}\right|$$

and from above by

$$\max_{A\in\mathrm{M}(\mathrm{Ext}(\mathbb{K}))}\left|\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid A \cap B \neq \emptyset\}\right|$$

Proof. Let $\mathbb{K}$ be a formal context for which the nominal scaling dimension is defined. We can infer from Proposition 35 that the nominal scaling dimension of $\mathbb{K}$ is equal to the minimum number of partitions $\mathcal{N}_1, \ldots, \mathcal{N}_d \subseteq \mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ the union of which is equal to $\mathrm{M}(\mathrm{Ext}(\mathbb{K}))$.

[≥] : Distinct sets $A_1, \ldots, A_k \in \mathrm{Ext}(\mathbb{K})$ that have a common element $x \in A_1 \cap \cdots \cap A_k$ need to be in different partitions N_i. Thus, there have to be at least $\max_{x\in G}\left|\{A \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid x \in A\}\right|$ many partitions.

[≤] : In the following, let $\mathrm{ord}\big(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\big) := \max_{A\in\mathrm{M}(\mathrm{Ext}(\mathbb{K}))}\left|\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid A\cap B\neq\emptyset\}\right|$. For the upper bound, we first show that we can build a partition N_1 such that the remaining $\mathrm{ord}\left(\mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus N_1\right)$ is less than or to $\mathrm{ord}\left(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\right) - 1$. Let A_1 be a meet-irreducible extent of $\mathbb{K}$ with $\left|\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid A_1 \cap B \neq \emptyset\}\right| = \mathrm{ord}\left(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\right)$. Now select elements $A_2, \ldots, A_k$ from $\mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \{A_1\}$ in an iterative manner, such that (1) for each A_i the equality $|\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid A_i \cap B \neq \emptyset\}| = \mathrm{ord}\left(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\right)$ holds and (2) for all $1 \leq j < i$ it holds that $A_i \cap A_j = \emptyset$. We repeat this selection until there is no $D \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \{A_1, \ldots, A_k\}$ left that satisfies (1) and (2). In the case that there is no element D in $\mathrm{M}(\mathrm{Ext}(\mathbb{K})) \setminus \{A_1, \ldots, A_k\}$ with $\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid D \cap B \neq \emptyset\} = \mathrm{ord}\big(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\big)$ left, we have shown the decrease of $\mathrm{ord}\big(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\big)$ by at least 1. Suppose there is such an element D, then there is at least one selected A_i with (2) $A_i \cap D \neq \emptyset$. Thus, the inequality $|\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K}))\setminus\{A_1, \ldots, A_k\} \mid D\cap B \neq \emptyset\}| \leq |\{B \in \mathrm{M}(\mathrm{Ext}(\mathbb{K})) \mid D\cap B \neq \emptyset\}|-1$ holds. Per recursion follows that all elements of $\mathrm{M}(\mathrm{Ext}(\mathbb{K}))$ are covered after at most $\mathrm{ord}\left(\mathrm{M}(\mathrm{Ext}(\mathbb{K}))\right)$ selection steps. □

10.2.4 Interordinal Scaling Dimension

A more challenging problem is to determine the interordinal scaling dimension of a context $\mathbb{K}$. We investigate this with the help of the following definition.

Extent ladders and interordinal scales

Definition 42. *An **extent ladder** of $\mathbb{K}$ is a set $\mathcal{R} \subseteq \text{Ext}(\mathbb{K})$ of non-empty extents that satisfies:*

i) the ordered set $(\mathcal{R}, \subseteq)$ has width ≤ 2, i.e., $\mathcal{R}$ does not contain three mutually incomparable extents, and

ii) $\mathcal{R}$ is closed under complementation, i.e., when $A \in \mathcal{R}$, then also $G \setminus A \in \mathcal{R}$.

Note that a (non-empty) extent ladder is the disjoint union of two chains of equal cardinality, for the following reason: Consider a minimal extent E in the ladder. Any other extent must either contain E or be contained in the complement of E, because otherwise there would be three incomparable extents. The extents containing E must form a chain, and so do their complements, which are all contained in the complement of E.

Interordinal scaling dimension

Based on the definition of extent ladders, we are able to derive a characterization for the interordinal scaling dimension $\text{ISD}(\mathbb{K})$ of a context $\mathbb{K}$.

Theorem 8 (Interordinal Scaling Dimension). *The interordinal scaling dimension of a formal context $\mathbb{K}$, if it exists, is equal to the smallest number of extent ladders, the union of which contains all meet-irreducible extents of $\mathbb{K}$.*

Proof. Let $\mathbb{K}$ be a formal context with interordinal scaling dimension d. WLoG we may assume that $\mathbb{K}$ was derived by plain interordinal scaling from a complete many-valued context $\mathbb{D}$ with d many-valued attributes. We have to show that the irreducible attribute extents of $\mathbb{K}$ can be covered by d extent ladders, but not by fewer.

To show that d extent ladders suffice, note that the extents of an interordinal scale form a ladder, and so do their pre-images under a scale-measure. Thus, Proposition 8 provides an extent ladder for each of the d scales, and every extent is an intersection of those. Meet-irreducible extents cannot be obtained from a proper intersection and therefore must all be contained in one of these ladders.

For the converse assume that $\mathbb{K}$ contains l ladders covering all meet-irreducible extents. From each such ladder $\mathcal{R}_i$ we define a formal context $\mathbb{R}_i$, the attribute extents of which are precisely the extents of that ladder, and note that this context is an interordinal scale (up to clarification). Define a many-valued context with l many-valued attributes m_i. The attribute values of m_i are the minimal non-empty intersections of ladder extents, and the incidence is declared by the rule that an object g has the value V for the attribute m_i if $g \in V$. The formal context derived from this many-valued context by plain interordinal scaling with the scales $\mathbb{R}_i$ has the same meet-irreducible extents as $\mathbb{K}$, and therefore the same interordinal scaling dimension. Thus $l \geq d$. □

Ordinal and interordinal scaling dimension

With Proposition 34 we found that the ordinal scaling dimension is equal to the width of meet-irreducible elements. By the following proposition we show the relation of the interordinal scaling dimension to this quantity.

Proposition 37 (Width and Interordinal Scaling Dimension). *Let w denote the width of the ordered set of meet-irreducible extents of the formal context $\mathbb{K}$. The interordinal scaling dimension of $\mathbb{K}$, if defined, is bounded below by $w/2$ and bounded above by w.*

$\mathbb{D}$	m_1	m_2
g_1	1	d
g_2	2	c
g_3	3	b
g_4	4	a

Figure 10.2: Example many-valued context where the attribute values are ordinally pre-scaled by $1 < 2 < 3 < 4$ and $a < b < c < d$. The interordinal scaling dimension of the interordinally scaled context from $\mathbb{D}$ is one and the ordinal scaling dimension of the ordinally scaled context from $\mathbb{D}$ is two.

Proof. An extent ladder consists of two chains, and w is the smallest number of chains covering the meet-irreducible extents. So at least $w/2$ ladders are required.

Conversely from any covering of the irreducible extents by w chains a family of w ladders is obtained by taking each of these chains together with the complements of its extents. □

The last inequality of this proposition, i.e., $\mathrm{OSD}(\mathbb{K}) \leq 2 \cdot \mathrm{ISD}(\mathbb{K})$, results from using each two chains of the extent ladders as ordinal scales, where $\mathrm{OSD}(\mathbb{K})$ denotes the ordinal scaling dimension and $\mathrm{ISD}(\mathbb{K})$ the interordinal scaling dimension of $\mathbb{K}$. This results in an upper bound for the ordinal scaling dimension and a lower bound for the interordinal scaling dimension. A context where $\mathrm{OSD}(\mathbb{K}) \neq 2 \cdot \mathrm{ISD}(\mathbb{K})$ is depicted in the next section in Figure 10.4.

Example

Another inequality that can be found in terms of many-valued contexts. For a many-valued context $\mathbb{D}$ and its ordinal scaled context $\mathbb{O}(\mathbb{D})$ and interordinal scaled context $\mathbb{I}(\mathbb{D})$ is the ISD of $\mathbb{I}(\mathbb{D})$ is in general not equal to the OSD of $\mathbb{O}(\mathbb{D})$. Consider for this the counter example given in Figure 10.2. The depicted many-valued context has two ordinally pre-scaled attributes that form equivalent interordinal scales.

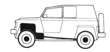
10.3 Small Case Study

Drive concepts data

To consolidate the understanding of the notions and statements on the (interordinal) scaling dimension, we provide an explanation based on a small case example taken from the *drive concepts* [80] data set. This data set is a many-valued context (see Figure 10.3) consisting of five objects, which characterize different ways of arranging the engine and drive chain of a car, and seven many-valued attributes that measure quality aspects for the driver, e.g., *economy of space*. The data set is accompanied by a scaling that consists of a mixture of bi-ordinal scalings of the quality (attribute) features, e.g., *good* < *excellent* and *very poor* < *poor*, and a nominal scaling for categorical features, e.g., for the *steering behavior*. The concept lattice of the scaled context consists of twenty-four formal concepts and is depicted in Figure 10.4.

Scaling dimensions

First, we observe that the concept lattice of the example meets the requirements to be derivable from interordinal scaling (Theorem 5). All objects are annotated to the atom concepts and the complement of every attribute extent is an extent as well. The interordinal scaling dimension of the scaled *drive concept* context is three which is much smaller than the original seven many-valued attributes. Using the extent ladder characterization provided in Theorem 8 we highlighted three extent ladders in color in the concept lattice diagram (see Figure 10.4). The first and largest extent ladder (highlighted in red) can be inferred from the outer most concepts and covers sixteen out of twenty-four concepts. The remaining two extent ladders have only two elements and are of dichotomic scale.

Conventional	Front-wheel	Rear-wheel	Mid-engine	All-wheel
				

$\mathbb{D}$	De	Dl	R	S	E	C	M
Conventional	poor	good	good	understeering	good	medium	excellent
Front-wheel	good	poor	excellent	understeering	excellent	very low	good
Rear-wheel	excellent	excellent	very poor	oversteering	poor	low	very poor
Mid-engine	excellent	excellent	good	neutral	very poor	low	very poor
All-wheel	excellent	excellent	good	understeering/neutral	good	high	poor

$\mathbb{S}_{De}, \mathbb{S}_{Dl}, \mathbb{S}_{R}, \mathbb{S}_{E}, \mathbb{S}_{M}$	++	+	-	- -
excellent	×	×		
good		×		
poor			×	
very poor			×	×

$\mathbb{S}_{S}$	u	o	n	u/n
u	×			
o		×		
u			×	
u/n				×

$\mathbb{S}_{C}$	vl	l	m	h
vl	×	×		
l		×		
m			×	
h			×	×

$\mathbb{K}$	De				Dl				R				S				E				C				M			
	++	+	-	- -	++	+	-	- -	++	+	-	- -	u	o	n	u/n	++	+	-	- -	vl	l	m	h	++	+	-	- -
Conventional			×			×				×			×					×					×		×	×		
Front-wheel		×					×		×	×			×				×	×			×	×				×		
Rear-wheel	×	×			×	×					×	×		×					×			×				×		
Mid-engine	×	×			×	×				×					×				×	×		×					×	×
All-wheel	×	×			×	×				×						×		×					×	×			×	

Figure 10.3: The driving concepts (first column) many-valued context (second column), scales for each attribute (third column) and the derived context (forth column) from Ganter and Wille [80].

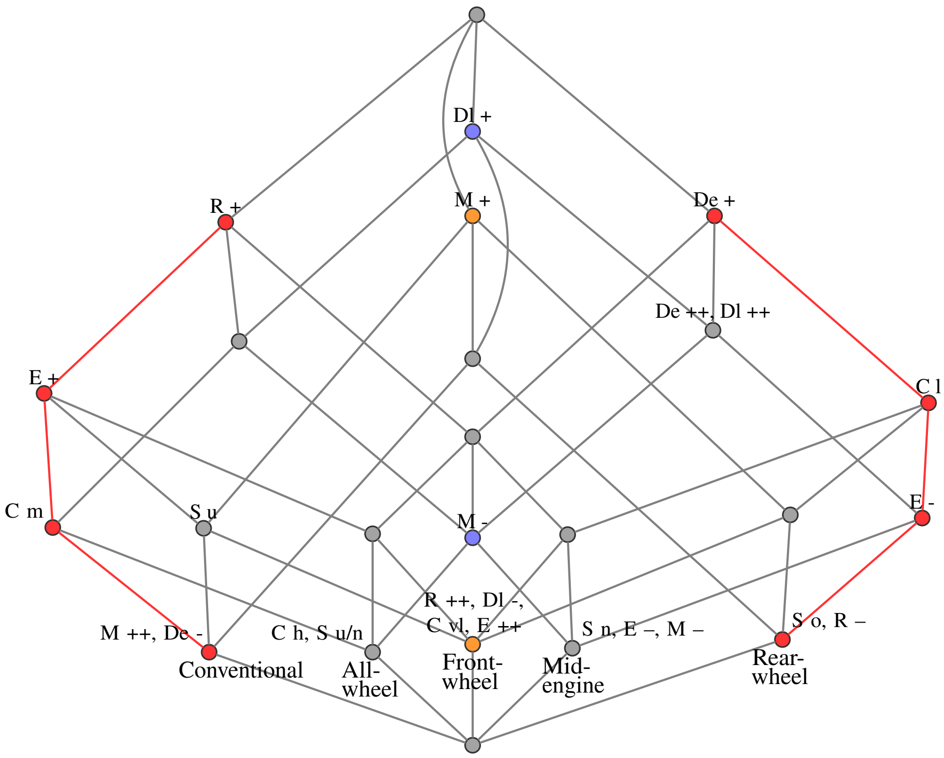


Figure 10.4: Concept lattice for the context of drive concepts (cf. Figure 1.13 and 1.14 in Ganter and Wille [80]). The extent ladders indicating the three interordinal scales are highlighted in color. The ordinal scaling dimension of this context as well as the order dimension of its concept lattice are four.

$\mathbb{D}$	$\mathbb{I}_1$	$\mathbb{I}_2$	$\mathbb{I}_3$
Conventional	M++, DE-;	Dl+ ;	M+ ;
Front-wheel	E+ ; Cl	; R++, Dl-, Cvl, E++	M+ ;
Rear-wheel	; So, R-	Dl+ ;	M+ ;
Mid-engine	R+ ; E-	Dl+ ;	; M-
All-wheel	Cm ; De+	Dl+ ;	; M-

$\hat{\mathbb{D}}$	$\mathbb{I}_1$	$\mathbb{S}_{\mathbb{I}_1}$								
Conventional	M++, DE-;	M++, DE-;	×	×	×	×				
Front-wheel	E+ ; Cl	Cm ; De+		×	×	×	×			
Rear-wheel	; So, R-	E+ ; Cl			×	×	×	×		
Mid-engine	R+ ; E-	R+ ; E-				×	×	×	×	
All-wheel	Cm ; De+	; So, R-					×	×	×	×

Figure 10.5: Equivalent driving concepts data table (top) with compressed features and the largest interordinal scale (bot).

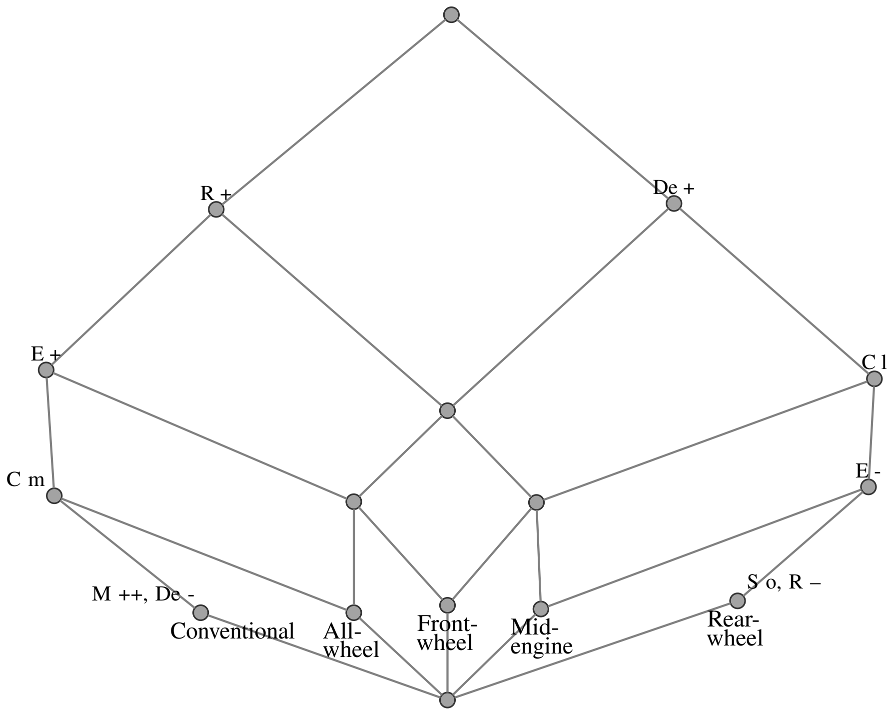


Figure 10.6: Conceptual view of the drive concepts data set on the largest scale with respect to inverse scaling.

Compressed representation

Based on the inverse scaling as given in Proposition 9 we are able to compute a complete many-valued context $\mathbb{D}$ from which the drive concepts data set is derived. In our experiment, we found that the interordinal scaling dimension equals three. Thus, there is a context $\mathbb{D}$ with three many-valued attributes. This is in contrast to the original seven many-valued attributes and provides a compressed representation of the data. In Figure 10.5 (top) we provide such a context. The attribute values of this data set encode the derivation of an object with respect to the two chains of the extent ladders. The elements of the tuple are split by a **;** symbol instead of the tuple notation for a simpler visualization. The scales for each many-valued attribute are the interordinal scales generated by the extent ladders as seen in Figure 10.5 (bot).

Principle components

The computed scales allows for a decomposition of $\mathbb{K}$ into *conceptual components*, i.e., scale contexts. This is similar in spirit to methods like principle component analysis. By selecting the largest scales we result in a conceptual view with reduced complexity. We depict in Figure 10.6 the reduction of the data on to the largest scale. Due to the given analogy, we call this approach **principle ordinal component analysis**.

10.4 Discussion

Applications in data science

The presented results on the scaling dimension have a number of interfaces and correspondences to classical data science methods. A natural link to investigate would be comparing the scaling dimension with standard correlation measures. Two features that correlate perfectly, e.g., Figure 10.1, induce an equivalent conceptual scaling on the data. An analog of the scaling dimension in this setting would be the smallest number of independent features. Or, less strict, the smallest number of features such that these features do not correlate more than some parameter. This obvious similarity of both methods is breached by a key advantage of our approach. In contrast to correlation measures, our method relies solely on ordinal properties [201] and does not require the introduction of measurements for distances or ratios.

Principle components

Another application for of the scaling dimension and the inverse plain scaling are given by our example for a *conceptual principle component analysis* in Section 10.3. While our findings seem very promising for data reduction, it remains to be seen if they can be used in other applications like machine learning.

Global explanations of concept lattices

Our contribution to inverse conceptual scaling extend our studies on ordinal motifs and explanations of concept lattices (cf. Chapter 9). With a full scale-measure into the semi-product of standard scales, we derive global full explanations of concept lattices. The new introduced notion of dimensionality can be used to asses the complexity of concept lattices with respect to their explainability.

Other notions of dimensions

Proposition 34 has already shown that there is a relationship between an aspect of the scaling dimension of a formal context and the order dimension of its concept lattice. The assumption that further such relationships may exist is therefore reasonable. An investigation on how the scaling dimension relates to other measures of dimension within the realm of FCA [94, 212] is therefore deemed future work.

Computational aspects

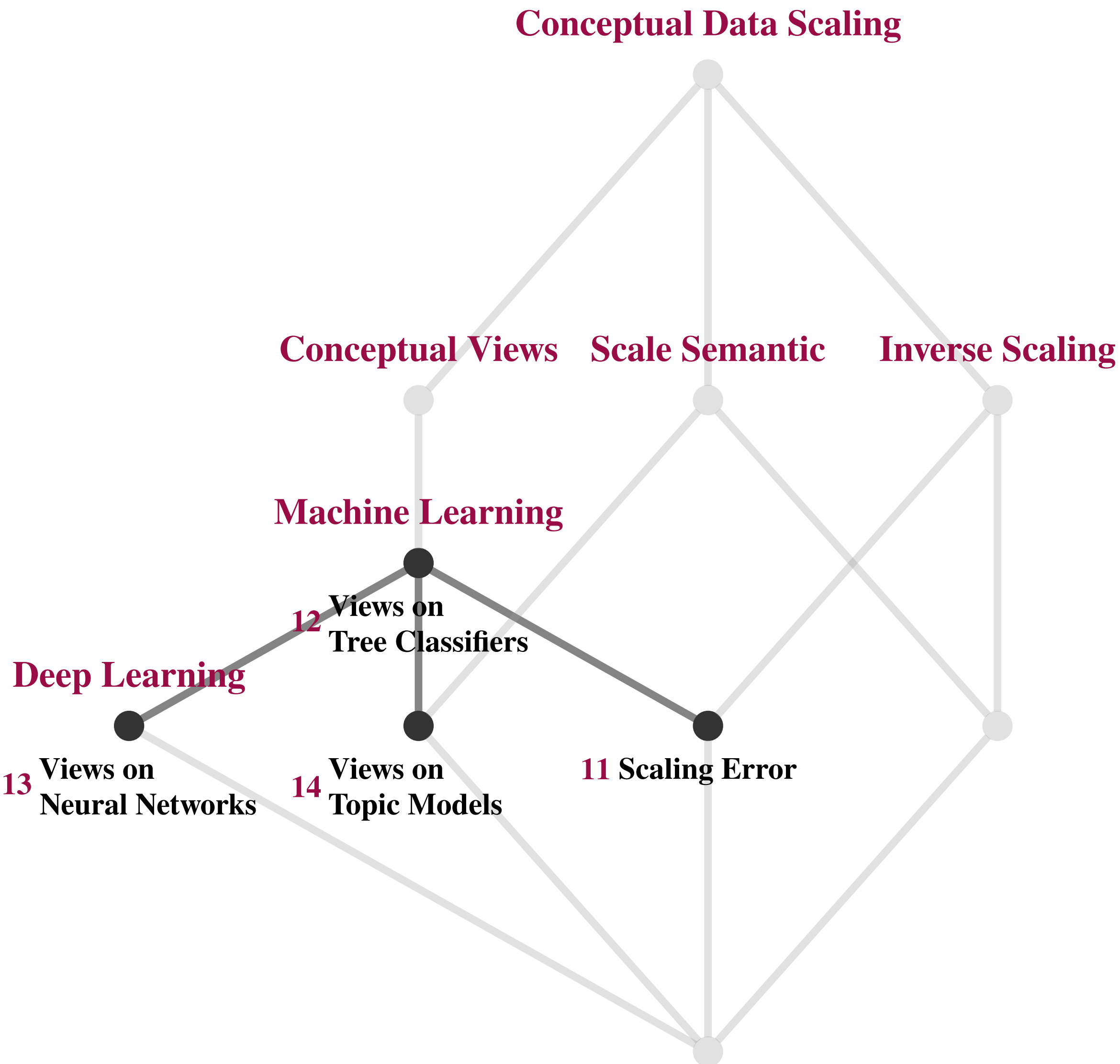
Despite the complexity of deciding the scaling dimension, efficient algorithms for real-world data that compute the scaling dimension and its specific versions, i.e., ordinal, interordinal, nominal, etc, need to be developed. In addition to that, so far it is unknown if an approximation of the scaling dimension, e.g., with respect to some degree of *conceptual scaling error* Chapter 11 or other bounds, is tractable. If computationally feasible, such an approximation could allow larger data sets to be handled.

Improved drawing algorithms

Another line of research that can be pursued in future work is how the scaling dimension can be utilized to derive more readable line diagrams. We can envision that diagrams of concept lattices that are composed of fewer scales, i.e., have a lower scaling dimension, are more readable even if they have slightly more concepts.

Part III

Conceptual Views in Machine Learning



11

Conceptual Scaling Error in Dimensionality Reduction

The analysis of large and complex data sets is presently a challenge for many data driven research fields. This is especially true when using sophisticated analysis and machine learning methods due to their computational complexity. One aspect of largeness and complexity is the explicit data dimension, e.g., number of features, of a data set. Therefore, a variety of methods have been developed to reduce exactly this data dimension to a computable size, such as *Latent Semantic Analysis* [46, 62], *principle component analysis* or *binary matrix factorization* [22, 155, 230], or other embedding techniques like *multidimensional scaling* [152, 157].

Numeric reduction methods in FCA

These methods are mainly developed to compute lower dimensional representations that preserve a notion of distance/similarity of objects or allow for the reconstruction of the original data set. This optimization criterion is often sufficient for numeric matrix representations and statistic analysis methods. This however does not need to be the case when analyzing patterns in the data such as formal concepts.

Error in conceptual data reduction

To assess the quality of data reduction methods with respect to conceptual data reduction (Figure 7.1, bottom right), we introduce a new notion of error. This is called *conceptual scaling error* and is based on our definition for consistent conceptual data reduction from Chapter 8. The conceptual scaling error is the set of all closed object sets that contradict the (conceptual) continuity of the underlying reduction.

Experiment

In a small case study, we evaluate the newly introduced measure based on *Boolean factor analysis* (BFA) [230] of formal contexts and compare it to standard evaluation scores. BFA studies the factorization of binary matrices with values 0 (false $\bot$) or 1 (true $\top$). For example, given the binary data set matrix K, the application of a BFA yields two binary data matrices S, H of lower dimension, such that $S \cdot H$ approximates K with respect to a previously selected norm $|\cdot|$. The product $\cdot$ is the usual matrix product with $1 +_b 1 = 1$ ($\top \vee \top = \top$). The factor S can be considered as a lower dimensional representation of K, i.e., a data reduction of K. The connection between the scaling features of S and the original

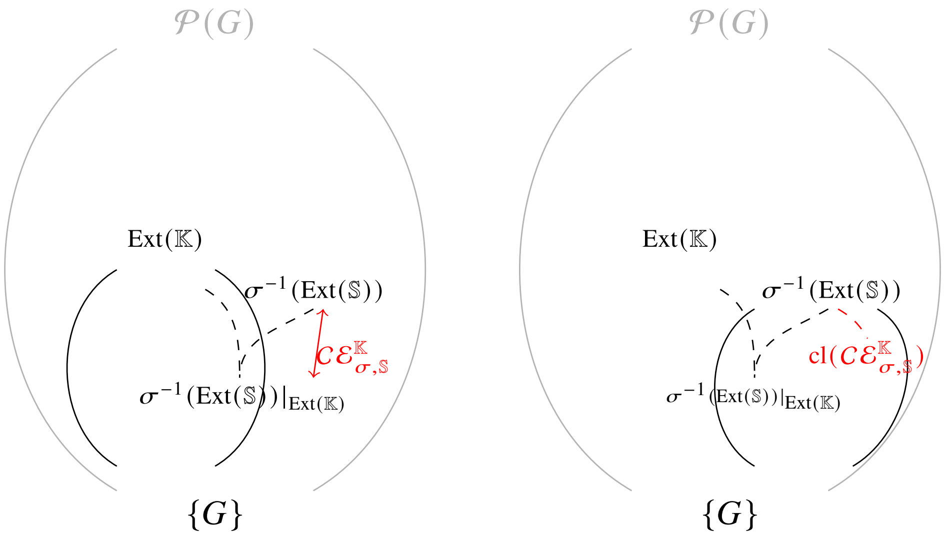


Figure 11.1: The conceptual scaling error and the consistent part of $(\sigma, \mathbb{S})$ in $\underline{\mathfrak{S}}(\mathbb{K})$ (left). The right represents both parts as scale-measures of $\mathbb{S}$.

data features of K is represented by H.

Results

Our experiment shows that even for reductions with seemingly good matrix similarities [53, 128], the conceptual scaling error can be quite high. This shows that – in FCA – one should not solely rely on such evaluation measures but also compare the resulting concept lattices. An advantage of our method is that it is agnostic to the used reduction method. On top of that it is not limited to compare contexts of same size, i.e., equal set of objects and attributes. The only requirement is that we have a map σ from the objects from $\mathbb{K}$ to $\mathbb{S}$. Usually this map is equal to the identity map in numeric matrix reductions.

11.1 Conceptual Scaling Error

In Section 7.2 we characterized conceptual data reductions that respect the conceptual structure of a context $\mathbb{K}$ with scale-measures. The context $\mathbb{S}$ satisfies this criterion iff its extentional structure is entailed in $\mathbb{K}$, i.e., $\sigma^{-1}(\text{Ext}(\mathbb{S})) \subseteq \text{Ext}(\mathbb{K})$. The extents in $\mathbb{S}$ give rise to a natural notion of error in the data reduction.

Definition 43 (Conceptual Scaling Error). *Let $\mathbb{K}, \mathbb{S}$ be contexts and $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$. The* conceptual scaling error *of $(\sigma, \mathbb{S})$ with respect to $\mathbb{K}$ is the set*

$$\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}} := \sigma^{-1}(\text{Ext}(\mathbb{S})) \setminus \text{Ext}(\mathbb{K}).$$

Conceptual scaling error

The conceptual scaling error $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ consists of all pre-images of closed object sets in $\mathbb{S}$ that are not closed in the context $\mathbb{K}$, i.e., the object sets that contradict the scale-measure criterion. Hence, $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}} = \emptyset$ iff $(\sigma, \mathbb{S}) \in \mathfrak{S}(\mathbb{K})$. By $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})} := \sigma^{-1}(\text{Ext}(\mathbb{S})) \cap \text{Ext}(\mathbb{K})$ we denote the set of consistently reflected closed object sets of $\mathbb{S}$ by σ. This set can be represented as the intersection of two closure systems and is thereby a closure system as well. Using this notation together with the canonical representation Proposition 10 we can provide a scale-measure that reflects exactly the consistent part of a conceptual data reduction.

View of consistent concepts

Corollary 6 (View of Consistent Concepts). *For* $\mathbb{K},\mathbb{S}$ *and* $\sigma : G_{\mathbb{K}} \mapsto G_{\mathbb{S}}$*, there exists a scale-measure* $(\psi,\mathbb{T}) \in \mathfrak{S}(\mathbb{K})$ *with* $\psi^{-1}(\text{Ext}(\mathbb{T})) = \sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$.

Scaling error is not a closure system

The conceptual scaling error $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ does not constitute a closure system on G, since it lacks the top element G. Moreover, the meet of elements $A, D \in \mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ can be closed in $\mathbb{K}$ and thus $A \wedge D \notin \mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$. The relation between the data reduction $(\sigma,\mathbb{S})$ of a $\mathbb{K}$ and the conceptual scaling error is visualized Figure 11.1 with respect to the scale-hierarchy of $\mathbb{K}$ within the lattice of all closure systems on G. On the right we depicted by $\text{cl}(\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}})$ the smallest closure system that entails the conceptual scaling error. The map cl is the closure operator in $\underline{\mathfrak{S}}(\mathbb{S})$

Attributes that cause scaling error

To pinpoint the cause of inconsistencies in conceptual data reduction we may investigate the scale's attributes using Proposition 7.

Proposition (Attribute Scale-Measures). *Let* $\mathbb{K} = (G, M, I)$ *and* $\mathbb{S} = (G_{\mathbb{S}}, M_{\mathbb{S}}, I_{\mathbb{S}})$ *be two formal contexts and* $\sigma : G \to G_{\mathbb{S}}$*, then TFAE:*

i) σ *is an* $\mathbb{S}$*-measure of* $\mathbb{K}$

ii) σ *is a* $\big(G_{\mathbb{S}}, \{n\}, I_{\mathbb{S}} \cap (G_{\mathbb{S}} \times \{n\})\big)$*-measure of* $\mathbb{K}$ *for all* $n \in M_{\mathbb{S}}$

Attribute scaling error

Based on this result, we can decide if $(\sigma,\mathbb{S})$ is a scale-measure of $\mathbb{K}$ and thus if $|\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}| = 0$ solely based on the attribute extents of $\mathbb{S}$. In turn this enables us to determine the particular attributes n that cause conceptual scaling errors, i.e. $\sigma^{-1}(\{n\}^{I_{\mathbb{S}}}) \notin \text{Ext}(\mathbb{K})$.

Definition 44 (Attribute Scaling Error). *Let* $\mathbb{K},\mathbb{S}$ *be contexts and* $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$*. The* ***attribute scaling error*** *of* $(\sigma,\mathbb{S})$ *with respect to* $\mathbb{K}$ *is the set*

$$\mathcal{AE}^{\mathbb{K}}_{\sigma,\mathbb{S}} \coloneqq \{m \in M_{\mathbb{S}} \mid \sigma^{-1}(\{m\}^{I_{\mathbb{S}}}) \notin \text{Ext}(\mathbb{K})\}.$$

Consistent attribute view

An advantage of the attribute scaling error is that it is easier to compute compared to the conceptual scaling error. The consistent part with respect to the attribute scaling error yields the following scale-measure.

Corollary 7 (View of Consistent Attributes). *For two formal contexts* $\mathbb{K},\mathbb{S}$ *and map* $\sigma : G_{\mathbb{K}} \to G_{\mathbb{S}}$ *let the set* $N = \{m \in M_{\mathbb{S}} \mid \sigma^{-1}(\{m\}^{I_{\mathbb{S}}}) \in \text{Ext}(\mathbb{K})\}$*. Then is the pair* $\big(\sigma, \mathbb{S}[G_{\mathbb{S}}, N]\big)$ *a scale-measure of* $\mathbb{K}$.

Proof. Follows directly from applying Proposition 7. □

Attribute and conceptual scaling error

The by Corollary 7 constructed scale-measure does not necessarily reflect all extents in $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$. For this, consider the example $\mathbb{N}_3 \coloneqq (\{1,2,3\},\{1,2,3\},=)$ with $\mathbb{B}_3 \coloneqq (\{1,2,3\},\{1,2,3\},\neq)$ and the map $\iota_{\{1,2,3\}}$ from $\mathbb{N}_3$ to $\mathbb{B}_3$. The error set is equal to $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}} = \binom{\{1,2,3\}}{2}$. Hence, none of the scale-attributes $M_{\mathbb{S}} = \{1,2,3\}$ fulfills the scale-measure property. By omitting the whole set of attributes $M_{\mathbb{S}}$, we result in the context $(G,\{\},\{\})$ whose set of extents is equal to $\{G\}$. The set $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$ however is equal to $\big\{\{\},\{1\},\{2\},\{3\},\{1,2,3\}\big\}$.

11.1.1 Representation and Structure of Conceptual Scaling Errors

So far, we apprehended $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ as the set of erroneous pre-images. However, the conceptual scaling error may be represented as a part of a scale-measure. In the following, we present three approaches on how to deal with conceptual scaling error.

Highlight error

i) The first approach is to analyze the extent structure of $\sigma^{-1}(\text{Ext}(\mathbb{S}))$. The conceptual scaling error $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ is a subset of the reflected extents of $(\sigma, \mathbb{S})$ and can be highlighted in the concept lattice diagram.

Split error from consistent part

ii) The second approach is based on our result in Corollary 6. The conceptual scaling error $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ cannot be represented as scale-measure of $\mathbb{K}$. However, since $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}} \subseteq \sigma^{-1}(\text{Ext}(\mathbb{S}))$ there is a coarsest scale-measure of $\mathbb{S}$ that reflects $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ (right, Figure 11.1). Such a scale-measure can be computed using the canonical representation of scale-measures as highlighted by $\text{cl}(\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}})$ in Figure 11.1. Since the scale-hierarchy is join-pseudocomplemented (see Proposition 17), we can compute smaller representations of $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$ and $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$. In detail, for any $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$ there exists a least element in $\underline{\mathfrak{S}}(\mathbb{S})$ whose join with $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$ yields $\sigma^{-1}(\text{Ext}(\mathbb{S}))$. Due to its smaller size, the so computed join-pseudocomplement can be more human comprehensible than $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$.

Split attribute error from consistent part

iii) The third option is based on splitting the scale context according to its consistent attributes, see Corollary 7. Both split elements are then considered as scale-measures of $\mathbb{S}$. This results in two smaller, potentially more comprehensible, concept lattices.

In addition to that, all discussed scale-measures can be given in conjunctive normalform.

11.1.2 Computational Tractability

Computational aspects

The first thing to note, with respect to the computational tractability, is that the size of the concept lattice of $\mathbb{S}$, as proposed in i) (above) is larger compared to the split approaches, as proposed in ii) and iii). This difference results in potentially smaller order dimensions for the split elements compared to $\underline{\mathfrak{B}}(\mathbb{S})$ (cf. Proposition 31). The approach in ii) splits the scale $\mathbb{S}$ according to the conceptual scaling error $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$, a potentially exponentially sized problem with respect to $\mathbb{S}$. The consecutive computation of the join-pseudocomplement involves computing all meet-irreducibles in $\sigma^{-1}(\text{Ext}(\mathbb{S}))$, another computationally expensive task. In contrast, approach iii) splits $\mathbb{S}$ based on consistent attributes and takes therefore polynomial time in the size of $\mathbb{S}$. However, as shown in the example after Corollary 7, approach iii) may lead to less accurate representations.

11.2 Small Case Study

Data sets

To provide practical evidence for the applicability of the just introduced conceptual scaling error, we conducted an experiment on eleven data sets. In those, we compared the classical errors, such as Frobenius norm, to the conceptual scaling error. The data sets were, if not otherwise specified, nominally scaled to a binary representation. Six of them are available through the UCI[1] [60] data sets repository: i) *Diagnosis* [50] with *temperature* scaled in intervals of [35.0 37.5) [37.5 40.0) [40.0 42.0], ii) *Hayes-Roth* iii) *Zoo* iv) *Mushroom*

[1] i) https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations,
ii) https://archive.ics.uci.edu/ml/datasets/Hayes-Roth,
iii) https://archive.ics.uci.edu/ml/datasets/zoo,
iv) https://archive.ics.uci.edu/ml/datasets/mushroom,
v) https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage,
vi) https://archive.ics.uci.edu/ml/datasets/Plants and https://plants.sc.egov.usda.gov/java/,

Table 11.2: Quantifying the Conceptual Scaling Error for approximations $\widetilde{\mathbb{K}} = \mathbb{S} \circ \mathbb{H}$ of data sets $\mathbb{K}$ by Binary Matrix Factorization. Cells with '-' where not computed due to computational intractability. Density (D), Attribute Error ($\mathcal{AE}$), Conceptual Scaling Error ($\mathcal{CE}$), Hemming distance between $I_{\mathbb{K}}$ and $I_{\widetilde{\mathbb{K}}}$ relative to $|G|\cdot|M|$ (H%), Frobenius measure between $I_{\mathbb{K}}$ and $I_{\widetilde{\mathbb{K}}}$.

	Context $\mathbb{K}$				Approximated Context $\widetilde{\mathbb{K}} = \mathbb{S} \circ \mathbb{H}$					Respective Scale $\mathbb{S}$				
	$\|G\|$	$\|M\|$	D	$\|\underline{\mathfrak{B}}\|$	Frob	H%	$\|\underline{\mathfrak{B}}\|$	$\mathcal{AE}$	$\mathcal{CE}$	$\|M\|$	D	$\|\underline{\mathfrak{B}}\|$	$\mathcal{AE}$	$\mathcal{CE}$
Diagnosis	120	17	0.471	88	13.04	8.3	26	6	7	4	0.250	6	0	0
Hayes-Roth	132	18	0.218	215	16.40	11.3	33	8	26	4	0.350	12	3	8
Domestic	41	55	0.158	292	8.49	3.2	148	14	68	10	0.183	34	6	15
Zoo	101	43	0.395	4579	15.52	5.5	442	13	347	7	0.315	25	2	4
Chess	346	683	0.473	3211381	82.01	2.8	229585	246	224673	26	0.767	4334	24	4280
Mushroom	8124	119	0.193	238710	243.86	6.2	10742	48	10598	11	0.277	139	7	116
HIV-1PC	6590	162	0.055	115615	221.38	4.6	303	32	229	13	0.154	330	12	236
Plant-Habitats	34781	68	0.127	-	322.16	4.4	-	68	-	8	0.128	256	8	255
Airbnb-Berlin	22552	145	0.007	-	130.29	0.5	-	0	0	12	0.057	8	1	1
UFC-Fights	5144	1915	0.001	-	101.43	0.1	2	0	1	44	0.932	2	44	1
Recipes	178265	58	0.057	-	492.8	2.3	-	7	-	8	0.236	256	4	187

v) *HIV-1ProteaseCleavage* [181] and vi) *Plant-Habitats* four kaggle[2] data sets vii) *Top-Chess-Players* with *rating,rank,games,bith_year* ordinally scaled, viii) neighbourhood data from the *Airbnb-Berlin* data sets, ix) *A_fighter* and *B_fighter* from the *UFC-Fights* data sets and x) *Recipes* [137]. The eleventh data set is generated from the Wikipedia list of *Domesticated Animals*.[3] This data set is also used for a qualitative analysis. We summarized all data sets in Figure 11.2, first and second major column.

Binary matrix factorization

As dimension reduction method, we employ the *binary matrix factorization* [230] of the Nimfa [233] framework. Their algorithm is an adaption of the *non-negative matrix factorization* (NMF). In addition to the regular NMF a penalty and a thresholding function are applied to binarize the output. To apply BMF we use for a formal context $\mathbb{K} := (G, M, I)$ the inverse scaling to $K := (G, M, \{0, 1\}, J)$ where $m(g) = 1$ iff $(g, m) \in I$ and $m(g) = 0$ otherwise. The output of the BMF algorithm are two binary matrices S, H with $K \approx S \cdot H$. The product used here is that from binary spaces $\{0, 1\}^k$ and the usual operations with the exception of $1 +_b 1 = 1$ (cf. Chapter 6).

Inverse scaling

Scaling

The resulting factors $\widetilde{K} = S \cdot H$ are scaled to formal contexts $\mathbb{S} := (G, M_S, I_S)$ with $(g, n) \in I_S$ iff $S_{g,n} = 1$, $\mathbb{H} := (M_S, M, I_H)$ with $(n, m) \in I_S$ iff $H_{n,m} = 1$ and $\widetilde{\mathbb{K}} := (G, M, \widetilde{I})$ with $(g, m) \in \widetilde{I}$ iff $\widetilde{K}_{n,m} = 1$. We often write $\mathbb{S} \circ \mathbb{H}$ due to the equality $\widetilde{\mathbb{K}} = (G, M, I_{\mathbb{S}} \circ I_{\mathbb{H}})$.

Parameter

The BMF factorization algorithm takes several parameters, such as convergence λ_w, λ_h, which we left at their default value of 1.1. We increased the maximum number of iterations to 500 to ensure convergence and conducted ten runs, of which we took the best fit. The target number of attribute (features) in $|M_{\mathbb{S}}|$ was set approximately to $\sqrt{|M_{\mathbb{K}}|}$ to receive a data dimension reduction of one magnitude.

Experiment

We depicted the results, in particular the quality of the factorizations, in Figure 11.2 (major column three and four). Our investigation considers standard measures, such as Frobenius norm (*Frob*) and relative Hamming distance (*H%*), as well as the proposed conceptual scaling error ($\mathcal{CE}$) and attribute scaling error ($\mathcal{AE}$). For the large data sets, i.e., the last four in Figure 11.2, we omitted computing the number of concepts due to its computational intractability, indicated by '-'. Therefor, we were not able to compute the conceptual scaling errors of the approximate data sets $\widetilde{\mathbb{K}}$. However, the conceptual scaling error of the related scales $\mathbb{S}$ is independent of the computational tractability of $\mathcal{CE}$ of $\widetilde{\mathbb{K}}$.

Results

We observe that the values for Frob and for H% differ vastly among the different data sets. For Example H% varies from 0.1 to 11.3. We find that for all data sets $|\mathfrak{B}(\mathbb{K})|$ is substantially larger than $|\mathfrak{B}(\widetilde{\mathbb{K}})|$, independently of the values of Frob and H%. Hence, BMF leads to a considerable loss of concepts. When comparing the conceptual and attribute scaling error to Frob and H%, we observe that the novel conceptual errors capture different aspects than the classical matrix norm differences. For example, *Domestic* and *Chess* have similar values for H%, however, their error values with respect to attributes and concepts differ significantly. In detail, the ratio of $|\mathcal{CE}|/|\mathfrak{B}(\widetilde{\mathbb{K}})|$ is 0.98 for *Chess* and 0.46 for *Domestic*, and the ratio for $|\mathcal{AE}|/|M|$ is 0.36 for *Chess* and 0.25 for *Domestic*.

Large contexts

While we do not know the number of concepts for *Airbnb-Berlin*, we do know that conceptual scaling error of the related $\widetilde{\mathbb{K}}$ is 0 due to $\mathcal{AE}$ being 0 and Proposition 7. The factorization of the *UFC-Fights* produced an empty context $\widetilde{\mathbb{K}}$. Therefore, all attribute derivations in $\widetilde{\mathbb{K}}$ are the empty set, whose pre-image is an extent of $\mathbb{K}$, hence, $\mathcal{AE}$ is 0. We suspect that BMF is unable to cope with data sets that exhibit a very low density. It is noteworthy that we cannot elude this conclusion from the value of the Frob and H%. By

[2] vii) https://www.kaggle.com/odartey/top-chess-players and https://www.fide.com/,
viii) https://www.kaggle.com/brittabettendorf/berlin-airbnb-data/,
ix) https://www.kaggle.com/rajeevw/ufcdata,
x) https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions,

[3] xi) https://en.wikipedia.org/w/index.php?title=List_of_domesticated_animals, 25.02.2020

Attributes of $\mathbb{K}$

clearing land (CL), draft (Dr), dung (Du), education (Ed), eggs (Eg), feathers (Fe), fiber (F), fighting (Fi), guarding (G), guiding (Gu), herding (He), horns (Ho), hunting (Hu), lawn mowing (LM), leather (Le), manure (Ma), meat (Me), milk (Mi), mount (Mo), narcotics detection (ND), ornamental (O), pack (Pa), pest control (PC), pets (Pe), plowing (Pl), policing (Po), racing (Ra), rescuing (R), research (Re), service (Se), show (Sh), skin (Sk), sport (Sp), therapy (Th), truffle harvesting (TH), vellum (V), weed control (WC), working (W)

$\mathbb{S}$	0	1	2	3	4	5	6	7	8	9
brahman cattle			×					×		
european cattle			×					×	×	×
guppy					×					
alpaca						×			×	×
bactrian camel				×						
bali cattle			×					×		
barbary dove										×
canary					×					
cat					×					×
chicken		×					×			
dog	×								×	×
donkey				×		×		×		
dromedary				×						
duck		×								
fancy mouse					×					
fancy rat					×					
ferret							×			
fuegian dog									×	
gayal			×							
goat			×			×			×	×
goldfish						×				
goose		×								
guinea pig					×	×				×
guineafowl		×								
hedgehog										
horse				×		×		×		
koi							×			
lama						×			×	×
mink					×					
muscovy duck		×								
pig					×	×	×		×	
pigeon							×			
rabbit					×	×				×
sheep			×		×	×	×		×	
silkmoth					×					
silver fox					×					
society finch					×					
striped skunk					×					
turkey		×								
water buffalo			×					×		×
yak				×				×	×	

Figure 11.3: Attributes of the *Domestic* data set (top) and a factor (bottom).

investigating the binary factor $\mathbb{S}$ using the conceptual scaling error and the attribute error, we are able to detect the occurrence of this phenomenon. In detail, we see that 44 out of 44 attributes are inconsistent.

Scaling error and matrix differences

We can take from our investigation that low H% and Frob values do not guaranty good factorizations with respect to preserving the conceptual structure of the original data set. In contrast, we claim that the proposed scaling errors are capable of capturing such error to some extent. On a final note, we may point out that the conceptual scaling errors enable a quantifiable comparison of a conceptual data reduction $\mathbb{S}$ to the original data set $\mathbb{K}$, despite different dimensionality.

11.2.1 Qualitative Analysis

Data set

The *domestic* data set includes forty animals as objects and fifty-five purposes for their domestication as attributes, such as *pets, hunting, meat, etc*. The resulting $\mathbb{K}$ has a total of 2255 incidences and the corresponding concept lattice has 292 formal concepts. We applied the BMF algorithm as before, which terminated after 69 iterations with the scale depicted in Figure 11.3. The incidence relation of $\widetilde{\mathbb{K}}$ has seventy-three wrong incidences, i.e., wrongfully present or absent pairs, which results in H% of 3.2. The corresponding concept

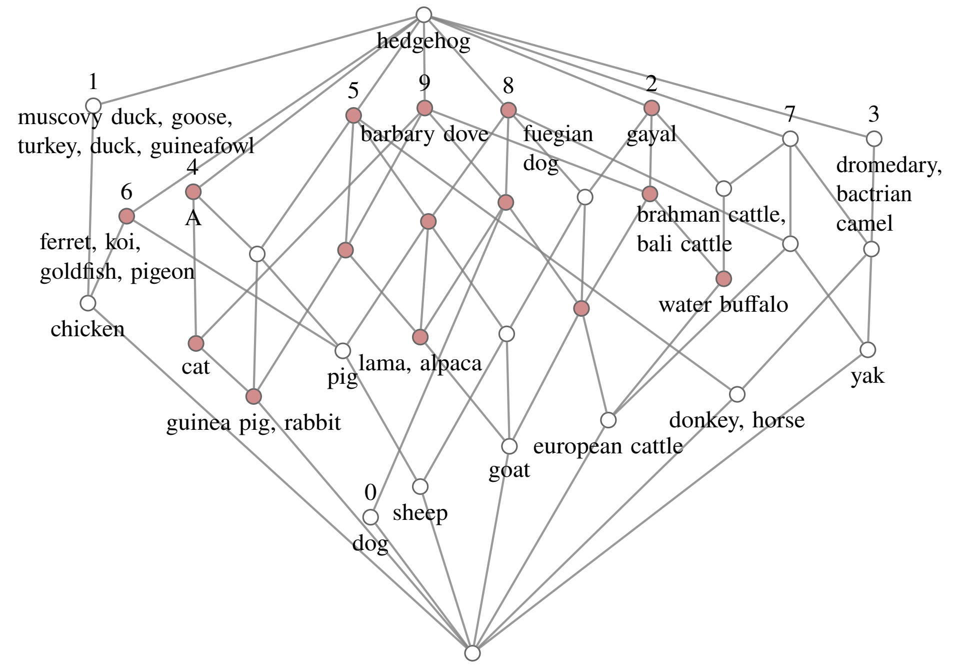


Figure 11.4: Concept lattice of the *Domestic* scale context. $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ is indicated in red. A={society finch, silkmoth, fancy mouse, mink, fancy rat, striped skunk, guppy, canary, silver fox}

lattice of $\widetilde{\mathbb{K}}$ has 148 concepts, which is nearly half of $\mathfrak{B}(\mathbb{K})$. Furthermore, out of these 148 concepts there are only 80 correct, i.e., in $\text{Ext}(\mathbb{K})$. This results in a conceptual scaling error of 68, which is especially interesting in the light of the apparently low H% error.

Highlight error

To pinpoint the particular errors, we employ i)-iii) from Section 11.1.1. The result of the first approach is visualized in Figure 11.4 and displays the concept lattice of $\sigma^{-1}(\text{Ext}(\mathbb{S}))$ in which the elements of $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ are highlighted in red. First, we notice in the lattice diagram that the inconsistent extents $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$ are primarily in the upper part. Seven out of fifteen are derivations from attribute combinations of 9, 8, and 5. This indicates that the factorization was especially inaccurate for those attributes. The attribute extents of 6, 4, and 2 are in $\mathcal{CE}^{\mathbb{K}}_{\sigma,\mathbb{S}}$, however, many of their combinations with other attributes result in extents of $\text{Ext}(\mathbb{K})$.

Split error from consistent part

The resulting lattices of applying approach ii) are depicted in Figure 11.5, the consistent lattice of $\sigma^{-1}(\text{Ext}(\mathbb{S}))|_{\text{Ext}(\mathbb{K})}$ is at the top and its join-pseudocomplement is at the bottom. The consistent part has nineteen concepts, all depicted attributes are in conjunctive normalform. The join-pseudocomplement consists of twenty-two concepts are colored.

Based on this representation, we can see that twenty out of the forty-one objects have no associated attributes. These include objects like *lama, alpaca* or *barbary dove*, which we have also indirectly identified by i) as derivations of 5, 8, 9. Furthermore, we see that thirteen out of the fifty-five attributes of $\mathbb{K}$ are not present in any conjunctive attributes. These attributes include domestication purposes like *tusk, fur*, or *hair*. Out of our expertise we suppose that these could form a meaningful cluster in the specific data realm. In the join-pseudocomplement, we can identify the attributes 5,8,9 as being highly inconsistently scaled, as already observed in the paragraph above.

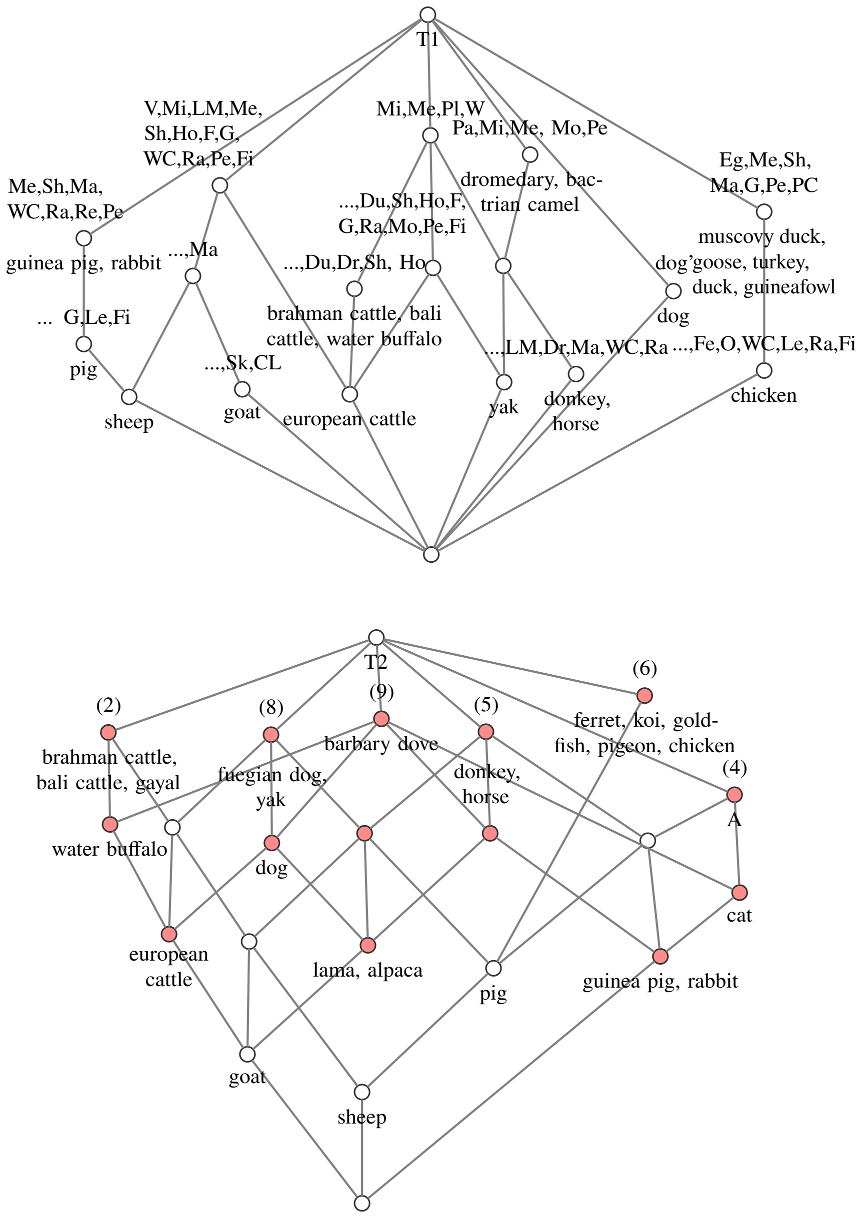


Figure 11.5: The concept lattice of all valid extents of the scale $\mathbb{S}$ (left, in conjunctive normalform in $\underline{\mathfrak{S}}(\mathbb{K})$) and its join-pseudocomplement (right) of the *Domestic* data set. Extents in the lattice drawing of the join-pseudocomplement that are extents not in the *Domestic* context are highlighted in red. dog'={Ed,TH,Pa,Gu,He,Dr,Sp,Me,Sh,Po,F,W,Re,-G,ND,Le,Ra,Hu,Se,Th,Pe,PC,Fi} A={society finch, silkmoth, fancy mouse, mink, fancy rat, striped skunk, guppy, canary, silver fox}, T1={fuegian dog, lama, ferret, alpaca, society finch, silkmoth, fancy mouse, koi, hedgehog, mink, fancy rat, striped skunk, goldfish, barbary dove, guppy, canary, pigeon, silver fox, cat, gayal}, T2={dromedary, muscovy duck, bactrian camel, goose, turkey, hedgehog, duck, guineafowl}

Split by consistent attributes

The third approach results in a scale $\mathbb{S}[G,N]$ of four consistent attributes N and a scale $\mathbb{S}[G,\hat{N}]$ of six non-consistent attributes $\hat{N}$. The scale $\mathbb{S}[G,N]$ has seven concepts and the scale $\mathbb{S}[G,\hat{N}]$ has twenty-two. We may note that the concept lattice of $\mathbb{S}[G,\hat{N}]$ is equivalent to the join-pseudocomplement of the previous approach. In general this is not the case. While $\underline{\mathfrak{B}}(\mathbb{S}[G,N])$ misses some of the consistent extents,we claim that the combination of $\mathbb{S}[G,N]$ and $\mathbb{S}[G,\hat{N}]$ still provides a good overview of the factorization shortcomings.

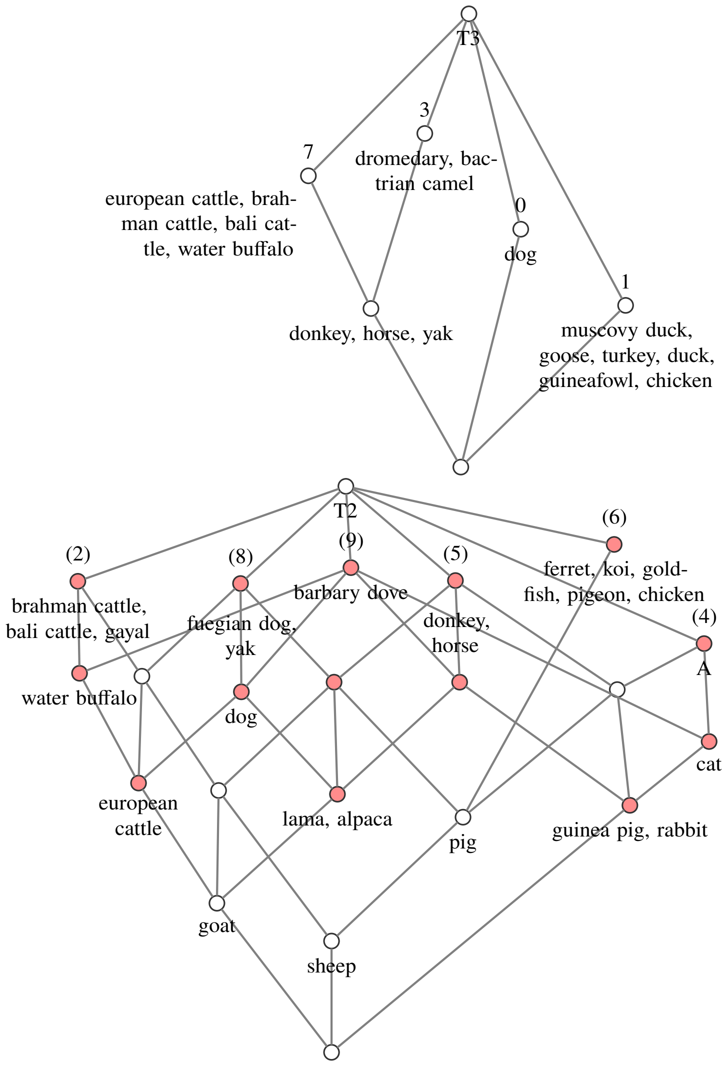


Figure 11.6: The concept lattice of all valid (top) and invalid (bottom) attributes of the *Domestic* scale-measure. Extents in the lattice drawing of the invalid attributes that are not extents in the *Domesticated Animals* context are marked in red. A={society finch, silkmoth, fancy mouse, mink, fancy rat, striped skunk, Guppy, canary, silver fox} T3= {fuegian dog, lama, sheep, ferret, pig, alpaca, society finch, goat, silkmoth, fancy mouse, koi, guinea pig, rabbit, hedgehog, mink, fancy rat, striped skunk, goldfish, barbary dove, Guppy, canary, pigeon, silver fox, cat, gayal} T2={dromedary, muscovy duck, bactrian camel, goose, turkey, hedgehog, duck, guineafowl}

11.3 Related Work

Binary matrix factorization

To cope with large data sets, a multitude of methods were introduced to reduce the dimensionality. With binary matrix factorization [230] we studied one of them with respect to its capability of preserving formal concepts. A comparison of several BMF methods with respect to their capability of preserving incidences can be found in Belohlávek, Outrata, and Trnecka [20]. Identifying the best BMF method with respect to the conceptual scaling error is outside the scope of this work, since it is neither limited to nor by BMF. A comprehensive study of several state-of-the-art data reduction methods for conceptual data reduction across different types of methods is deemed future work.

Evaluation scores

Other evaluations of BMF, besides $|K - S \cdot H|$, have been considered [128]. They investigate the quality of implications in $\mathbb{S} \circ \mathbb{H}$ for some classification task. Additionally, they use different measures [53], e.g., *fidelity* and *descriptive loss*. Other statistical approaches often find *euclidean loss*, *Kullback-Leibler* divergence, *Residual Sum of Squares*, adequate.

Qualitative evaluation

All previously mentioned evaluation criteria do not account for the complete conceptual structure of the resulting data set. Moreover, they are not intrinsically able to pinpoint to the main error portions of the computed data reduction. Furthermore, approaches based on the computation of implications are infeasible for larger data set. An advantage of our approach is the polynomial estimation of the conceptual error in the size of $\mathbb{S}$ and $\mathbb{K}$ through the attribute scaling error.

11.4 Discussion

Contributions

With this chapter, we have presented a new approach to evaluate conceptual data reduction methods in particular and dimension reduction in general. The proposed conceptual scaling error can be computed agnostic to reduction method, which makes it applicable in various machine learning settings. Beyond the quantification of the conceptual error, we have succeeded in presenting a method to explicitly represent the error generated by the dimension reduction, and to visualize it with the help of conceptual lattices.

Which method is best for conceptual data reduction?

Our small case study showed that simple matrix similarity measures do not guarantee a low conceptual scaling error. So far it is unclear which reduction methods are best suited for conceptual data reduction. For this a comprehensive comparison of several state of the art data reduction methods should be studied. Such an investigation would include reduction methods of several types and possible recommendations for parameters.

Repair reductions

A very useful result for BMF is Theorem 1 from Belohlávek and Vychodil [22]. It states that the attributes in $\mathbb{S}$ and objects in $\mathbb{H}$ respectively can be replaced by formal concepts such that for the new $\hat{\mathbb{S}}, \hat{\mathbb{H}}$ we find that $\mathbb{S} \circ \mathbb{H} = \hat{\mathbb{S}} \circ \hat{\mathbb{H}}$. An advantage of this substitution is that $\text{Ext}(\hat{\mathbb{S}}) \subseteq \text{Ext}(\mathbb{K})$ and thus the conceptual scaling error is equal to zero. This opens the question if there is a method to *repair* arbitrary data reductions such that they are consistent to the conceptual structure of $\mathbb{K}$.

Approximately correct concepts

The notion of conceptual scaling error and the evaluation of consistency of extents is discrete. There are concept lattices that seem very similar with a high conceptual scaling error. We can envision that a more relaxed, possibly continuous, notion may be helpful to express the similarity of closure systems. A promising notion for this is the accuracy measure for approximately correct implicational basis [27, Defintion 1]. One can use the connection between closure systems and implicational basis (cf. Section 5.4) to transfer this measure to conceptual data reductions. Such an investigation is deemed future work.

12

Conceptual Views on Tree Classifiers

Explainable decision trees

Decision trees are among the most popular explainable machine learning models. That is why they are often used as surrogates for other, less transparent machine learning models. Even though they are very expressing [153], their training procedures often do not perform as well as more contemporary classification procedures. Furthermore, decision trees cannot naturally cope with missing data (without further help or data preparation).

Tree ensembles

A popular class of classifiers, tree ensembles, do remedy these disadvantages. They employ multiple tree structures simultaneously (*boosting*) and make potentially use of individual *baggings* of the data, e.g., Random Forests or Gradient Boosted Trees [29, 73]. While these methods are capable of high classification performance, they do not possess the same interpretability as decision trees. This fact is due to the dispersion of information into a large number of incomparable parallel branches from differently rooted trees. There are many attempts to rectify this problem for explainability. One approach to explain tree ensembles is to merge all trees into a single decision tree [203]. Even though the resulting tree structure can be called (more) "explainable" (because it is a tree), it tends to grow incomprehensibly large and also loses the ability to handle missing values.

Lattices instead of trees

With the present work we propose a novel method for translating tree ensembles into a data structure that is interpretable by design, while allowing for parallelism to cope with missing information. With conceptual scaling, we derive concept lattices (views) from a tree ensemble instead of a single (surrogate) tree [217]. While in a tree there is always a unique path from the root to a node, lattices allow for multiple (parallel) paths in a single structure, leading to the same conclusion element. This results in a better handling of missing values.

Dealing with large views

One issue that is common to conceptual views and tree surrogates is their potentially incomprehensible size. However, for concept lattices FCA is equipped with a lot of data reduction methods [22, 90, 93, 134, 208] to achieve representations of human comprehensible size. One class of methods is presented in Chapter 8 which computes conceptual views reflecting specific pre-defined aspects of the classifier.

Views

Our work is not the first to take the step from tree ensembles to concept lattices [61], however, it drives this research in two aspects: first, we present established and novel views on tree classifiers in the language of conceptual views. Second, the newly proposed views achieve an unprecedented expressiveness and thus explainability of tree ensemble classifiers. Third, we provide a formal interpretation framework on how to understand a classifier through the lens of conceptual views. This allows for both local and global explanations with varying levels of detail, as we will demonstrate. In particular, the global scope of the derived explanations for tree classifiers is new and of special importance for the research field of explainable AI.

Experiment

In our experimental study, we demonstrate our explanation method and its applicability on a real world example from the openml CC18 [24] classification benchmark data set, namely, *car* [60] (binary class version[1]). The analyzed Random Forest was trained with realistic parameter assumptions, i.e., we used 100 trees and did not limit their individual depths.

12.1 Tree Based Classifiers

In this section, we formally introduce decision trees as data structures (cf. Chapter 2). The notions are illustrated based on a small example of a decision tree for the *tennis play* [158] data set (see Figure 12.1).

Decision tree structure

A **decision tree** of a many-valued context is a data structure $\underline{\mathcal{T}} := (\mathcal{T}, \mathcal{P}_{\mathcal{T}}, \leq_{\mathcal{T}}, \varphi_{\mathcal{T}})$, where: 1) $(\mathcal{T}, \leq_{\mathcal{T}})$ is an ordered set of at least three nodes that constitutes a proper binary tree, i.e., a join-semilattice in which every node has either two or no direct lower neighbor (**children**) and the order filter of every node is a linear order. The elements with no lower neighbor are called **leaves**. For $m \leq n$ we say that n lies on the unique path from m to the root. 2) $\mathcal{P}_{\mathcal{T}}$ is a set of predicates where for each $P \in \mathcal{P}_{\mathcal{T}}$ it holds that $\neg P \in \mathcal{P}_{\mathcal{T}}$. 3) $\varphi_{\mathcal{T}}$ is a right-unique relation where every, but the **root** r (top element), node is annotated by a **predicate** $\varphi(n)$ that *splits* the data, so that every data object either satisfies $\varphi(n)$ or its negation $\neg\varphi(n)$. For three nodes $l, r, t \in \mathcal{T}$ with $l, r \prec t$ and $l \neq r$ we have $\varphi(l) \cong \neg\varphi(r)$.

Example

The example in Figure 12.1 shows in its upper part a many-valued context with fourteen data objects, numbered $0, \ldots, 13$. There are four attributes, *overlook*, *temperature*, *humidity*, and *windy*. The last column *play* contains the classification outcome. Each attribute is ordinally pre-scaled:

rainy < overcast < sunny,
cool < mild < hot,
normal < high, and
not windy < windy.

Interordinal predicates

Although decision trees can work with any type of predicates, most decision tree implementations require that the attributes have a linear ordinal pre-scaling $(W(m), \leq_m)$. The resulting predicates compare values from $W(m)$ to specified threshold values $t \in W(m)$ with $\leq$ and $\geq$. Thus, the decision tree has an interordinal interpretation of the data. We represent the predicates in the form *attribute≤threshold* and *attribute≥threshold* where $w \models \varphi(n)$ with $\varphi(n) = m \leq t$ iff $w \in W(m)$ and $w \leq_m t$.

Classification procedure

For the classification, an object $g \in G$ is threaded through the tree from the root to a leaf node b such that it satisfies all predicates along the path $r, \ldots, b$. We call b the **decision leaf** and the path $r, \ldots, b$ the **decision path** of g. Finally there is a second mapping, associating to every leaf the classification outcome, usually *yes* or *no*.

[1] `https://www.openml.org/d/991`

$\mathbb{D}$	overlook	temperature	humidity	windy	play
0	sunny	hot	high	False	no
1	sunny	hot	high	True	no
2	overcast	hot	high	False	yes
3	rainy	mild	high	False	yes
4	rainy	cool	normal	False	yes
5	rainy	cool	normal	True	no
6	overcast	cool	normal	True	yes
7	sunny	mild	high	False	no
8	sunny	cool	normal	False	yes
9	rainy	mild	normal	False	yes
10	sunny	mild	normal	True	yes
11	overcast	mild	high	True	yes
12	overcast	hot	normal	False	yes
13	rainy	mild	high	True	no

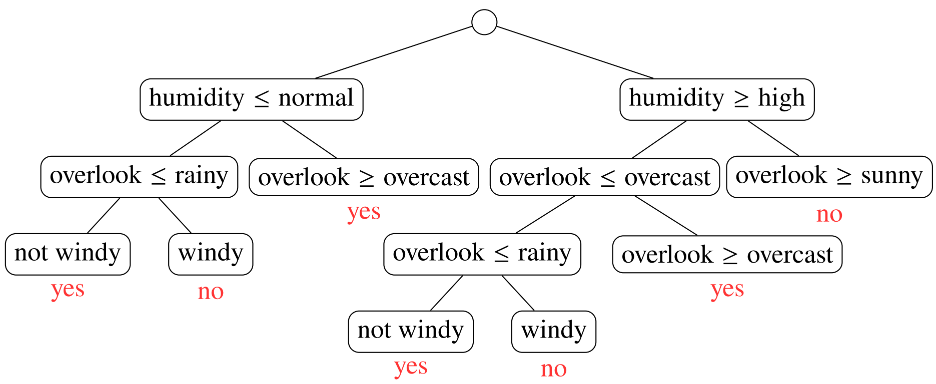


Figure 12.1: Decision tree for the tennis data set. Each data object follows the path from the root (the top node) along the predicates it fulfills until it reaches a leaf node. The decision for *play* is annotated in red.

Overlook	≤ *rainy*	≤ *overcast*	≥ *overcast*	≥ *sunny*
rainy	×	×		
overcast		×	×	
sunny			×	×

Temperature	≤ *cold*	≤ *mild*	≥ *mild*	≥ *hot*
cold	×	×		
mild		×	×	
hot			×	×

Humidity	≤ *normal*	≥ *high*
normal	×	
hot		×

Wind	*windy*	*not windy*
True	×	
False		×

$\mathbb{I}(\mathbb{D})$	*Overlook* ≤ *rainy*	*Overlook* ≤ *overcast*	*Overlook* ≥ *overcast*	*Overlook* ≥ *sunny*	*Temperature* ≤ *cool*	*Temperature* ≤ *mild*	*Temperature* ≥ *mild*	*Temperature* ≥ *hot*	*Humidity* ≤ *normal*	*Humidity* ≥ *high*	*windy*	*not windy*
0			×	×			×	×		×		×
1			×	×			×	×		×	×	
2		×	×				×	×		×		×
3	×	×				×	×			×		×
4	×	×			×	×			×			×
5	×	×			×	×			×		×	
6		×	×		×	×			×		×	
7			×	×		×	×			×		×
8			×	×	×	×			×			×
9	×	×				×	×		×			×
10			×	×		×	×		×		×	
11		×	×			×	×			×	×	
12		×	×				×	×	×			×
13	×	×				×	×			×	×	

Figure 12.2: The interordinal scaling of the tennis data set (Figure 12.1). Each incidence is indicated by a × in the cross-tables. The concept lattice of the derived context has 108 concepts.

Interordinal scaling

The context derived from the tennis many-valued context with interordinal scaling is shown in Figure 14.3 and has 108 formal concepts, which is considerably larger than the size of the decision tree in Figure 12.1. Attributes with empty and full incidences where omitted for readability reasons.

12.1.1 Training a decision tree on many-valued contexts

Training

In order to train a decision tree on a many-valued context $\mathbb{D} := (G, M, W, I)$, we need to know the classification labels for all $g \in G$. The tennis data set in Figure 12.1 shows an example. The last column contains the classification labels, while the rest of the table represents a many-valued context. However, this example is tiny compared to realistic datasets from real life. For such one needs fast implementations like the C4.5 algorithm [176]. At each step the algorithm extends the tree at a node n that has no children. The algorithm chooses the pair of predicates $P, \neg P$ from $\mathcal{P}_{\mathcal{T}}$ that separates the set of objects $H \subseteq G$, whose values are model to all predicates along the path from n to the root, best with respect to some measure of information entropy. After the selection of $P, \neg P$ two nodes l, r are added below n annotated with P and $\neg P$. The result is a decision tree $\underline{\mathcal{T}}$, as described in Section 12.1.

Explanation tasks

For a given decision tree $\underline{\mathcal{T}}$ and its training data set $\mathbb{D}$, two main explanatory tasks can be formulated. The first addresses the question of how adequately $\underline{\mathcal{T}}$ represents the training data, in particular the objects $g \in G$, and which general explanations can be inferred from $\underline{\mathcal{T}}$ using G. These explanations range from local ones, i.e., why was an object g classified to a particular class from C, and global ones, such as, which predicate combination describe a class. The second task is to understand the *view* of $\underline{\mathcal{T}}$ on a so far unknown set of objects $\check{G}$ whose values match $\mathbb{D}$ (Section 7.1.1). Again, these views can be locally and globally.

12.2 Concept Lattices from Tree Classifiers

In the following we introduce different conceptual views on tree classifiers. All proposed views are conceptual views of the underlying interordinal scaling of the data (cf. Figure 14.3). We first revisit three approaches from the literature and derive a unified representation for them in the language of Formal Concept Analysis and conceptual data scaling. We will consider these methods as baselines for our two novel approaches in the next section.

12.2.1 Approaches from the literature

Partition view

The first approach to investigate is given by the `RandomTreesEmbedding` from sklearn[2] and simply reflects the clustering of the data objects induced by the leaf nodes of the decision tree [26, 144, 159].

Definition 45 (Conceptual Leaf View on $\underline{\mathcal{T}}$). *Given a mv-context $\mathbb{D} := (G, M, W, I)$ and a decision tree $\underline{\mathcal{T}}$ that was trained on $\mathbb{D}$, we define the* contextual leaf view on $\underline{\mathcal{T}}$ *as*

$$\mathbb{N}(G, \underline{\mathcal{T}}) := (G, \mathcal{L}(\underline{\mathcal{T}}), J), \text{ where } (g, l) \in J \text{ iff } g \models \varphi(n) \text{ for all } n \geq_{\mathcal{T}} l.$$

*The corresponding concept lattice $\mathfrak{B}(\mathbb{N}(G, \underline{\mathcal{T}}))$ is called the **conceptual leaf view on** $\underline{\mathcal{T}}$. A slight generalization allows to include not only the training data, but also other data objects with the same attributes and attribute values, i.e., objects that match the $\mathbb{D}$ (cf. Section 7.1.1).*

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomTreesEmbedding.html`

G may thus be replaced by $\check{G}$. The set $\mathcal{L}(\underline{\mathcal{T}})$, which is used as set of attributes here, is the set of leaves of the decision tree.

Interpretation

The conceptual leaf view on a set of objects $\check{G}$ enables to view said objects through the classification leaves of $\underline{\mathcal{T}}$. This means, for any two objects $g_1, g_2 \in \check{G}$ that are classified by the same leaf l, we have that $\{g_1\}^J = \{g_2\}^J$, and therefore the object concepts $(\{g_1\}^{JJ}, \{g_1\}^J)$ and $(\{g_2\}^{JJ}, \{g_2\}^J)$ are equal. Informally said, g_1 and g_2 are *clustered* in the same concepts. We may note at this point, that this view is limited to this fact. Hence, objects that are classified by different leaf nodes do not share any attributes and are therefore incomparable. Figure 12.3 shows the contextual view for our running example.

Limitation

More general, the conceptual leaf view is very simple. It is just an antichain plus a top and a bottom element. This is due to the fact that $\underline{\mathfrak{B}}(\mathbb{N}(\check{G}, \underline{\mathcal{T}}))$ is of nominal scale, that is, it has exactly one concept per leaf, and all these concepts are pairwise incomparable. Altogether, this view is coarse and does not exhibit hierarchical information, i.e, there are no concepts in sub-concept relation, apart from those involving the top $(\check{G}, \check{G}^J)$ or bottom $(\mathcal{L}(\underline{\mathcal{T}})^J, \mathcal{L}(\underline{\mathcal{T}}))$ concepts.

The tree order

The second baseline view accounts for the whole order structure of the decision tree $\underline{\mathcal{T}}$ [26]. The corresponding concept lattice is an isomorphic representation to the one proposed by previous work [61].

Definition 46 (Conceptual Tree View on $\underline{\mathcal{T}}$). *Given a mv-context $\mathbb{D} := (G, M, W, I)$ and a decision tree $\underline{\mathcal{T}}$ that was trained on $\mathbb{D}$, we define the **contextual tree view on $\underline{\mathcal{T}}$** by taking the tree nodes as attributes. A node is incident with a data object if and only if it was used for classifying that object.*

$$\mathbb{T}(G, \underline{\mathcal{T}}) := (G, \underline{\mathcal{T}}, J), \text{ where } (g, t) \in J \text{ iff } t \text{ is on the decision path of } g \text{ in } \underline{\mathcal{T}}.$$

*The corresponding concept lattice $\underline{\mathfrak{B}}(\mathbb{T}(G, \underline{\mathcal{T}}))$ is called the **conceptual tree view on $\underline{\mathcal{T}}$**. As in Definition 45 above, G may be replaced by a more general set $\check{G}$.*

Comparison to leaf view

In contrast to the contextual leaf view on $\underline{\mathcal{T}}$ the contextual tree view accounts for all nodes of $\underline{\mathcal{T}}$. Hence, objects that are classified by different leaf nodes may have common nodes in their decision paths. The more their respective decision paths overlap, the more attributes they have in common.

The concept lattice $\underline{\mathfrak{B}}(\mathbb{T}(G, \underline{\mathcal{T}}))$ is order-isomorphic to the decision tree with an added smallest element, i.e., to $(\mathcal{T} \cup \{\bot\}, \leq)$ where for all $n \in \mathcal{T} : \bot \leq n$. The conceptual tree view on $\underline{\mathcal{T}}$ can thus be considered as an almost one-to-one translation of the decision tree order relation into the realm of Formal Concept Analysis.

View on unseen objects

For a given arbitrary object sets $\check{G}$, only parts of the order structure $(\mathcal{T} \cup \{\bot\}, \leq)$ are reached. More precise, since for all elements of $\check{G}$ there is an element of G having the same decision path, we can conclude that there is a (unique up to context clarification) isomorphism from $\mathbb{T}(\check{G}, \underline{\mathcal{T}})$ into a sub-context of $\mathbb{T}(G, \underline{\mathcal{T}})$. Thus, $\text{Int}(\mathbb{T}(\check{G}, \underline{\mathcal{T}})) \subseteq \text{Int}(\mathbb{T}(G, \underline{\mathcal{T}}))$. However, for the rest of our work, we will not explore this relationship further.

The third and last baseline is the view reflecting the complete interordinal scaling $\mathbb{I}(\mathbb{D})$. It is the most expressive in terms of formal concepts and, in contrast to the other two views, solely depends on the many-valued data set $\mathbb{D}$.

Commonalities

All just introduced views have in common, that their respective concept lattices are atomistic. Moreover, in all lattices are the atoms given by the set of all object concepts. This observation depends on two assumptions, a) there are no missing values for any of the objects viewed by a scaling (cf. complete many-valued context) and b) every leaf node is supported by an object $g \in G$.

$\mathbb{P}(\underline{\mathcal{T}})$	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n_8	n_9	n_{10}	n_{11}	n_{12}
0		×				×						
1		×				×						
2					×					×		
3		×			×				×		×	
4	×		×				×					
5	×		×					×				
6	×			×								
7		×				×						
8	×			×								
9	×		×				×					
10	×			×								
11					×					×		
13		×			×				×			×

$\mathbb{N}(G,\underline{\mathcal{T}})$	l_0	l_1	l_2	l_3	l_4	l_5	l_6
0	×						
1	×						
2		×					
3			×				
4						×	
5							×
6					×		
7	×						
8					×		
9						×	
10					×		
11		×					
12					×		
13				×			

Figure 12.3: The contextual leaf view on $\underline{\mathcal{T}}$ (right in Figure 12.1) for the running example (left in Figure 12.1). Its concept lattice contains 55 concepts. The contextual predicate view $\mathbb{P}(\underline{\mathcal{T}})$ of $\underline{\mathcal{T}}$ is shown on the right.

12.2.2 Predicate Views

Views based on the annotated predicates

The so far presented approaches for views on tree-based classifiers do not use the predicates as attribute sets. Yet, these predicates are essential for human interpretation. Hence, we introduce in the following two novel conceptual views, i.e., conceptual scalings, that can also be extracted from the decision tree. However, in contrast to Definitions 45 and 46, the attribute set will be comprised of the predicates from $\mathcal{P}_{\mathcal{T}}$ instead of the tree nodes.

Towards predicate based views

In order to do this, we introduce an intermediate structure, in detail a formal context, which takes the place of the annotation function φ of $\underline{\mathcal{T}}$. This context is defined by $\mathbb{P}(\underline{\mathcal{T}}) := (\mathcal{T}, \mathcal{P}_{\mathcal{T}}, J)$ where $(n, P) \in J$ iff there exists a $h \geq_{\mathcal{T}} n$ with $\varphi(h) = P$. That is, a node n of the tree $\underline{\mathcal{T}}$ is in incidence with a predicate $P \in \mathcal{P}_{\mathcal{T}}$ if and only if P is annotated to n, or a predecessor of n. In the following this context is called the **predicate view of** $\underline{\mathcal{T}}$. We want to hint why this structure enables a (formal) interpretation of $\underline{\mathcal{T}}$ by means of the predicates. For any node n that is on the decision path of an object g, we have that $g \models \{n\}^J$, i.e., g is a model for all predicates that are incident with n. Moreover, for the leaf node l on the decision path of g, the set $\{l\}^J$ is exactly the set of predicates that were used by $\underline{\mathcal{T}}$ to classify the object g. Hence, we can interpret the classification for any object in terms of $\mathcal{P}_{\mathcal{T}}$.

Tree Predicate View

Definition 47 (Conceptual Tree Predicate View). *For a mv-context $\mathbb{D} := (G, M, W, I)$ and a decision tree $\underline{\mathcal{T}}$ (trained on $\mathbb{D}$), we define the **contextual tree predicate view** on $\underline{\mathcal{T}}$ by*

$$\mathbb{T}_{\mathcal{P}}(G, \underline{\mathcal{T}}) := (G, \mathcal{P}_{\mathcal{T}}, I_{\mathbb{T}(G,\underline{\mathcal{T}})} \circ I_{\mathbb{P}(\underline{\mathcal{T}})}).$$

*Analogously to Definition 45 we say **conceptual tree predicate view** on $\underline{\mathcal{T}}$ to $\underline{\mathfrak{B}}(\mathbb{T}_{\mathcal{P}}(G, \underline{\mathcal{T}}))$. Likewise, this view can be applied to unknown data $\check{G}$.*

The *is classified by* view

This definition of a view differs slightly from the ones given in the last section. First of all, the attribute set of the tree predicate view is comprised of the predicates of $\underline{\mathcal{T}}$. Moreover, the incidence relation of $\mathbb{T}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}})$ is implicitly given by the relation product $\circ$ of the incidences from the tree view and the predicate view. Hence the name tree predicate view. Our reasoning here is that we want to link objects to predicates via tree nodes. For example, if $g \in \check{G}$ is incident with node $n \in \underline{\mathcal{T}}$, and again n is incident with some predicate $P \in \mathbb{P}(\underline{\mathcal{T}})$, then g is incident with P in $I_{\mathbb{T}(\check{G},\underline{\mathcal{T}})} \circ I_{\mathbb{P}(\underline{\mathcal{T}})}$. Based on this construction, an object g is in incidence with a predicate P iff P is used to classify g.

Tree and tree predicate view

In contrast to the tree view, objects that have a disjoint decision path may still have predicates in common, and therefore common incidences in $I_{\mathbb{T}_{\mathcal{P}}(\check{G},\underline{\mathcal{T}})}$. The concept lattice of the tree predicate view $\mathbb{T}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}})$ is not necessarily tree shaped, since additional concepts may emerge from the meet of predicates that were annotated multiple times. In particular for the case where $\check{G} = G$, we find that

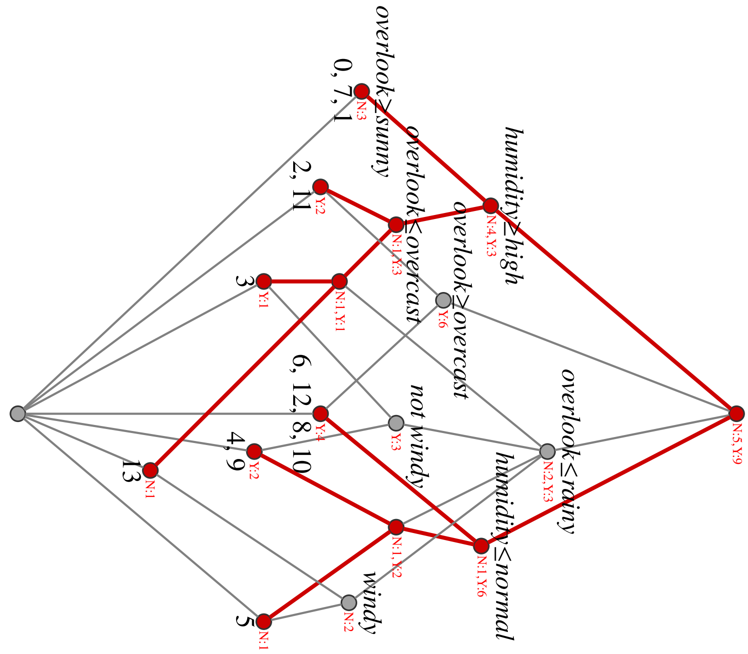
$$\mathrm{Ext}(\mathbb{T}(G, \underline{\mathcal{T}})) \subseteq \mathrm{Ext}(\mathbb{T}_{\mathcal{P}}(G, \underline{\mathcal{T}})), \tag{12.1}$$

since for all $n \in \underline{\mathcal{T}}$ we have that

$$n^{I_{\mathbb{T}(G,\underline{\mathcal{T}})}} = \{\varphi(m) \mid m \geq n\}^{I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}.$$

Classification in the tree predicate view

To see why this is true, we refer the reader to Proposition 38. Another useful property we prove in Proposition 38 for $\mathbb{T}_{\mathcal{P}}(\underline{\mathcal{T}})$ is that its concept lattice is atomistic and its object extents are equal to those of $\mathbb{T}(G, \underline{\mathcal{T}})$. A natural consequence of this fact is, that conceptual



$\mathbb{T}_{\mathcal{P}}(G,\mathcal{T})$	0	1	2	3	4	5	6	7	8	9	10	11	12	13
humidity≤normal					×	×	×		×	×	×		×	
overlook≤overcast			×	×								×		×
overlook≤rainy				×	×	×				×				×
windy						×								×
humidity≥high	×	×	×	×				×				×		×
overlook≥sunny	×	×						×						
overlook≥overcast			×				×		×		×	×	×	
not windy				×	×					×				

Figure 12.4: The contextual and conceptual tree predicate view on $\mathcal{T}$ (right in Figure 12.1) for the running example (left in Figure 12.1). The original decision tree $\mathcal{T}$ is highlighted in red in the concept lattice diagram.

$\mathbb{I}_{\mathcal{P}}(\underline{\mathcal{T}})$	*humidity≤normal*	*overlook≤overcast*	*overlook≤rainy*	*windy*	*humidity≥high*	*overlook≥sunny*	*overlook≥overcast*	*not windy*
0					×	×	×	×
1				×	×	×	×	
2		×			×		×	×
3		×	×		×			×
4	×	×	×					×
5	×	×	×	×				
6	×	×		×			×	
7					×	×	×	×
8	×					×	×	×
9	×	×	×					×
10	×			×		×	×	
11		×		×	×		×	
12	×	×					×	×
13		×	×	×	×			

Figure 12.5: The contextual interordinal predicate view on $\underline{\mathcal{T}}$ (right in Figure 12.1) for the running example (left in Figure 12.1). Its concept lattice contains 55 concepts.

view's concept lattice can classify every object $g \in G$ in the same way as the tree classifier would. That is, for any $g \in G$ the closure of g in the tree predicate view is equal to the closure in the leaf view, i.e., the set of objects that are classified by the same leaf. What is more important, the tree predicate view exhibits concepts that are not related to a node of $\underline{\mathcal{T}}$, however, they explain how different nodes of $\underline{\mathcal{T}}$ are related in terms of common predicates. This is the reason for the super set relation in Equation (12.1).

In the case that all predicates are annotated exactly once in the tree is the annotate function injective. Here, we find that $\text{Ext}(\mathbb{T}(\check{G}, \underline{\mathcal{T}})) = \text{Ext}(\mathbb{T}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}))$.

Interordinal Predicate View

Model relation view

The just introduced tree predicate view is capable of reflecting the predicates that are important for the specific classification of an object g. However, the incidences of g are limited to those predicates annotated to the decision path of g. In the view to be introduced in a moment we want to lift this restriction by extending the incidence relation to all predicates $P \in \mathcal{P}_{\mathcal{T}}$ for which g is a model, i.e., $g \models P$.

Definition 48 (Conceptual Interordinal Predicate View). *For a many-valued context* $\mathbb{D} := (G, M, W, I)$ *and a decision tree* $\underline{\mathcal{T}}$ *(trained on* $\mathbb{D}$*), we define the* ***contextual interordinal predicate view*** *on* $\underline{\mathcal{T}}$ *by*

$$\mathbb{I}_{\mathcal{P}}(G, \underline{\mathcal{T}}) := (G, \mathcal{P}_{\mathcal{T}}, J), \text{ where } (g, P) \in J \text{ iff } g \models P.$$

Analogously to all previous definitions we say ***conceptual interordinal predicate view*** *on* $\underline{\mathcal{T}}$ *to* $\underline{\mathfrak{B}}(\mathbb{I}_{\mathcal{P}}(G, \underline{\mathcal{T}}))$*. Likewise, this view can be applied to previously unknown data* $\check{G}$*.*

The predicate views

Since the interordinal predicate view is defined on the same set of objects and attributes as the tree predicate views, both views are related. More precisely, we find that the incidence relation of the tree predicate view is a subset of the interordinal predicate view, i.e., $I_{\mathbb{T}_{\mathcal{P}}(\check{G},\underline{\mathcal{T}})} \subseteq I_{\mathbb{I}_{\mathcal{P}}(\check{G},\underline{\mathcal{T}})}$. With the following proposition we show how the classical and our novel views are related to each other with respect to their training data set, i.e., $\check{G} = G$. From this we can infer how, or to which extent, the views can be employed for the explanation of tree based classifiers.

Proposition 38 (Views on Tree Classifiers). *Let $\mathbb{D}$ be a complete many-valued context whose attribute domains are linearly ordered and let $\underline{\mathcal{T}}$ be a decision tree, which was trained on $\mathbb{D}$, such that every leaf node of $\underline{\mathcal{T}}$ is supported. Then the following statements hold:*

i) in every view $\mathbb{N}(G,\underline{\mathcal{T}}), \mathbb{T}(G,\underline{\mathcal{T}}), \mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}}), \mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})$, and $\mathbb{I}(\mathbb{D})$ we find that the object concepts are the atoms of their respective concept lattice,

ii) $\mathbb{N}(G,\underline{\mathcal{T}}) \lesssim \mathbb{T}(G,\underline{\mathcal{T}}) \lesssim \mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ and
$\mathbb{N}(G,\underline{\mathcal{T}}) \lesssim \mathbb{T}(G,\underline{\mathcal{T}}) \lesssim \mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}}) \lesssim \mathbb{I}(\mathbb{D})$ where $\lesssim$ is the **is view of** *relation,*

iii) $I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})} \subseteq I_{\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})}$,

iv) the object extents of $\mathbb{N}(G,\underline{\mathcal{T}}), \mathbb{T}(G,\underline{\mathcal{T}})$ and $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ are equal.

Proof. i) For nominal and interordinal scales it is a known fact the set of atoms is comprised of object concepts. Since the introduced leaf view is of nominal scale we can infer the statement to be true. In case of the tree view $\mathbb{T}(G,\underline{\mathcal{T}})$ we know that $\{g\}^J$ is equal to the set of nodes from $\underline{\mathcal{T}}$ that are on the (complete) decision path of g. If we assume that there is a formal concept below $(\{g\}^{JJ}, \{g\}^J)$, then there must exist a $h \in G$ with $\{g\}^J \subseteq \{h\}^J$ and $\{g\}^J \neq \{h\}^J$. However, this would imply that the decision path of g can be extended, which is a contradiction. We may note that in case of an incomplete many-valued context $\mathbb{D}$, this argument does not hold. For the tree predicate view we can apply the same argument.

For the interordinal predicate view, assume there are two objects $g, h \in G$ with $\{g\}' \subseteq \{h\}'$ and $\{g\}' \neq \{h\}'$, then $\{h\}'$ is not an atom in $\underline{\mathfrak{B}}(\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}}))$. Note, the derivation $(\cdot)'$ is taken with respect to $\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})$. Hence, there is a predicate $P \in \{h\}'$ with $h \models P$ and $g \not\models P$. Due to the completeness that arises from the complete mv-context, we can infer that $g \models \neg P$. Therefore, $\neg P \in \{g\}'$, which contradicts the assumption.

ii) All contexts are defined on the same set of objects. Thus, it is to show that for $\mathbb{S}_1 \lesssim \mathbb{S}_2$ it holds that $\mathrm{Ext}(\mathbb{S}_1) \subseteq \mathrm{Ext}(\mathbb{S}_2)$. [Case $\mathbb{N}(G,\underline{\mathcal{T}}) \lesssim \mathbb{T}(G,\underline{\mathcal{T}})$] The set of leaf nodes $\mathcal{L}(\underline{\mathcal{T}})$ is a subset of all nodes in $\underline{\mathcal{T}}$. From this fact we can infer that $\mathbb{N}(G,\underline{\mathcal{T}})$ is an attribute induced sub-context of $\mathbb{T}(G,\underline{\mathcal{T}})$ and thus $\mathrm{Ext}(\mathbb{N}(G,\underline{\mathcal{T}})) \subseteq \mathrm{Ext}(\mathbb{T}(G,\underline{\mathcal{T}}))$, see context apposition [80]. [Case $\mathbb{T}(G,\underline{\mathcal{T}}) \lesssim \mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$] For a nonempty extent $A \subseteq G$ of $\mathbb{T}(G,\underline{\mathcal{T}})$ we know its derivation in said view is a path in the decision tree from the root up to some node n. For a decision tree, this path is uniquely identified by the annotated predicates, since the predicates $Q \subseteq \mathcal{P}_{\mathcal{T}}$ of the split in the root node cannot be annotated twice. Otherwise, this would lead to an unsupported leaf, which contradicts the requirements of the proposition. The derivation of Q in $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ is equal to A, since all objects that pass through node n are exactly those that are a model of Q. It remains to be shown that the empty set is an extent on both views, which follow from i) and the fact that the decision tree $\underline{\mathcal{T}}$ has at least one split.[3] Thus $\mathrm{Ext}(\mathbb{T}(G,\underline{\mathcal{T}})) \subseteq \mathrm{Ext}(\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}}))$. [Case $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}}) \lesssim \mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})$] The same arguments as in the tree predicate view case apply here.

For any path from the root node to some other node $n \in \mathcal{T}$ let $Q \subseteq \mathcal{P}_{\mathcal{T}})$ be the set of annotated predicates. Then the derivation of Q in $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ is equal to the derivation

[3] We made this requirement for all decision trees that are considered in this work in Section 12.1.

of Q in $\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})$. This is true since the set of objects that passes through the node n, i.e., $Q^{I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}$, is given by the set of objects that models Q, i.e., $Q^{I_{\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}$. Thus $\text{Ext}(\mathbb{T}(G,\underline{\mathcal{T}})) \subseteq \text{Ext}(\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}}))$. [Case $\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}}) \lesssim \mathbb{I}(G)$] This property follows directly from the fact that $\mathbb{I}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ is an induced sub-context of $\mathbb{I}(G)$.

iii) Follows directly from their definitions.

iv) Any object extent $A \subseteq G$ of $\mathbb{N}(G,\underline{\mathcal{T}})$ has the property that there is a unique leaf node $l \in \underline{\mathcal{T}}$ such that A is the set of objects that is classified by l.

From i) we can infer that the object extents of $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ are the atoms in the concept lattice $\mathfrak{B}(\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}}))$. For any object extent A of $\mathbb{N}(G,\underline{\mathcal{T}})$, let $g \in G$ be a generator of A, i.e., $\{g\}^{I_{\mathbb{N}(G,\underline{\mathcal{T}})}I_{\mathbb{N}(G,\underline{\mathcal{T}})}} = A$.

We can find the associated set of predicates in the tree predicate view by computing the derivation of g in $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$. In detail, we can compute $\{g\}^{I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}$ by projecting

$$\{(g,n) \mid n \text{ is on the dec. path of } g\} \circ \{(n,P) \mid \exists m \in \underline{\mathcal{T}} : n \leq_{\mathcal{T}} m \text{ and } \varphi(m) = P\}$$

on the second element. This set is equal to the set of predicates annotated to the decision path of g. The second application of the derivation operation in $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ yields the set of objects that are model of $\{g\}^{I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}$. Hence, all elements of $\{g\}^{I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}I_{\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})}}$ are classified by the same leaf as g. Thus, the object extents of $\mathbb{N}(G,\underline{\mathcal{T}})$ and $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ are equal. The rest of the statements follows directly from ii).

□

Explaining the view relation results

From this proposition we can draw essential consequences for the conceptual interpretation of (or view on) tree classifiers. From ii) we can infer that the extent structure of the decision tree is entailed in both the tree predicate view and the interordinal predicate view. Hence, the whole decision tree structure is captured by both views. To demonstrate this within the scope of our running example, we depicted the decision tree within the tree predicate view of the training data G in Figure 12.4 (right). In addition to the tree structure of $\underline{\mathcal{T}}$, we can observe in $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ multiple predicate combinations that span across different tree branches. Moreover, are all but the tree predicate view in conceptual view relation with the context derived from plain interordinal scaling. The tree predicate view is in general not in said view relation. This can be seen from the following characterization of P': An object $g \in G$ is in incidence with P iff it is in model relation to the disjunction of all logical expressions A, where A is the conjunction of all predicates along a path $(n,\ldots,r)$ in $\underline{\mathcal{T}}$ with $\varphi_{\mathcal{T}}(n) = P$. The tree predicate view can be considered as a view of a context derived from logical scaling based on the just described logical expressions. From iii) we can infer that the tree predicate view is a sub-context of the interordinal predicate view. From iv) we can follow that the views $\mathbb{N}(G,\underline{\mathcal{T}})$, $\mathbb{T}(G,\underline{\mathcal{T}})$ and $\mathbb{T}_{\mathcal{P}}(G,\underline{\mathcal{T}})$ have equivalent decision nodes in their concept lattices.

Decision leafs

In the next subsection we show how all introduced views lead to interesting insights into and explanations of decision trees.

12.2.3 Explaining Decision Trees

Explanation tasks

The theoretical findings from the last section lead to several approaches for the interpretation of decision trees. In particular we derive five methods that cover different explanation aspects. We want to introduce and discuss these using the tennis example.

Alternative Leaf Descriptions: Our method can generate alternative descriptions for leaf nodes in the predicate language of $\underline{\mathcal{T}}$. For example, the leaf l that classifies object 13 in Figure 12.4 contains the predicates *humidity≥high, overlook≤overcast,* *overlook≤rainy, windy*. This leaf has an upper neighbor within the concept lattice of the tree predicate view having the attribute *windy*. There is no node within $\underline{\mathcal{T}}$ representing this concept. Despite that, we can use this concept to construct an alternative combination of predicates that generates the concept of l. The leaf can be represented by the meet $\bigwedge\{$*windy, humidity≥high*$\}$. This is in fact a **minimal generator** for the intent of the concept associated to l.

In order to interpret a given decision tree $\underline{\mathcal{T}}$, one can generate all minimal generators for all leaf concepts, and use these as shorter descriptions to comprehend the classification structure of the tree. The thereby obtained shorter explanations are potentially more comprehensible. This is in particular useful, when decision trees are large, for example, when trained on large data sets having a many attributes.

Explaining Leaf Sets: For any set of leafs $L \subseteq \mathcal{T}$, we can compute in the tree predicate view their (conceptual) join $(A, B) := \bigvee L$. The formal concept $(A, B) \in \underline{\mathfrak{B}}(\mathbb{T}_{\mathcal{P}}(G, \underline{\mathcal{T}}))$ is not necessarily associated to a node of $\underline{\mathcal{T}}$, for example, take the join of the leafs having objects 13 and 5 in Figure 12.4. From this we learn that both leafs share the predicate *windy*. When we follow the lattice towards the top concept, we find that the leafs 13 and 5 also share the predicate *overlook≤rainy*. In contrast, within $\underline{\mathcal{T}}$, the decision paths of both leafs have only the root node in common. Hence, our method is capable of expressing commonalities of the set of leafs L, that are inexpressible within the structure of $\underline{\mathcal{T}}$.

Control for Missing Data: The alternate descriptions from the previous two items allow for coping with missing attributes. For example, when classifying an object $g \in \check{G}$ that has no value for the attribute *overlook*, the predicate tree view can map g to the leaf having object 13 using the attributes *windy* and *humidity*, as discussed in the previous items.

Global Influence of a Predicate: A common method for interpreting and explaining decision trees is to identify attributes that are used first or second in the tree. Yet, as the tree predicate view reveals, there are other structurally important predicates, e.g., *overlook≤rainy* and *overlook≥overcast*, as there is no upper neighbor for the associated concepts besides the root node. We say a formal concept (A, B) *is dominated by* another concept (C, D) iff $(A, B) \leq (C, D)$. Based on this notion, we can say that a predicate P is dominated by another predicate Q iff the attribute concept of P is a lower neighbor of the attribute concept of Q, i.e., $P^J \subseteq Q^J$.

Leaf Coverage of Predicates: Another measure of importance for a predicate P within a decision tree $\underline{\mathcal{T}}$ is the number of leafs that P is involved with. It is not surprising that a predicate which is used first in the decision tree will be involved in many leaves. However, as we can infer from the lattice diagram in Figure 12.4, the predicate *overlook≤rainy* is involved in four leaves while its predecessor in the tree is only involved in three leaves. Hence, the conceptual view on $\underline{\mathcal{T}}$ allows for structurally identifying important predicates. Moreover, one may easily select a subset of predicates that covers all leafs of a tree $\underline{\mathcal{T}}$.

The methodology just presented for the analysis and interpretation of decision trees can be applied to the interordinal tree view in an analogous way, as Proposition 38 ii) points out. In general, given the set inclusion on the extent sets, one can consider the views $\mathbb{N}(G, \underline{\mathcal{T}}), \mathbb{T}(G, \underline{\mathcal{T}})$ as coarse, the views $\mathbb{T}_{\mathcal{P}}(G, \underline{\mathcal{T}}), \mathbb{I}_{\mathcal{P}}(G, \underline{\mathcal{T}})$ as intermediate, and the view $\mathbb{I}(\mathbb{D})$ as a fine scaling of $\underline{\mathcal{T}}$.

12.2.4 Tree Ensemble Views

From trees to forests

In the last section, we introduced all tree views in a common language. This enables us to compare their different *views* on a decision tree, as we have seen. This comparability provides the cornerstone for a comprehensive approach to the interpretation of tree ensembles. In particular, we present a principle approach for the interpretation of families of trees, as they are used in common supervised machine learning procedures, such as Random Forests or AdaBoost. In the following we will use the notation $\mathbb{S}(\check{G}, \underline{\mathcal{T}})$ as a general name for any of the views introduced in the last sections.

Definition 49 (Forest View). *For a many-valued context $\mathbb{D} := (G, M, W, I)$, a family of decision trees $\mathfrak{T} = (\underline{\mathcal{T}}_i)_{i \in F}$ that were trained on $\mathbb{D}$, and a conceptual view for each tree $\mathbb{S}(G, \underline{\mathcal{T}}_i)$, we define their* ***forest view*** *to be*

$$\mathbb{S}(G, \mathfrak{T}) := \bigcup_{i \in F} \mathbb{S}(G, \underline{\mathcal{T}}_i), \text{ where } \mathbb{S}_1 \cup \mathbb{S}_2 := (G_1 \cup G_2, M_1 \cup M_2, I_1 \cup I_2).$$

Likewise, this view can be applied to previously unknown data $\check{G}$.

We want to elaborate on the reasoning behind this definition. First we may note, that for our views the sets G_1, G_2, G are equal, hence, their union results in G. This is by design, since we want our method to interpret a forest given a particular set of objects G, analogously to the decision tree views. In contrast, the attribute sets and incidences can be different. Hence, the ability of the forest view to provide an interpretation is directly connected to the union of all attributes from the set of tree views. This definition preserves the relations ii) and iii) presented in Proposition 38.

Special notes on bagging

Random Forests are constructed using two essential techniques, **bagging** and an empirical variant of **boosting**. The term bagging, also known as **bootstrap aggregation**, describes the procedure to draw samples uniformly from the training data set, with replacement. The common approach is to sample for any tree in an ensemble its own training data set [30]. In the language of our formal contexts, each tree is constructed using a random induced many-valued sub-context $\mathbb{H} \leq \mathbb{D}$. This modeling does not take into account the possibility that the same object can be drawn twice, however, this can obviously dealt with by creating copies of objects.

12.2.5 Explaining Random Forests

Interpretation of forest views

The forest view enables us to apply the in Section 12.2.3 derived explanation approaches for decision trees to Random Forests. In the following, we derive explanation approaches for forest views based on each of the four introduced conceptual views on trees, i.e., leaf view, tree view, tree predicate view and interordinal predicate view. The following descriptions are abstract extensions to the explanation approaches proposed in Section 12.2.3. In Section 12.4.2 we provide an in-depth analysis for a comprehensibly sized data set.

Leaf View: In this view the attribute set is comprised of the set of leafs $\mathcal{L}(\underline{\mathcal{T}})$ for any tree $\underline{\mathcal{T}}$ in the ensemble $\mathfrak{T}$. For any two trees $\underline{\mathcal{T}}_1$ and $\underline{\mathcal{T}}_2$, we consider their leaf sets $\mathcal{L}(\underline{\mathcal{T}}_1)$ and $\mathcal{L}(\underline{\mathcal{T}}_2)$ to be disjoint sets, i.e., $\mathcal{L}(\underline{\mathcal{T}}_1) \cap \mathcal{L}(\underline{\mathcal{T}}_2) = \emptyset$.

This implies that any forest view on leaf scaled trees is equal to the context apposition of the set of leaf views $|_{i \in F}\ \mathbb{S}(\check{G}, \underline{\mathcal{T}}_i) = \bigcup_{i \in F} \mathbb{S}(\check{G}, \underline{\mathcal{T}}_i)$. From the context apposition we can deduce that the set of its extents, i.e., $\mathrm{Ext}(|_{i \in F}\ \mathbb{S}(\check{G}, \underline{\mathcal{T}}_i))$, is equal to the set of intersections of all subsets of extents from all tree views, i.e., $\{\bigcap \mathcal{A} \mid \mathcal{A} \subseteq \bigcup_{i\in F} \text{Ext}(\mathbb{S}(\check{G}, \underline{\mathcal{T}}_i))\}$.

Tree View: Analogously to the modeling by the leaf view, we consider the nodes for any two trees $\underline{\mathcal{T}}_1, \underline{\mathcal{T}}_2$ to be disjoint, i.e., $\mathcal{T}_1 \cap \mathcal{T}_2 = \emptyset$. Hence, the forest view is equal to the apposition of all tree views, having the same consequence on its extents as shown in the last item.

Tree Predicate View: Forest views that are based on the tree predicate view are more complicated than the previous two. For example, given two trees $\underline{\mathcal{T}}_1, \underline{\mathcal{T}}_2$, an object $g \in \check{G}$ and a predicate $P \in \mathcal{P}_{\mathcal{T}1} \cap \mathcal{P}_{\mathcal{T}2}$, the case may arise that $(g, P) \in I_{\mathbb{T}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}_1)}$ but $(g, P) \notin I_{\mathbb{T}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}_2)}$. Therefore, the tree predicate view based forest view is not a simple apposition of its individual tree predicate views. Hence, it is possible that this forest view can come up with extents that were not simple intersection of already known extents. An advantage of employing tree predicate view based forest views is that their resulting context representation is smaller. This is due to the fact that any two trees might share predicates, i.e., $\mathcal{P}_{\mathcal{T}1} \cap \mathcal{P}_{\mathcal{T}2} \neq \emptyset$.

Interordinal Predicate View: In contrast to the last view, we can state for the interordinal predicate view on Random Forests that given two trees $\underline{\mathcal{T}}_1, \underline{\mathcal{T}}_2$, an object $g \in \check{G}$ and a predicate $P \in \mathcal{P}_{\mathcal{T}1} \cap \mathcal{P}_{\mathcal{T}2}$ we find $(g, P) \in I_{\mathbb{I}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}_1)}$ iff $(g, P) \in I_{\mathbb{I}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}_2)}$. From this we can infer that the forest view is almost the apposition of the set of interordinal predicate views $\mathbb{I}_{\mathcal{P}}(\check{G}, \underline{\mathcal{T}}_i)$, with the exception that any predicate P that occurs in more than one view is not duplicated by coloring. Since clarification of attributes, i.e., the removal of duplicate attributes in a formal context, does not affect the set of extents, we can apply the same reasoning as shown for the leaf and tree view.

The just introduced views allow for a variety of practical applications for the explainability of tree ensembles. The main advantage of the different notions of views on trees is that they integrate all trees of a Random Forest in a unified (lattice) structure. Their capability to explain the Random Forest increases with respect to the chosen underlying view Proposition 38, ii). At the same time we increase the capability, we also increase the complexity of the necessary computations. We want to refer the reader for a detailed discussion of these aspects to Section 12.4.2.

Applications

In the next section, we show how our approach for forest views can be applied to large data sets. For this, we recall methods from the literature that deal with applying FCA to large data sets. Moreover, we introduce two new methods specifically designed for, yet not limited to, applying forest views to Random Forests. We demonstrate both on a comprehensibly sized data set in Section 12.4.2.

Notes on the view sizes

12.3 Dealing with Large Conceptual Views

We discussed in Sections 12.2.2 and 12.2.4 the utility and applicability of the different conceptual views. Yet, for many examples of real-world sized data these views are potentially incomprehensibly large. Thus, in order to derive human-comprehensible selections and aggregations of the conceptual views, we introduce the following methods. These methods are based on common data reduction procedures for formal contexts, however, adapted for conceptual views.

Dealing with large sizes

Projection

Object or Attribute Selection The first class of methods are selection methods to compute induced sub-context of contextual views. Selecting a subset of the object set will result in a coarser closure system on the set of attributes. A selection of attributes of the contextual view has the same effect on the objects (cf. Corollary 4). There are numerous ways on how to select relevant attributes from formal contexts [19, 64, 93]. In our experiments (Section 12.4), we employ the feature importance scores that are provided by the Random Forest models. Furthermore, one may apply `KMedoid` [194, 195] clustering to identify representative objects, which we call **center objects**. The advantage of `KMedoid` compared to other popular methods, such as `kmeans`, is that the cluster centers are existing objects of the data set. Thus, the clustering can be interpreted as computing an induced sub-context with a subset of the original object set.

Local views

Structure based Object Selection A particular method for selecting objects can be based on the structural position of an object within the concept lattice. For a given object $g \in G$, a natural approach would be to compute the order filter $\uparrow\{g\}^{I_{\mathbb{S}}I_{\mathbb{S}}} \subseteq \underline{\mathfrak{B}}(\mathbb{S})$ for a contextual view $\mathbb{S}$. This results in a *local conceptual view* that allows for deriving explanations for individual objects. A second approach additionally includes neighboring concepts of $\uparrow\{g\}^{I_{\mathbb{S}}I_{\mathbb{S}}}$. The resulting local conceptual view enables more comprehensive explanations of the structural position of g within $\underline{\mathcal{T}}$, and its dependence from different attribute values. In particular, this allows for investigations that are comparable to **partial dependence plots** [95], i.e., it enables the study of attribute value perturbations.

Neighboring concepts can be added using covering elements of the set $\uparrow\{g\}^{I_{\mathbb{S}}I_{\mathbb{S}}}$, i.e., $\{A \in \underline{\mathfrak{B}}(\mathbb{S}) \mid A \prec B \vee B \prec A \text{ for } B \in \uparrow\{g\}^{I_{\mathbb{S}}I_{\mathbb{S}}}\}$. Those elements can be enumerated recursively using the `next_neighbor` algorithm [143].

Importance measures

Concept Selection Methods To reduce the number of formal concepts, it is common to apply different criteria for their importance. The FCA literature provides a multitude of measures [132]. In our experimental work, we select concepts based on their support [208] (TITANIC), i.e., the number of objects that are contained in an extent divided by the number of all objects. This procedure results in a subset of the concepts of a conceptual view. This set constitutes a join-semilattice, i.e., the iceberg concept lattice.

Decomposition

Composition Methods Another approach is to split a conceptual view into multiple parts based on a given partition of the object or attribute set. The original concepts of a conceptual view can be retrieved from the individual parts by combining them using the meet and joins operations. Reasonable partitions of the object set can be derived using their class labels. Hence, from this one can compute a drawing per class label. A meaningful choice for a partition of the attributes is to draw on their semantics. For example, employing ontological background knowledge. Furthermore, one may restrict a view to a particular *order direction* of the threshold values, i.e., $\leq$ and $\geq$. This procedure can be considered as an ordinal factors with respect to the context apposition operation [80].

Clustering

Attribute Aggregation A reason for why the number of concepts of a conceptual view gets large is the number of different predicates derived during the training. Aggregating different predicates by clustering them may lead to a significant reduction in the number of concepts. For this, one should account for the different attribute value distributions on which the predicates are based on. A clustering using *grades* as aggregated values, e.g., *low≤med≤high*, can be especially comprehensible to human readers.

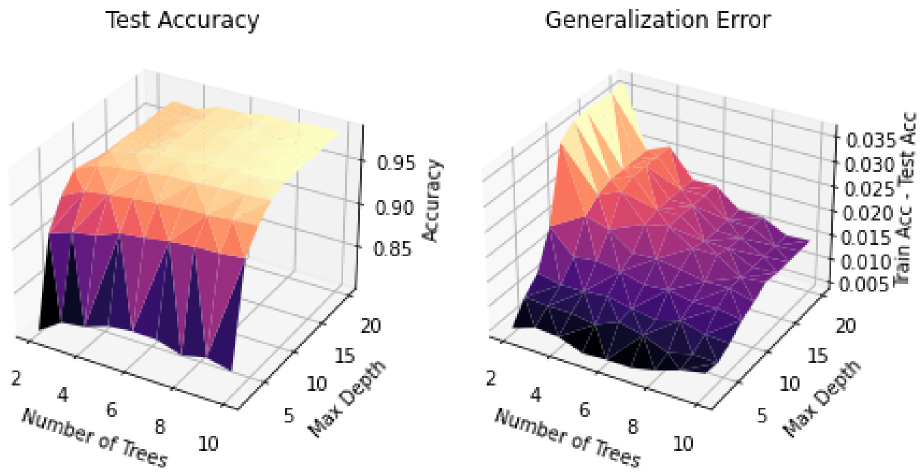


Figure 12.6: Visualization of distribution of the performance measures accuracy (ACC) and generalization error, of trained Random Forest classifiers using the hyper parameters `md` and `nt`. Mean values are reported.

12.4 Experimental Study

Research questions

The following experiments shall support our theoretical findings with respect to two practical research questions. First, *is the size of conceptual views manageable with respect to human-comprehensibility and to what extent depends its size on the choice of hyper parameters of the tree training algorithm?* Second, *are explanations derived from conceptual views meaningful for human-understanding?*

Data set

To answer these questions we conduct two experiments using Random Forests. As for a data set, we choose the well-known *car* data set [69, ID:991], which is comprised of 1728 objects on seven (many-valued) attributes. This dataset presents a binary classification problem, using the class labels `positive` and `negative`.

12.4.1 Sizes of Conceptual Views: a Parameter Study

We investigate the first research question by means of a parameter study. The two most important hyperparameters of the Random Forest procedure are the *number of trees* (`nt`) and their *maximal depth* (`md`). Other parameters, such as attributes per tree, purity, split criterion, etc, also have a significant influence, however, not on the size of the resulting conceptual view.

Experimental setup

For our study, we trained different Random Forest classifiers using $2 \leq \texttt{nt} \leq 10$ and $2 \leq \texttt{md} \leq 20$. We completed ten runs for each parameter combination, using ten different initial random seeds. In Figure 12.6 (left) we report the classification performance using the average accuracy and in Figure 12.6 (right) the generalization error. The latter is comprised of subtracting the accuracy on the test data set from the accuracy that was achieved on the train data set, i.e., $\text{Error}_{\text{Generalization}} := \text{ACC}_{\text{Train}} - \text{ACC}_{\text{Test}}$. This value allows us to estimate the amount to which our trained Random Forest classifier is prone to overfitting. In all our experiments, we conducted four fold cross-validation, however, we observed stable results.

Classification performance

From a supervised-learning point of view, we notice that for the model accuracy the

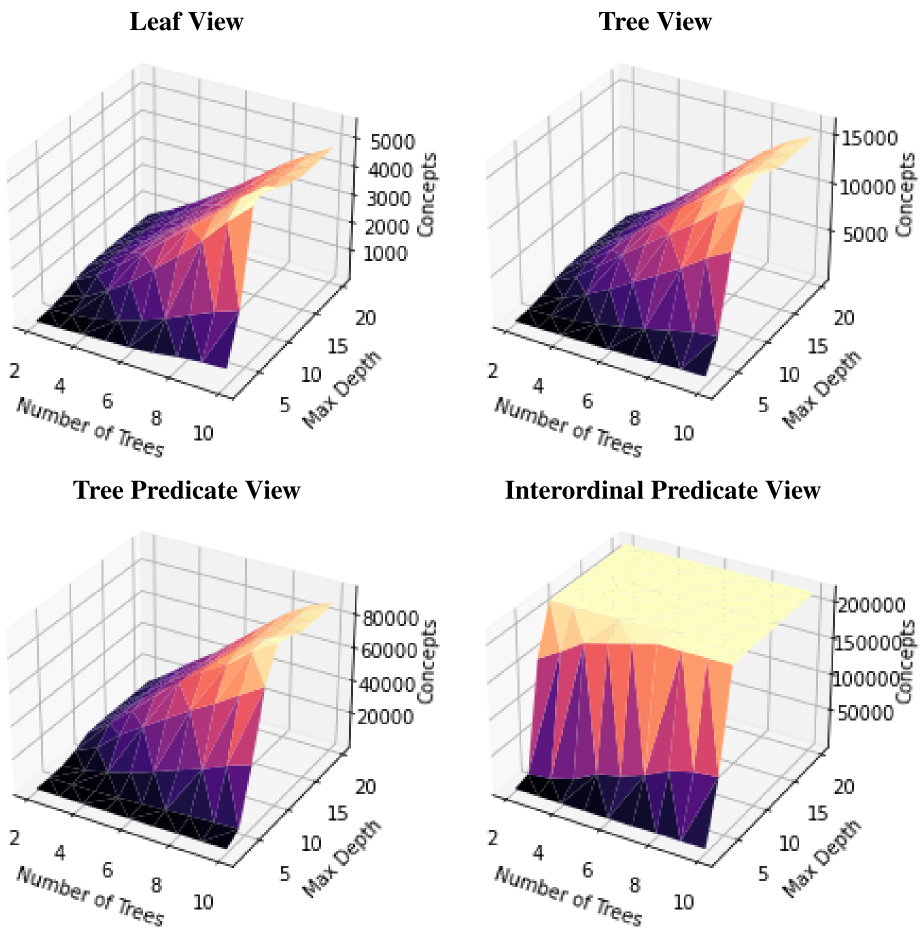


Figure 12.7: Distribution of the number of formal concepts for different hyperparameter combinations and different conceptual views: Leaf view (top left), tree view (top right), tree predicate view (bottom left), and interordinal predicate view (bottom right).

maximum depth md has a greater impact than the number of trees nt. However, based on the generalization error (Figure 12.6, right), we find that the number of trees is instrumental to prevent overfitting. At this point, we feel justified in stating that very good classifiers exist for nt $\geq$ 8 and md $\geq$ 15. With that, we can turn to the conceptual views and their capabilities of explaining the Random Forest classifiers.

View sizes

For this, we first examine the influence of the parameters on the number of concepts of each respective conceptual view. In this experiment, the Random Forest classifier are trained on the entire data set where no cross-validation was applied. Afterwards, we computed the different views $\mathbb{S}(G, \mathfrak{T})$, also using the entire data set, and depicted the number of formal concepts per view and parameter combination in Figure 12.7.

First, we notice that the visual shapes for three different conceptual views are similar to some extent. The exception is the interordinal predicate view, which increases more quickly with increasing md. This observation is expected, since the occurrence of split predicated increases with the depth of the trees, and, in contrast to the tree predicate view, the object-predicate incidences are independent of the location of said predicates in the trees. For all plots we can report that increasing the depth beyond ten has no noticeable impact. The performance measurements in Figure 12.6 behaved analogously, yet, since we applied different training sets, we should refrain from a direct comparison.

Which parameters to choose

In terms of the absolute number of formal concepts, we find that the leaf view generates the smallest amount (5,000), followed by the tree view (15,000), the tree predicate view (80,000) and the interordinal predicate view (200,000). This observation is expected due to our theoretical findings in Section 12.2. Obviously, due to the observed number of formal concepts, all views elude from a direct human-comprehension. Hence, consecutive data reduction methods, as proposed in Section 12.3, are required.

Views are large

12.4.2 Deriving Meaningful Conceptual Explanations

For our final take on explaining Random Forest classifiers using conceptual views, we choose extreme hyperparameters in order to show the viability of our approach. In detail, we set the number of trees to 100, which is a commonly accepted default value [174]. For the maximum depth of the trees we set no limitation, i.e., the training algorithm splits nodes until class purity is achieved. The resulting conceptual views, more precisely their number of formal concepts, naturally rises to the amount as seen in the last section and more. Although the computation of such and larger sets of formal concepts is not a challenge for algorithms from the field of Formal Concept Analysis, human comprehensibility now requires the application of the selection and aggregation methods presented in Section 12.3.

Dealing with large views

We want to start with combining the composition method with object, attribute, and concept selection procedures. For this, we first employ `KMedoids` clustering from sklearn to select a smaller number of representative objects. We determine the parameter k, i.e., the number of medoids, to be nineteen, by trial and error and evaluating the **silhouette score** [183] on the results within the range $2 \leq k \leq 50$. In a second step, we restrict the set of view attributes (i.e., predicates) in the following way. We computed for all seven many-valued attributes of the car data set their significance for the classification using the notion of permutation importance [6]. The result allows us to select the most important ones. For the rest of our study, we stick to four. From these many-valued attributes, we can derive a subset of important predicates, i.e., view attributes.

Selection based on medoid clustering

Starting from this state, we have applied various other methods for selection. In Figure 12.8 we depict the result for the tree predicate view when additionally applying a) composition, more specific, we partition the object set using the related class labels, and b) the TITANIC algorithm [208].

Decomposition and TITANIC

The Tree Predicate View

The top diagram in Figure 12.8 is comprised of the objects bearing the positive class label, and the bottom diagram is comprised of objects bearing the negative class label. The respective values for minimum support are five and three, i.e., all concepts in the view using the positive labels have an extent size of at least five, and analogously three for the negative part. These values were chosen such that the resulting iceberg concept lattices is of comprehensible size. The particular values three and five seem to reflect the imbalance of the class labels to some extent, however, this observation is not essential. Both diagrams are annotated in the usual way. In addition to that, we annotated on the right to each concept node the class purity in this concept, i.e., the number of positive and negative labeled objects

Parameters

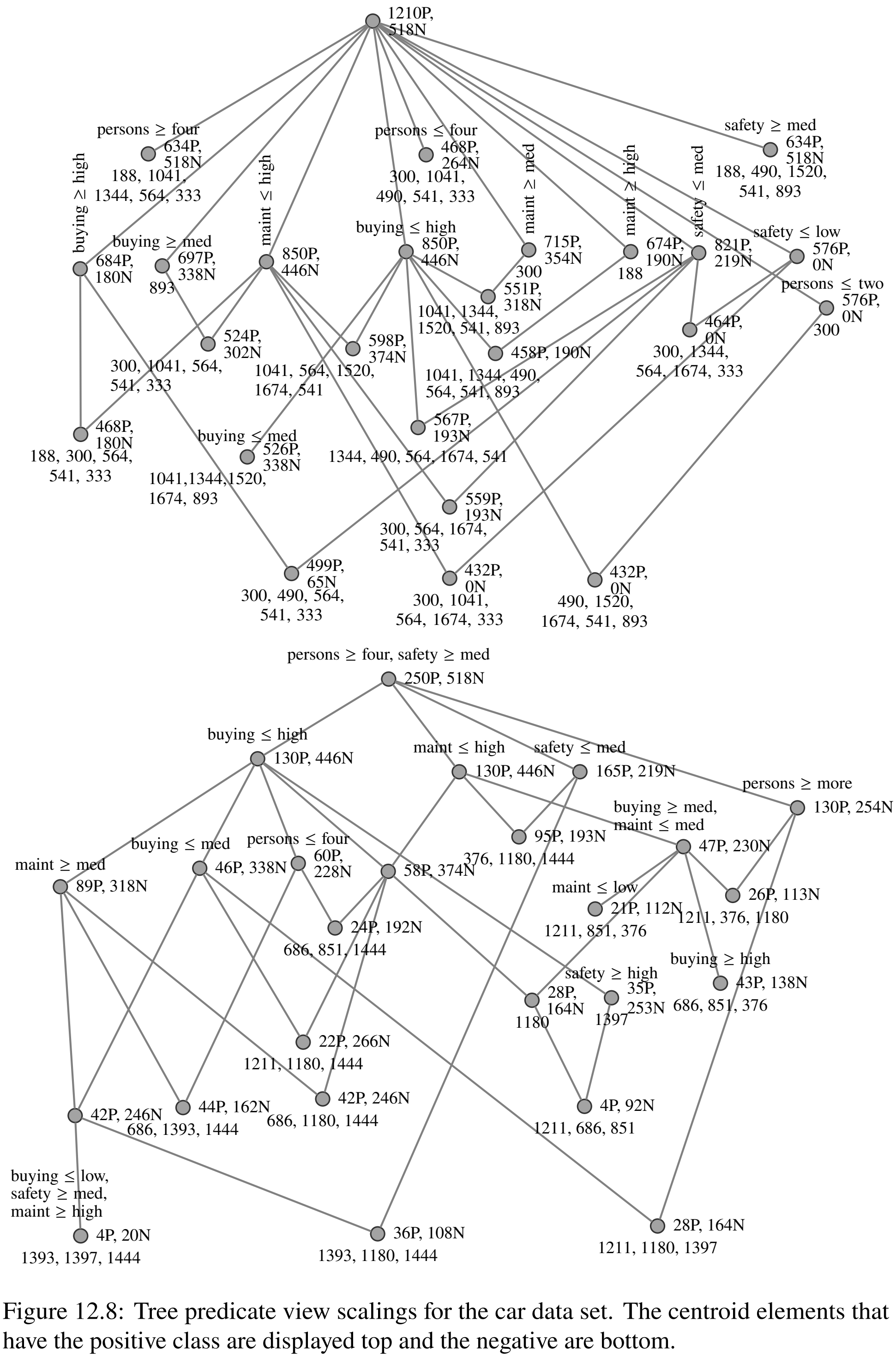


Figure 12.8: Tree predicate view scalings for the car data set. The centroid elements that have the positive class are displayed top and the negative are bottom.

of the data set.

First of all, we observe structural differences between the iceberg concept lattices of the positive (PICL) and negative (NICL) center objects, although both have twenty-five formal concepts. PICL has twelve co-atoms while NICL has four co-atoms. NICL has a longest chain of five elements while PICL's is three. We claim that PICL is easier to comprehend than NICL due to its smaller depth. At the same time, NICL implies that the description of the negative class is more difficult and demanding with respect to the number of attributes, i.e., intent sizes. More generally, in both diagrams we can infer descriptions of the positive and negative class from the concepts lowest in the diagrams. Even though there are methods to explain the influence of single attributes on the classification, the iceberg concept lattice allows to easily comprehend the influence of arbitrary attribute combinations. For example, the concept with extent label `300, 490, 564, 541, 333` is a result of the attribute combination *buying* $\geq$ *high* and *safety* $\geq$ *med*. Furthermore, the conceptual structures allows to identify attributes with a high global influence on the classification. For example, *maint* $\leq$ *high* has five direct lower neighbors whereas *persons* $\leq$ *two* has one direct lower neighbor.

Views for the positive and negative class

Depth

Description lengths

Study attribute combinations

Dominant attributes in RF classifiers

A particularly interesting observation in the NICL diagram is the presence of attributes that support all objects. Hence, these are essential for classification of all objects with the negative class label. This conclusion is especially easy to infer from the conceptual structure compared with analyzing all hundred trees of the underlying Random Forest. Finally, the iceberg concept lattice of NICL reveals redundant attributes. For example, the concept annotated with the object extent *1393, 1397, 1444* has three annotated attributes of which only one is needed to identify this concept.

Redundant attributes

A more general inference about the Random Forest is that the tree predicate view allows for an identification of *"costly"* objects. By this we mean objects whose classification required a large number of (potentially redundant) threshold value tests. For example, we refer the reader to the concept bearing the objects *1393,1397,1444* within NICL. On one hand, all these objects required redundant testing of attributes, namely *maint*$\geq$*med* and *maint*$\geq$*high*. On the other hand, the composition of the intent includes four many-valued attributes, i.e., *buying,safety,maint*, and *persons*.

Identifying costly objects

Ordinal Factors

Interordinal scalings are, in general, more complex for human readers. The reason for this is that in interordinal scaling an interval of threshold values has to be considered instead of only order ideals based on $\leq$ thresholds. For example, in Figure 12.8 (bottom) we find the concept *c* with the extent *686,1180,1444* that is a lower neighbor to two concepts bearing the attributes *maint*$\geq$*med* and *maint*$\leq$*high* respectively. Thus, the human reader has to consider the interval [*med,high*] within the linear order of threshold values for *maint*, which is *low, med, high, vhigh*. Moreover, *c* has the attributes *buying* $\leq$ *high*, *persons* $\geq$ *four*, and *safety* $\geq$ *med*. Thus, a human reader has to comprehend different *directions* of order, i.e., $\geq$ and $\leq$, at the same time. With this in mind, we focused on the $\geq$-ordering in Figure 12.9. Of course, this limits the expressiveness of possible explanations based on the views. However, as illustrated above, the comprehensibility increases. Furthermore, this approach results in fewer concepts in general. Hence, it allows us to use lower support values for the iceberg concept lattice. We present in Figure 12.9 the iceberg concept lattice using the support values of one for the negative class and three for the positive class. This approach altogether can be considered as an ordinal factor approach with respect to the context operation apposition [80].

From intervals to chains

In addition to the advantages discussed above, one may apply all analysis deductions to

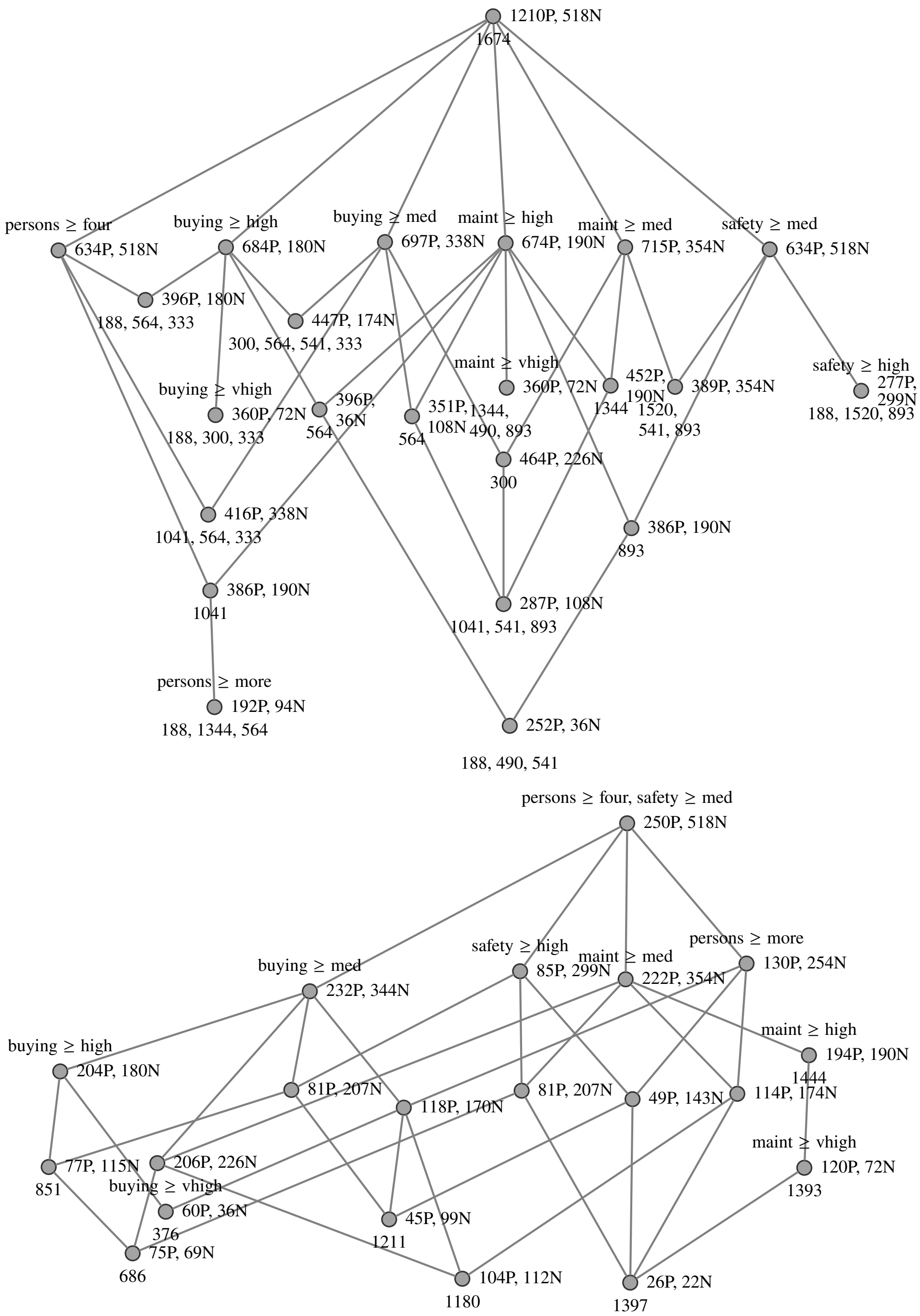


Figure 12.9: Forest view based on individual tree predicate views. We restricted the predicates to the ones using expressions with ≥. The diagram at the top shows the related iceberg concept lattice for the center objects bearing the positive class label, whereas the bottom diagram shows the analogue for the negative class labels.

the diagrams in Figure 12.9 that were explained for Figure 12.8.

The Interordinal Predicate View

As for our last analysis example we present an ordinal factor of the interordinal predicate view on the Random Forest. As in Figure 12.9, we choose ≥ and the same parameter for support. In contrast to the examples based on the tree predicate view, the lattices shown in Figure 12.10 encode a different kind of information for explaining the Random Forest. More precisely, the interordinal predicate view represents the model relationship between objects and the predicates of the Random Forest.

Implications

A distinctive feature of the interordinal predicate view is that it reflects implications between attribute thresholds values that are enforced by their order relation. For example, we find in Figure 12.10 that *vhigh*→*high*→*med* for the *buying* attribute of the data. Moreover, one can easily read the corresponding chains from the diagram.

Furthermore, one can infer from interordinal predicate view all valid attribute implications between values of different attributes of the data. For example, one can find in Figure 12.10 that the attribute value *safety* ≥ *med* implies *maint* ≥ *med*. Although there are also implications present in the iceberg concept lattice of the tree predicate view, we may note that those are not necessarily implications within the data. This is due to the fact that the tree predicate view is not guaranteed to be a view of the interordinally scaled context (cf. Proposition 38).

Implicational theory

We would like to conclude the analysis of the interordinal predicate view by emphasizing two important facts. First, the particular attribute threshold values were derived by the training procedure (i.e., Random Forest) and do therefore represent the *view* of the trained classifier function on the data. Hence, when revealing threshold value implications by means of the interordinal predicate view, we actually find implications that are valid within the data when viewed through the scaling of the Random Forest. Second, the set of all valid implications with respect to all data objects bearing the same class label is the implicational theory of this class as *"seen"* by the Random Forest. Thus, by computing both implicational theories, i.e., for both class labels, one can compare both theories for similarities and differences.

12.5 Related Work

Tree classifiers

There are a multitude of classification methods using trees and tree-ensembles, e.g., decision tree [31], Random Forest [29], or decision stumps [109], to name a few. The most important property of a single decision tree classifier is its human interpretability. For example, the visualization of such a tree provides insights to the classification process and, at the same time, presents a scaled view on the data set. Unfortunately, the latter approach received only very little research attention, so far. Methods that address these *scalings* are `RandomTreesEmbedding`, as implemented in *sklearn*, and *tree views* [61]. The first method extracts a partition of the data set objects depending on the tree leafs that classify them. This partition view, however, is a very coarse scaling of the data set and makes very little use of the hierarchical tree structures. The second method analyses the order structure of the trees through a concept lattice. While the authors provide a novel translation of the tree structures into the realm of Formal Concept Analysis, they do not elaborate how those can be utilized for interpretation. Furthermore, they solely reflect the order structure and its hierarchy induced on the data set. Proceeding in this manner does not account for the used tree predicates, which are essential for human comprehension. Nonetheless, the

Views

View interpretation

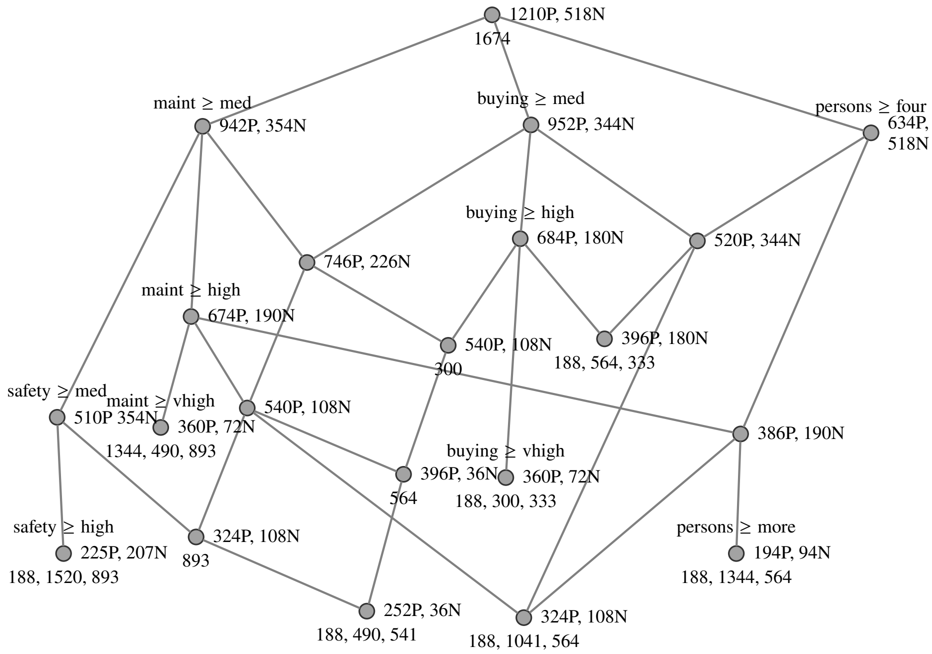


Figure 12.10: Forest view based on individual interordinal predicate views. We restricted the predicates to the ones using expressions with ≥. The diagram at shows the related iceberg concept lattice for the center objects bearing the positive class label.

translation approach itself is fruitful since this enables the application of FCA based post processing methods and therefor explanations. For example, scale-measures (cf. Chapter 8), TITANIC [208], core structures [90] or importance measures [132].

Tree merging

Other methods that try to achieve a unified view on tree ensembles combine all trees into a new tree through merging [203]. Yet, there are two main disadvantages of these approaches. The first is that their output is again a tree, which in contrast to a lattice order allows only for linear paths for each node. Thus, they loose the ability to cope with missing information and do not cover the concurrency of tree ensemble. The second disadvantage regards the interpretation of the output: the sole goal of the outputted merged tree is to induce the same partition on the data set, as the ensemble would. This, however, omits the internal representations of the trees.

FCA based classifiers

A different, yet related line of research is the construction classifiers from concept lattices [16, 175]. For example, one may compute the concept lattice in a top-to-bottom fashion with class purity, used as stopping criterion [17], and then select a tree from the generated partial ordered set. Although mathematically elegant, these approaches are outperformed by methods like Random Forest.

12.6 Discussion

Important attribute combinations

Obviously, our approach is capable of identifying important combinations of attribute threshold values and their influence on the classification results. Certainly, there is a wide range of combinatorial methods to identify interesting and meaningful combinations.

However, the major advantage of the proposed conceptual method is that it provides a structured mathematical way to directly and efficiently identify the important combinations and, at the same time, their semantic interpretation [80].

Future FCA research on decision trees

At this point, we would like to conclude our study on tree based classifiers by pointing out that the developed conceptual structures allow for the possibility for the application of a variety of other conceptual methods. For example, as outlined in earlier in this section, an analysis of implication structures between attribute threshold values within a class can reveal new insights into a Random Forest. Likewise, the conceptual views allow to compare different trained Random Forest classifiers for their implicational differences and similarities. A detailed investigation of these questions is planned as future work.

Parameter studies

In our work, we did not focus on technical details of the tree ensembles, especially hyperparameter studies for computing trees and Random Forests, since our approach emphasizes explaining a *given* forest with respect to known, and potentially unknown, data. Nonetheless, a detailed study investigating the relationship between the hyperparameters and the resulting forests and their different conceptual views could provide deeper insights into the training process of Random Forests. Another closely related topic we did not dive into is that the introduced views in combination with the classifier can be employed for automatically scaling of many-valued data sets. Furthermore, we could also envision applications for enumerating distinct decision trees, as they are lattice ordered [185]. Also, surrogate-based approaches [121] for explaining black-box classifier functions may profit from in-depth explanations based on our conceptual views. The same applies to bandit-based approaches [5].

Supervised automatic scaling

Surrogates

13

Conceptual Views on Neural Networks

Neural networks (NN) are known for their great performance in solving machine learning problems. However, these excellent results are almost always achieved at the price of human explainability. This problem is addressed in research and practice from different standpoints. There are calls to refrain from using NN for important problems and to rely on explainable methods, even if they give worse results in terms of accuracy [184]. The second major direction is to develop methods for explaining NN models. Such explanations can be classified as *local explanations*, i.e., why a particular data point was treated in a specific manner [178], and *global* explanations, i.e., approaches for explaining the whole NN model. The latter can be achieved, e.g., by mapping the NN to an explainable surrogate. A common approach for locally explaining NN models is to highlight activation at some hidden layer [70] or, if possible, project this inversely. For flat data, e.g., images, this is a viable approach since an essential explanatory component, the human, can be integrated into the process. This is not the case for high-dimensional or complex data. Global approaches are more difficult, in particular for high-dimensions, and therefore less frequent. A typical idea is to find an (explainable) surrogate for a NN, e.g., symbolic regression [3].

Explain neural networks

The difficulty of explaining complex data

We contribute to the growing interest in *global explanations procedures for NN models* by extracting conceptual views from a neural network. We demonstrate how NN models can be represented through views and how surrogate training, e.g., with decision trees, can profit from this. We further demonstrate how to compare NN models, e.g., when derived from diverse architectures, using Gromov-Wasserstein [154] distance within the views. Moreover, we show by an application of subgroup discovery how human-comprehensible propositional statements can be derived from NN models with the use of background knowledge and conceptual views. This allows us to extract global rules in form of propositional statements using the neurons of the NN.

With views towards global explanations

13.1 Related Work

Input highlighting methods

Several approaches aim to provide insights or explanations into neural networks. Many of them highlight parts of the input that were relevant for a particular prediction [178], so called *local explanations*. Those however, rely on the user's capability to comprehend input data representations. Hence, this approach is infeasible for problems with higher dimensional inputs. To overcome this limitation, the SOTA is to apply neuro→symbolic [188] methods to derive explanations based on *symbolic concepts*, i.e., explainable (symbolic) binary attributes. For example, Mao et al. [150], Asai and Fukunaga [11] and Fong and Vedaldi [70] introduce methods which classify the inputs of a model to pre-defined concepts. Hence, they require manually created input representations for all pre-defined concepts, in contrast of extracting them automatically. Particularly successful is *TCAV* [120], which predicts the importance of user-defined concepts. The above are complemented by methods that automatically detect concepts for a given set of input/output pairs through identifying similar patterns of input samples at a given layer, e.g., *ACE* [81]. So far these methods do detect only particularly outstanding concepts. Recent works try to estimate to which extent a detected set of concepts is capable to approximate the model [227]. This approach, however, emphasizes classification performance and not explainability, i.e., concepts that are important for explanations may be omitted. This is in general true for surrogate based procedures that were not designed towards human comprehensibility [3].

Concept based explanations

Automatic *Concept* based explanations

The problem with initial layers

Moreover, a recent study shows that the translation of initial layers does often correlate with random layers or gradient detectors in the input [1]. The most crucial downside of the automatic detection methods above is that although they provide symbolic concepts, they do not have to be interpretable. The overall principle of our approach is based on the fact that a substantial portion of the input data is aggregated and represented in the last hidden layer [43, 124].

A global interpretation of the NN needs a decoding into a human comprehensible space. We contribute towards this problem by deriving interpretable (cf. Proposition 22) conceptual views (cf. Chapter 12).

13.2 The views of neural networks

First, we have to encode aspects of a neural network that we want to analyze in the form of (pre-scaled) many-valued contexts. To derive them, we first introduce the structure of neural networks (see Figure 13.1). We may note, that the following characterization may not be applicable to all but most types of neural networks, including feedforward neural networks.

Neural network structure

Let N be the set of neurons of the last hidden layer of a NN. We interpret an NN as a function that maps input objects $g \in G$, that are represented as $g = (v_1, \ldots, v_m) \in \mathbb{R}^m$, to outputs in $[0,1]^{|C|}$ for classes C. The parameter m specifies the number of input features (see Figure 13.1). Naturally, we can interpret each neuron $n \in N$ as a function by itself from the input layer up to the activation of n, i.e., $n : \mathbb{R}^m \to \mathbb{R}$. The output neurons can be characterized analogously by a map $c : \mathbb{R}^{|N|} \to \mathbb{R}$. With $w_{i,j}$ we address the weights connecting the output neuron $c_i \in C$ with hidden neuron $n_j \in N$.

Many-valued views

Definition 50 (Many-Valued Views of a NN). *Let NN be a neural network, C its output classes and $N = \{n_1, \ldots, n_h\}$ the neurons of the last hidden layer. We define the* ***many-valued view*** *as $\mathcal{V}_{\mathbb{D}} = (\mathbb{O}_{\mathbb{D}}, \mathbb{W}_{\mathbb{D}})$, where $\mathbb{O}_{\mathbb{D}} := (G, N, \mathbb{R}, O)$ is a many-valued context with $n_j(g_i)$ equal to the activation of neuron n_j when NN receives g as input. The other many-valued*

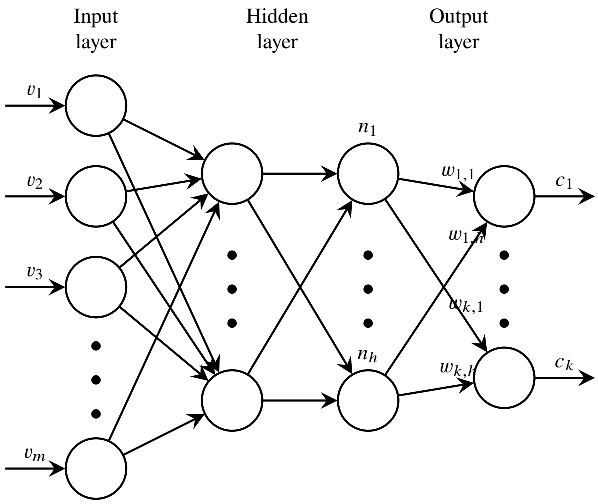


Figure 13.1: The structure of (feedforward) neural networks with annotated notions.

view is defined as a many-valued context $\mathbb{W}_{\mathbb{D}} := (C, N, \mathbb{R}, W)$ where $n_j(c_i)$ is equal to the weight $w_{i,j}$ that connect neuron n_j with the output for c_i. We call $\mathbb{O}_{\mathbb{D}}$ the ***many-valued object view*** *and $\mathbb{W}_{\mathbb{D}}$ the* ***many-valued class view*** *of NN.*

In some settings, we interpret these views to be numeric matrices, i.e., $\mathbb{W}_{\mathbb{D}} \in \mathbb{R}^{|C|\times|N|}$ and $\mathbb{O}_{\mathbb{D}} \in \mathbb{R}^{|G|\times|N|}$.

Motivation

To give a short motivation: With the many-valued object view $\mathbb{O}_{\mathbb{D}}$, we want to study the activation of the neurons N given an object g. Complementary, with the many-valued class view $\mathbb{W}_{\mathbb{D}}$, we investigate the relation of the neurons N to the outputs $c \in C$ by their corresponding weights $w_{i,j}$. In Figure 13.2 we depict example views for a neural network. In this, we find that $n_k(o_t)$ is greater than $n_1(o_t)$, from which we infer that the relation of o_t to n_k is greater than n_1. We want to employ the just introduced views to comprehend the complete classification that is captured by a NN model. We can represent any object g as a row in the many-valued object view matrix, i.e., $O(g) := (n_1(g), \ldots, n_h(g))$. Analogously, we can represent any class c_i as a row in the many-valued class view matrix, i.e., $W(c_i) := (w_{i,1}, \ldots, w_{i,h})$. The outputs of the NN for class c_i follow from the term $O(g) \cdot W(c_i) + b$, where b is a bias. This can be rewritten as $|O(g)| \cdot |W(c_i)| \cos(O(g), W(c_i)) + b$ where $\cos(O(g), W(c_i))$ is the cosine value of the angle between $O(g)$ and $W(c_i)$. Thus, to understand the inner representation of the classes C within the NN, it may be reasonable to grasp the objects and classes in the same space and classify objects using similarity measures. Using this approach, we can introduce an **object-class distance map** $d_{\mathcal{V}} : G \times C \to \mathbb{R}, (g, c) \mapsto d(O(g), W(c))$, where a sensible choice for d is the cosine similarity or the Euclidean distance (see Chapter 6). We will investigate both in Section 13.3.1. Hence, using $d_{\mathcal{V}}$ and similar distance maps for $G \times G$ and $C \times C$, one can derive a pseudo-metric space (cf. Definition 32) $(G \cup C, \hat{d}_{\mathcal{V}})$. From this representation of G and C one can infer a simple classification map, e.g., by applying **1-NN classification**.

Classification with views

Both views in a single space

Similarity of neural networks

Conceptual views enable a direct comparison of NNs. One can employ the **Gromov-Wasserstein distance** [154], as experimentally demonstrated in Section 13.3.2. We contrast our results with a baseline of model fidelity. We may note two important facts. First,

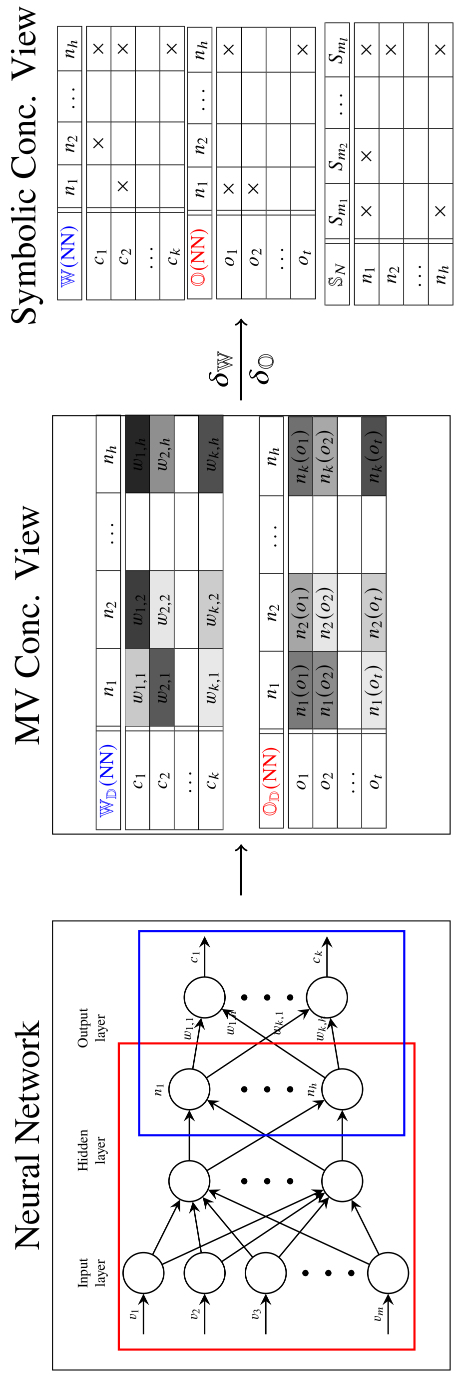


Figure 13.2: A simplified neural network drawing (left), its many-valued views (middle) and its conceptual views (right).

our approach for similarity is comparable to the recent idea of relating neural networks to particular **kernel spaces** [136, 197]. This enables us to study how objects are hierarchically clustered in such a space. We may stress that our notion does not consider how objects are mapped into this (kernel) space, but rather investigates the space itself. Second, the used GW distance is invariant with respect to permutations of the many-valued conceptual views.

13.2.1 (Symbolic) Conceptual Views

Scaling of many-valued views of NNs

The next step is to derive conceptual views from the many-valued views, i.e., mapping objects into a symbolic space. Given the ordinal pre-scaling of the many-valued views and dichotomy of class assignments, we decided to use interordinal scales $\mathbb{I}_2$. These scales split the value domains into two parts based on a single threshold value. In our approach, we use two separate thresholds for the many-valued object view $\delta_{\mathbb{O}}$ and the many-valued class view $\delta_{\mathbb{W}}$ (see Figure 13.2). As a final remark before we introduce the conceptual view on NN we want to point out a simple but powerful observation. The to be employed relational structure is invariant with respect to row- or column permutations in the related many-valued conceptual view (Definition 50).

Conceptual views

Definition 51 (Conceptual Views on Neural Networks). *Let* $\mathcal{V} = (\mathbb{O}, \mathbb{W})$ *the many-valued conceptual view of a NN and let* $\delta_{\mathbb{O}}, \delta_{\mathbb{W}}$ *be threshold values. We define the* ***conceptual views*** $\mathcal{V} = (\mathbb{O}, \mathbb{W})$ *by*

$$\mathbb{O} := (G, N \cup \bar{N}, I_{\mathbb{O}}),\ \text{with } (g, n_j) \in I_{\mathbb{O}} :\Longleftrightarrow n_j(g) > \delta_{\mathbb{O}} \text{ and} \qquad \text{(Object View)}$$
$$(g, \bar{n}_j) \in I_{\mathbb{O}} :\Longleftrightarrow n_j(g) \leq \delta_{\mathbb{O}}$$
$$\mathbb{W} := (C, N \cup \bar{N}, I_{\mathbb{W}}),\ \text{with } (c_i, n_j) \in I_{\mathbb{W}} :\Longleftrightarrow w_{i,j} > \delta_{\mathbb{W}} \text{ and} \qquad \text{(Class View)}$$
$$(c_i, \bar{n}_j) \in I_{\mathbb{W}} :\Longleftrightarrow w_{i,j} \leq \delta_{\mathbb{W}}.$$

We introduced with $\bar{N} := \{\bar{n} \mid n \in N\}$ *a set of artificial symbols and use them as defined above.*

This definition enables investigations of the representation of a neural network with methods from formal concept analysis. Moreover, we are able to construct human comprehensible explanations for a NN given a background ontology, e.g., in form of human annotations of the objects or classes. We exemplify that in Figure 13.2 using a formal context $\mathbb{S}_N$ that employs interpretable features $S_{m_1}, \ldots, S_{m_l}$ as background knowledge to decode the symbols in the conceptual views in $\mathbb{S}_N$. We provide more details in Section 13.4. Suitable threshold values $\delta_{\mathbb{W}}, \delta_{\mathbb{O}}$ depend on the architecture of the to be analyzed NN model. For example, if the activation function is ReLu, the neuron's co-domain is positive. Thus, it becomes difficult to determine a reasonable δ for negative symbols $\bar{N}$, as studied in Section 13.3.3.

Decode symbols with background knowledge

Good threshold values

13.3 Experimental Study

Experiments

We support our theoretical modeling of conceptual views, in particular the derived formal contexts, through an experimental study using common and well known data sets and NN models. First, we evaluate the suitability of the many-valued views through the in Section 13.2 introduced pseudo-metric space and a classification task. Second, we show how one may compare many-valued views, possibly from different NN models. Third, we demonstrate how to derive a human comprehensible representation for a NN model, that we can employ for explanations in Section 13.4.

Table 13.1: The average weights $w_{i,j}$, object values $n_i(g)$ and the number of neurons $|N|$ and activation function f of the last hidden layer of tensorflow imagenet models.

Model	$\mathbb{W}$ - values	Bias	$\mathbb{O}$ - values	$\|N\|$	f
VGG16	-5.359e-07 ± 0.008	1.404e-06 ± 0.191	0.679 ± 1.514	4096	ReLu
VGG19	-6.707e-07 ± 0.008	-1.287e-05 ± 0.192	0.613 ± 1.402	4096	ReLu
IncV3	-3.808e-05 ± 0.034	-0.0099 ± 0.308	6.025 ± 15.13	2048	ReLu
DenseNet121	2.139e-08 ± 0.049	-1.014e-07 ± 0.012	1.731 ± 4.603	1024	ReLu
DenseNet169	1.456e-08 ± 0.039	-1.038e-07 ± 0.012	1.675 ± 5.529	1664	ReLu
DenseNet201	1.019e-08 ± 0.036	-1.178e-07 ± 0.011	1.146 ± 4.167	1920	ReLu
MobilNetV1	-0.0001 ± 0.081	-0.005 ± 0.744	0.435 ± 0.838	1024	ReLu
MobilNetV2	-3.138e-05 ± 0.041	0.0002 ± 0.319	0.358 ± 0.747	1280	ReLu
NasNetLarge	-2.080e-07 ± 0.026	4.424e-05 ± 0.040	0.198 ± 0.533	4032	ReLu
NasNetMobile	-3.336e-07 ± 0.039	0.0001 ± 0.066	0.382 ± 4.389	1056	ReLu
ResNet50	3.774e-07 ± 0.033	-4.881e-08 ± 0.009	0.546 ± 0.871	2048	ReLu
ResNet101V2	6.668e-06 ± 0.027	0.0016 ± 0.292	39.97 ± 167.8	2048	ReLu
ResNet152V2	1.038e-05 ± 0.026	0.0016 ± 0.287	94.08 ± 187.4	2048	ReLu
ResNet50V2	8.014e-07 ± 0.028	0.0011 ± 0.292	19.91 ± 74.65	2048	ReLu
IncResNetV2	-3.060e-05 ± 0.037	-0.0012 ± 0.230	106.8 ± 124.9	1536	ReLu
XCeption	-3.246e-06 ± 0.055	0.0008 ± 0.281	2.974 ± 13.41	2048	ReLu
EffB0	-7.495e-05 ± 0.068	-5.143e-05 ± 0.058	0.065 ± 0.321	1280	Swish
EffB1	-5.647e-05 ± 0.063	-4.343e-05 ± 0.045	0.056 ± 0.313	1280	Swish
EffB2	-7.152e-05 ± 0.059	-4.153e-05 ± 0.054	0.019 ± 0.260	1408	Swish
EffB3	-6.323e-05 ± 0.054	-3.547e-05 ± 0.046	0.010 ± 0.252	1536	Swish
EffB4	-3.106e-05 ± 0.050	-3.138e-05 ± 0.057	-0.039 ± 0.194	1792	Swish
EffB5	-2.043e-05 ± 0.049	-2.738e-05 ± 0.055	-0.036 ± 0.170	2048	Swish
EffB6	-8.656e-06 ± 0.046	-2.691e-05 ± 0.071	-0.043 ± 0.135	2304	Swish
EffB7	-9.562e-06 ± 0.041	-2.441e-05 ± 0.060	-0.041 ± 0.136	2560	Swish

Table 13.2: The fidelity between twenty-four neural networks and their many-valued object/class view using 1-NN for classification.

Model	Euclidean	Cosine	Model	Euclidean	Cosine
VGG16	0.945	0.841	ResNet101V2	0.995	0.466
VGG19	0.942	0.842	ResNet152V2	0.999	0.314
IncV3	0.990	0.753	IncResNetV2	0.999	0.983
DenseNet121	0.978	0.737	XCeption	0.977	0.792
DenseNet169	0.989	0.843	EffB0	0.944	0.933
DenseNet201	0.972	0.728	EffB1	0.960	0.946
MobilNetV1	0.575	0.449	EffB2	0.969	0.957
MobilNetV2	0.947	0.925	EffB3	0.974	0.961
NasNetMobile	0.935	0.808	EffB4	0.981	0.972
NasNetLarge	0.880	0.831	EffB5	0.979	0.972
ResNet50	0.954	0.800	EffB6	0.982	0.976
ResNet50v2	0.995	0.734	EffB7	0.985	0.979

13.3.1 Many-Valued Views on ImageNet

Data sets

We demonstrate that many-valued views are capable of capturing a large share of a NN model. For this, we use all twenty-four[1] NN models from tensorflow that are trained on the ImageNet [52] data set. The object view is calculated using the test set, i.e., 100k images, of *ImageNet* used in the ILSVRC [186] challenge. In Table 13.1 we compiled basic statistics on these networks and our views. Although we report in columns two and four mean values and their standard deviation, we may stress that we do not consider the individual values to be normally distributed.

Evaluate ImageNet Views

To evaluate the quality of our views, we compare a one-nearest-neighbor (1-NN) classifier on the in Section 13.2 introduced pseudo-metric space $(G \cup C, \hat{d}_{\mathcal{V}_{\mathbb{D}}})$ directly with the NN classification function on all 100,000 test images. In detail, we use model fidelity, i.e., we count the instances where the 1-NN outputs the same class label as the NN and normalize this number by the cardinality of the test set. The results are depicted in Table 13.2. We differentiate in our experiments between using cosine similarity and Euclidean distance within $\hat{d}_{\mathcal{V}_{\mathbb{D}}}$.

Observations

We find that the view model is capable of achieving high fidelity (see Table 13.2). The MobilNetV1 model is the only exception. Moreover, we can state that using the Euclidean distance is superior to the cosine similarity in all instances. This is in particular true for the ResNet models, where the difference is up to 0.6. Furthermore, for the EfficientNets we notice that there is an almost monotone relation between the number of neurons N (last hidden layer) and the fidelity. All this together suggests that the many-valued conceptual view is meaningful and that classification functions that are based on the resulting pseudo metric space can be used as surrogates for the NN model.

13.3.2 Similarity of neural networks

Comparing many-valued views

Based on the many-valued views, we can derive for all NN models a pseudo-metric space as introduced in Section 13.2. Hence, given the theory about metric spaces there are different approaches for comparing them. For example, one could compute the **Gromov-Hausdorff distance** [154]. However, due to the vast number of data points, any direct computation of the GH distance is infeasible. A different approach, which is still costly, but can be performed for a subset of the data, is the Gromov-Wasserstein [154] distance. In Figure 13.3 (right) we depict the individual distances for all considered models with respect to the class and object view. We employed ten percent of the test data set and applied a uniform probability measure on the data points, i.e., a normalized counting measure. We compare our analysis to a baseline that is derived from the fidelity measure (Figure 13.3, left).

Parameters

Observations

From the pairwise fidelity diagram (Figure 13.3, left), we can infer that almost all models are very distinct with the exception of VGG16, VGG19, ResNet50 and the EfficientNet instances. In addition to that, we find that the later models become more similar with increasing number of neurons. The similarity plots for the views are different from the fidelity plot. We can visually identify clusters of models. These clusters do often correspond with similar networks architectures. For example in the object view we observe two clusters. In the class view this clustering is finer.

[1] July 2022

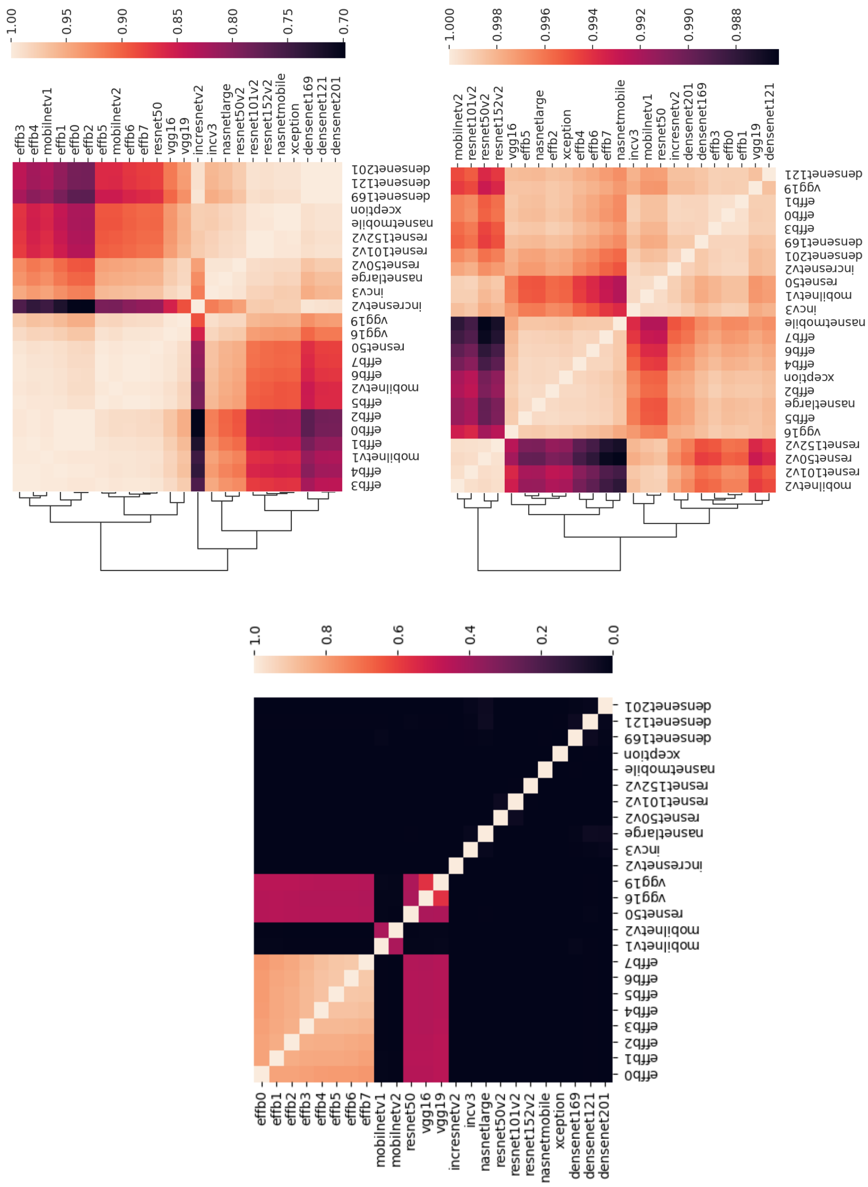


Figure 13.3: The similarity of twenty-four neural networks trained on the ImageNet data set. The base-line (left) is pair-wise fidelity between the employed models compared to a similarity using Gromov-Wasserstein distance on the object (top right) and class (bottom right) view.

13.3.3 Symbolic Conceptual View

Influence of the activation function

In this section we study thoroughly the activation functions and the number of neurons for a reasonable determination of threshold values in order to compute meaningful symbolic conceptual views.

Data set

To evaluate the influence of the choice of the activation function as well as the number of neurons, we trained one NN architecture several times on the *Fruits-360* [162] data set. The used data set contains 67,692 images of 131 types of fruits or vegetables. The test set contains an additional 22,688 images.

NN model

We train the architecture from the Fruits-360 experiment[2] using all procedure parameters from Mureşan and Oltean [162] and modified the last two hidden layers. For the (last) hidden layer N we vary the size 2^n between 2^4 and 2^9 with powers of two. For the layer before that we follow the common approach for smooth decrease in dimension, i.e., we chose $2^{\lfloor \frac{10+n}{2} \rfloor}$ with $4 \leq n < 10$ dependent on the last hidden layer. For activation functions, we studied the impact of **ReLu**, **Linear**, **Swish** and **TanH** in all layers. For each parameter setting we trained ten models and computed their respective conceptual views for the test data set, see their distributions in Figures 13.4 and 13.5. Statistics on the quality of the computed views can be seen in Figure 13.7. We tested these distributions against normalization of the column vectors in the views and can report that the reported results are invariant.

Observations

In general, we observe that the distributions for the object views differ quite largely among the different activation functions. From these we found that TanH causes the most notable seperation of positive and negative values. We depicted the results for all activation functions in Figures 13.4 and 13.5. Furthermore, we find that splitting with $\delta_{\mathbb{O}} = 0$ seems to be meaningful for all examples with respect to separation and symmetry. This split into two set of almost equal size. The same is true for $\delta_{\mathbb{W}}$. Apart from these values, we experimented in this and all following experiments with different approaches to determine thresholds, such as mean values, median values, median per neuron, as well as kernel-density estimation for bi-variate Gaussians. However, the split at 0 was favorable with respect to the achieved model fidelity. We report these scores for all activation functions in Figure 13.7. We conclude from our ablation study that the use of TanH is suggested as well as $\delta_{\mathbb{O}}, \delta_{\mathbb{W}} = 0$. We acknowledge that the used data set might influence this choice [14].

Conceptual Views of ImageNet

Based on the just found parameters we derive the conceptual views for the ImageNet models via $\mathbb{I}_2$ scaling. We report the results in Figure 13.6. We found that the classes are uniquely represented (class separation equals 1), thus a perfection classification procedure is theoretically possible using the symbolic view.

Inverse scaling

In the following, we apply methods designed for binary or numeric matrices to formal contexts. For this, we employ inverse scaling to yield binary data tables with values 0 or 1. Such an inverse scale is reasonable due to the dichotomy of the contexts, i.e., since $(g,n) \notin I_{\mathbb{O}} \iff (g,\bar{n}) \in I_{\mathbb{O}}$ and $(c,n) \notin I_{\mathbb{W}} \iff (c,\bar{n}) \in I_{\mathbb{W}}$. The resulting data set combines $n, \bar{n}$ into a single binary attribute.

[2] https://github.com/Horea94/Fruit-Images-Dataset

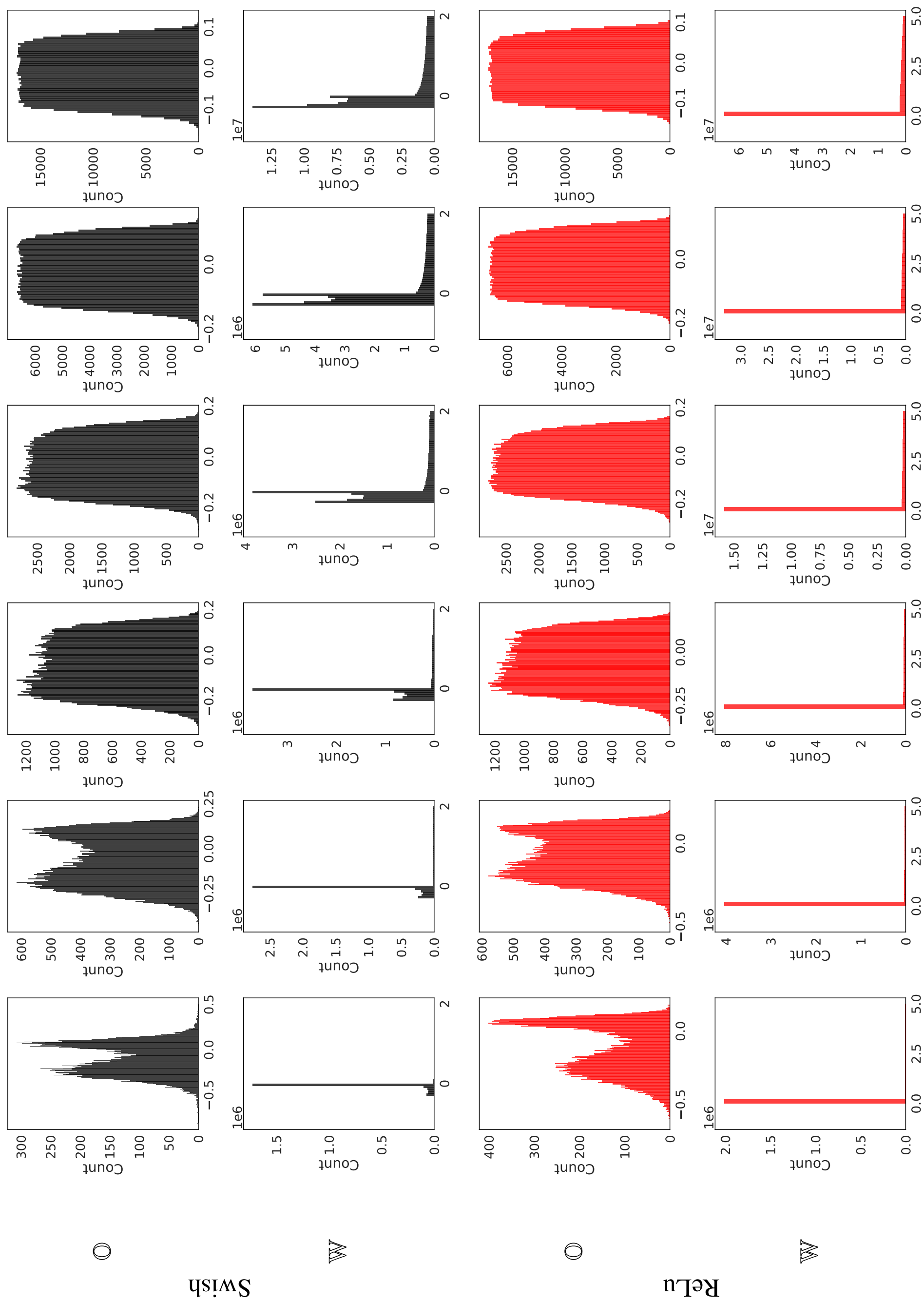


Figure 13.4: These are the value distributions for the object (𝕆) and class (𝕎) view for ten runs using the Fruits-360 data set and the swish and ReLu activation functions. The last hidden layer of size 2^4 (first column) to 2^9 (last column).

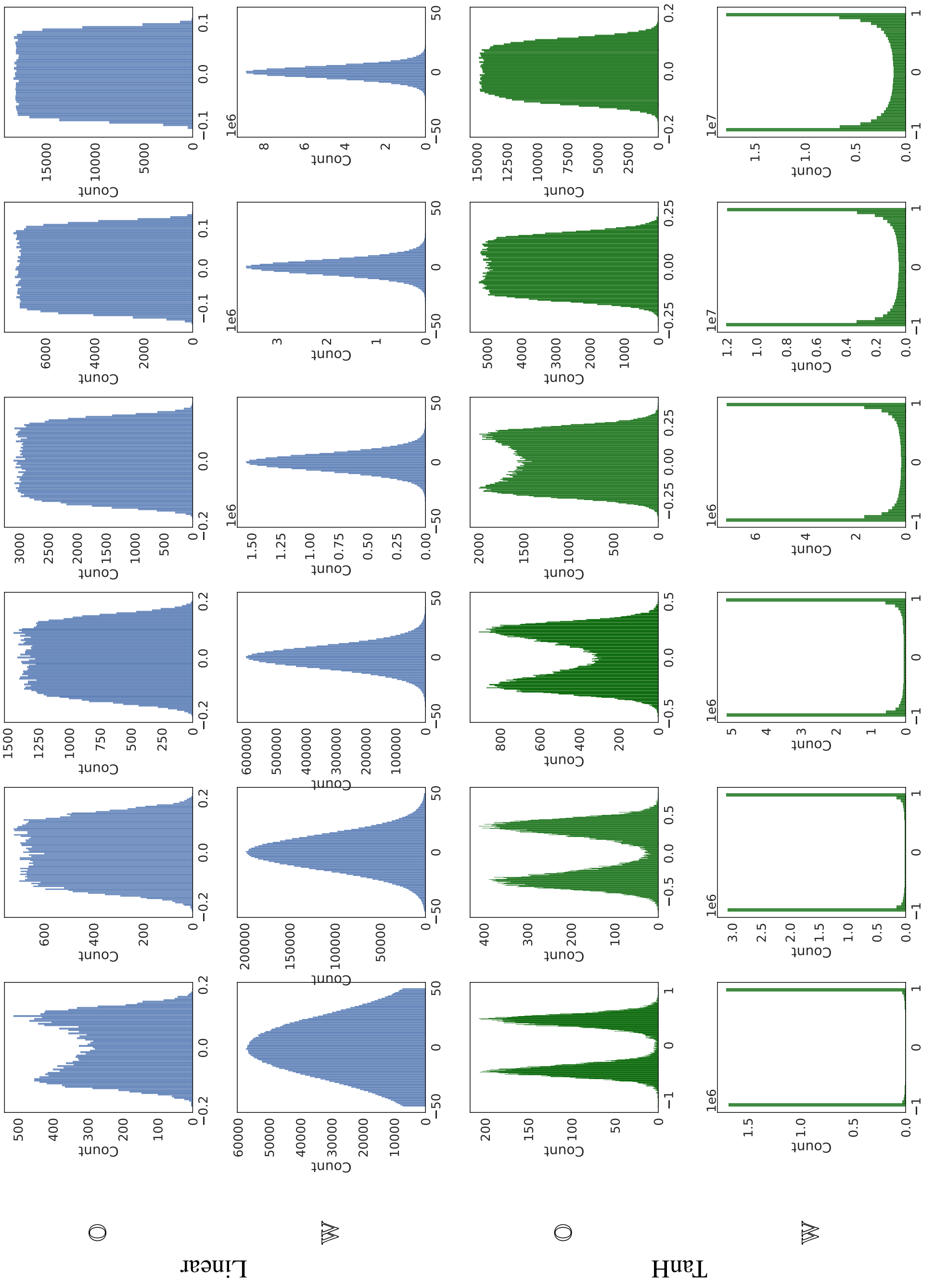


Figure 13.5: These are the value distributions for the object (𝕆) and class (𝕎) view for ten runs using the Fruits-360 data set and the linear and TanH activation functions. The last hidden layer of size 2^4 (first column) to 2^9 (last column).

Observations

We observe that a direct application of 1-NN procedure using the binary vectors that arise from the conceptual view does result in very poor classification performance for ReLu. In contrast to that, the Swish based models achieved from mediocre to good results. Especially the larger (EfficientNet) NN models resulted in better conceptual views. A reason for the unfavorable results with ReLu might esteem from its positive co-domain, which hinders the construction of negated attributes $\bar{N}$ in our approach. Some selected distributions can be found in Figure 13.8.

Conceptual views on Fruit-360

The twenty-four models in ImageNet employ ReLu and Swish activation functions only. Thus, we want to complement our experimental study on conceptual views with results for TanH activation, which we conduct again on the Fruits-360 data set. Hence, we trained five models, namely the base-line model from Mureşan and Oltean [162], VGG16, ResNet50, IncV3, and EffB0, the latter initialized with the original ImageNet weights. With exception of the baseline, we added to each model three dense layers (including dropout layers with p=0.2) on top, that are sized 1024, 256, and 32. To all models we also added an additional layer of size 16. This reduction in size added in order to enable human explainability. The baseline model as well as all added layers employ the TanH activation. The output (prediction) layer is a dense layer using *softmax* activation without bias. We used sparse categorical crossentropy as the loss function. All other relevant parameters for reproducing our results are drawn from the published baseline model. In the statistical experiments we

Model parameters

Table 13.6: The fidelity between twenty-four neural networks and their object/class view using nearest neighbor and cosine similarity for classification.

	1NN	Cos	Class Sep	Activation
VGG16	0.552	0.552	1.0	ReLu
VGG19	0.5672	0.5672	1.0	ReLu
IncV3	0.000	0.000	1.0	ReLu
DenseNet121	0.000	0.000	1.0	ReLu
DenseNet169	0.001	0.001	1.0	ReLu
DenseNet201	0.000	0.000	1.0	ReLu
MobilNetV1	0.036	0.036	1.0	ReLu
MobilNetV2	0.305	0.305	1.0	ReLu
NasNetMobile	0.004	0.004	1.0	ReLu
NasNetLarge	0.009	0.009	1.0	ReLu
ResNet50	0.007	0.000	1.0	ReLu
ResNet50V2	0.001	0.003	1.0	ReLu
ResNet101V2	0.012	0.012	1.0	ReLu
ResNet152V2	0.000	0.000	1.0	ReLu
IncResNetV2	0.000	0.000	1.0	ReLu
XCeption	0.220	0.220	1.0	ReLu
EffB0	0.758	0.758	1.0	Swish
EffB1	0.813	0.813	1.0	Swish
EffB2	0.869	0.869	1.0	Swish
EffB3	0.898	0.898	1.0	Swish
EffB4	0.929	0.929	1.0	Swish
EffB5	0.935	0.935	1.0	Swish
EffB6	0.951	0.951	1.0	Swish
EffB7	0.957	0.957	1.0	Swish

compare the out analysis on the above described architecture to models trained without the layer of size 16, i.e., only the three layers of size 1024, 256, and 32 were added.

Observations

The results in Figure 13.9 show that all models have high accuracy on the test data set, while the four transfer learned models outperform the baseline. We want to stress that our predictive results are only used to demonstrate that the model did fit to the classification problem. We find that both, Euclidean and cosine based 1-NN did perform well on the many-valued views as well as the conceptual views. In detail, we could not find significant difference between the representations. Moreover, we cannot identify significant differences in the classification performance with respect to the NN model. We also observed that an additionally trained decision tree classifier was unable to learn within the many-valued view representation. However, the same procedure applied to the conceptual view was capable of producing competitive classification results, a surrogate for the NN with very high fidelity.

Analyze views with FCA

Besides the investigations of the derived views with respect to classification tasks, we employ methods from FCA for their interpretation. This leads to *concept based explanations* [188] of neural networks through formal concepts in the respective views. The size of the concept lattices (see Section 13.3.3) serve as an upper bound for the number of learned concepts.

View sizes

We computed the concept lattices for Base, ResNet, VGG16, IncV3, and EffB0 and find, that their sizes vary between 126,487 (VGG16) and 134,100 (IncV3) for $|N| = 16$, and between 3,498,829 (VGG16) and 3,803,799 (ResNet50) for $|N| = 32$. If we restrict our computation to N, i.e., omitting the artificially introduced negations $\bar{N}$, we find the concept lattice sizes decrease by one magnitude. In detail, between 5,200 (VGG16) and 6,573 (EffB0) for $|N| = 16$, and 150,884 (EffB0) and 198,152 (IncV3) for $|N| = 32$. We compiled all values in Section 13.3.3. Overall, we observe that all are similar in size and therefore cannot be visualized using a line diagram. We note that formal concepts are composed of combinations of features. The minimum number of encoded features is present by the meet-irreducible elements. Hence, the number of meet-irreducible elements is bound by the number of attributes in the context, which serves as a lower bound the concepts captured by the NN model.

Shared concepts

Independent of the size does this translation to the realm of FCA enable the application of various knowledge-based methods, such as Description logic or Subgroup discovery, as investigated in Section 13.4. Within FCA we might consider to analyze cuts of the lattice, in particular those for which we suspect problems in the representation. As we discovered in previous experiments, that *Apple Red*, *Pink Lady*, *Plum* and *Cherry* are indistinguishable by some concept based representations (Figure 13.9), one might want to "zoom" into those. We did this statistically with Figure 13.11 . In the former one can identify formal concept based similarities among the selected fruits and the number of their shared concepts. For example, the fruits *Cherry* and *Plum* are indistinguishable in the views of IncV3 and *Pink Lady* and *Apple Red* in EffB0. From these and instances with high similarity, we can infer for which instances it is difficult for a model to distinguish fruits. Contrary to that, are *Cherry* and *Plum* in the ResNet model quite different due to the small number of shared concepts. Moreover, do the similarity plots indicate that the models extract different properties from the data.

Conceptual structure

From the concept lattice (see Figure 13.12) the reader can infer the hierarchical dependencies between the different fruits (objects).

Table 13.7: Study on the influence of the activation function and number of neurons on the quality of the computed views. The quality is measured in terms of fidelity of nearest neighbor classification in the many-valued views (MV-Fid) and conceptual views (V-fid).

	2^4	2^5	2^6	2^7	2^8	2^9
Swish						
$\delta_{\mathbb{O}} = 0$ Split	57.1/62.9	52.8/47.2	49.2/50.8	44.5/55.5	45.6/54.4	45.0/55.0
$\delta_{\mathbb{W}} = 0$ Split	65.4/44.6	65.0/35.0	62.0/38.0	60.1/39.9	57.2/43.8	56.5/43.5
Model Acc	93.5± 0.8	94.5± 0.5	95.3± 0.3	95.1± 0.3	95.4± 0.4	95.1± 0.5
MV-Fid	99.5± 0.4	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0
V-Fid	77.1± 9.2	88.8± 1.3	89.6± 1.6	88.8± 1.4	88.4± 0.9	86.7± 1.6
ReLu						
$\delta_{\mathbb{O}} = 0$ Split	55.1/44.9	55.1/44.9	54.7/45.3	53.3/46.7	55.3/44.7	54.1/45.9
$\delta_{\mathbb{W}} = 0$ Split	66.3/33.7	66.7/33.2	64.0/36.0	61.6/38.4	58.9/41.1	58.2/41.8
Model Acc	93.7± 0.4	94.5± 0.5	94.9± 0.5	94.9± 0.5	95.0± 0.4	94.8± 0.6
MV-Fid	99.7± 0.0	99.8± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0
V-Fid	79.9± 3.7	89.0± 1.2	90.0± 1.2	89.5± 1.1	89.0± 1.5	88.2± 1.5
Linear						
$\delta_{\mathbb{O}} = 0$ Split	49.8/50.2	49.6/50.4	49.5/50.5	49.9/50.1	49.9/50.1	49.9/50.1
$\delta_{\mathbb{W}} = 0$ Split	49.7/50.3	49.3/50.7	49.7/50.3	50.0/50.0	50.0/50.0	50.0/50.0
Model Acc	85.3± 0.6	88.8± 0.7	89.9± 1.0	92.0± 0.5	91.8± 0.9	91.5± 0.9
MV-Fid	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0	99.9± 0.0
V-Fid	55.5± 1.9	64.7± 1.4	68.3± 2.6	74.8± 1.7	78.2± 2.1	81.5± 1.1
TanH						
$\delta_{\mathbb{O}} = 0$ Split	49.7/50.3	49.7/50.3	49.8/50.2	49.9/50.1	50.0/50.0	49.9/50.1
$\delta_{\mathbb{W}} = 0$ Split	49.9/50.1	49.8/50.2	49.8/50.2	50.0/50.0	49.9/50.1	50.0/50.0
Model Acc	90.5± 0.8	94.3± 0.5	94.7± 0.5	94.9± 0.4	95.0± 0.4	94.8± 0.3
MV-Fid	98.3± 0.5	99.5± 0.1	99.7± 0.0	99.7± 0.0	99.8± 0.0	99.8± 0.0
V-Fid	94.3± 1.4	97.4± 0.4	97.7± 0.4	97.6± 0.1	97.8± 0.2	97.6± 0.2

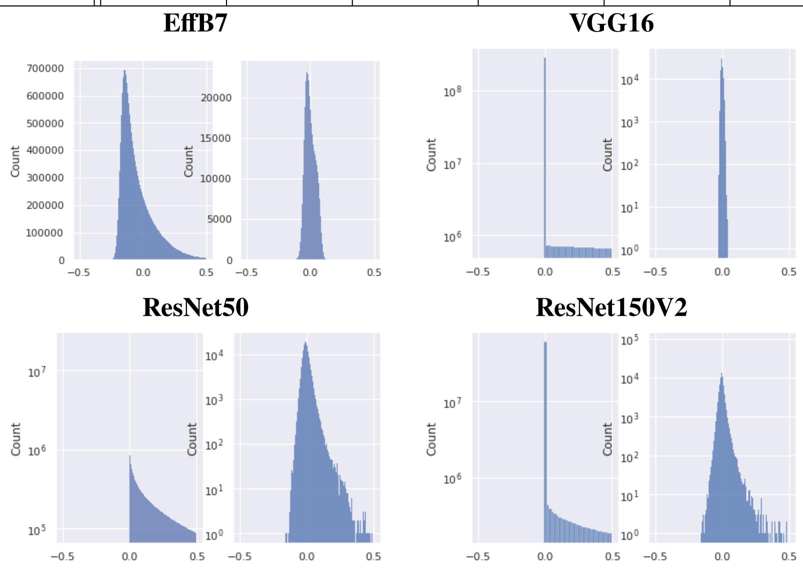


Figure 13.8: The value distributions for the object/class view for four models.

Table 13.9: Five neural networks (using TanH activation) and their (symbolic) conceptual views were captured by different surrogates (decision tree, 1-NN). We report their fidelity and accuracy. In symbolic conceptual view: IncV3 was unable to distinguish *Cherry 1* and *Plum*; EffB0 was unable to distinguish *Apple Red 1* and *Apple Pink Lady*.

Model	Model	DTree		Euclidean		Cos		
	ACC	ACC	Fid	ACC	Fid	ACC	Fid	Class Sep
$\|N\| = 16$								
Baseline	0.936	0.017	0.017	0.935	0.989	0.935	0.988	
VGG16	0.988	0.017	0.018	0.988	0.998	0.988	0.997	
ResNet50	0.989	0.018	0.018	0.989	0.998	0.989	0.997	
IncV3	0.983	0.013	0.013	0.983	0.999	0.984	0.999	
EffB0	0.984	0.007	0.007	0.984	0.998	0.983	0.984	
Symbolic								
Baseline	0.936	0.857	0.879	0.927	0.964	0.927	0.964	1.0
VGG16	0.988	0.972	0.977	0.988	0.994	0.988	0.994	1.0
ResNet50	0.989	0.952	0.957	0.988	0.996	0.988	0.996	1.0
IncV3	0.983	0.975	0.988	0.984	0.997	0.984	0.997	0.992
EffB0	0.984	0.938	0.934	0.984	0.996	0.984	0.996	0.992
$\|N\| = 32$								
Baseline	0.954	0.021	0.019	0.953	0.996	0.953	0.994	
VGG16	0.987	0.014	0.014	0.986	0.998	0.986	0.997	
ResNet50	0.991	0.024	0.024	0.991	0.999	0.990	0.997	
IncV3	0.989	0.010	0.010	0.989	0.999	0.989	0.999	
EffB0	0.987	0.021	0.022	0.987	0.999	0.986	0.997	
Symbolic								
Baseline	0.954	0.837	0.845	0.949	0.983	0.949	0.983	1.0
VGG16	0.987	0.873	0.876	0.987	0.996	0.987	0.996	1.0
ResNet50	0.991	0.946	0.949	0.990	0.996	0.990	0.996	1.0
IncV3	0.989	0.976	0.983	0.989	0.998	0.989	0.998	1.0
EffB0	0.987	0.904	0.906	0.986	0.993	0.986	0.993	1.0

13.4 Abductive learning of partial explanations

Adding background knowledge

Conceptual views enable the application of various logical methods to derive human-comprehensible (partial) explanations. We draw from this correspondence and construct a formal context $\mathbb{C} = (C, S_M, I_\mathbb{C})$, where $S_M = \{S_{m_1}, \dots, S_{m_l}\}$ is a set of human-interpretable features that are known about the classes C, i.e., *background knowledge*.

Definition 52 (Symbolic Interpretation of Views on Neural Networks). *Given the conceptual views $\mathcal{V} = (\mathbb{O}, \mathbb{W})$ of a NN, background knowledge $\mathbb{C}$, and a similarity relation $\sim$ on the classes $\mathcal{P}(C)$. Then is the formal context $\mathbb{S} = (N, S_M, R)$ with $(n, S_m) \in R :\Longleftrightarrow \{n\}^{I_\mathbb{W}} \sim \{S_m\}^{I_\mathbb{C}}$ the* ***symbolic interpretation*** *of the NN with respect to $\mathbb{C}$ and $\sim$.*

Decode neuron attributes

We require $\sim$ to be reflexive and symmetric but not necessarily transitive. The task for symbolic interpreting a NN is to deduce or infer $\sim$ using background knowledge, which is

Table 13.10: The size of the concept lattices of the conceptual views in Figure 13.9 (see $|N| = 16$) and for 32 neuron (see $|N| = 32$), for all (see All column) and only positive attributes (see Pos column).

Model	$\lvert N\rvert = 16$		$\lvert N\rvert = 32$	
	All	Pos	All	Pos
Base	130969	6517	3192044	155416
ResNet50	133130	5872	3803799	165009
VGG16	126487	5200	3498829	193516
IncV3	134100	5670	3782226	198152
EffB0	132403	6573	3767964	150884

done in the next section. Given a symbolic interpretation for a NN we are able to express neurons in terms of the human-interpretable features S_M by applying the incidence relation in $\mathbb{S}$, i.e., for all $n \in N$ one can compute $\{n\}^R$. Furthermore, if $\mathbb{S}$ is additionally equipped with propositional logic $F[S_M, \{\vee, \wedge, \neg\}]$ then FCA [91] also provides the means for expressing neurons in terms of propositional statements.

Background knowledge for Fruit-360

We want to motivate how symbolic interpretations can be used to interpret neurons in terms of (human-comprehensible) features S_M and vice versa. To demonstrate both cases, we analyze the symbolic view of the Fruit-360. For the attributes S_M we use visual features

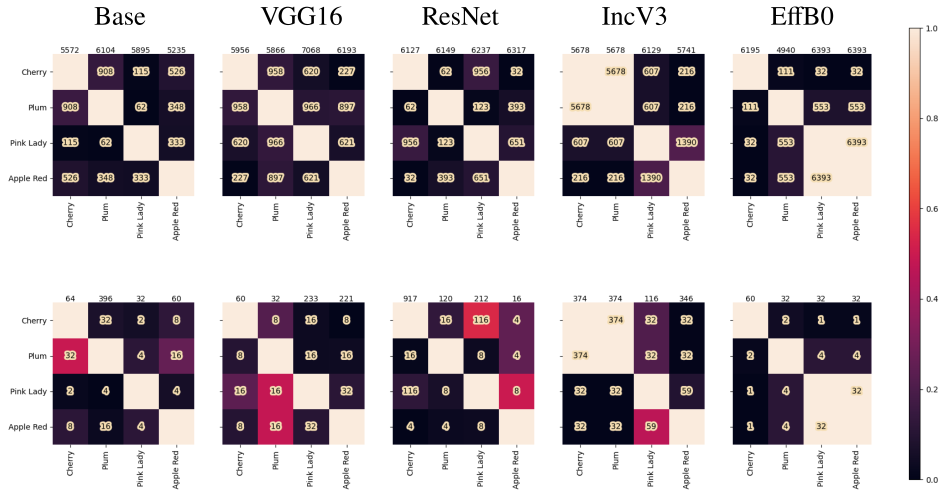


Figure 13.11: FCA results for *Apple Red*, *Pink Lady*, *Plum* and *Cherry* using views on all attributes $N \cup \bar{N}$ (top heatmaps) and only *positive* N attributes only (bottom heatmap). The columns labels of each heatmap displays the number of formal concepts. The number within a cell is the number of shared concepts between the related row/column fruits. Heat indicates its fraction.

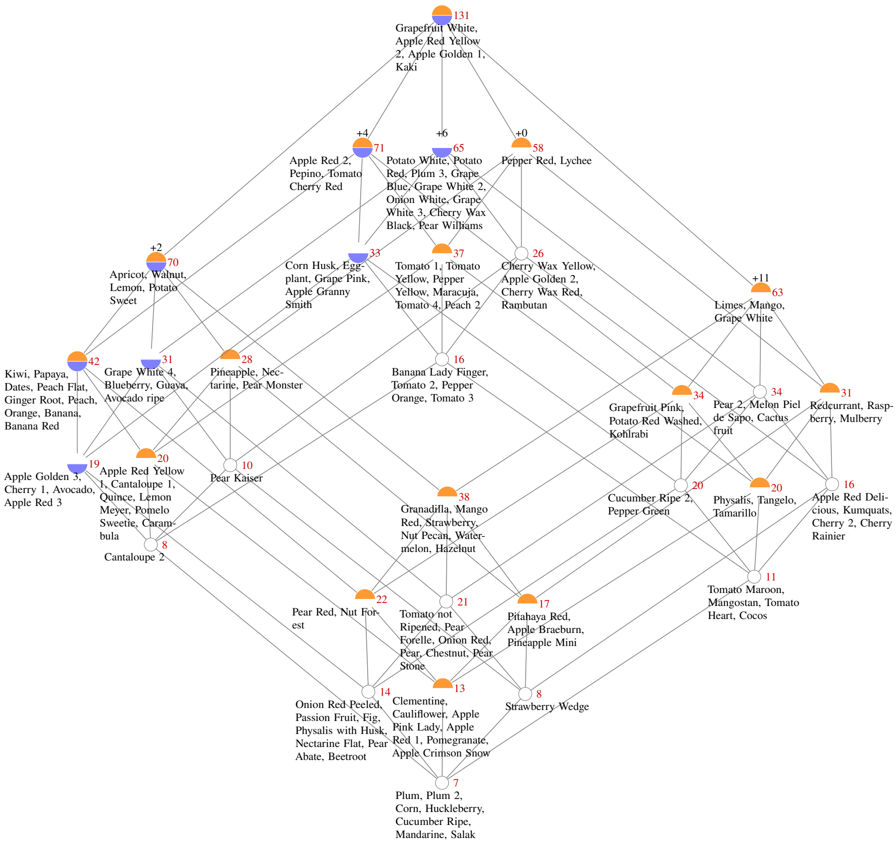


Figure 13.12: Concept lattice representation of the selected fruits by the VGG16 model (without attributes $\bar{N}$). Formal concepts containing *Apple Pink Lady* and *Apple Red 1* are highlighted in Orange and *Cherry 1* is highlighted in blue.

(1), such as shapes or colors, and the *Scientific classification* taxonomy (2) published in Wikipedia[3] for each fruit/vegetable. We combined the German and English Wikipedia articles in order to derive a data set as complete as possible. We infer the similarity relation in our experiment using *subgroup detection* [96], as implemented in `pysubgroup`.

Evaluate the symbolic interpretation

We depict four example results in Figure 13.13, where the taxons are given in the respective diagram titles, e.g., Apple, Orange. Diagrams on the left depict subgroups in terms of neurons, and, vice versa, on the right in terms of interpretable features. For both sides we find on the abscissa propositional statement combining the respective features.

From the high share (see ordinate) of the respective subgroups we can infer that the propositional statements using the neurons or S_M features describe the taxons from adequately up to very good. In particular for the latter case (right) we see that the subgroups are pure, yet, not complete. To give two concrete statements: 1) Fruits that are not brown, not stained, not orange and not star shaped will use neuron $\bar{13} \in \bar{N}$. 2) If the neuron $\bar{13}$, $\bar{14}$ and $\bar{9}$ are used by the NN we can infer that the fruit is orange with confidence about 0.54. Using this method one can infer the similarity relation $\sim$ and provide an explanation framework.

[3]See for example `https://en.wikipedia.org/wiki/Apple` in the right box.

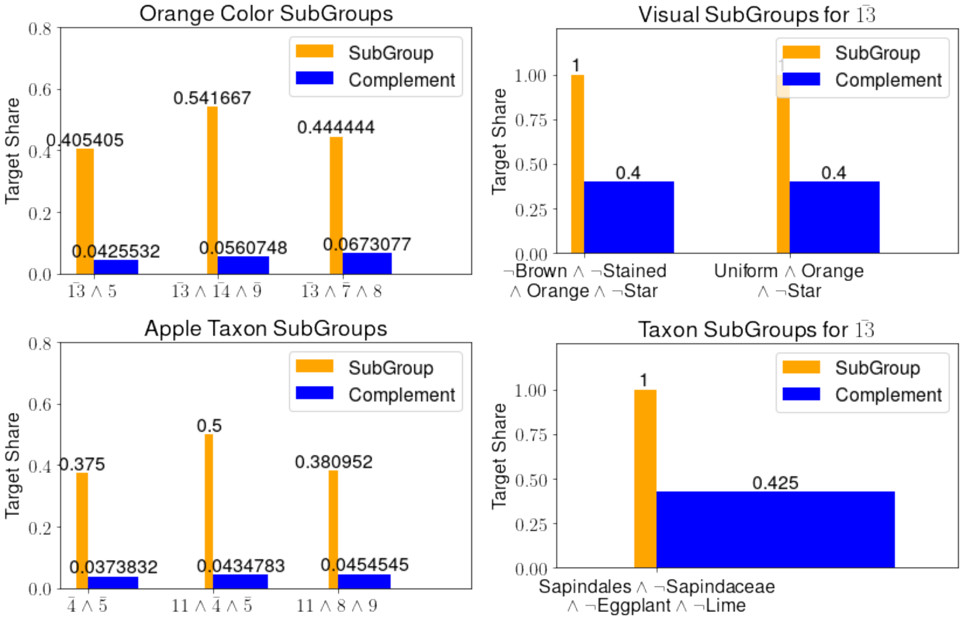


Figure 13.13: Exemplary results of the subgroup detection. The width of the orange bar indicates the size of the subgroup and the height the ratio of elements in the subgroup that have the target attribute. The analogue applies to the complement of the subgroup in blue color.

13.5 Discussion

Novelty

With the presented chapter we have shown how conceptual views can be used to interpret neural networks. Compared to other explanation methods, our approach is novel and different to former ideas with respect to three properties: first, we do not employ further hardly explainable methods, such as autoencoders. Second, our method is global by design. Third, conceptual views, as introduced in our work, do not require pre-defined concepts and their related input representations. We accomplished this by decoding both, the weights of all output neurons and the activations of the last hidden layer.

Sparsity regularization

For future work, we can envision that an investigation on the influence of regularization, e.g., sparsity of hidden neurons N, may lead to smaller concept lattices and therefor more interpretable views.

Limitations with respect to outputs

Our approach is limited by the necessary existence of multiple outputs. However, there are common approaches for splitting single outputs. Yet, a more significant limitation concerns the restriction to non-recursive architectures. Adapting our approach to such settings is probably possible, but requires a substantial adjustment to the definition of the views. Finally, our method requires for human-comprehensible explanations the existence of domain-specific background knowledge.

Future applications

Apart from this we envision that the presented link of NN models to FCA using conceptual views allows for both the explainability of NNs as well as increasing the performance of NN surrogate learning procedures. Therefore, this beneficial research line should be further investigated and tested.

14

The Geometric Structure of Topic Models

The strengths of topic models

Topic models are a popular tool for clustering and analyzing textual data. They discover groups or hierarchies of topics in text corpora using various techniques. Using these topics one may characterize new text samples within the space the topics span. However, this topic space also makes it possible to better understand the original text corpus used to find the topics. Thus, they are an excellent tool for organizing and structuring large text corpora and for extracting knowledge about various entities that are contained in the text or can be derived from it.

Applications of topic models

The number of application domains for topic models is vast. Prominent domains are recommendation systems [38, 59, 113], sentiment analysis [216, 228, 232], and text summarization [72, 118, 182, 200]. A particular interesting application for the present work is analyzing and mapping entities from large text corpora [51, 189, 191, 211]. Besides that, they have been shown to be very useful in feature-heterogeneous domains, e.g., social network analysis combined with large corpora based topic spaces [139, 190].

Interpretable topic models

The vast majority of topic models encode relationships between topics and the terms (i.e., word or n-grams of these words) of which they are composed. This property makes them comparatively easy to interpret and allows for topic representations of individual documents or term representations of individual topics. This is in particular true for the topic modeling procedures non-negative matrix factorization (NMF) by Lee and Seung [135], as it enforces all components of a topic to be additive.

The problem with numerical interpretations

Methods that aim at explaining and interpreting the relation between sets of documents, topics, and terms, do often rely on vector space models. In order to derive human-readable visualizations, embeddings into two or three dimensional real-valued Euclidean spaces are computed. However, the resulting diagrams are limited (and might be) distorted by the employed maps. Since these maps are often non-linear in nature, the resulting visualizations are difficult to interpret. In particular, it does not make sense to relate proximity in the diagram to similarities between topics.

Relational approaches

A fundamentally different line of research aims at explaining topic models with methods rooted in ordinal data analysis [35, 40, 207, 209]. For this purpose, the natural term-topic and document-topic relationships of a topic model are implicitly understood as (geometric) incidence structures. These relate multiple topics or documents to each other in an interpretable manner.

Geometric interpretation of topic models

With this chapter, we build on this view and make explicit use of the geometric character of these relationships. In detail, we present how to derive *geometric structures* from the topic-term and document-topic relationships. We show how these structures allows for rich interpretations of the topic model and how to extract explainable patters from them. The latter contribution is based on ordinal motifs (cf. Chapter 9). Moreover, we point out how to comprehensively visualize the geometric character of a topic model based on said ordinal motifs.

Application

We demonstrate our approach based on a well researched topic model that was derived from a large corpus of scientific works within the realm of machine learning [189]. We show that our method is capable of capturing insights about authors and research venues extracted from the corpus data and how our method can be used to track changes in their individual topic distribution over time. Finally, we visually depict the interplay between terms and their temporal dependency for topics.

14.1 Related Work

Vector space interpretation

Several methods to interpret topic models have been proposed. Some of them represent the document-topic relation through a vector space models [189, 191]. This allows the reader to visually interpret this relation through proximity in two or three dimensional diagrams. Common methods applied here are multidimensional scaling [152] or t-distributed stochastic neighborhood embedding [215].

Relational approaches

Other explanation approaches follow a more relational approach. A very simple method is to compute a relation based on topic-topic correlation. This is, for example, computed using cosine similarities between topics in a vector space model [189]. Some topic models, like the *Correlation Topic Model* (CTM) [25], allow for directly inferring this relation from the model and presented in a graph structure. These approaches allow for one-to-one comparisons of topics. Extensions to allow for hierarchical interpretation of this relation has been proposed based on clustering techniques [12].

Incidence relations

With this chapter, we contribute to the explainability of topic models based on relations that are defined on the data, i.e., documents-topics and term-topics relations. These are often weighted [177] and are *scaled* to binary relations [74]. A basic investigation of the derived (binary) relations are visualizations using bipartite graphs [49]. Based on these graphs, simple (tree shaped) hierarchies can be inferred by applying hierarchical clustering methods [2]. A benefit of hierarchical interpretations is that one can infer topic-topic relations of higher arity from them. On top of that, they allow for assessing the overall *global* structure of a data. However, tree structures are very limited (cf. Chapter 12) in their expressiveness. For instance, there is only a single path connection two nodes.

Formal Concept Analysis

A hierarchical structure that is not limited by this property can be computed using *Formal Concept Analysis* [80] (FCA). With FCA, we compute from the bipartite graph the set of all maximal bi-cliques. These exhibit a natural order relation which results in a lattice structure, i.e., the concept lattice. An application of concept lattices on documents-topics and term-topics relations has been shown to be useful for organizing discussion forums in educational software [56]. We revisit this method after an introduction of incidence relations in topic models (Section 14.3.2).

Ordinal motifs in topic models

With our approach, we expand on FCA based methods by transferring new explanation approaches for concept lattice (cf. Chapter 9) into the realm of topic modeling. These allow for a novel global interpretation of the topic model structure.

Related lines of research

A different, yet related line of research is the computation of hierarchical topic models [40, 83, 138, 231]. Multiple visualization techniques have been proposed for their explanation [198].

14.2 Topic Models and their Interpretations

Towards topic models

Topic Models are statistical models that describe documents and words in terms of the *"topics"* they cover. There are a variety of machine learning methods proposed for this task.

Input data

Starting with a set of *textual* documents D (*corpus*) a vector representation is computed. The two most commonly applied representations are *bag-of-words* (BoW) and *tf-idf*. The BoW model splits a textual document $d \in D$ into words called *terms* and counts the number of occurrences of each term in d. Formally, for the set of all terms S in the corpus D a document $d \in D$ is mapped to $\mathbb{N}^{|S|}$ where $\text{BoW}(d) := (s_1, \ldots, s_k)$ and s_i is equal to the number of occurrences of s_i in d. The tf-idf model builds on this representation by including a measure of importance based on the rarity of terms. The *inverted document frequency* (idf) of term $s \in S$ in D equals the logarithm of the inverse of the document occurrences of s, i.e., $\text{idf}(s; D) := \log({|D|}/{|\{d \in D | s \in d\}|})$. The *term frequency* (tf) of s in d is equal to the normalized BoW representation of s in d, i.e., $\text{tf}(s; d) := {\text{BoW}(d)_s}/{|\text{BoW}(d)|}$ where $\text{BoW}(d)_s$ equals the value of $\text{BoW}(d)$ for term s and $|\text{BoW}(d)|$ equals the sum of all values. The tf-idf is defined as $\text{tf-idf}(d) := \big(\text{tf-idf}(s_1; d, D), \ldots, \text{tf-idf}(s_k; d, D)\big)$ where is defined as $\text{tf-idf}(s_1; d, D) := \text{tf}(s; d) \cdot \text{idf}(s; D)$. Since we do not further elaborate on these mappings, we assume that D is given in either vector representation for simplicity reasons.

Topic models

A topic model TM is a machine learning model that maps $d \in D$, usually in vector representation, into a *topic space* $\mathbb{R}^n$. Each dimension of this space represents a topic. For easier comprehension the topic space is often $[0, 1]^n$ and document representations are normalized to one. From this, one can derive ratios of membership for each document to the topics. A key difference between topic models and regular embeddings, i.e., mappings into lower dimensional spaces, is that the dimensions of the target space as well as the resulting images $\text{TM}(d)$ are interpretable. This is typically achieved by representing a topic $t \in T$ by a list of terms.

14.2.1 Topic Model Visualization

The analyzed model

With our method, we contribute towards a richer interpretation of documents in the topic space. We compare our method to three commonly used topic model visualizations based on the `SSH21` topic model from Schaefermeier, Stumme, and Hanika [189]. This topic model is computed on machine learning research papers [7] and has twenty-two topics. The topic model itself is discussed in more detail in Section 14.4.

Similarity heatmaps

The first visualization of `SSH21` is given by a similarity heatmap in Figure 14.1. This plot depicts in each cell the output of a similarity measure between two topics. In this case a cosine similarity based on the in Section 14.4.1 discussed term-topic relations was used. This plot is great at visualizing 1-to-1 relations between topics or possibly inferring small clusters of topics. However, it is difficult to infer n-to-n relations. In addition to that, one is not able to explain the similarity of topics in terms of term-topic or document-topic relations solely based on this visualization.

Vector space visualizations

Another commonly used visualization of topic models are embeddings into the $\mathbb{R}^2$ or (rarely) to $\mathbb{R}^3$. These visualizations aim at visualizing the proximity of topics. A popular

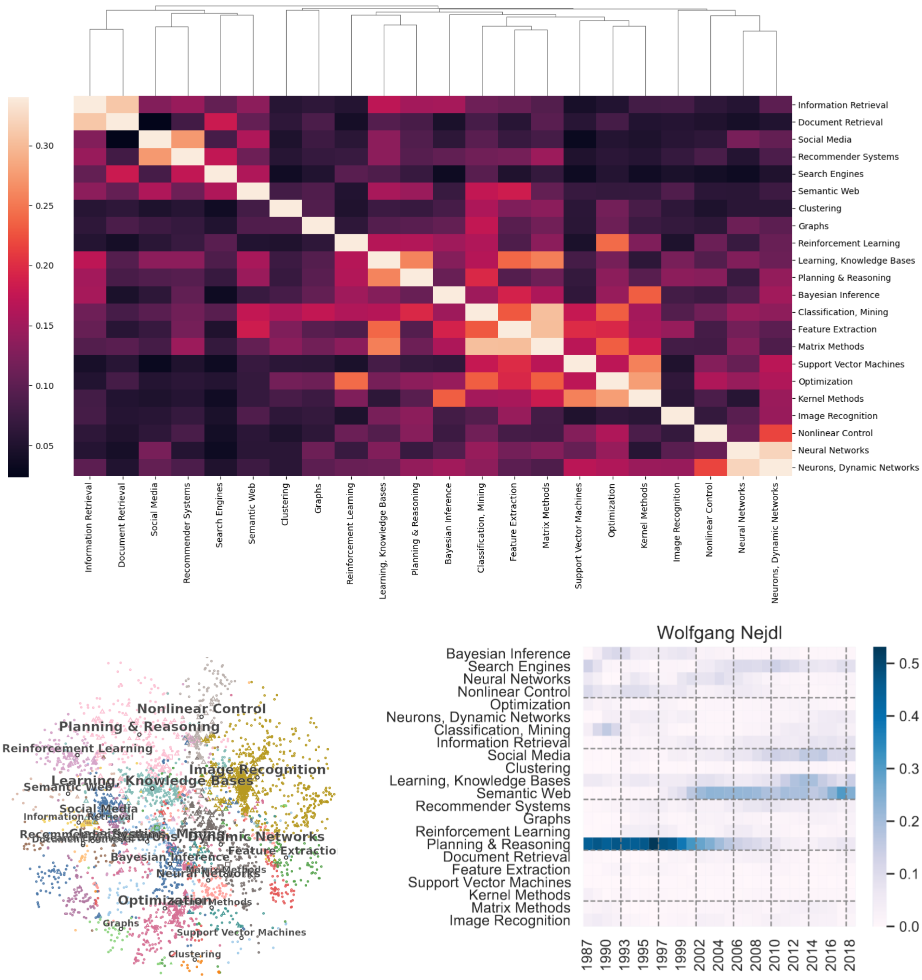


Figure 14.1: Visualizations of the `SSH21` topic model from the literature. The similarity heatmap (top) is from Figure 4.1 in Schaefermeier, Stumme, and Hanika [189], the vector space representation (bottom left) is from `https://sci-rec.org/maps` and Schäfermeier, Stumme, and Hanika [191] and the bottom right heatmap is from Figure 2 in [191].

method for this task is t-SNE [215] as depicted in Figure 14.1 (bottom left). The t-SNE method computes a non-linear embedding that is, in theory, capable of computing good visualizations [10]. However, methods like t-SNE are hard to explain due to their non-linearity. However, in case of non-linear mappings it is very difficult to relate the distance of topics in such a diagram with the distance of topics in the topic space. This makes it difficult to assess topic-topic relations from the resulting diagrams. A similar problem arises when interpreting cluster shapes which possibly are artifacts of the non-linear mappings. In contrast to that, linear models often fail to separate classes [10] or lack performance for very low dimensions [8].

Development over time

The third visualization is dedicated to the topic representations of individual entities like authors or venues that are in relation to a subset of the documents $H \subseteq D$. The depicted

$D \backslash T$	t_1	t_2	...	t_n
d_1	$w_{1,1}$	$w_{1,2}$		$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$		$w_{2,n}$
...				
d_l	$w_{l,1}$	$w_{l,2}$		$w_{l,n}$

$S \backslash T$	t_1	t_2	...	t_n
s_1	$\hat{w}_{1,1}$	$\hat{w}_{1,2}$		$\hat{w}_{1,n}$
s_2	$\hat{w}_{2,1}$	$\hat{w}_{2,2}$		$\hat{w}_{2,n}$
...				
s_k	$\hat{w}_{k,1}$	$\hat{w}_{k,2}$		$\hat{w}_{k,n}$

Figure 14.2: The weighted term-topic and document-topic relations.

heatmap displays for each year (column) the topic representation[1] of the author *Wolfgang Nejdl*. From this visualization we can infer the topic distribution of an author at track his/her development over time. Similar to the first diagram, this visualization fails to provide document-topic or term-topic relations. Moreover, it is not clear how a column is distributed over the documents in that year, e.g., it is not clear if there are documents on *Semantic Web* and *Planning & Reasoning* or if these topics are unrelated.

Our method

With our approach, we use methods from (order-)relational data analysis which do not rely on numerical embeddings. Instead, our methods extracts hierarchical (n-to-n) document-topic and term-topic relations with well defined semantic meaning. The employed method and related approaches are discussed in Section 14.3. In the following sections, we provide further comparisons of our method to the visualizations in Figure 14.1.

14.3 Conceptual Views on Topic Models

Hierarchical topic model interpretations

In contrast to the methods in the last section, from now on, we want to discuss hierarchical approaches based on FCA. For this, we require a proper definition of the incidences that naturally result from topic models. Based on this we will review previous research on topic modeling with FCA. We will extend some of these approaches in the next main section and consecutively develop a novel theory for representing and visualizing topic models on a global scale.

14.3.1 Incidence Relations in Topic Models

Relations in topic models

The goal of this section is to present a principled approach to extract relational structures from topic models. Every topic model TM exhibits at least two fundamental relations, i.e., the topics for any document $d \in D$ and the terms for a given topic $t \in T$. Often the relations of TM are weighted, e.g., a document might have topic t_1 with a weight of $w \in [0, 1]$. We depict these relations in Figure 14.2 via a document topic matrix (left) and a topic term matrix (right). The first matrix represents the documents in a lower dimensional topic space. The second matrix provides an interpretation of the topic space and is used for state-of-the-art topic model evaluation measures [108, 119] like `npmi` [179].

Extracting these relations

The document topic matrix can be extracted by embedding each document $d \in D$ by the topic model TM into the topic space, i.e., computing $\mathrm{TM}(d_i) = (w_{i,1}, \ldots, w_{i,n})$. For the computation of the term topic matrix there are multiple options depending on the used topic model method. The first is to embed a document that is only composed of term t_j into the topic space. This results in a row in the term topic matrix. The second option is applicable to methods that have an additional decoder map from the topic space to the document space,

[1]Each column depicts the mean topic representation of all documents published in a given year.

e.g., auto-encoder or NMF. In this case we can map the vector $(0, 0, \dots, 1, \dots, 0)$ that has a zero for all topics but t_j by the decoder. The result is a column in term topic matrix.

Conceptual scaling

Based on both matrices, we can infer incidence relations, i.e., binary relations $\mathcal{D} \subseteq D \times T$ and $\mathcal{S} \subseteq S \times T$, in the following way. One natural approach is to apply threshold values to the weights. We apply this to the document topic matrix by selecting a value $\delta \in [0, 1]$. This results in the **document-topic-incidence** $\mathcal{D}_\delta$ with $(d, t) \in \mathcal{D}_\delta$ iff the weight for (d, t) in the document topic matrix is greater than or equal to δ. We pursue a different path for the **term-topic-incidence** $\mathcal{S}$. We extract for each topic the top-n terms, i.e., the top n entries in the respective column in the term topic matrix. For a term $s \in S$ and a topic $t \in T$ is $(s, t) \in \mathcal{S}_n$ iff the weight of (s, t) is among the top-n greatest weights in column t in the term topic matrix. We denote the respective contexts by $\mathbb{T}(\mathcal{D})_\delta := (D, T, \mathcal{D}_\delta)$ and $\mathbb{T}(\mathcal{S})_n := (S, T, \mathcal{S}_n)$.

Contextual views

14.3.2 Analyzing Topic Models with Formal Concept Analysis

Related methods

The state-of-the-art on analyzing topic models with FCA is to derive a formal context from the document-topic $\mathcal{D}$ and the term-topic relation $\mathcal{S}$ [56]. This is done by applying thresholds to both relations as discussed in Section 14.4.1.

However, for the term-topic relation $\mathcal{S}$, we use the top-n term relation. This relation is used in evaluation measures, like `npmi` [179], which correlate with human interpretation of topics [108, 119]. So far, the resulting concept lattices have mainly been used as a means to navigate between forum entries.

14.4 Conceptual Views

Ordinal motifs in topic model relations

We now extend the just introduced procedures by means of the novel ordinal motifs approach (see Chapter 9). We want to introduce and demonstrate our novel approach based on an already published and extensively discussed topic model [189] which we call in the following `SSH21`.

Topic model parameters

It was build based on a document corpus on 35,200 scientific publications from the realm of machine learning research. It was consecutively evaluated on a corpus of about 350,000 documents from the same domain. All documents were retrieved via the *Semantic Scholar Open Research Corpus* (S2ORC[2]) [7]. The employed topic modeling technique is *non-negative matrix factorization* [135], a widely used procedure that has several advantages with respect to explainability [189]. We may remark at this point that our notion is agnostic with respect to the topic modeling technique. The topic model consists of twenty-two [189, Table 4.1] topics which were manually assigned based on the top ten terms per topic. The training corpus and therefore the resulting topic model has 14,828 terms.

14.4.1 Computing Incidence Relations

How to choose δ

For computing the document-topic-incidence $\mathcal{D}$ we have to set a threshold δ. Since the respective document topic matrix is (or can be) row normalized we have to choose a value from $[0, 1]$. The goal for any choice of δ is to derive a sparse [142] document topic incidence. Thereby documents are mainly represented by their most important topics. Moreover, this leads to a comprehensibly sized concept lattice, which fosters the overall understanding of the results. However, at the same time increasing the values for δ to much may lead to loosing

[2] `https://github.com/allenai/s2orc`

	0.1	0.2	0.25	0.3	0.4	0.5
Sebastian Thrun	398	89	51	31	18	13
Bernhard Schölkopf	640	144	80	51	20	17
Dieter Fox	342	66	42	28	15	10
Wolfgang Nejdl	387	94	58	39	19	14
ECML	568	173	108	74	37	20
RecSys	389	83	60	47	21	11
NeurIPS	2366	405	212	142	66	22
Neural Networks	1368	261	155	107	43	21

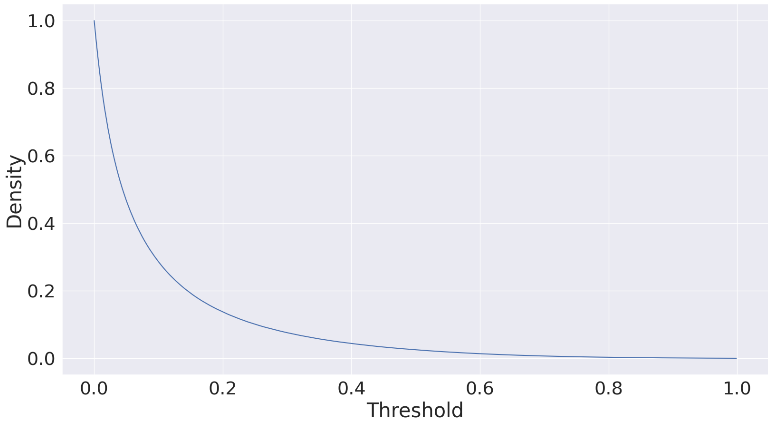


Figure 14.3: The density of the document-topic relation for given thresholds (bottom) and the concept lattice sizes for given entities and thresholds (top).

substantial parts of the concept lattice structure. We decided to determine δ based on the resulting density of $\mathcal{D}_\delta$. That is $|\mathcal{D}|_\delta/|D \times T|$, which is depicted on the left in Figure 14.3.

Author and venue views

We also want to propose a threshold estimation method tailored for investigating particular entities of a document corpus, such as authors or publication venues. This requires background knowledge about the topic corpus, e.g., which documents belong to a particular author or which documents were published at a certain venue. To address the implicit goal for achieving a comprehensible number of formal concepts in these cases, we also computed their number for selected values of δ and four different authors as well as four different venues, see Figure 14.3 top. Computing the number of concepts for a particular scientist a or venue v means that we considered only documents that were co-authored by a or published at v. The result is an induced sub-context of the $\mathbb{T}(\mathcal{D})_\delta$ formal context. For the set of venues we decided to look into *ECML, RecSys, NeurIPS* and *Neural Networks* as they were extensively discussed for the `SSH21` [189]. As for the set of authors we chose *Thrun, Schölkopf, Fox* and *Nejdl* since their individual publication trajectories were extensively discussed in a follow-up paper [191].

Determine δ

Based on the results that we achieved and reported in Figure 14.3, we decided for the threshold $\delta = 0.25$. We acknowledge that the number of formal concepts is still high in some cases. We depict the resulting number of concepts as well as the sizes of the induced sub-contexts in the first four columns of Figure 14.4.

Dealing with larger views

Formal Concept Analysis has a rich tool-set of data reduction methods. A particular feature of these tools is that they allow for controlling the (conceptual) error (see Chapter 11).

Table 14.4: The table displays the number of objects, attributes, the density and the number of concepts of the context derived from $\mathcal{D}$ for four authors and four venues. The sixth column depicts the number of concepts in the 2, 8-core and the last column the number of concepts after applying TITANIC with a minimum support value of three percent.

	documents (objects)	topics (attributes)	density	concepts	core concepts	view concepts
Sebastian Thrun	266	22	0.055	51	11	11
Bernhard Schölkopf	522	22	0.059	80	50	22
Dieter Fox	244	22	0.057	42	16	15
Wolfgang Nejdl	387	22	0.059	58	28	21
ECML	639	22	0.061	108	52	25
RecSys	856	22	0.058	60	29	23
NeurIPS	6233	22	0.060	212	202	25
Neural Networks	3521	22	0.061	155	144	21

The threshold for the $\mathcal{S}$ relation will be discussed in Section 14.4.6.

14.4.2 Conceptual Data Reduction

Data reduction methods

In order to reduce the size of the just computed incidence relations we rely on two established methods from FCA, TITANIC [208] and pq-cores [90]. The overall goal is to compute hierarchical representations of comprehensible size. We consider diagrams of size up to thirty or in some cases up to fifty concepts to have diagrams that are presentable in a human comprehensible way.

The core of views

***pq*-cores** The technique pq-cores computes the densest part of $\mathbb{T}(\mathcal{D})$. That is, the largest subset of documents $H \subseteq D$ and topics $S \subseteq T$ such that for each document $d \in H$ has at least p topics and every topic $t \in S$ has at least q documents in $\mathcal{D}$. This method can easily be restricted to a proper subset of the documents, such as all documents belonging to an author a or a venue v. We then call the result the *core topics* of an author or a venue respectively. The pq-core of a formal context is an induced sub-context $\mathbb{S} \trianglelefteq \mathbb{T}(\mathcal{D})$. Thus, the identity map ι on H yields a local scale-measure of $\mathbb{T}(\mathcal{D})$.

Core parameters

The proper choice of parameters p, q is supported by an importance measure. A parameter pair is considered to be interesting with respect to the data if every increase in p or q causes a large reduction in the number of formal concepts. In our study this lead to the pair $p = 2$ and $q = 8$.

Iceberg views

TITANIC The `TITANIC` algorithm computes the hierarchy of formal concepts in a top down fashion, with respect to a pre-defined importance parameter. For this parameter one can choose the value of a monotonous function on the set of concept intents. That is monotonous with respect to set inclusion. A commonly used function for this task is the support function, i.e., $\text{supp}_{\mathbb{K}} : \text{Int}(\mathbb{K}) \to [0, 1]$, where $\text{supp}(B) := {}^{|B'|}/_{|G|}$. In other words, the support of an intent B reflects the relative number of objects that have all attributes from B. Based on this, the `TITANIC` algorithm computes the hierarchy of all formal concepts that satisfy a minimum threshold value $c \in [0, 1]$. The result is called **iceberg concept lattice**, i.e., a join-semilattice. The main advantage of the `TITANIC` algorithm is that it computes concept hierarchies of readable size. In our case study on the `SSH21` topic model we found the value $c = 0.03$ to be sufficient for in Section 14.4.1 computed sub-contexts.

The TITANIC parameter

The reduction in terms of formal concepts by both methods are reported in the last two columns of Figure 14.4.

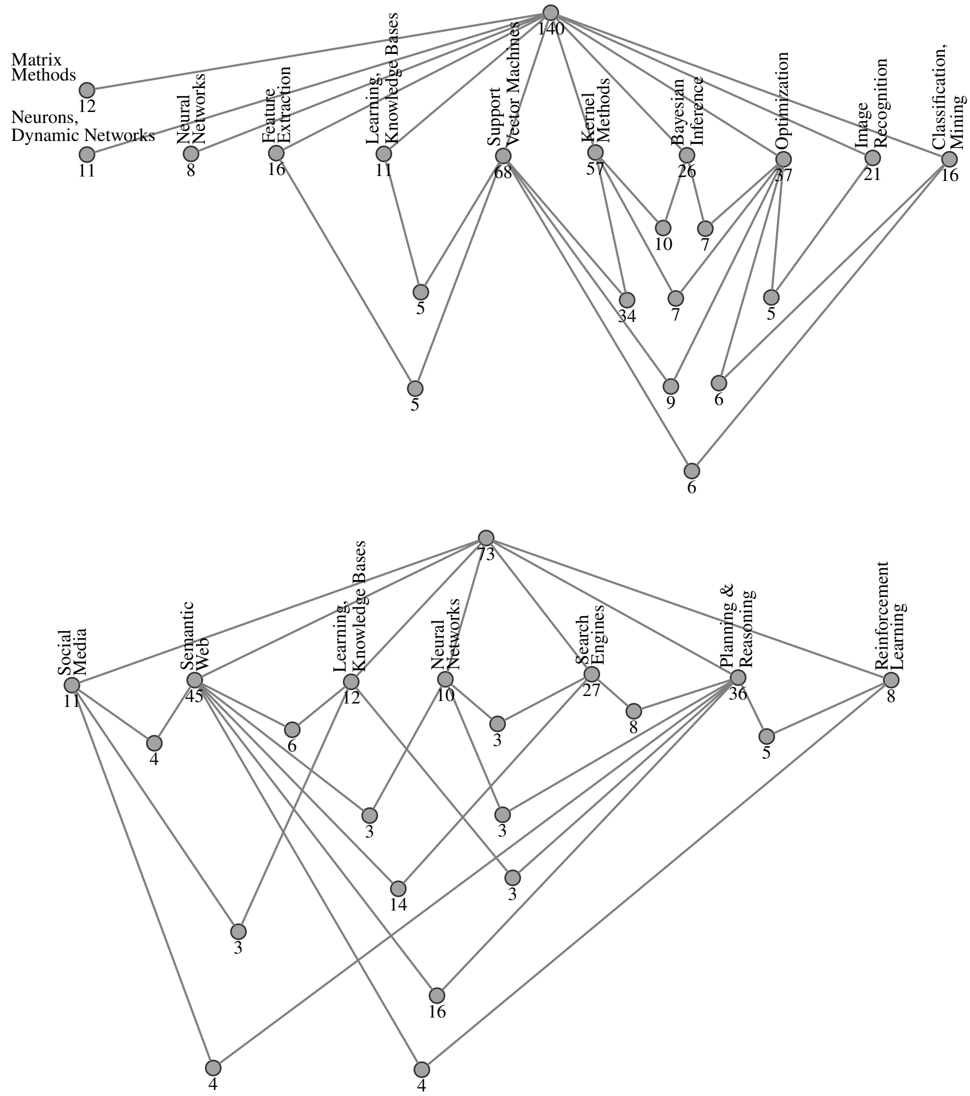


Figure 14.5: The concept lattice for *Bernhard Schölkopf* (top) and *Wolfgang Nejdl* (bottom).

14.4.3 The Resulting Conceptual View and Interpretation

Author views

In Figure 14.5 we depict the iceberg concept lattice for the researcher Bernhard Schölkopf[3] $\underline{\mathfrak{B}}_{\text{BS}}$ (top) and Wolfgang Nejdl[4] $\underline{\mathfrak{B}}_{\text{WN}}$ (bot). We employ order diagrams with in short-hand notation. Due to the large quantity of documents, we annotated extent sizes below concepts, e.g., $|$*Support Vector Machines*$'| = 68$.

Wolfgang Nejdl view

As a first observation, we can read from $\mathfrak{B}_{\text{WN}}$ that only seven topics out of twenty-two were identified as core topics of *Wolfgang Nejdl* by the combined method. The most frequent topic is *Semantic Web* which occurs in forty-five documents. Out of these, sixteen also have

[3] https://dblp.org/pid/97/119.html
[4] https://dblp.org/pid/n/WolfgangNejdl.html

the *Planning & Reasoning* topic and fourteen are also associated to *Search Engines*. The second most frequent topic is *Planning & Reasoning*, which occurs in thirty-six documents. Overall, the $\mathfrak{B}_{\mathrm{WN}}$ iceberg concept lattice has twenty-one concepts.

The benefit of structure based reductions

This novel approach for a comprehensive analysis of a topic model with respect to an entity allows for several new insights. By combining the core approach with TITANIC, we identified those topics for an entity that are not only frequent but also strongly interconnected [90]. With this structural approach, we overcome the limitations of commonly used methods that are based on filtering topics by frequency. For example, selecting the strongest signals in Figure 14.1 would result in a total ordered ranking without structural insights. In particular, one cannot infer how the topics are connected in terms of shared documents. Another advantage of our approach is that infrequent and isolated topics are omitted.

Bernhard Schölkopf view

We present the same analysis for $\underline{\mathfrak{B}}_{\mathrm{BS}}$. For *Bernhard Schölkopf* we can identify twenty-two concepts comprised of eleven core topics. The most supported and structurally important topics are *Support Vector Machines* with sixty-eight documents and *Kernel Methods* with fifty-seven documents. The topics coincide in thirty-four documents. The large overlap of these topics is not surprising due to the close connection between SVMs and kernel methods.

Comparison of views

Apart from the individual analysis of an entity's topic structure, our novel approach does also enable an in-depth cross entity comparisons. Comparing $\underline{\mathfrak{B}}_{\mathrm{BS}}$ and $\underline{\mathfrak{B}}_{\mathrm{WN}}$ we observe that both authors have two topics in common, namely *Learning Knowledge Bases* and *Neural Networks*.

Despite that, these topics are completely differently interconnected within their respective research. While Wolfgang Nejdl studies *Learning Knowledge Bases* in the context of *Semantic Web, Social Media* and *Planning & Reasoning*, Bernhard Schölkopf studies *Learning Knowledge Bases* in the context of *Support Vector Machines*. An analog differentiation can be found for *Neural Networks*.

Ordinal motifs

Ordinal Motifs in Lattices: The geometric aspects of our novel analysis method allow the application of ordinal motifs. This method extracts sub-structures within the concept hierarchy and provides visual *geometric* interpretations. We discuss the three types of ordinal motifs that occurred in our data. It is quite possible that for other data sets other ordinal motifs might occur. Nonetheless, our method can be applied analogously. Similar to our experiment in Section 9.6, we study the ordinal motifs on the set of attributes, i.e., via scale-measures on the dual contexts.

The studied motifs

In Figure 14.6 we show in the first two columns the nominal ordinal motif on 3 + 1 elements, crown ordinal motif on 10 elements and contranominal ordinal motif on 3 elements in context and concept lattice representation. For the contranominal ordinal motif we depict additionally the *inner* concepts in a different layout. By inner concepts we refer to the non-top and non-bottom concepts. We call this layout the **tulip layout**. This layout results in more readable drawings for join-semilattices, since it is free of edge crossings. The last column is discussed in greater detail in Section 14.5. In short, it employs a novel geometric technique for drawing lattices.

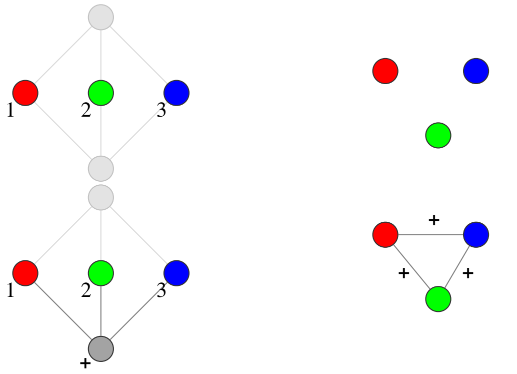
The $\mathbb{N}_{3^+}$ ordinal motif

The nominal ordinal motif is a simple structure that reflects the incomparability (Figure 14.6) of the related elements (topics). A slightly more expressive structure results by adding an additional (artificial) + object as the meet of all elements of the motif. This motif can be observed in real world data, for example, it occurs in Figure 14.12 (see concept with the term *learning*). Equipped with the + version of the ordinal motif, we will demonstrate a novel geometric drawing technique Section 14.5.

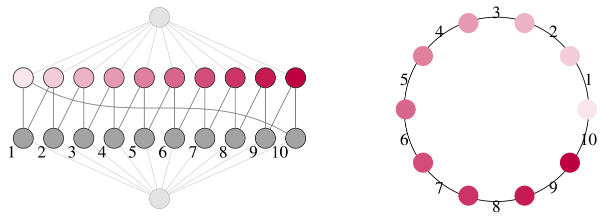
Crowns in topic views

The crown motif can be identified by its *zig-zag* pattern and reflects that there is a *round-trip* or a cycle along topics and documents (see Figure 14.6, second row). We were able to identify many crown ordinal motifs in $\underline{\mathfrak{B}}_{\mathrm{BS}}$ and $\underline{\mathfrak{B}}_{\mathrm{WN}}$. The largest crown ordinal

$\mathbb{N}_{3^+}$	r	g	b
1	×		
2		×	
3			×
+	×	×	×



$\mathbb{C}_{10}$	a	b	c	...	i	j
1	×	×				
2		×	×			
3			×			
...						
9					×	×
10	×					×



$\mathbb{B}_3$	r	g	b
c		×	×
p	×		×
y	×	×	

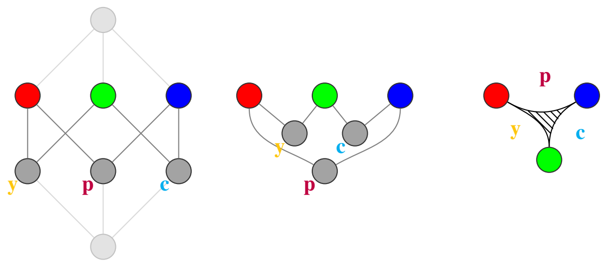


Figure 14.6: The nominal (top) crown (middle) and contranominal (bottom) ordinal motif in context (left), concept lattice (middle) and geometric drawing style (right) representation. The nominal context $\mathbb{N}_3$ has an optional + object and two different display styles based on the existence of this element. The contranominal ordinal motif has an additional layout, namely the tulip layout, for its inner concepts.

motif in $\underline{\mathfrak{B}}_{\mathrm{BS}}$ is a cycle over the topics (*SVM*, *KM*, *BI*, *O*, *CLass*, *SVM*). In $\underline{\mathfrak{B}}_{\mathrm{WN}}$ we find many cycles on four elements. For example, one of them is (*SemW*, *LKB*, *PR*, *SM*, *SemW*). The occurrence of such a motif may reflect a topic based cycle within the research history of an author. Hence, these cycles constitute interesting candidates for a temporal topical analysis. In any case, they are a useful tool to guide readers through an author's research documents.

Finally, the contranominal ordinal motif can be visually identified using the tulip layout in the lattice diagrams. This motif reflects that there is a unique set of documents for any combination of topics. Moreover, this type of motif represents a densely explored area

Contranominal ordinal motifs in topic views within the topic space. By explored, we refer to the research activities of the respective entity, e.g., an author. For example, within $\underline{\mathfrak{B}}_{\mathrm{WN}}$ we find that *Social Media*, *Semantic Web* and *Learning Knowledge Base*, or *Semantic Web*, *Neural Networks* and *Search Engines* constitute a contranominal structure. Furthermore, we find within $\underline{\mathfrak{B}}_{\mathrm{BS}}$ the contranominal motifs *Support Vector Machines*, *Kernel Methods* and *Optimization*, or *Support Vector Machines*, *Optimization*, *Classification*. The involved topics are structurally important within $\underline{\mathfrak{B}}_{\mathrm{BS}}$, i.e., within the research of *Bernhard Schölkopf*. Moreover, one may deduce from such structural results that the occurring topics are highly related within the machine learning domain. At least they represent a candidate for an important topic subset. Beyond the occurrence of ordinal motifs, the absence of such substructures also carries information.

A note on almost motifs

Particularly important are cases where a motif *almost* occurs, i.e., adding a few incidences results in a motif. These may reflect missing lines of research for future investigations. In the same way, almost occurring motifs may indicate that important data is missing or has been filtered in the process. For example while processing the corpus data with pq-core and TITANIC, we may have removed motifs with low support.

Towards a novel geometric drawing

Concluding this analysis, we want to motivate our novel approach (Section 14.5) by providing a different geometric interpretation of the motifs. For example, the crown ordinal motifs reflect a cycle shape of objects (documents) in the topic space. Hence, one should consider a drawing that reflects this shape directly. Analogously, the contranominal ordinal motif reflects a hyperball of documents in the topic space. Their importance was addressed in the text above. Yet, these structures cannot be (visually) recognized easily in the lattice diagram. Therefore, we represent them in our novel geometric representation in a unique shape, i.e., a filled n-polygon where n is the dimension of the hyperball.

Venue views

Ordinal Motifs in Lattices — Venue Analysis: Analogous to the analysis above, we present an ordinal motif analysis for the *Recommender Systems* (*RecSys*), *NeurIPS* and *Neural Network* venues on the same reduction parameters. We depict their iceberg concept lattices in Figure 14.7.

Recommender Systems venue

First, we observe, that the concept hierarchies differ in their ordinal structure. In particular, we see that the diagram for *RecSys* is the only one where we attribute annotations on sub-concepts: *Classification, Mining*, *Learning, Knowledge Bases* and *Matrix Methods* only occur in concepts where *Recommender Systems* occurs. This may be interpreted as *Recommender Systems* dominating the other topics. This relation constitutes a so far not discussed ordinal motif, called ordinal ordinal motif [sic] (see Figure 14.16). We acknowledge that the dominating role of the *Recommender Systems* topic is not surprising. Yet, we may point out that this fact was discovered in an unsupervised fashion, without background information. A majority of papers involve this topic, which is not surprising given the title of the conference. Striking the same chord, we find that the *Recommender Systems* topic occurs in the most number of documents. There are numerous nominal ordinal motifs of size two. The *Search Engines* topic gives rise to nine nominal motifs without the + element. We can conclude from this, that the topic *Search Engines* is isolated in this concept lattice structure. All other nominal ordinal motifs are in relation to the *Recommender Systems* topic, e.g., *Social Media* and *Recommender Systems*, or *Semantic Web* and *Recommender Systems*. The latter, are in fact nominal + motifs within the lattice structure, since their meet is present. We observe no (non-trivial) contranominal or crowns ordinal motifs.

Dominating topic

Nominal ordinal motifs

Discussion of the results

We want to summarize the novelty of the ordinal approach with respect to the *RecSys* data. Of course, the identification of *Recommender Systems* as the most important topic is a simple question of counting documents and does not require the ordinal approach. Schaefermeier, Stumme, and Hanika [189] enabled with their topic space trajectories (i.e., heatmaps of topic distributions over time) the identification of co-occurring topics. Yet,

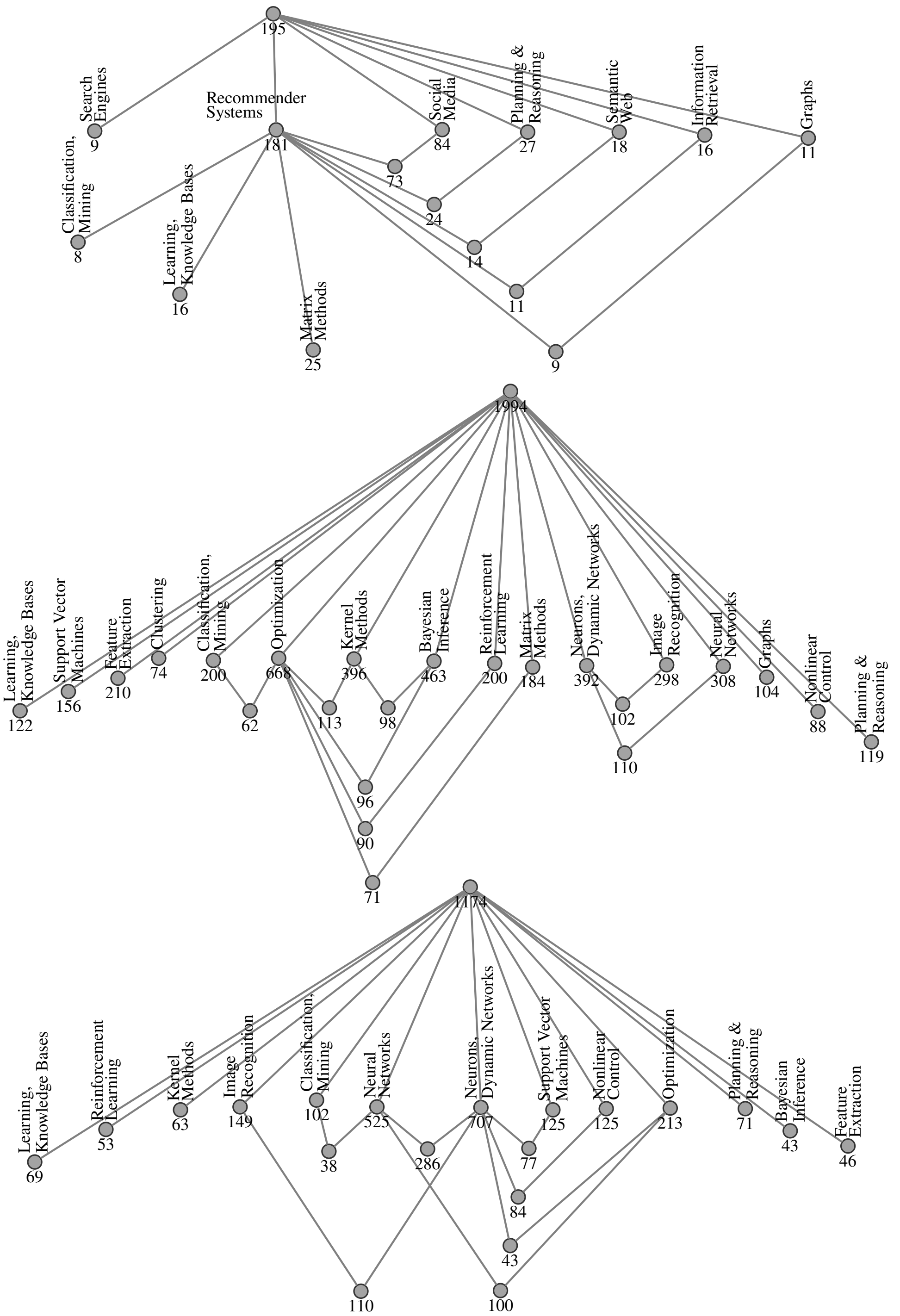


Figure 14.7: The iceberg concept lattice for the *RecSys* (top), *NeurIPS* (middle) and *Neural Networks* (bottom) venues.

without the conceptual hierarchy the relation between different topics is unknown. For example, is *Search Engines* either (1) dominated by, (2) incomparable to, or (3) coinciding with the topic *Recommender Systems*? Given the lattice diagram (cf. Figure 14.7) we can answer this question. The analytical building blocks *ordinal motifs* allow for a structured approach to answering the question above. Moreover, they enable an automatic extraction of *dominating* topics, *incomparable* topics, and *incomparable and meet coinciding* topics.

Neural network venues

For the other two examples entities, i.e., *NeurIPS* and *Neural Networks*, we observe different results. In short, different motifs occur, many non-coinciding topics and there are no dominating topics. Remarkable is the occurrence of contranominal ordinal motifs. For the *NeurIPS* entity, we find a contranominal ordinal motif of *Optimization*, *Kernel Methods* and *Bayesian Inference* and for the *Neural Networks* entity, we find the contranominal ordinal motif *Neural Network*, *Neurons, Dynamic Networks* and *Optimization*. Both indicate that there is a strong connection within the respective topics, i.e., every subset combination of topics occurs. However, all three contranominal topics do not occur at the same time.

Width of views

A more global observation is that the conceptual structures of *NeurIPS* and *Neural Networks* have a larger width compared to *RecSys*. In case of *RecSys* the width is nine while *Neural Networks* has a width of thirteen and *NeurIPS* of sixteen.

Topic focus

While *Neural Networks* and *NeurIPS* have similar frequent topics, their conceptual structure looks quite different. There are more frequent combinations that involve the topics *Neural Network* or *Neural Dynamic Networks* within the view of the *Neural Networks* entity. From this observation, we can infer that both entities have a different topic focus.

14.4.4 Conceptual Views on Topic Models over Time

Views over time

A particular feature of the just introduced methods is that they enable an investigations over time. We demonstrate this on two examples. First, we consider the conceptual structure of *Wolfgang Nejdl*, as depicted in Figure 14.8. In this figure, we show the diagram for three different time periods. These periods were chosen based on the following observation within the heatmap in Figure 14.1. We empirically identified three main periods in *Wolfgang Nejdls* research history. The first is from 1987-1999 where he researched mainly on *Planning & Reasoning*. Second, the period from 2000 to 2008 which seems to be a transition phase where both the *Planning & Reasoning* and *Semantic Web* topics are present. Lastly, there is the period starting with 2009 where he focused primarily on *Semantic Web*.

Temporal snapshots

Annotations in views over time

For each period, we annotated at concepts their support value, i.e., the relative number of research articles for the corresponding topics. We highlighted the supported concepts (or toned down the non-supported concepts) to make the resemblance to the original structure more clear. It is important to note that we encountered the case that a set of topics is not closed for the period 1987-1999. Nonetheless, we stick to the common conceptual representation, but highlighted non-closed topic sets in red (cf. Chapter 11).

First period

Based on the support values, we see that *Planning & Reasoning* is the most important topic in the first period. Even more, as we can infer from the non-red nodes, all topics coincide with *Planning & Reasoning*. The second most frequent topics are *Search Engines* and *Reinforcement Learning*. Another observation we can draw is that a majority of the topic combinations are not supported or closed at this stage. In the second time period, we see that almost all topic combinations are supported. The exception is the combination of *Social Media* and *Semantic Web*. The *Planning & Reasoning* topic is less supported. The *Semantic Web* topic has the highest support value. In the last period the activity on *Planning & Reasoning* declines again. Five of eight topic combinations involve the *Semantic Web* topic. Compared to the second diagram, fewer topic combinations (eight compared to

Second period

Third period

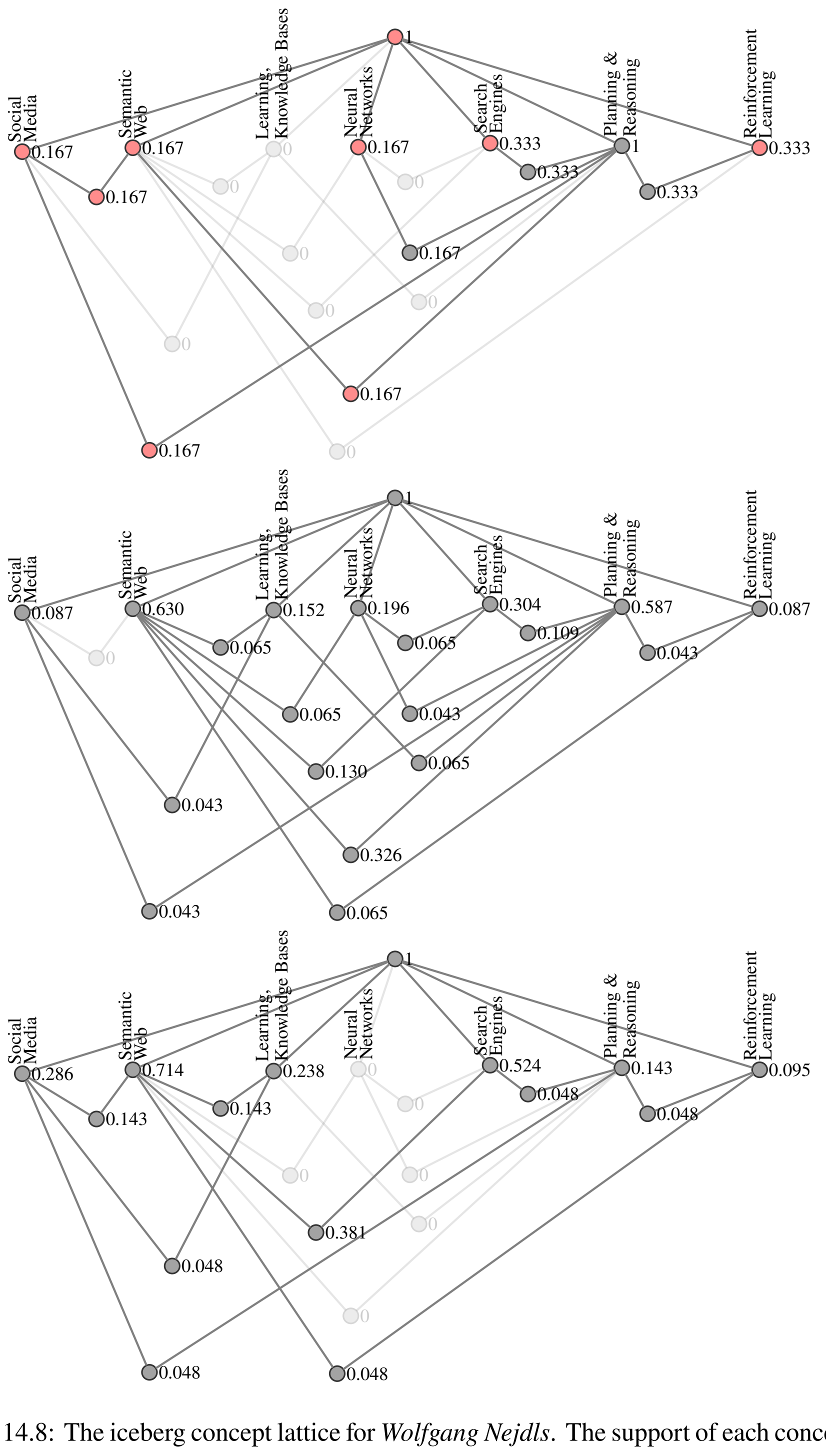


Figure 14.8: The iceberg concept lattice for *Wolfgang Nejdls*. The support of each concept for the years 1987-1999, 2000-2008 and 2009-2020 is annotated next to each concept. Attribute sets that are not closed in a given time interval are highlighted in red.

twelve) are supported. Overall we see a shift in activity from concepts depicted on right to concepts on the left.

RecSys over time

We approach *RecSys* in the same way (see Figure 14.9). Here halved the time lifespan of the venue into the time periods 2007-2014 and 2015-2020. In contrast to *Wolfgang Nejdl*, we do not observe notable conceptual differences over time. One may take this for evidence that *RecSys* has a stable focus.

14.4.5 Association Rules in Conceptual Topic Views

Association rules in entity views

The introduced conceptual structures allow for the extraction of rules between topics. In this section, we study such rules via the Luxenburger basis [146] (see Section 5.4). That is, a basis for the set of all rules that satisfy a minimum support η and a minimum confidence γ. For computing the Luxenburger basis for the entities of $\mathcal{D}$ (see Section 14.4.2) we chose a minimum support of three percent. This is the same parameter as for the TITANIC algorithm in Section 14.4.2. Hence, the computed rules are reflected by the iceberg concept lattices diagrams (cf. Figures 14.5 and 14.7). For γ we chose fifty percent in order to find

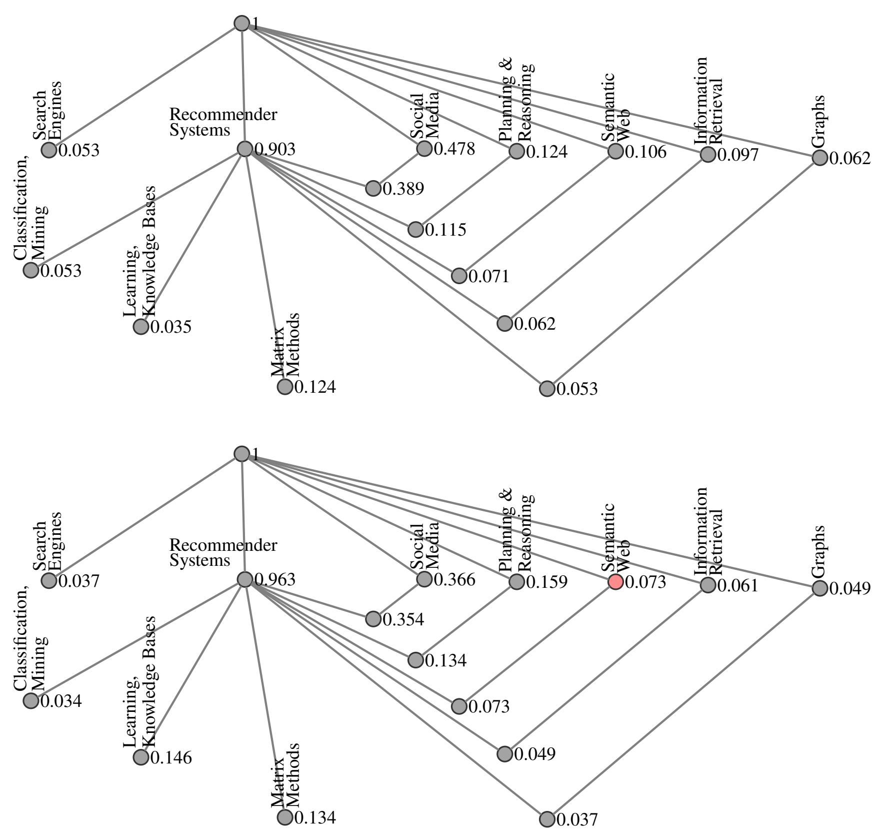


Figure 14.9: The iceberg concept lattice for the *RecSys* venue. The support of each concept for the years 2007-2014 and 2015-2020 is annotated next to each concept. Attribute sets that are not closed in a given time interval are highlighted in red.

meaningful rules.

The resulting bases are depicted in Figure 14.10. In the following, we discuss the results for the considered entities. For *Bernhard Schölkopf* the rules identify a strong inter-dependence between *Support Vector Machines* and *Kernel Methods*. This is consistent to our findings in Section 14.3. For *Wolfgang Nejdl*, we first note the overall importance of the *Semantice Web* topic. Four out of five rules have *Semantice Web* in their head. Yet, this topic never occurs in the body of a rule. This is in contrast to the topics of *Bernhard Schölkopf*. For *RecSys*, the found rules confirm our results in Section 14.3, since all rules have the topic *Recommender Systems* in their head. The same applies to *Neural Networks*, where all rules have *Neurons Dynamic Networks* in their head. *NeurIPS* on the other hand does not have any rules in the given basis for the given parameters. This may indicate that the NeurIPS is a topic diverse venue within the field of machine learning. Results

We may note that the support values of the given topic combinations can also be read directly from the concept diagrams (see Section 14.4.3). However, the computed rules allow for a more comprehensive representation of the most confident topic dependencies.

Table 14.10: The luxenburger basis for the document topic relation and entities from Section 14.4.2 for a minimum support of three percent and minimum confidence of fifty percent.

Luxenburger Basis			Support	Confidence
Bernhard Schölkopf				
Support Vector Machines	→	Kernel Methods	0.48	0.50
Kernel Methods	→	Support Vector Machines	0.40	0.59
Wolfgang Nejdl				
$\emptyset$	→	Semantic Web	1.00	0.61
Reinforcement Learning	→	Planning & Reasoning	0.10	0.62
Search Engines	→	Semantic Web	0.36	0.51
Reinforcement Learning	→	Semantic Web	0.10	0.50
Learning Knowledge Bases	→	Semantic Web	0.16	0.50
RecSys				
$\emptyset$	→	Recommender Systems	1.00	0.92
Information Retrieval	→	Recommender Systems	0.08	0.68
Planning & Reasoning	→	Recommender Systems	0.13	0.88
Semantic Web	→	Recommender Systems	0.09	0.77
Social Media	→	Recommender Systems	0.43	0.86
Graphs	→	Recommender Systems	0.05	0.81
Neural Networks				
$\emptyset$	→	Neurons Dynamic Networks	1.00	0.60
Image Recognition	→	Neurons Dynamic Networks	0.12	0.73
Support Vector Machines	→	Neurons Dynamic Networks	0.10	0.61
Neural Networks	→	Neurons Dynamic Networks	0.44	0.54
Nonlinear Control	→	Neurons Dynamic Networks	0.10	0.67
NeurIPS				

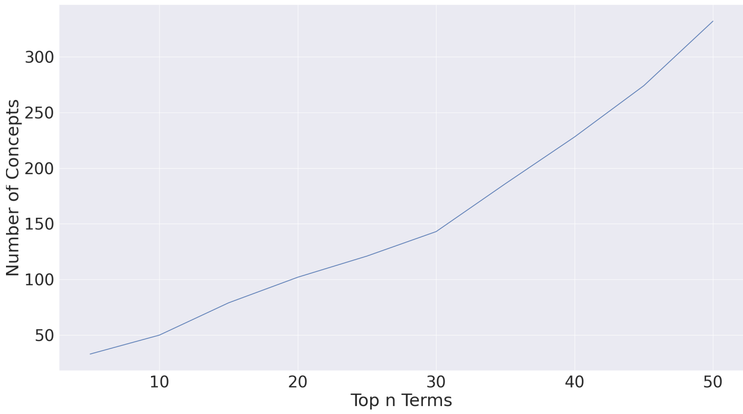


Figure 14.11: The number of formal concepts of $\mathcal{S}$ depending on the number n of top terms parameter.

14.4.6 The Conceptual Term-Topic Structure

Scaling of the term-topic relation

The topic-term relation $\mathcal{S}$ (see Figure 14.2) entails important information on the SSH21 topic model. It allows us to explain the topics of SSH21 via terms $s \in S$. As discussed in Section 14.4.1, we derive an incidence structure $\mathcal{S}_n$ from $\mathcal{S}$. This parameter is the number of top-n terms per topic. Our choice for the $n \in \mathbb{N}$ depends on the corresponding number of formal concepts, as depicted in Figure 14.11. From the plot, we infer that the parameter of $n = 10$ (see $\mathcal{S}_{10}$ Figure 14.15) is reasonable, as it results in about fifty formal concepts. This parameter choice is common in the literature [56, 179, 189], independently of our requirement on a low number of concepts.

Towards topic model explanations

We depict the concept lattice of $\mathcal{S}_{10}$ in Figure 14.12. We omitted to present the bottom concept, since it was not supported. An advantage of this structure is that we can explain found topic dependencies in terms of their shared terms. For example, the topics *Kernel Methods* and *Support Vector Machines* are connected via the term *kernels*. Analogously, the topics *Neural Networks* and *Graphs* are connected via the term *nodes*.

Ordinal motifs

We highlighted several contranominal and crown ordinal motifs using different colors. For example, the topics *Search Engines*, *Semantic Web* and *Social Media* are of contranominal structure. For this, the terms *web*, *user* and *content* are responsible (highlighted in orange). Another example is the set of topics *Planning & Reasoning*, *Matrix Methods* and *Leaning Knowledge Bases*, which are also of contranominal structure. For this motif, the terms *domain*, *learning* and *knowledge* are responsible (highlighted in blue). Hence, these terms are pair-wise differently used in the SSH21 topic model, yet they are very similar. One may deduce from this observation that this is also true for the research corpus D. An example crown motif is given by the sub-structure highlighted in red (right). This motif spans from the topic *Classification* over the topics *Neural Networks*, *Neurons and Dynamic Networks* to *Reinforcement Learning* and back to the *Classification* topic. A larger crown is depicted in purple on the left in the diagram.

Discussion

The proposed method allows for meaningful and structural investigations of the SSH21 topic model. This distinguishes our method from other approaches, such as those presented in Figure 14.1 and Figure 14.15. In particular, our method is capable of identifying

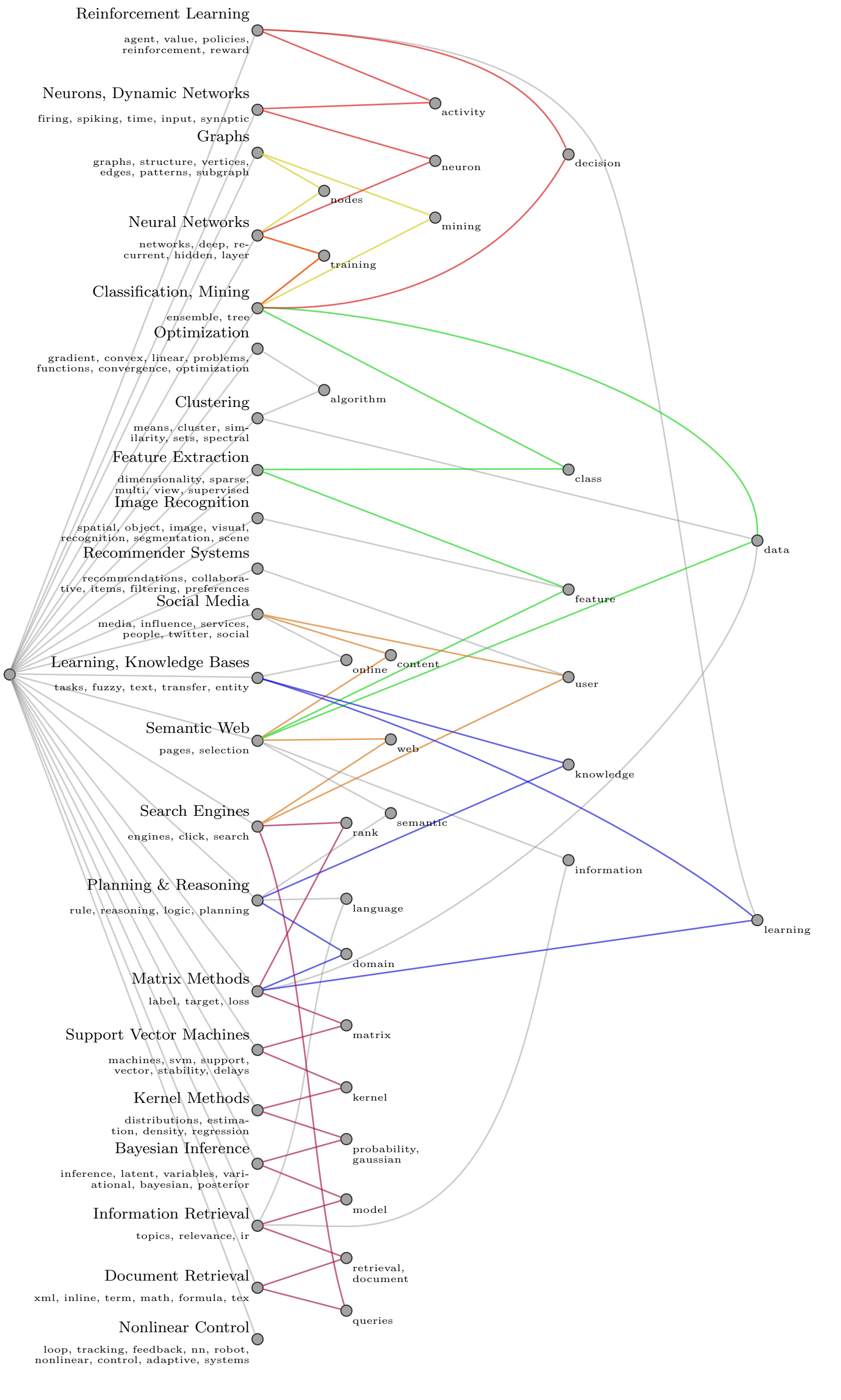


Figure 14.12: The concept lattice of the term-topic relation of the SSH21 topic model.

dependencies in the topic space, such as the cycle between topics that emerges from the crown motif. In summary, the conceptual structure of $\mathcal{S}_{10}$ allows for a global and deep investigation of the SSH21 topic model.

14.4.7 Zoom-In on topics

More terms

The limitation of $n = 10$ can be softened by focussing on particular topics of interest. We call this *zoom-in on topics*. For example, in our experiment, we are interested in concepts on *Neural Networks*. With the same reasoning as in the last section, i.e., small number of concepts, we found $n = 30$ to be appropriate. The direct approach to compute all concepts that contain *Neural Networks* is to first compute all concepts of $\mathcal{S}_{30}$ and consecutively filter them for the topic in question. In Figure 14.13 we depicted the result. Again, we omitted the unsupported bottom element. Out of 157 concepts do twenty-six include the *Neural Network* topic.

Implications within neural networks

We see in the diagram that some topics are drawn as lower neighbors of other topics. Restricted to the zoom-in on *Neural Networks*, our method identifies several implications. For example topic terms of *Kernel Methods* are also topic terms of *Matrix Methods*. Another example is that topic terms of *Optimization* are also topic terms of *Reinforcement Learning*. At this point we want to note two important points about the interpretation of these implications. First, the computed implications are valid within the analyzed topic model. Hence, any logical conflicts with respect to real world observations (or expert assessment) may indicate flaws of the topic model. Second, for any implication the inverse is not necessarily true.

Ordinal motifs

We can identify several ordinal motifs in the zoomed-in structure. For example, (1) *Classification*, *Matrix Methods* and *Optimization*, and (2) *Clustering*, *Learning Knowledge Base* and *Optimization*. Both are contranominal ordinal motifs.

Term use

Synonyms are important for training and applying topic models. Within the scope of *Neural Networks* we can draw from the structure questions, such as: What differentiates the terms *algorithm* and *method*? In which topics are they used as synonyms? The same questions can be formulated for *train* and *learn*.

14.5 The Geometric Structure

Towards global explanations

Important aspects of data and their interpretation are captured through geometric properties. This is in particular true for incidence geometries. The study of ordinal motifs allows for further geometric interpretation of the sub-structures within the topic space. In the last section, we analyzed the topic model with singular ordinal motifs at a time. The goal now is to employ the comprehensive geometric structure, i.e., the set of all ordinal motifs. The *geometric structure* of a (contextual) data set is a multi-relational hypergraph structure. In this hypergraph every hyperedge relation encodes one type of ordinal motif $\mathcal{S}_1, \ldots, \mathcal{S}_n$.

Definition 53 (Geometric Structure). *For a context $\mathbb{K} := (G, M, I)$ and families of ordinal motifs $\mathcal{S}_1, \ldots, \mathcal{S}_n$ is the* ***geometric structure*** *of $\mathbb{K}$ with respect to $\mathcal{S}_1, \ldots, \mathcal{S}_n$ a multi-hypergraph $(M, E_1, \ldots, E_n)$ where*

$$E_i := \{N \subseteq M \mid N \text{ is of ordinal motif type } \mathcal{S}_i \text{ in } \mathbb{K}\}.$$

The ordinal motif problem

In the present chapter, we study ordinal motifs with respect to attribute sets. Therefore, all notions are translated to their dual counterparts with respect to the formal context. The question *"Is $\mathbb{K}[G, N]$ of ordinal motif type?"* is a formal decision problem called the

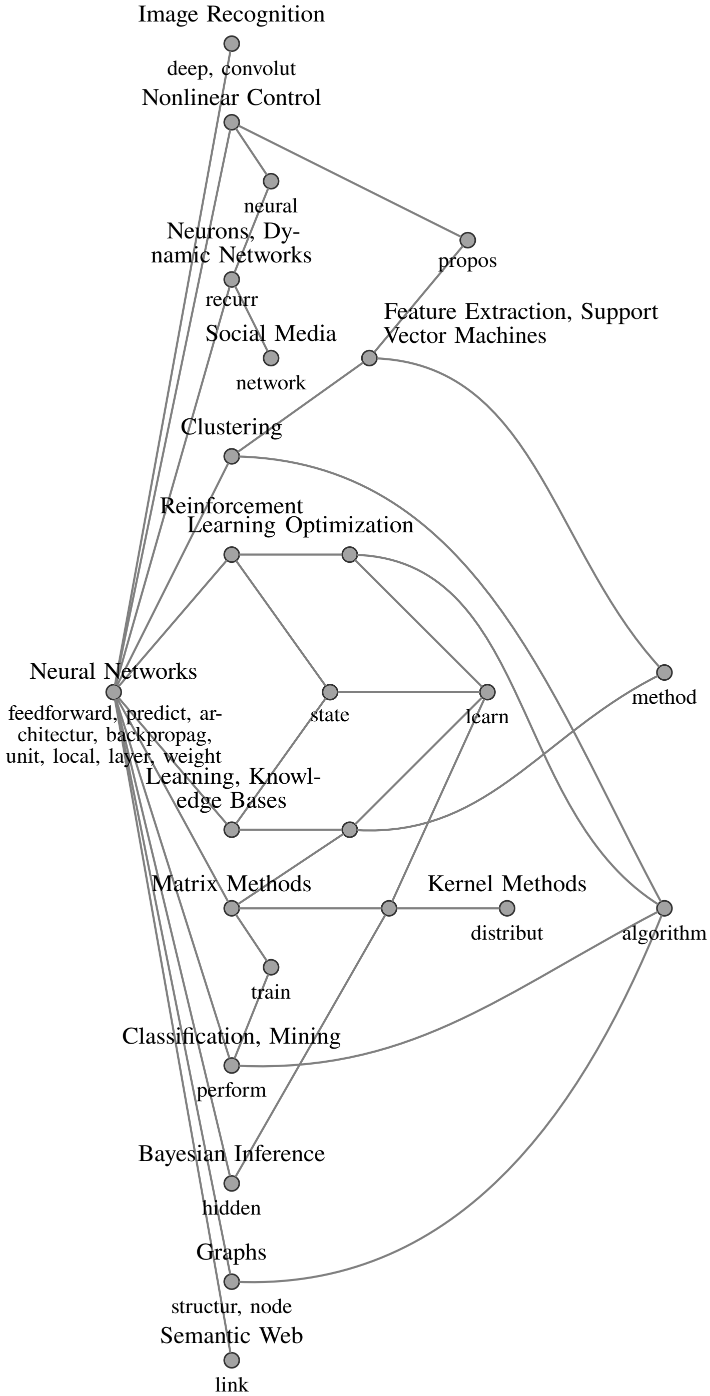


Figure 14.13: Zoom in on the *Neural Network* topic of the term-topic relation for the `SSH21` topic model. The top term parameter is set to thirty.

ordinal motif problem (see Problem 15.2). Formally, we employ the surjective local full scale-measure ordinal motif problem. The corresponding computational complexities were investigated in Chapter 9. All instances of the ordinal motif problem that we investigate in this chapter are in P (see Proposition 25).

Selection of ordinal motif types

We may note that we do not impose any particular choice of $\mathcal{S}_i$. The selection of suitable ordinal motifs [101] is up to the analyst.

14.5.1 Geometric Structure Diagram

Drawing the geometric structure

To make the geometric structure human accessible, we decided for a diagrammatic presentation. We define a set of drawing rules for each ordinal motif type. After that, we apply our method to the SSH21 topic model. We call the resulting figure the *geometric drawing* of $\mathbb{K}$. This method is inspired by the geometric representations introduced by Wille for drawing concept lattices [222].

Ordinal motif drawing styles

In this geometric drawing, every attribute $m \in M$ (i.e., topic for SSH21) is represented by a node. The hyperedges of an ordinal motif type are drawn as connections between nodes. Each ordinal motif type has its own drawing style to ensure that they can be distinguished. The connection lines are annotated by the objects $g \in G$ (i.e., terms for SSH21) that induce the respective ordinal motif. This allows for deriving explanations of ordinal motifs both in terms of the attributes they connect and the objects they entail.

We present the drawing rules for each type of ordinal motifs in greater detail:

Nominal There are two cases of nominal ordinal motifs that we distinguish with respect to the objects they entail. In case there is an object $g \in G$ that supports all attributes N, i.e., $g \in N'$, we draw an edge between all pairs of attributes $n_1, n_2 \in N$ and annotate the connecting lines by g. Otherwise, no edge is drawn. A prototypical example for both cases is shown in Figure 14.6 (top right).

Crown Crown ordinal motifs do not require their own drawing rule. Instead, they can be read from (closed) cycles of nominal ordinal motifs. Yet, two conditions need to be satisfied: (1) Objects must not occur more than once along a cycle. (2) No other edges, apart from the cycle edges connect cycle attributes. A prototypical example for the crown ordinal motif on ten elements is depicted in Figure 14.6 (middle right).

Contranominal A contranominal ordinal motif encodes that any subset $D \subseteq N$ is supported by a unique set of objects $H \subseteq G$. Therefore, they reflect a densely connected part within the conceptual structure. Our drawing shall reflect this by applying the drawing rule: contranominal hyperedges of size n are drawn by a filled n-polygon that connects the attributes N. The edge between two attributes $n_1, n_2 \in N$ then annotated by all objects that they have in common, i.e., by the objects $\{n_1, n_2\}'$. Prototypical examples for contranominal ordinal motifs of size three and four can be found in Figure 14.6 (bot right) and Figure 14.16 (bot right).

Ordinal Ordinal motifs that are of ordinal[5] type encode rankings among the attributes. In such a motif, the greatest element subsumes all the incidences of a smaller attribute. We reflect this in the diagram by the drawing rule: an attribute (node) is drawn such that it overlaps the next lower ranked attribute (node). For this kind of motif, objects are annotated next to nodes. At each attribute (node) we annotate all objects such

[5]We remind the reader that *ordinal motif* is a defined class of objects and ordinal type addresses a particular sub-class.

that there is no lower ranked attribute that is in incidence with this object. This procedure is in accordance to the short-hand notation of concept lattice line diagrams. A prototypical example is depicted in Figure 14.16 (top right).

Interordinal The interordinal ordinal motif encodes two ordinal motifs of ordinal type, whose rankings on the attributes are complementary[6] to each other. We have depicted an example on four elements in Figure 14.16 (middle right). To display an interordinal ordinal motif one should draw an hyperedge that encloses the motif's attributes. The objects are annotated next to the attribute nodes based on the two rankings and the ordinal drawing rule. Objects from the same ranking have to be drawn on the same side of the hyperedge.

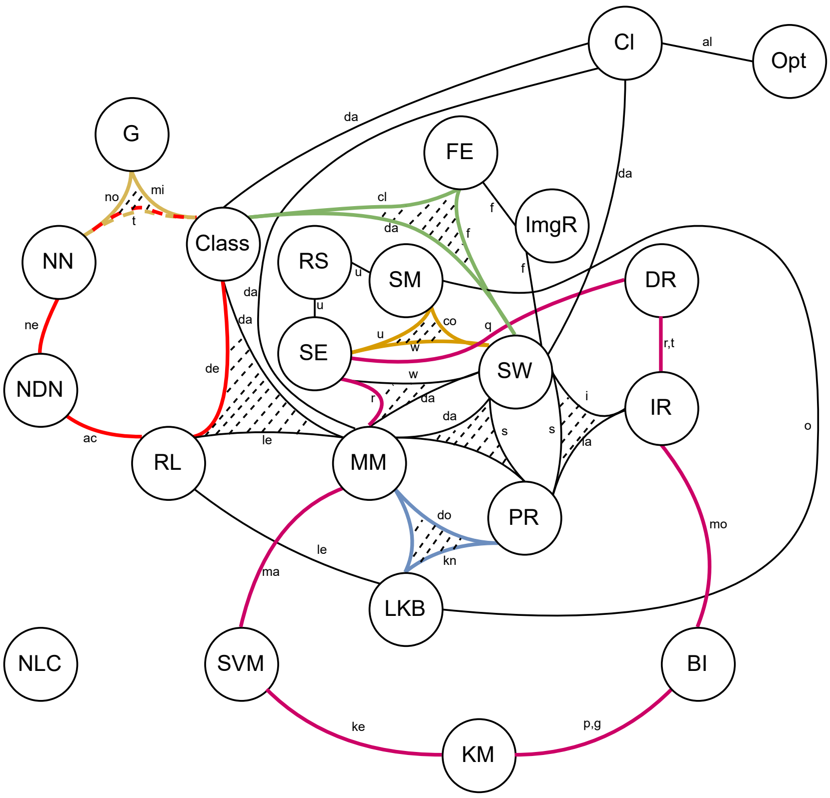


Figure 14.14: The geometric drawing of the SSH21 topic model for top-ten terms (cf. Figure 14.11). Topic and term names are abbreviated for better readability. Their full length names can be inferred from Figure 14.15.

[6]A natural occurring example for this are the *"x is hotter than y"* and *"y is colder than x"* relation.

14.5.2 The Geometric Structure of SSH21

The selection of ordinal motifs

The resulting geometric drawing for SSH21 is depict in Figure 14.14. The drawing includes nominal, contranominal and crown ordinal motifs. There are no non-trivial ordinal or interordinal types in the topic model structure. This fact is, however, not surprising, since the topics within a topic model are optimized to be independent of each other. For readability reasons, we abbreviated the terms and topics. Their un-abbreviated versions are listed in Figure 14.15.

The geometric drawing

In order to increase the readability and comparability to the last section (cf. Figure 14.11), we have highlighted ordinal motifs in the same color. The representation of the relations in the geometric drawing is fundamentally different to line diagrams. By design, in the geometric drawing it is easier to identify ordinal motifs. For example, we can easily read from the diagram that there are eight contranominal ordinal motifs. Out of the twenty-two topics, *Semantic Web* (SW) occurs in five contranominal ordinal motifs and is therefore structurally very important for within the topic model. This is followed by the *Matrix Methods* (MM) topic which occurs in four contranominal ordinal motifs. In contrast to line diagrams, this information is easy to infer from the geometric drawing. The *Non-Linear Control* (NLC) topic is very isolated and does not exhibit any (non-trivial) connection to other topics.

Important topics

Crown motifs

Crown ordinal motifs can easily be read from the structure as (closed) cycles. For example, we find *NN – Class – RL – NDN –NN* (orange) and *SW – IR – DR – SE – SW*. Both crowns identified in the line diagram Figure 14.11 are highlighted in the geometric drawing in the same color.

Discussion

We invite the reader to compare the geometric drawing to classical approaches such as topic-topic heatmaps and t-SNE embeddings, as depicted in Figure 14.1. Based on this comparison, we argue that geometric drawings of topic models allow for a non-flat analysis of the inter-topic relation and their respective terms.

14.6 Limitations & Conclusion

Applications

In this chapter, we proposed a comprehensive approach for analyzing and visualizing high dimensional topic models. In principle, this method is applicable to arbitrary matrix shaped data sets. We have shown that our method is capable of capturing insights about researchers and venues from the realm of machine learning research. Moreover, we demonstrated how conceptual structures can be used to track the change in their topics. For our analysis, we employed ordinal patterns which occurred frequently in the data. These sub-structures allow for a rich interpretation of the topic model. In particular, the inter-topic and term-topic relation.

Interpretability of data sets

This interpretability, of course, depends on the overall understanding of the terms of the topic model. Hence, although our method is applicable to arbitrary matrix shaped data sets, meaningful interpretations are limited by the available background knowledge. Another limitation of our method is the number of concepts one can visualize in a readable fashion. This number is dependent on the number of topics, documents and selected top terms per topic. To compensate for this limitation, we proposed the use of (graph) core structures.

Hierarchical topic models

As the present chapter has established a robust link between topic models and their conceptual analysis, we envision several directions for future work. First, the absence of (non-trivial) ordinal and interordinal motifs within the analyzed topic model is not surprising. This due to the fact that topic models optimize to compute independent (non-nesting) topics.

Yet, this is not true in the case of hierarchical topic modeling. The logical next step is to apply our methods to these models, e.g., HLDA or PAM [40, 83, 138, 231].

Human-centered design

Second, within the research field of human-computer interaction, we propose to conduct a user study in order to gather statistical evidence. Moreover, this may reveal new insights into the developed geometric drawings and potentially their visual optimization. Third, in order to conduct a study, as proposed above, a difficult algorithmic task has to be solved. Although, the geometric drawings are well-defined, their algorithmic computation is an open problem.

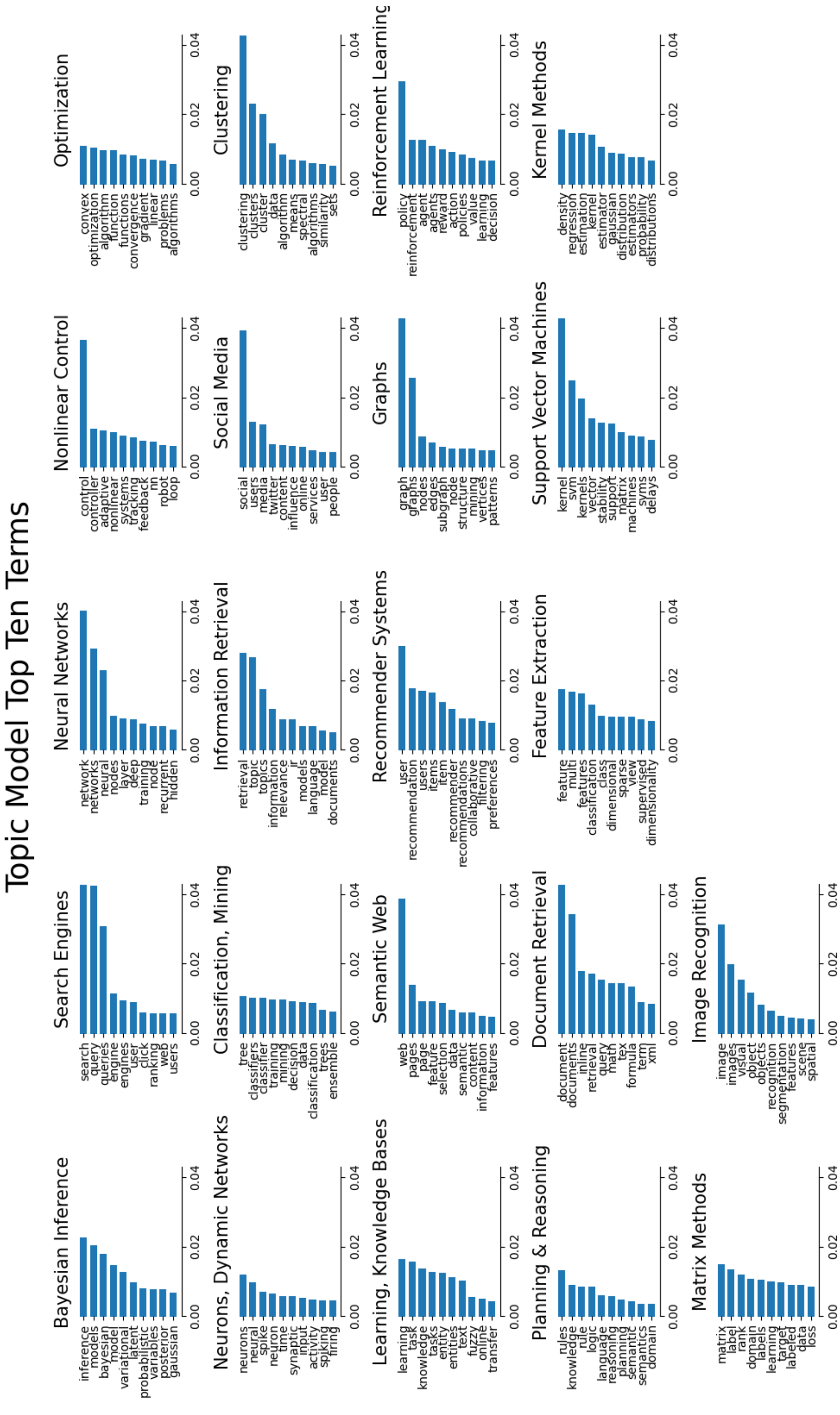
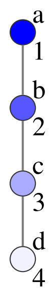
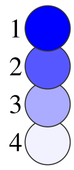


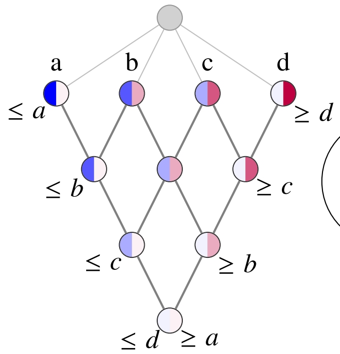
Figure 14.15: The top ten terms for each topic of the `SSH21` topic model from Schaefermeier, Stumme, and Hanika [189].

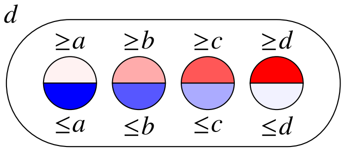
$\mathbb{O}_4$	a	b	c	d
1	×			
2	×	×		
3	×	×	×	
4	×	×	×	×



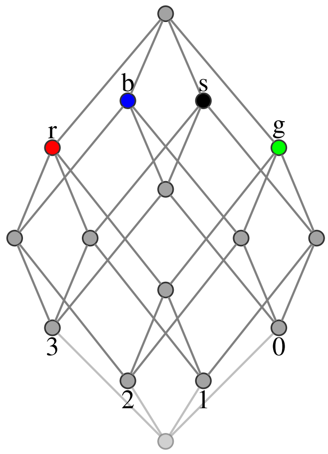


$\mathbb{I}_4$	a	b	c	d
$\leq a$	×			
$\leq b$	×	×		
$\leq c$	×	×	×	
$\leq d$	×	×	×	×
$\geq a$	×	×	×	×
$\geq b$		×	×	×
$\geq c$			×	×
$\geq d$				×





$\mathbb{B}_4$	r	g	b	s
1		×	×	×
2	×		×	×
3	×	×		×
4	×	×	×	



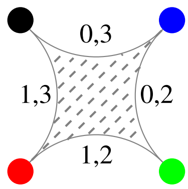
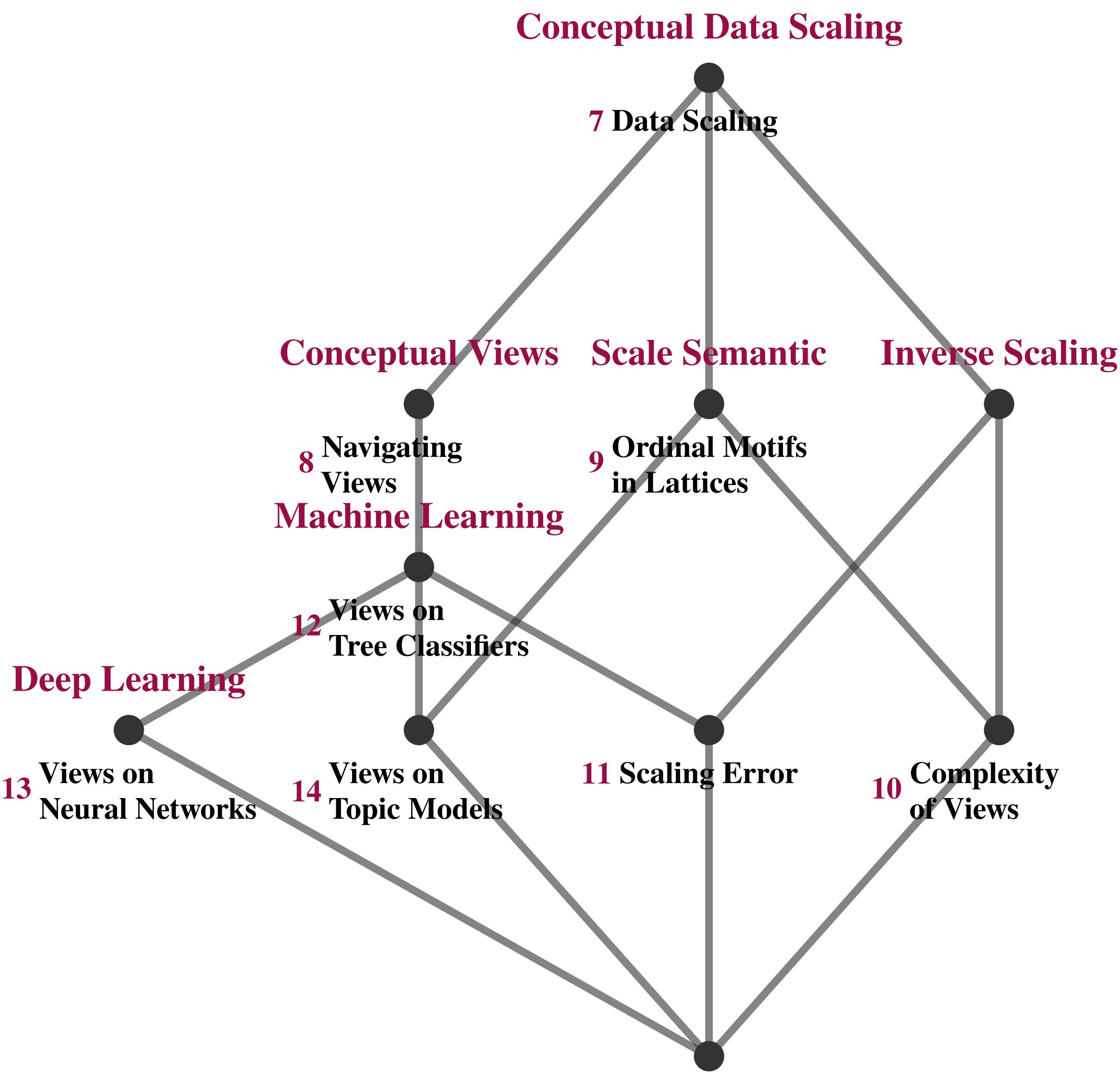


Figure 14.16: The formal context, concept lattice and geometric drawing style of the ordinal, interordinal and contranominal ordinal motif.

Part IV

Summary



15

Outlook and Open Problems

With our work, we have expanded the capabilities of conceptual data scaling in more than one way. In Part II, we have enhanced the understanding of the four scaling tasks, i.e., conceptual scaling, inverse conceptual scaling, conceptual data reduction and conceptual data compatibility. In Part III, we have developed multiple scaling methods for data representations in the realm of machine learning and showed how they contribute to the field of explainable artificial intelligence. In the following we recall from this work three open problems that we find to be the most important and promising ideas for future work.

Principle Ordinal Component Analysis Our contribution to the identification of standard scales provide foundational methods that can be used for a new type of concept lattice decomposition. This decomposition should extract for a given formal context $\mathbb{K}$ the minimum number of standard scales $\mathcal{S}$ such that $\mathbb{K}$ can be fully measured by $\mathcal{S}$. Such a principle ordinal component analysis can be seen as generalization of existing decomposition and factorization methods and may lead to fewer and more interpretable parts.

Ordinal Motifs as Tool for Data Interpretation We have shown in many applications that ordinal motifs are very useful to interpret data and extract higher level relations on the conceptual level. We can envision that ordinal motifs can establish themselves as standard tool of analysis alongside implications and concept lattices. For this we provided the theoretical and analytical framework and showed connections between ordinal motifs and implications. Despite that, there is lots of room for improvement in terms of efficient algorithms, user-friendly software, data visualizations and applications. In particular, the study of textual explanations and the geometric structure may benefit from human-centered approaches.

Topological data analysis With the geometric structure we have presented a new structure to derive global explanations for (machine learning) data representations. Besides that, we have shown that the set of hereditary ordinal motifs are simplicial complexes. These allow for applications of methods from the realm of topological data analysis. This not only connects two fields of research, but – in combination – can improve the overall state-of-the-art on global data explanations.

16
Conclusion

With this thesis, we have explored the use of conceptual structures as human comprehensible knowledge representations in many applications. We focused on two core tasks within the realm of conceptual data analysis. The first deals with the problem on how to represent and interpret data through conceptual structures. This is done by a process called conceptual scaling which defines (ordinal) conceptual scales that encode how the data is to be interpreted on the ordinal level. Here we extended the state-of-the-art with several methods on how to derive meaningful conceptual structures from machine learning data representations and how to interpret them. The second task deals with the size of the derived conceptual structure and how to compute data reductions that are explainable and consistent to the conceptual interpretation of the data, i.e., the defined scales. Here, we introduced a theory on how to identify consistent conceptual data reductions, how to derive them and how to explain them. On top of that, we developed navigation methods to enable users a self-determined exploration of the data and its relations.

For a more detailed overview on our contributions we recall the in Figure 16.1 derived overview on the four problems within the realm of conceptual data scaling. These four

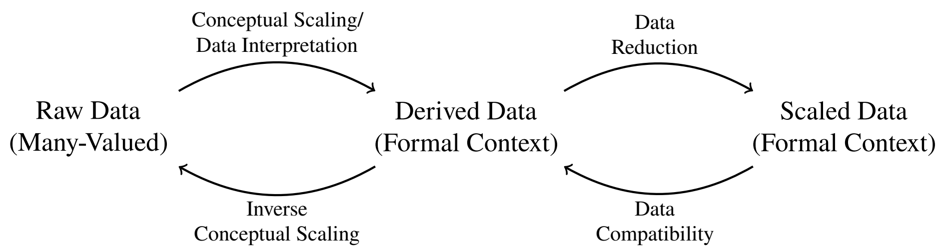


Figure 16.1: An overview of the four main problems in conceptual scaling.

problems represent the two tasks given above, as well as their inverse. For scaling, we have conceptual data scaling and inverse conceptual scaling, i.e., determining from what data set a conceptual structure is derived from. For the second task we deal with consistent conceptual data reduction and the identification of error introduced by dimensionality reduction methods. In Chapter 7, we gave a detailed overview on the state-of-the-art methods. Open problems and future work is outlined in Chapter 15.

Conceptual Scaling With Chapters 12 to 14, we contributed to the problem of conceptual scaling by proposing and studying data scalings for representations in machine learning. In particular, we derived scalings for tree based classifiers, e.g., Random Forest or Gradient Boosted Trees, latent spaces of neural networks and topic models. We studied their meaningfulness with respect to structural properties of the conceptual views, the machine learning model and the classification task. In an experimental setting, we showed how the resulting conceptual structures can be used to interpret and explain the underlying machine learning model. Thereby we yield a new concept based explanation method for machine learning models. Beyond explaining a single model, the derived conceptual structures can be used to compare the representations of different model types.

Inverse Conceptual Scaling In Chapter 10, we studied how to invert conceptual scaling, i.e., what data set is a conceptual structure derived from. This is achieved by extracting conceptual scales that are contained in a conceptual structure. In this field, we introduced a new notion of dimensionality, which addresses the size and complexity of the original data set in terms of the scales needed to cover the conceptual structure. We studied this notion in greater detail for two families of standard scales and showed how inverse conceptual scaling can be used for feature compression. The latter is achieved when the scaling dimension of the data is smaller than the number of features in the original data. Moreover, we described a data reduction method that reduces the data features to only the largest scales resulting in a principal ordinal component analysis method.

Conceptual Data Reduction In Chapter 7 we contributed to the problem of conceptual data reduction. First, we introduced scale-measures, i.e., extent continuous maps, as a formalism to characterize consistent data reductions. In Chapter 8, we introduced several methods to derive conceptual data reductions including a semi-automatic recommendation algorithm. On top of that, we introduced several methods to combine and navigate between data reductions to derive finer and coarser views on the data. This enables users to exploration the data and its relations in a self-determined fashion. Moreover, we showed how to explain data reductions with respect to logical expressions. This explanation method is agnostic to the reduction method which makes it applicable in many settings. This is especially useful for black box models. In addition to that, showed how to capture information that is neglected by the reduction procedure using the join-pseudocomplement in the hierarchy of conceptual data reductions.

In Chapter 9, we introduced with ordinal motifs a new approach to derive explanations on a local and global level. This allowed us to identify frequent recurring ordinal patterns for the analysis of large and complex structures. We introduced a new method on how ordinal motifs can be used to automatically derive textual explanations of conceptual structures in an unsupervised setting. The derived explanations account for principles from human computer interaction for explainable artificial intelligence. For global explanations we introduced in Chapter 14 the geometric structure of concept lattices, which generates a global view on the data based on ordinal patterns contained in the data. We demonstrate the applicability of this approach by providing explanations for a topic model on scientific documents.

Conceptual Data Compatibility With the conceptual scaling error, we introduced in Chapter 11 a new notion to quantify inconsistencies in dimensionality reduction methods with respect to the conceptual structure. This notion is agnostic to the reduction method resulting in many possible applications. Moreover, does the conceptual scaling error solely dependent on the conceptual structure and is thereby independent of the features that create them. Thus, it can be applied to data representations of different dimensions, which is in contrast to common matrix distance measures. In an experimental setting, we showed that data reduction methods with seemingly good performance (reconstruction error) can have high conceptual scaling error. This suggests that applications working with formal concepts or similar patterns should consider this new notion. In addition to the quantification of error, we provided methods to qualitatively study the conceptual scaling error through concept lattices.

Bibliography

[1] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, and B. Kim. “Sanity Checks for Saliency Maps.” In: *NeurIPS*. Ed. by S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. 2018, pp. 9525–9536.

[2] N. Akhtar, H. Javed, and T. Ahmad. “Hierarchical Summarization of Text Documents Using Topic Modeling and Formal Concept Analysis”. In: *Data Management, Analytics and Innovation* (2018).

[3] A. M. Alaa and M. van der Schaar. “Demystifying Black-box Models with Symbolic Metamodels.” In: *NeurIPS*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché Buc, E. B. Fox, and R. Garnett. 2019, pp. 11301–11311.

[4] A. Albano and B. Chornomaz. “Why concept lattices are large: extremal theory for generators, concepts, and VC-dimension”. In: *International Journal of General Systems* 46 (2017), pp. 440–457.

[5] T. Aldinucci, E. Civitelli, L. di Gangi, and A. Sestini. *Contextual Decision Trees*. 2022. DOI: 10.48550/ARXIV.2207.06355.

[6] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. “Permutation importance: a corrected feature importance measure”. In: *Bioinformatics* 26.10 (2010), pp. 1340–1347.

[7] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha, R. Kinney, S. Kohlmeier, K. Lo, T. Murray, H.-H. Ooi, M. E. Peters, J. Power, S. Skjonsberg, L. L. Wang, C. Wilhelm, Z. Yuan, M. van Zuylen, and O. Etzioni. “Construction of the Literature Graph in Semantic Scholar.” In: *CoRR* abs/1805.02262 (2018).

[8] F. Anowar, S. Sadaoui, and B. Selim. “Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne)”. In: *Computer Science Review* 40 (2021), p. 100378.

[9] W. W. Armstrong. “Dependency structures of data base relationships”. In: *IFIP congress*. Vol. 74. Geneva, Switzerland. 1974, pp. 580–583.

[10] S. Arora, W. Hu, and P. K. Kothari. “An Analysis of the t-SNE Algorithm for Data Visualization.” In: *CoRR* abs/1803.01768 (2018).

[11] M. Asai and A. Fukunaga. “Classical Planning in Deep Latent Space: Bridging the Subsymbolic-Symbolic Boundary.” In: *AAAI*. Ed. by S. A. McIlraith and K. Q. Weinberger. AAAI Press, 2018, pp. 6094–6101.

[12] M. Aznag, M. Quafafou, and Z. Jarir. “Leveraging Formal Concept Analysis with Topic Correlation for Service Clustering and Discovery”. In: *2014 IEEE International Conference on Web Services*. 2014, pp. 153–160. DOI: 10.1109/ICWS.2014.33.

[13] M. A. Babin and S. O. Kuznetsov. "Recognizing Pseudo-intents is coNP-complete." In: *CLA*. Ed. by M. Kryszkiewicz and S. A. Obiedkov. Vol. 672. CEUR Workshop Proceedings. CEUR-WS.org, 2010, pp. 294–301.

[14] I. Bellido and E. Fiesler. "Do Backpropagation Trained Neural Networks have Normal Weight Distributions?" In: *ICANN '93*. Ed. by S. Gielen and B. Kappen. London: Springer London, 1993, pp. 772–775. ISBN: 978-1-4471-2063-6.

[15] R. Belohlavek and J. Macko. "Selecting important concepts using weights". In: *Formal Concept Analysis: 9th International Conference, ICFCA 2011, Nicosia, Cyprus, May 2-6, 2011. Proceedings 9*. Springer. 2011, pp. 65–80.

[16] R. Belohlávek, B. D. Baets, J. Outrata, and V. Vychodil. "Characterizing Trees in Concept Lattices". In: *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 16.Supplement-1 (2008), pp. 1–15. DOI: 10.1142/S0218488508005212.

[17] R. Belohlávek, B. D. Baets, J. Outrata, and V. Vychodil. "Inducing Decision Trees via Concept Lattices." In: *CLA*. Ed. by P. W. Eklund, J. Diatta, and M. Liquiere. Vol. 331. CEUR Workshop Proceedings. CEUR-WS.org, 2007.

[18] R. Belohlávek and J. Konecny. "Scaling, Granulation, and Fuzzy Attributes in Formal Concept Analysis." In: *FUZZ-IEEE*. IEEE, 2007, pp. 1–6.

[19] R. Belohlávek, J. Outrata, and M. Trnecka. "Impact of Boolean factorization as preprocessing methods for classification of Boolean data." In: *Annals of Mathematics and Artificial Intelligence* 72.1-2 (2014), pp. 3–22.

[20] R. Belohlávek, J. Outrata, and M. Trnecka. "Toward quality assessment of Boolean matrix factorizations." In: *Inf. Sci.* 459 (2018), pp. 71–85.

[21] R. Belohlávek and M. Trnecka. "Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications". In: *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*. Ed. by F. Rossi. IJCAI/AAAI, 2013, pp. 1233–1239.

[22] R. Belohlávek and V. Vychodil. "Discovery of optimal factors in binary data via a novel method of matrix decomposition." In: *J. Comput. Syst. Sci.* 76.1 (Jan. 11, 2010), pp. 3–20.

[23] G. Birkhoff. *Lattice theory*. Vol. 25. American Mathematical Soc., 1940.

[24] B. Bischl, G. Casalicchio, M. Feurer, F. Hutter, M. Lang, R. G. Mantovani, J. N. van Rijn, and J. Vanschoren. "OpenML Benchmarking Suites". In: *arXiv:1708.03731v2 [stat.ML]* (2019).

[25] D. Blei and J. Lafferty. "A correlated topic model of Science". In: *Annals of Applied Statistics* 1 (2007), pp. 17–35.

[26] H. Blockeel, L. D. Raedt, and J. Ramon. "Top-Down Induction of Clustering Trees". In: *ArXiv* cs.LG/0011032 (1998).

[27] D. Borchmann, T. Hanika, and S. Obiedkov. "On the Usability of Probably Approximately Correct Implication Bases." In: *ICFCA*. Ed. by K. Bertet, D. Borchmann, P. Cellier, and S. Ferré. Vol. 10308. Lecture Notes in Computer Science. Springer, 2017, pp. 72–88. ISBN: 978-3-319-59271-8.

[28] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi. "Counting graph homomorphisms". In: *Topics in Discrete Mathematics: Dedicated to Jarik Nešetřil on the Occasion of his 60th Birthday*. Springer, 2006, pp. 315–371.

[29] L. Breiman. "Random Forests". English. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 0885-6125. DOI: 10.1023/A:1010933404324.

[30] L. Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.

[31] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth, 1984. ISBN: 0-534-98053-8.

[32] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. "Geometric deep learning: going beyond euclidean data". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 18–42.

[33] P. Burmeister and R. Holzer. "Treating Incomplete Knowledge in Formal Concept Analysis." In: *Formal Concept Analysis*. Ed. by B. Ganter, G. Stumme, and R. Wille. Vol. 3626. Lecture Notes in Computer Science. Springer, 2005, pp. 114–126. ISBN: 3-540-27891-5.

[34] N. Caspard and B. Monjardet. "The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey". In: *Discrete Applied Mathematics* 127.2 (Apr. 2003), pp. 241–269.

[35] A. Castellanos, J. Cigarran, and A. Garcia-Serrano. "Formal concept analysis for topic detection: a clustering quality experimental analysis". In: *Information Systems* 66 (2017), pp. 24–42.

[36] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins. "What are ontologies, and why do we need them?" In: *IEEE Intelligent Systems and their applications* 14.1 (1999), pp. 20–26.

[37] G. Chao. "Human-computer interaction: process and principles of human-computer interface design". In: *2009 International Conference on Computer and Automation Engineering*. IEEE. 2009, pp. 230–233.

[38] J. Chen, J. Chen, S. Zhao, Y. Zhang, and J. Tang. "Exploiting word embedding for heterogeneous topic model towards patent recommendation". In: *Scientometrics* 125.3 (2020), pp. 2091–2108.

[39] N. R. Chrisman. "Rethinking Levels of Measurement for Cartography". In: *Cartography and Geographic Information Systems* 25.4 (1998), pp. 231–242.

[40] J. M. Cigarrán, Á. Castellanos, and A. M. García-Serrano. "A step forward for Topic Detection in Twitter: An FCA-based approach". In: *Expert Systems with Applications* 57 (2016), pp. 21–36.

[41] P. Cimiano, A. Hotho, G. Stumme, and J. Tane. "Conceptual knowledge processing with formal concept analysis and ontologies". In: *Concept Lattices: Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23-26, 2004. Proceedings 2*. Springer. 2004, pp. 189–207.

[42] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, and S. Melacci. "Logic explained networks". In: *Artificial Intelligence* 314 (2023), p. 103822.

[43] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. "What Does BERT Look at? An Analysis of BERT's Attention". In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*. Ed. by T. Linzen, G. Chrupala, Y. Belinkov, and D. Hupkes. Association for Computational Linguistics, 2019, pp. 276–286. DOI: 10.18653/v1/W19-4828.

[44] E. F. Codd, S. B. Codd, and C. T. Salley. *Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate*. E. F. Codd and Associates. 1993.

[45] E. F. Codd. “A relational model of data for large shared data banks”. In: *Communications of the ACM* 13.6 (1970), pp. 377–387.

[46] V. Codocedo, C. Taramasco, and H. Astudillo. “Cheating to achieve Formal Concept Analysis over a Large Formal Context.” In: *CLA*. Ed. by A. Napoli and V. Vychodil. Vol. 959. CEUR-WS.org, 2011, pp. 349–362.

[47] P. Colomb, A. Irlande, and O. Raynaud. “Counting of Moore Families for n=7”. In: *Formal Concept Analysis*. Ed. by L. Kwuida and B. Sertkaya. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 72–87. ISBN: 978-3-642-11928-6.

[48] S. A. Cook. “The complexity of theorem proving procedures”. In: *Proceedings of the Third Annual ACM Symposium*. ACM. New York, 1971, pp. 151–158.

[49] P. J. Crossno, A. T. Wilson, T. M. Shead, and D. M. Dunlavy. “Topicview: Visually comparing topic models of text collections”. In: *2011 ieee 23rd international conference on tools with artificial intelligence*. IEEE. 2011, pp. 936–943.

[50] J. Czerniak and H. Zarzycki. “Application of rough sets in the presumptive diagnosis of urinary system diseases”. In: *Artificial Intelligence and Security in Computing Systems*. Ed. by J. Sołdek and L. Drobiazgiewicz. Boston, MA: Springer US, 2003, pp. 41–51. ISBN: 978-1-4419-9226-0.

[51] S. Daenekindt and J. Huisman. “Mapping the scattered field of research on higher education. A correlated topic model of 17,000 articles, 1991–2018”. In: *Higher Education* 80.3 (2020), pp. 571–587.

[52] J. Deng, R. Socher, L. Fei-Fei, W. Dong, K. Li, and L.-J. Li. “ImageNet: A large-scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*. Vol. 00. June 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[53] S. M. Dias and N. J. Vieira. “Reducing the Size of Concept Lattices: The JBOS Approach”. In: (2010).

[54] R. Diaz-Bone. *Soziologie der Konventionen: Grundlagen einer pragmatischen Anthropologie*. Campus-Verlag, 2011.

[55] R. Diestel. *Graph Theory*. 5th ed. Graduate Texts in Mathematics. Springer Berlin, Heidelberg, 2016. ISBN: 978-3-662-53621-6.

[56] D. Distante, A. Fernandez, L. Cerulo, and A. Visaggio. “Enhancing online discussion forums with topic-driven content search and assisted posting”. In: *Knowledge Discovery, Knowledge Engineering and Knowledge Management: 6th International Joint Conference, IC3K 2014, Rome, Italy, October 21-24, 2014, Revised Selected Papers 6*. Springer. 2015, pp. 161–180.

[57] F. Distel and B. Sertkaya. “On the complexity of enumerating pseudo-intents”. In: *Discrete Applied Mathematics* 159.6 (2011), pp. 450–466. ISSN: 0166-218X. DOI: https://doi.org/10.1016/j.dam.2010.12.004.

[58] S. Doerfel, T. Hanika, and G. Stumme. “Clones in Graphs”. In: *Foundations of Intelligent Systems*. Ed. by M. Ceci, N. Japkowicz, J. Liu, G. A. Papadopoulos, and Z. W. Raś. Cham: Springer International Publishing, 2018, pp. 56–66. ISBN: 978-3-030-01851-1.

[59] Y. Du, X. Meng, and Y. Zhang. "CVTM: A content-venue-aware topic model for group event recommendation". In: *IEEE Transactions on Knowledge and Data Engineering* 32.7 (2019), pp. 1290–1303.

[60] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017.

[61] E. Dudyrev and S. O. Kuznetsov. "Decision concept lattice vs. decision trees and random forests". In: *International Conference on Formal Concept Analysis*. Springer. 2021, pp. 252–260.

[62] S. T. Dumais. "Latent semantic analysis." In: *ARIST* 38.1 (2004), pp. 188–230.

[63] D. Dürrschnabel, M. Koyda, and G. Stumme. "Attribute Selection Using Contranominal Scales". In: *Graph-Based Representation and Reasoning*. Ed. by T. Braun, M. Gehrke, T. Hanika, and N. Hernandez. Cham: Springer International Publishing, 2021, pp. 127–141. ISBN: 978-3-030-86982-3.

[64] D. Dürrschnabel, M. Koyda, and G. Stumme. "Attribute Selection Using Contranominal Scales". In: *Graph-Based Representation and Reasoning: 26th International Conference on Conceptual Structures, ICCS 2021, Virtual Event, September 20–22, 2021, Proceedings*. Springer. 2021, pp. 127–141.

[65] P. W. Eklund, B. Groh, G. Stumme, and R. Wille. "A Contextual-Logic Extension of TOSCANA". In: *Conceptual Structures: Logical, Linguistic and Computational Issues: 8th International Conference on Conceptual Graphs, ICCS'99*. LNAI 1867. see Goda project description for details of this collaboration. Springer Verlag, 2000, pp. 453–467.

[66] M. Erné. "Categories of contexts". In: *arXiv preprint arXiv:1407.0512* (2014).

[67] L. Euler. "Solutio problematis ad geometriam situs pertinentis". In: *Commentarii Academiae Scientiarum Imperialis Petropolitanae* 8 (1736), pp. 128–140.

[68] M. Felde and G. Stumme. "Interactive Collaborative Exploration using Incomplete Contexts." In: *CoRR* abs/1908.08740 (2019).

[69] M. Feurer, J. N. van Rijn, A. Kadra, P. Gijsbers, N. Mallik, S. Ravi, A. Mueller, J. Vanschoren, and F. Hutter. "OpenML-Python: an extensible Python API for OpenML". In: *arXiv* 1911.02490 (2020).

[70] R. Fong and A. Vedaldi. "Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks." In: *CVPR*. IEEE Computer Society, 2018, pp. 8730–8738.

[71] S. Fortunato. "Community detection in graphs". In: *Physics reports* 486.3-5 (2010), pp. 75–174.

[72] L. Frermann and A. Klementiev. "Inducing document structure for aspect-based summarization". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 6263–6273.

[73] J. H. Friedman. "Stochastic gradient boosting". In: *Computational Statistics & Data Analysis* 38.4 (Feb. 2002), pp. 367–378.

[74] B. Ganter and R. Wille. "Conceptual scaling". In: *Applications of combinatorics and graph theory to the biological and social sciences*. Ed. by F. Roberts. Springer-Verlag, 1989, pp. 139–167.

[75] B. Ganter. "Attribute Exploration with Background Knowledge". In: *Theor. Comput. Sci.* 217.2 (1999), pp. 215–233. DOI: 10.1016/S0304-3975(98)00271-0.

[76] B. Ganter. "Two Basic Algorithms in Concept Analysis". English. In: *Formal Concept Analysis*. Ed. by L. Kwuida and B. Sertkaya. Vol. 5986. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, pp. 312–340. ISBN: 978-3-642-11927-9. DOI: 10.1007/978-3-642-11928-6_22.

[77] B. Ganter, T. Hanika, and J. Hirth. "Scaling Dimension". In: *Formal Concept Analysis - 17th International Conference, ICFCA 2023, Kassel, Germany, July 17-21, 2023, Proceedings*. Ed. by D. Dürrschnabel and D. López-Rodríguez. Vol. 13934. Lecture Notes in Computer Science. Springer, 2023, pp. 64–77.

[78] B. Ganter and S. A. Obiedkov. *Conceptual Exploration*. Springer, 2016, pp. 1–315. ISBN: 978-3-662-49291-8.

[79] B. Ganter, J. Stahl, and R. Wille. "Conceptual measurement and many–valued contexts". In: *Classification as a tool of research*. Ed. by W. Gaul and M. Schader. Amsterdam: North–Holland, 1986, pp. 169–176.

[80] B. Ganter and R. Wille. *Formal concept analysis: mathematical foundations*. Berlin; New York: Springer, 1999.

[81] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim. "Towards Automatic Concept-based Explanations." In: *NeurIPS*. Ed. by H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. B. Fox, and R. Garnett. 2019, pp. 9273–9282.

[82] J. Goguen. "What is a concept?" In: *Conceptual Structures: Common Semantics for Sharing Knowledge: 13th International Conference on Conceptual Structures, ICCS 2005, Kassel, Germany, July 17-22, 2005. Proceedings 13*. Springer. 2005, pp. 52–77.

[83] T. Griffiths, M. Jordan, J. Tenenbaum, and D. Blei. "Hierarchical topic models and the nested Chinese restaurant process". In: *Advances in neural information processing systems* 16 (2003).

[84] B. Groh and P. W. Eklund. "Algorithms for Creating Relational Power Context Families from Conceptual Graphs." In: *ICCS*. Ed. by W. M. Tepfenhart and W. R. Cyre. Vol. 1640. Lecture Notes in Computer Science. Springer, 1999, pp. 389–400. ISBN: 3-540-66223-5.

[85] J.-L. Guigues and V. Duquenne. "Familles minimales d'implications informatives résultant d'un tableau de données binaires". In: *Mathématiques et Sciences humaines* 95 (1986), pp. 5–18.

[86] M. Habib and L. Nourine. "The number of Moore families on n=6". In: *Discrete Mathematics* 294.3 (2005), pp. 291–296. ISSN: 0012-365X. DOI: https://doi.org/10.1016/j.disc.2004.11.010.

[87] T. Hanika and J. Hirth. "Conceptual views on tree ensemble classifiers". In: *International Journal of Approximate Reasoning* 159 (2023), p. 108930. ISSN: 0888-613X. DOI: https://doi.org/10.1016/j.ijar.2023.108930.

[88] T. Hanika and J. Hirth. "Conexp-Clj - A Research Tool for FCA." In: *ICFCA (Supplements)*. Ed. by D. Cristea, F. L. Ber, R. Missaoui, L. Kwuida, and B. Sertkaya. Vol. 2378. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 70–75.

[89] T. Hanika and J. Hirth. "Exploring Scale-Measures of Data Sets". In: *Formal Concept Analysis - 16th International Conference, ICFCA 2021, Strasbourg, France, June 29 - July 2, 2021, Proceedings*. Ed. by A. Braud, A. Buzmakov, T. Hanika, and F. L. Ber. Vol. 12733. Lecture Notes in Computer Science. Springer, 2021, pp. 261–269.

[90] T. Hanika and J. Hirth. “Knowledge cores in large formal contexts”. In: *Annals of Mathematics and Artificial Intelligence* 90 (2022), pp. 537–567.

[91] T. Hanika and J. Hirth. “On the lattice of conceptual measurements”. In: *Information Sciences* 613 (2022), pp. 453–468.

[92] T. Hanika and J. Hirth. “Quantifying the Conceptual Error in Dimensionality Reduction”. In: *Graph-Based Representation and Reasoning - 26th International Conference on Conceptual Structures, ICCS 2021, 2021, Proceedings*. Ed. by T. Braun, M. Gehrke, T. Hanika, and N. Hernandez. Vol. 12879. Lecture Notes in Computer Science. Springer, 2021, pp. 105–118.

[93] T. Hanika, M. Koyda, and 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*. Ed. by D. Endres, M. Alam, and D. Sotropa. Vol. 11530. Lecture Notes in Computer Science. Springer, 2019, pp. 102–116.

[94] T. Hanika, F. M. Schneider, and G. Stumme. “Intrinsic dimension of geometric data sets”. In: *Tohoku Mathematical Journal* (2018).

[95] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[96] F. Herrera, C. Carmona, P. González, and M. del Jesus. “An overview on subgroup discovery: foundations and applications”. In: *Knowledge and Information Systems* (2010), pp. 1–31. ISSN: 0219-1377. DOI: 10.1007/s10115-010-0356-2.

[97] A. Higuchi. “Lattices of closure operators”. In: *Discrete Mathematics* 179.1 (1998), pp. 267–272. ISSN: 0012-365X. DOI: https://doi.org/10.1016/S0012-365X(97)00099-X.

[98] J. Hirth and T. Hanika. *Formal Conceptual Views in Neural Networks*. 2022. arXiv: 2209.13517 [cs.LG].

[99] J. Hirth and T. Hanika. *The Geometric Structure of Topic Models*. 2024. DOI: 10.48550/arxiv.2403.03607. arXiv: 2403.03607 [cs.AI].

[100] J. Hirth, V. Horn, G. Stumme, and T. Hanika. “Automatic Textual Explanations of Concept Lattices”. In: *Graph-Based Representation and Reasoning*. Ed. by M. Ojeda-Aciego, K. Sauerwald, and R. Jäschke. Cham: Springer Nature Switzerland, 2023, pp. 138–152.

[101] J. Hirth, V. Horn, G. Stumme, and T. Hanika. “Ordinal Motifs in Lattices”. In: *Information Sciences* (2023), p. 120009.

[102] W. Hodges. *Model Theory*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1993. DOI: 10.1017/CBO9780511551574.

[103] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. d. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. “Knowledge graphs”. In: *ACM Computing Surveys (CSUR)* 54.4 (2021), pp. 1–37.

[104] P. W. Holland and S. Leinhardt. “Local structure in social networks”. In: *Sociological methodology* 7 (1976), pp. 1–45.

[105] P. W. Holland and S. Leinhardt. *The statistical analysis of local structure in social networks*. 1974.

[106] A. Horn. "On sentences which are true of direct unions of algebras". In: *Journal of Symbolic Logic* 16 (1951), pp. 14–21.

[107] V. Horn, J. Hirth, J. Holfeld, J. H. Behmenburg, C. Draude, and G. Stumme. "Disclosing Diverse Perspectives of News Articles for Navigating between Online Journalism Content". In: *Nordic Conference on Human-Computer Interaction*. NordiCHI 2024. Uppsala, Sweden: Association for Computing Machinery, 2024. ISBN: 9798400709661. DOI: 10.1145/3679318.3685414.

[108] A. Hoyle, P. Goel, D. Peskov, A. Hian-Cheong, J. Boyd-Graber, and P. Resnik. *Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence*. 2021. arXiv: 2107.02173 [cs.CL].

[109] W. Iba and P. Langley. "Induction of one-level decision trees". In: *Machine Learning Proceedings 1992*. Elsevier, 1992, pp. 233–240.

[110] T. Ihringer. *Allgemeine Algebra: mit einem Anhang „Abstrakte Datentypen "*. Teubner Studienbücher Mathematik. Vieweg+Teubner Verlag Wiesbaden, 2013. DOI: https://doi.org/10.1007/978-3-322-92676-0.

[111] P. Jackson. "Introduction to expert systems". In: (1986).

[112] R. M. Karp. "Reducibility Among Combinatorial Problems". In: *Complexity of Computer Computations*. Ed. by R. E. Miller and J. W. Thatcher. Plenum Press, 1972, pp. 85–103.

[113] M. Kawai, H. Sato, and T. Shiohama. "Topic model-based recommender systems and their applications to cold-start problems". In: *Expert Systems with Applications* 202 (2022), p. 117129.

[114] M. Kaytoue, S. Duplessis, S. O. Kuznetsov, and A. Napoli. "Two fca-based methods for mining gene expression data". In: *ICFCA*. Vol. 5548. Springer Heidelberg. 2009, pp. 251–266.

[115] M. Kaytoue, S. O. Kuznetsov, J. Macko, and A. Napoli. "Biclustering meets triadic concept analysis." In: *Ann. Math. Artif. Intell.* 70.1-2 (2014), pp. 55–79.

[116] D. Kelly and I. Rival. "Crowns, fences, and dismantlable lattices". In: *Canadian Journal of Mathematics* 26.5 (1974), pp. 1257–1271.

[117] K. Kersting, J. Peters, and C. Rothkopf. "Was ist eine Professur fuer Kuenstliche Intelligenz?" In: *arXiv preprint arXiv:1903.09516* (2019).

[118] S. A. Khanam, F. Liu, and Y.-P. P. Chen. "Joint knowledge-powered topic level attention for a convolutional text summarization model". In: *Knowledge-Based Systems* 228 (2021), p. 107273.

[119] M. Khodorchenko, N. Butakov, and D. Nasonov. "Towards Better Evaluation of Topic Model Quality". In: *2022 32nd Conference of Open Innovations Association (FRUCT)*. 2022, pp. 128–134. DOI: 10.23919/FRUCT56874.2022.9953874.

[120] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. sayres. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2668–2677.

[121] J. Kim and S. Choi. "On Uncertainty Estimation by Tree-based Surrogate Models in Sequential Model-based Optimization". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. Ed. by G. Camps-Valls, F. J. R. Ruiz, and I. Valera. Vol. 151. Proceedings of Machine Learning Research. PMLR, Mar. 2022, pp. 4359–4375.

[122] M. Klimushkin, S. Obiedkov, and C. Roth. "Approaches to the Selection of Relevant Concepts in the Case of Noisy Data". In: *ICFCA 2010*. Ed. by L. Kwuida and B. Sertkaya. Springer Berlin, 2010, pp. 255–266.

[123] W. Kollewe, M. Skorsky, F. Vogt, and R. Wille. "TOSCANA - ein Werkzeug zur begrifflichen Analyse und Erkundung von Daten". In: *Begriffliche Wissensverarbeitung-Grundfragen und Aufgaben*. Ed. by R. Wille and M. Zickwolff. Mannheim: B. I. -Wissenschaftsverlag, 1994, pp. 267–288.

[124] B. Korbar, A. M. Olofson, A. P. Miraflor, C. M. Nicka, M. A. Suriawinata, L. Torresani, A. A. Suriawinata, and S. Hassanpour. "Looking Under the Hood: Deep Neural Network Visualization to Interpret Whole-Slide Image Analysis Outcomes for Colorectal Polyps". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. July 2017.

[125] M. Koyda and G. Stumme. "Boolean Substructures in Formal Concept Analysis". In: *Formal Concept Analysis - 16th International Conference, ICFCA 2021, Strasbourg, France, June 29 – July 2, 2021, Proceedings*. Lecture Notes in Computer Science. Springer, 2021, pp. 38–53. ISBN: 978-3-030-77866-8.

[126] M. Krötzsch, P. Hitzler, and G.-Q. Zhang. "Morphisms in Context." In: *Proceedings of the 13th International Conference on Conceptual Structures (ICCS 2005)*. Ed. by F. Dau, M.-L. Mugnier, and G. Stumme. Vol. 3596. Lecture Notes in Computer Science. Springer, 2005, pp. 223–237. ISBN: 3-540-27783-8.

[127] C. A. Kumar and S. Srinivas. "Concept lattice reduction using fuzzy K-means clustering". In: *Expert systems with applications* 37.3 (2010), pp. 2696–2704.

[128] C. A. Kumar, S. M. Dias, and N. J. Vieira. "Knowledge reduction in formal contexts using non-negative matrix factorization." In: *Math. Comput. Simul.* 109 (2015), pp. 46–63.

[129] S. Kuznetsov. "On the intractability of computing the Duquenne-Guigues base". In: *Journal of Universal Computer Science* 10.8 (2004), pp. 927–933.

[130] S. O. Kuznetsov. "On Computing the Size of a Lattice and Related Decision Problems." In: *Order* 18.4 (2001), pp. 313–321.

[131] S. O. Kuznetsov. "On stability of a formal concept". In: *Annals of Mathematics and Artificial Intelligence* 49.1 (Apr. 2007), pp. 101–115. ISSN: 1573-7470. DOI: 10.1007/s10472-007-9053-6.

[132] S. O. Kuznetsov and T. P. Makhalova. "On interestingness measures of formal concepts". In: *Inf. Sci.* 442-443 (2018), pp. 202–219. DOI: 10.1016/j.ins.2018.02.032.

[133] S. O. Kuznetsov and S. Obiedkov. "Some decision and counting problems of the Duquenne–Guigues basis of implications". In: *Discrete Applied Mathematics* 156.11 (2008). In Memory of Leonid Khachiyan (1952 - 2005), pp. 1994–2003. ISSN: 0166-218X. DOI: https://doi.org/10.1016/j.dam.2007.04.014.

[134] L. Kwuida, R. Missaoui, A. Balamane, and J. Vaillancourt. "Generalized pattern extraction from concept lattices." In: *Ann. Math. Artif. Intell.* 72.1-2 (2014), pp. 151–168.

[135] D. D. Lee and H. S. Seung. "Learning the parts of objects by non-negative matrix factorization." In: *Nature* 401.6755 (Oct. 1999), pp. 788–791. ISSN: 0028-0836. DOI: 10.1038/44565.

[136] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[137] S. Li. *Food.com Recipes and Interactions*. 2019. DOI: 10.34740/KAGGLE/DSV/783630.

[138] W. Li and A. McCallum. "Pachinko allocation: DAG-structured mixture models of topic correlations". In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 577–584.

[139] Z. Liao, Z. Wu, Y. Li, Y. Zhang, X. Fan, and J. Wu. "Core-reviewer recommendation based on Pull Request topic model and collaborator social network". In: *Soft Computing* 24 (2020), pp. 5683–5693.

[140] L. Libkin. "Direct Decompositions of Atomistic Algebraic Lattices". In: *Algebra Universalis* 33.1 (1995), pp. 127–135.

[141] L. Libkin. "Direct product decompositions of lattices, closures and relation schemes." In: *Discret. Math.* 112.1-3 (1993), pp. 119–138.

[142] T. Lin, W. Tian, Q. Mei, and H. Cheng. "The dual-sparse topic model: mining focused topics and focused terms in short text." In: *WWW*. Ed. by C.-W. Chung, A. Z. Broder, K. Shim, and T. Suel. ACM, 2014, pp. 539–550. ISBN: 978-1-4503-2744-2.

[143] C. Lindig. "Fast concept analysis". In: *Working with Conceptual Structures – Contributions to ICCS 2000*. Shaker Verlag, 2000, pp. 152–161.

[144] B. Liu, Y. Xia, and P. S. Yu. "Clustering through decision tree construction". In: *International Conference on Information and Knowledge Management*. 2000.

[145] R. D. Luce, D. H. Krantz, P. Suppes, and A. Tversky. *Foundations of Measurement – Representation, Axiomatization, and Invariance*. Vol. 3. Academic Press, 1990.

[146] M. Luxenburger. "Implications partielles dans un contexte". In: *Mathématiques, Informatique et Sciences Humaines* 29.113 (1991), pp. 35–55.

[147] M. Mahn. *Gewürze: das Standardwerk*. Ed. by G. Müller-Wallraf. München: Christian, 2014, p. 319. ISBN: 9783862446773.

[148] J. A. Makowsky. "Why Horn Formulas Matter in Computer Science: Initial Structures and Generic Examples (Extended Abstract)." In: *TAPSOFT, Vol.1*. Ed. by H. Ehrig, C. Floyd, M. Nivat, and J. W. Thatcher. Vol. 185. Lecture Notes in Computer Science. Springer, 1985, pp. 374–387. ISBN: 3-540-15198-2.

[149] T. I. Mamun, K. Baker, H. Malinowski, R. R. Hoffman, and S. T. Mueller. "Assessing collaborative explanations of AI using explanation goodness criteria". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Vol. 65. 1. SAGE Publications Sage CA: Los Angeles, CA. 2021, pp. 988–993.

[150] J. Mao, C. Gan, P. Kohli, J. B. Tenenbaum, and J. Wu. “The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision.” In: *ICLR*. OpenReview.net, 2019.

[151] E. Margolis and S. Laurence. “The ontology of concepts-abstract objects or mental representations?” In: *Noûs* 41.4 (2007), pp. 561–593.

[152] A. Mead. “Review of the development of multidimensional scaling methods”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 41.1 (1992), pp. 27–39.

[153] D. P. Mehta and V. Raghavan. “Decision Tree Approximations of Boolean Functions.” In: *COLT*. Ed. by N. Cesa-Bianchi and S. A. Goldman. Morgan Kaufmann, 2000, pp. 16–24. ISBN: 1-55860-703-X.

[154] F. Mémoli. “Gromov-Wasserstein Distances and the Metric Approach to Object Matching.” In: *Foundations of Computational Mathematics* 11.4 (2011), pp. 417–487.

[155] P. Miettinen, T. Mielikainen, A. Gionis, G. Das, and H. Mannila. “The Discrete Basis Problem”. In: *Knowledge and Data Engineering, IEEE Transactions on* 20.10 (Oct. 2008), pp. 1348–1362. ISSN: 1041-4347. DOI: 10.1109/TKDE.2008.53.

[156] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. “Network motifs: simple building blocks of complex networks”. In: *Science* 298.5594 (2002), pp. 824–827.

[157] M. Minea, C. Dumitrescu, and I. Chiva. “Unconventional Public Transport Anonymous Data Collection employing Artificial Intelligence”. In: *2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*. 2019, pp. 1–6.

[158] T. M. Mitchell. *Machine learning*. New York, NY: McGraw-Hill, 2010.

[159] F. Moosmann, B. Triggs, and F. Jurie. “Fast Discriminative Visual Codebooks using Randomized Clustering Forests”. In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press, 2006.

[160] J. Morgado. “Note on the central closure operators of complete lattices”. In: *Proc. Koninkl. nederl. akad. wet. A*. Vol. 67. 1964, pp. 467–476.

[161] J. Morgado. “Note on the distributive closure operators of a complete lattice”. In: *Portugaliae mathematica* 23.1 (1964), pp. 11–25.

[162] H. Mureşan and M. Oltean. “Fruit recognition from images using deep learning”. In: *arXiv preprint arXiv:1712.00580* (2017).

[163] M. E. J. Newman. “The structure of scientific collaboration networks”. In: *Proceedings of the National Academy of Sciences* 98.2 (2001), pp. 404–409. DOI: 10.1073/pnas.98.2.404. eprint: http://www.pnas.org/content/98/2/404.full.pdf+html.

[164] H. Nguyen and T. Maehara. “Graph homomorphism convolution”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7306–7316.

[165] L. Nourine and O. Raynaud. “A fast algorithm for building lattices”. In: *Information processing letters* 71.5-6 (1999), pp. 199–204.

[166] T. Pattison, M. Enciso, Á. Mora, P. Cordero, D. Weber, and M. Broughton. "Scalable Visual Analytics in FCA". In: *Complex Data Analytics with Formal Concept Analysis*. Ed. by R. Missaoui, L. Kwuida, and T. Abdessalem. Cham: Springer International Publishing, 2022, pp. 167–200. ISBN: 978-3-030-93278-7. DOI: 10.1007/978-3-030-93278-7_8.

[167] D. C. Penn, K. J. Holyoak, and D. J. Povinelli. "Darwin's mistake: Explaining the discontinuity between human and nonhuman minds". In: *Behavioral and Brain Sciences* 31 (2008), pp. 109–130.

[168] J. Pfanzagl. *Theory of Measurement*. Heidelberg: Physica, 1971.

[169] S. Pollandt and R. Wille. "Functorial scaling of ordinal data". In: *Discret. Appl. Math.* 147 (2005), pp. 101–111.

[170] S. Prediger and G. Stumme. "Theory-Driven Logical Scaling". In: *Proc. 6th Intl. Workshop Knowledge Representation Meets Databases (KRDB'99)*. Ed. by E. F. et al. Vol. CEUR Workshop Proc. 21. Also in: P. Lambrix et al (Eds.): Proc. Intl. WS on Description Logics (DL'99). CEUR Workshop Proc. 22, 1999. 1999.

[171] S. Prediger. "Logical Scaling in Formal Concept Analysis". In: *Conceptual Structures: Fulfilling Peirce's Dream, Fifth International Conference on Conceptual Structures, ICCS '97, Seattle, Washington, USA, August 3-8, 1997, Proceedings*. Ed. by D. Lukose, H. S. Delugach, M. Keeler, L. Searle, and J. F. Sowa. Vol. 1257. Lecture Notes in Computer Science. Springer, 1997, pp. 332–341. DOI: 10.1007/BFb0027881.

[172] S. Prediger. "Logical Scaling in Formal Concept Analysis." In: *ICCS*. Ed. by D. Lukose, H. S. Delugach, M. Keeler, L. Searle, and J. F. Sowa. Vol. 1257. Lecture Notes in Computer Science. Springer, 1997, pp. 332–341. ISBN: 3-540-63308-1.

[173] S. Prediger and R. Wille. "The Lattice of Concept Graphs of a Relationally Scaled Context." In: *ICCS*. Ed. by W. M. Tepfenhart and W. R. Cyre. Vol. 1640. Lecture Notes in Computer Science. Springer, 1999, pp. 401–414. ISBN: 3-540-66223-5.

[174] P. Probst, M. N. Wright, and A.-L. Boulesteix. "Hyperparameters and tuning strategies for random forest". In: *Wiley Interdisciplinary Reviews: data mining and knowledge discovery* 9.3 (2019), e1301.

[175] O. Prokasheva, A. Onishchenko, and S. Gurov. "Classification methods based on formal concept analysis". In: *FCAIR 2012–Formal Concept Analysis Meets Information Retrieval* (2013), p. 95.

[176] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN: 1-55860-238-0.

[177] J. Ramos et al. "Using tf-idf to determine word relevance in document queries". In: *Proceedings of the first instructional conference on machine learning*. Vol. 242. 1. Citeseer. 2003, pp. 29–48.

[178] M. T. Ribeiro, S. Singh, and C. Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." In: *KDD*. Ed. by B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi. ACM, 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2.

[179] M. Röder, A. Both, and A. Hinneburg. "Exploring the Space of Topic Coherence Measures." In: *WSDM*. Ed. by X. Cheng, H. Li, E. Gabrilovich, and J. Tang. ACM, 2015, pp. 399–408. ISBN: 978-1-4503-3317-7.

[180] J. M. Rodríguez-Jiménez, P. Cordero, M. Enciso, and Á. Mora. "Negative Attributes and Implications in Formal Concept Analysis". In: *International Conference on Information Technology and Quantitative Management*. 2014.

[181] T. Rögnvaldsson, L. You, and D. Garwicz. "State of the art prediction of HIV-1 protease cleavage sites". In: *Bioinformatics* 31.8 (2015), pp. 1204–1210.

[182] R. K. Roul. "Topic modeling combined with classification technique for extractive multi-document text summarization". In: *Soft computing* 25 (2021), pp. 1113–1127.

[183] P. J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: `https://doi.org/10.1016/0377-0427(87)90125-7`.

[184] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215. ISSN: 25225839. DOI: `10.1038/s42256-019-0048-x`.

[185] S. Ruggieri. "Enumerating Distinct Decision Trees". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 2960–2968.

[186] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: `10.1007/s11263-015-0816-y`.

[187] S. J. Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

[188] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler. "Neuro-symbolic artificial intelligence". In: *AI Communications* 34.3 (Mar. 2022), pp. 197–209. DOI: `10.3233/aic-210084`.

[189] B. Schaefermeier, G. Stumme, and T. Hanika. "Topic space trajectories: A case study on machine learning literature". In: *Scientometrics* 126.7 (2021), pp. 5759–5795.

[190] B. Schäfermeier, J. Hirth, and T. Hanika. "Research Topic Flows in Co-Authorship Networks". In: *Scientometrics* (Oct. 2022).

[191] B. Schäfermeier, G. Stumme, and T. Hanika. *Mapping Research Trajectories*. 2022. DOI: `10.48550/ARXIV.2204.11859`.

[192] R. C. Schank and R. P. Abelson. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press, 2013.

[193] P. Scheich, M. Skorsky, F. Vogt, C. Wachter, and R. Wille. "Conceptual Data Systems." In: *Information and Classification*. Ed. by O. Opitz, B. Lausen, and R. Klar. Springer, Berlin-Heidelberg, 1993, pp. 72–84.

[194] E. Schubert and L. Lenssen. "Fast k-medoids Clustering in Rust and Python". In: *Journal of Open Source Software* 7.75 (2022), p. 4183. DOI: `10.21105/joss.04183`.

[195] E. Schubert and P. J. Rousseeuw. "Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms". In: *Information Systems* 101 (2021), p. 101804. ISSN: 0306-4379. DOI: `https://doi.org/10.1016/j.is.2021.101804`.

[196] G. Schwalbe and B. Finzel. "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts". In: *Data Mining and Knowledge Discovery* (2023), pp. 1–59.

[197] V. Shankar, A. Fang, W. Guo, S. Fridovich-Keil, J. Ragan-Kelley, L. Schmidt, and B. Recht. "Neural Kernels Without Tangents". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 8614–8623.

[198] A. Smith, T. Hawes, and M. Myers. "Hierarchie: Visualization for hierarchical topic models". In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 2014, pp. 71–78.

[199] D. Solans, C. Tauchmann, A. Farrell, K. Kappler, H.-H. Huber, C. Castillo, and K. Kersting. "Learning to Classify Morals and Conventions: Artificial Intelligence in Terms of the Economics of Convention". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 691–702.

[200] R. Srivastava, P. Singh, K. Rana, and V. Kumar. "A topic modeled unsupervised approach to single document extractive text summarization". In: *Knowledge-Based Systems* 246 (2022), p. 108636. ISSN: 0950-7051. DOI: `https://doi.org/10.1016/j.knosys.2022.108636`.

[201] S. S. Stevens. "On the Theory of Scales of Measurement". In: *Science* 103.2684 (1946), pp. 677–680. ISSN: 0036-8075.

[202] S. Strahringer and R. Wille. "Towards a Structure Theory for Ordinal Data". In: *Analyzing and Modeling Data and Knowledge*. Ed. by M. Schader. Berlin, Heidelberg: Springer Berlin Heidelberg, 1992, pp. 129–139. ISBN: 978-3-642-46757-8.

[203] P. Strecht. "A survey of merging decision trees data mining approaches". In: *Proc. 10th Doctoral Symposium in Informatics Engineering*. 2015, pp. 36–47.

[204] G. Stumme. "Hierarchies of Conceptual Scales". In: *Proc. Workshop on Knowledge Acquisition, Modeling and Management (KAW'99)*. Ed. by T. B. Gaines, R. Kremer, and M. Musen. Vol. 2. Banff, Aug. 1999, pp. 78–95.

[205] G. Stumme. "Boolesche Begriffe". Diplomarbeit. TH Darmstadt, 1994.

[206] G. Stumme. "Local Scaling in Conceptual Data Systems". In: *Conceptual Structures: Knowledge Representation as Interlingua Proc. ICCS'96*. Ed. by P. W. Eklund, G. Ellis, and G. Mann. Vol. 1115. LNAI. Heidelberg: Springer, 1996, pp. 308–320.

[207] G. Stumme, D. Dürrschnabel, and T. Hanika. *Towards Ordinal Data Science*. 2023. DOI: `10.48550/arXiv.2307.09477`. arXiv: `2307.09477`.

[208] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, and L. Lakhal. "Computing iceberg concept lattices with TITANIC". In: *Data & Knowledge Engineering* 42.2 (Aug. 2002), pp. 189–222. ISSN: 0169-023X. DOI: `10.1016/S0169-023X(02)00057-5`.

[209] F. Sun, Y. Li, and Z. Zhang. "A Tool for Visualizing Topic Evolution in Large Text Collections". In: *2013 IEEE 13th International Conference on Advanced Learning Technologies*. 2013, pp. 53–54. DOI: `10.1109/ICALT.2013.21`.

[210] P. Suppes, D. H. Krantz, R. D. Luce, and A. Tversky. *Foundations of Measurement – Geometrical, Threshold, and Probabilistic Representations*. Vol. 2. Academic Press, 1989.

[211] P. A. Takizawa. “Using a topic model to map and analyze a large curriculum”. In: *Plos one* 18.4 (2023), e0284513.

[212] N. Tatti, T. Mielikäinen, A. Gionis, and H. Mannila. “What is the Dimension of Your Binary Data?” In: *Sixth International Conference on Data Mining (ICDM’06)* (2006), pp. 603–612.

[213] N. Tatti, F. Moerchen, and T. Calders. “Finding Robust Itemsets under Subsampling.” In: *ACM Trans. Database Syst.* 39.3 (2014), 20:1–20:27.

[214] M. Uhlmann, J. Hirth, and V. Horn. *Jenseits der Logik der Empfehlung: Formale Begriffsanalyse als Grundlage für eine neue Variante zur Vermittlung von Nutzenden und journalistischen Inhalten*. Submitted. 2023.

[215] L. Van der Maaten and G. Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[216] M. Venugopalan and D. Gupta. “An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis”. In: *Knowledge-Based Systems* 246 (2022), p. 108668. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2022.108668.

[217] T. Vidal and M. Schiffer. “Born-Again Tree Ensembles”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 9743–9753.

[218] F. Vogt and R. Wille. “TOSCANA – A graphical tool for analyzing and exploring data.” In: vol. 894. LNCS. Springer, Heidelberg, 1995, pp. 226–233.

[219] Y. Wang, N. Wagner, and J. M. Rondinelli. “Symbolic regression in materials science”. In: *MRS Communications* 9.3 (2019), pp. 793–805.

[220] M. Wild. “The joy of implications, aka pure Horn formulas: mainly a survey”. In: *Theoretical Computer Science* 658 (2017), pp. 264–292.

[221] R. Wille. “Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies.” In: *Formal Concept Analysis*. Ed. by B. Ganter, G. Stumme, and R. Wille. Vol. 3626. Lecture Notes in Computer Science. Springer, 2005, pp. 1–33. ISBN: 3-540-27891-5.

[222] R. Wille. “Geometric representation of concept lattices”. In: *Conceptual and numerical analysis of data*. Ed. by O. Opitz. Berlin–Heidelberg: Springer–Verlag, 1989, pp. 239–255.

[223] R. Wille. “Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts”. English. In: *Ordered Sets*. Ed. by I. Rival. Vol. 83. NATO Advanced Study Institutes Series. Springer Netherlands, 1982, pp. 445–470. ISBN: 978-94-009-7800-3. DOI: 10.1007/978-94-009-7798-3_15.

[224] R. Wille. “The Basic Theorem of triadic concept analysis”. In: *Order* 12.2 (1995), pp. 149–158. ISSN: 0167-8094. DOI: 10.1007/BF01108624.

[225] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu. *ChatGPT is not Enough: Enhancing Large Language Models with Knowledge Graphs for Fact-aware Language Modeling*. 2023. arXiv: 2306.11489 [cs.CL].

[226] M. Yannakakis. “The Complexity of the Partial Order Dimension Problem”. In: *Siam Journal on Algebraic and Discrete Methods* 3 (1982), pp. 351–358.

[227] C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar. “On completeness-aware concept-based explanations in deep neural networks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 20554–20565.

[228] H. Yin, X. Song, S. Yang, and J. Li. “Sentiment analysis and topic modeling for COVID-19 vaccine discussions”. In: *World Wide Web* 25.3 (2022), pp. 1067–1083.

[229] V. N. Zemlyachenko, N. M. Korneenko, and R. I. Tyshkevich. “Graph isomorphism problem”. In: *Journal of Soviet Mathematics* 29 (1985), pp. 1426–1481.

[230] Z. Zhang, T. Li, C. Ding, and X. Zhang. “Binary matrix factorization with applications”. In: *7th IEEE int. conf. on data mining (ICDM 2007)*. IEEE. 2007, pp. 391–400.

[231] H. Zhao, L. Du, W. Buntine, and M. Zhou. “Inter and intra topic structure learning with word embeddings”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5892–5901.

[232] T. Zhou, K. Law, and D. Creighton. “A weakly-supervised graph-based joint sentiment topic model for multi-topic sentiment analysis”. In: *Information Sciences* 609 (2022), pp. 1030–1051.

[233] M. Zitnik and B. Zupan. “Nimfa: A Python Library for Nonnegative Matrix Factorization”. In: *JMLR* 13 (2012), pp. 849–853.

Research Data

Throughout this work we have exclusively used pre-existing data sets for our analysis. Data sets represented as formal contexts or concept lattices, including conceptual and contextual views, can be inferred from the respective diagrams. Besides these, we used *spices planner* from Mahn [147]; *Diagnosis* [50], *Hayes-Roth*, *Zoo*, *Mushroom*, *HIV-1ProteaseCleavage* [181] and *Plant-Habitats* obtained from the UCI repository[1] [60]; *Top-Chess-Players*, *Airbnb-Berlin*, *A_fighter* and *B_fighter* from the *UFC-Fights* and *Recipes* [137] from the kaggle[2] repository; *Domesticated Animals* from Wikipedia[3]; the *car* data set from OpenML [69, ID:991]; *ImageNet* [52] as well as all twenty-four[4] NN models from tensorflow that are trained on ImageNet; *Fruits-360* [162] and the *Semantic Scholar Open Research Corpus* [7] data set obtained on January 31st, 2019. From the latter, we used the subset given by thirty-two machine learning conferences [117] as also used in Schaefermeier, Stumme, and Hanika [189]. A detailed index on where we used each data set can be found in an accompanied index.

All methods proposed in Chapters 7 to 10 are implemented in `conexp-clj` [88], a research framework for FCA. The code for our experiments in Chapter 11,[5] Chapter 12,[6] Chapter 14[7] and Chapter 13[8] are published in public repositories.

[1] i) https://archive.ics.uci.edu/ml/datasets/Acute+Inflammations,
ii) https://archive.ics.uci.edu/ml/datasets/Hayes-Roth,
iii) https://archive.ics.uci.edu/ml/datasets/zoo,
iv) https://archive.ics.uci.edu/ml/datasets/mushroom,
v) https://archive.ics.uci.edu/ml/datasets/HIV-1+protease+cleavage,
vi) https://archive.ics.uci.edu/ml/datasets/Plants and https://plants.sc.egov.usda.gov/java/,

[2] vii) https://www.kaggle.com/odartey/top-chess-players and https://www.fide.com/,
viii) https://www.kaggle.com/brittabettendorf/berlin-airbnb-data/,
ix) https://www.kaggle.com/rajeevw/ufcdata,
x) https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions,

[3] xi) https://en.wikipedia.org/w/index.php?title=List_of_domesticated_animals 25.02.2020

[4] https://www.tensorflow.org/api_docs/python/tf/keras/applications/, July 2022

[5] https://github.com/hirthjo/Conceptual-Scaling-Error

[6] https://github.com/hirthjo/conceptual-views-on-tree-classifiers

[7] https://github.com/hirthjo/The-Geometric-Structure-of-Topic-Models

[8] https://github.com/FCA-Research/Formal-Conceptual-Views-in-Neural-Networks

List of Data Sets

List of Problems

List of Definitions and Statements

List of Figures and Tables