Hierarchies of Conceptual Scales

Gerd Stumme

Technische Universität Darmstadt, Fachbereich Mathematik,
Schloßgartenstr. 7, D–64289 Darmstadt; stumme@mathematik.tu-darmstadt.de

Abstract

Formal Concept Analysis allows to derive conceptual hierarchies from data tables. Formal Concept Analysis is applied in various domains, e. g., data analysis, information retrieval, and knowledge discovery in databases. In order to deal with increasing sizes of the data tables (and to allow more complex data structures than just binary attributes), conceptual scales have been developed. They are considered as metadata which structure the data conceptually. But in large applications, the number of conceptual scales increases as well. Techniques are needed which support the navigation of the user also on this meta-level of conceptual scales. In this paper, we attack this problem by extending the set of scales by hierarchically ordered higher level scales and by introducing a visualization technique called nested scaling. We extend the two-level architecture of Formal Concept Analysis (the data table plus one level of conceptual scales) to a many-level architecture with a cascading system of conceptual scales. The approach also allows to use representation techniques of Formal Concept Analysis for the visualization of thesauri and ontologies.

1 Introduction

Formal Concept Analysis (Wille 1982; Ganter, Wille 1999) is a mathematical theory which formalizes the understanding of 'concept' as a unit of thought consisting of two parts, its extension and its intension (Arnauld, Nicole 1668; Wagner 1973; DIN 1979). From descriptions of objects by attributes and attribute–value–pairs, Formal Concept Analysis generates a conceptual hierarchy which reflects the conceptual structure of the domain. During the last twenty years, Formal Concept Analysis has grown to a variety of methods for data analysis, information retrieval, knowledge acquisition, and knowledge discovery in databases.

In its most basic form, Formal Concept Analysis starts with a *formal context*, which is a binary relation between a set of *objects* and a set of *attributes*.[1] From the formal context, one derives formal concepts and a concept lattice: A *formal concept* consists of two parts, its extent and its intent. The *extent* is a subset of the set of objects, and the *intent* is a subset of the set of attributes such that each object in the extent belongs to each attribute in the intent, each attribute in the intent is common to all objects in the extent, and no further object and no further attribute can be added without violating one of these two conditions. The generic subconcept–superconcept–relation provides a lattice structure on the set of formal concepts, called *concept lattice*. The visualization of concept lattices by *line diagrams* is used to present the conceptual structure of the data to the user.

[1] Precise definitions of all words occuring in the introduction are given in the next section.

Conceptual scaling has been introduced as a technique for dealing with large contexts, and for dealing with so-called *many-valued contexts* which consist of attribute-value pairs. If the user is interested in analyzing the interrelationship between attributes of a large (one- or many-valued) context, he can choose among the conceptual scales those which contain the required attributes. The visualization of their apposition by a nested line diagram allows him to study the large concept lattice which is embedded in the direct product. Formal Concept Analysis, and especially conceptual scaling, provides structured conceptual meta-information about the data. It has been applied successfully in Information Retrieval because it allows users, who have only a vague idea of what they are looking for, a structured overview over a spectrum of related queries by providing a visualization of the resulting conceptual hierarchy.

In real-world applications, the number of conceptual scales becomes large. An efficient navigation is needed also on the meta-level, i. e., on the level of conceptual scales. In this paper, we introduce new *higher level attributes* and a taxonomy on these new attributes, from which we derive new, hierarchically ordered *higher level scales*. We obtain a cascading hierarchy of conceptual scales with increasing granularity. *Nested scaling* is introduced for visualizing the combination of higher level scales. It is based on the visualization technique of *local scaling* (Stumme 1996).

Higher level scales (which can be derived automatically from the taxonomy) provide information about the data on a more general level. They allow to observe global 'cross-scale' relationships that cannot be recognised easily otherwise. Hence higher level scales are an interesting technique for Knowledge Discovery in Databases (KDD; for an interplay between KDD and Formal Concept Analysis cf. also to (Stumme, Wille, Wille 1998; Pasquier, Bastide, Taouil, Lakhal 1999; Hereth, Stumme, Wille, Wille 1999; Stumme 1999b)), while they allow at the same time a drill-down to the original data as known from Online-Analytical Processing (OLAP; cf. also to (Stumme 1998)). In this paper, however, we focus on the use of higher level scales in an Information Retrieval setting.

In order to allow the navigation in ontologies and thesauri by representation techniques of Formal Concept Analysis, they have to be transformed into conceptual scales. Instead of a bottom-up approach as discussed above, one can here apply a top-down approach. Using the subconcept-superconcept-relation of the ontology or thesaurus, one obtains automatically a conceptual scale for each concept of the thesaurus by choosing as attributes of the scale its immediate subconcepts in the thesaurus.

In the next section, we briefly recall the notions of Formal Concept Analysis as far as they are needed in this paper. In Section 3, we extend the set of conceptual scales by higher level scales and discuss their use for a graphical user interface for thesauri. In Section 4 we introduce the visualization technique called nested scaling. Section 5 provides an evaluation and indicates open questions for further research.

2 Conceptual Scaling of Formal Contexts

2.1 Formal Contexts and Their Concept Lattices

The most basic data structure of Formal Concept Analysis is a formal context:

Definition: A *(formal) context* is a triple $\mathbb{K} := (G, M, I)$ where G and M are sets and I is a relation between G and M. The elements of G and M are called *objects* and *attributes*, respectively, and $(g, m) \in I$ is read *"object g has attribute m"*.

We illustrate the definition by an example. The example which we will use throughout the paper is a library retrieval system of the Center for Interdisciplinary Technology Studies (ZIT) of the Darmstadt University of Technology. The set G of objects consists of 1556 books of the library, the set of attributes of 377 catchwords. The binary relation I consists of ca. 50000 tuples, i.e, each book is assigned to 32 catchwords in average. The definition of the list of catchwords as well as the assignment to the books was a complex process. It is described in detail in (Rock, Wille 1999).

Definition: For $A \subseteq G$, we define $A' := \{m \in M \mid \forall g \in A\colon (g, m) \in I\}$. Dually, for $B \subseteq M$, we define $B' := \{g \in G \mid \forall m \in B\colon (g, m) \in I\}$. Now a *(formal) concept* is a pair (A, B) such that $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. (This is equivalent to A and B being each maximal with $A \times B \subseteq I$.) The set A is called the *extent* and the set B the *intent* of the concept (A, B).

For explaining how to determine the concepts of a formal context, we restrict the context of the library to just four catchwords: Germany*, GDR* (= German Democratic Republic), Federal Republic* (of Germany) and Eastern Germany*. A part of this subcontext is shown in Figure 1. In this context are in total nine formal concepts. One of them has Die intelligente Stadt (The intelligent city), Die Pendlergesellschaft (The commuter society), Jahrbuch Arbeit und Technik 1991 (Yearbook Work and Engineering 1991) and twelve more books (which are not shown in Figure 1) in its extent, and {Germany*, Eastern Germany*} as intent.

Definition: Each formal context $\mathbb{K}$ gives rise to a conceptual hierarchy, called *concept lattice* of $\mathbb{K}$, and denoted by $\underline{\mathfrak{B}}(\mathbb{K})$. The hierarchical subconcept–superconcept–relation on the concepts is formalized by

$$(A, B) \leq (C, D) :\Longleftrightarrow A \subseteq C \quad (\Longleftrightarrow B \supseteq D) \ .$$

The concept lattice of the context of Figure 1 is shown in Figure 2. Each circle stands for a formal concept, and the subconcept-superconcept hierarchy can be read by following ascending paths of straight line segments. The intent of each concept is given by all attributes reachable from that context by ascending paths of straight line segments, and its extent is given by all objects reachable by descending paths of straight line segments. At some concepts we have not listed the names of the objects, but just the number of objects which

	Germany*	GDR*	Federal Republic*	Eastern Germany*
Biographien bedeutender Physiker	×	×		
Das Bildungssystem der DDR	×	×		
Handbuch der kommunalen Verkehrsplanung				×
John von Neumann und Robert Wiener				×
Die intelligente Stadt	×			×
Die Pendlergesellschaft	×		×	×
Abrüstung und Konversion	×	×	×	
Jahrbuch Arbeit und Technik 1991	×	×	×	×
⋮	⋮	⋮	⋮	⋮

FIGURE 1: Part of a formal context about books dealing with 'Germany'

are attached to that concept. The concept mentioned above is the one labeled by the book title Die intelligente Stadt (The intelligent city). In the diagram, one can find above it the catchwords Germany* and Eastern Germany*, hence its intent. Its extent consists of the books listed below, i. e., Die intelligente Stadt plus the nine books listed in the lower right part of the diagram (which additionally have the catchword Federal Republic*) plus the five books at the bottom element (which have all four catchwords).

In line diagrams of concept lattices, one can see implications between the attributes. For instance, the catchword GDR* implies the catchword Germany*. This means that each book having the catchword GDR* also has the catchword Germany*. Federal Republic* implies Germany* as well, and GDR* together with Eastern Germany* imply the two remaining catchwords Federal Republic* and Germany*. I. e, there is no book in the ZIT library having GDR* and Eastern Germany* as catchwords, but not Federal Republic* and Germany*. After a short introduction, users of Conceptual Information Systems usually prefer this visual presentation of implications to long lists of implications written in a linear form, because it shows how the implications are interrelated.

2.2 Conceptual Scaling

Conceptual scaling has been introduced in order to deal with many-valued attributes. Often attributes are not one-valued as in the previous example, but allow a range of values. This is modeled by a *many-valued context*. A many-valued context is roughly equivalent to a relation of a relational database with one field being a primary key. As one-valued contexts are special cases of many-valued contexts, Conceptual scaling can also be applied to one-valued contexts in order to reduce the complexity of the visualization.

In this paper, we only deal with one-valued formal contexts. Readers who are interested in the exact definition of many-valued contexts and the use of conceptual scaling in this more general case are referred to (Ganter, Wille 1999). Applied to one-valued contexts, conceptual scales are used to determine the concept lattice which arises from one vertical 'slice' of a large context:

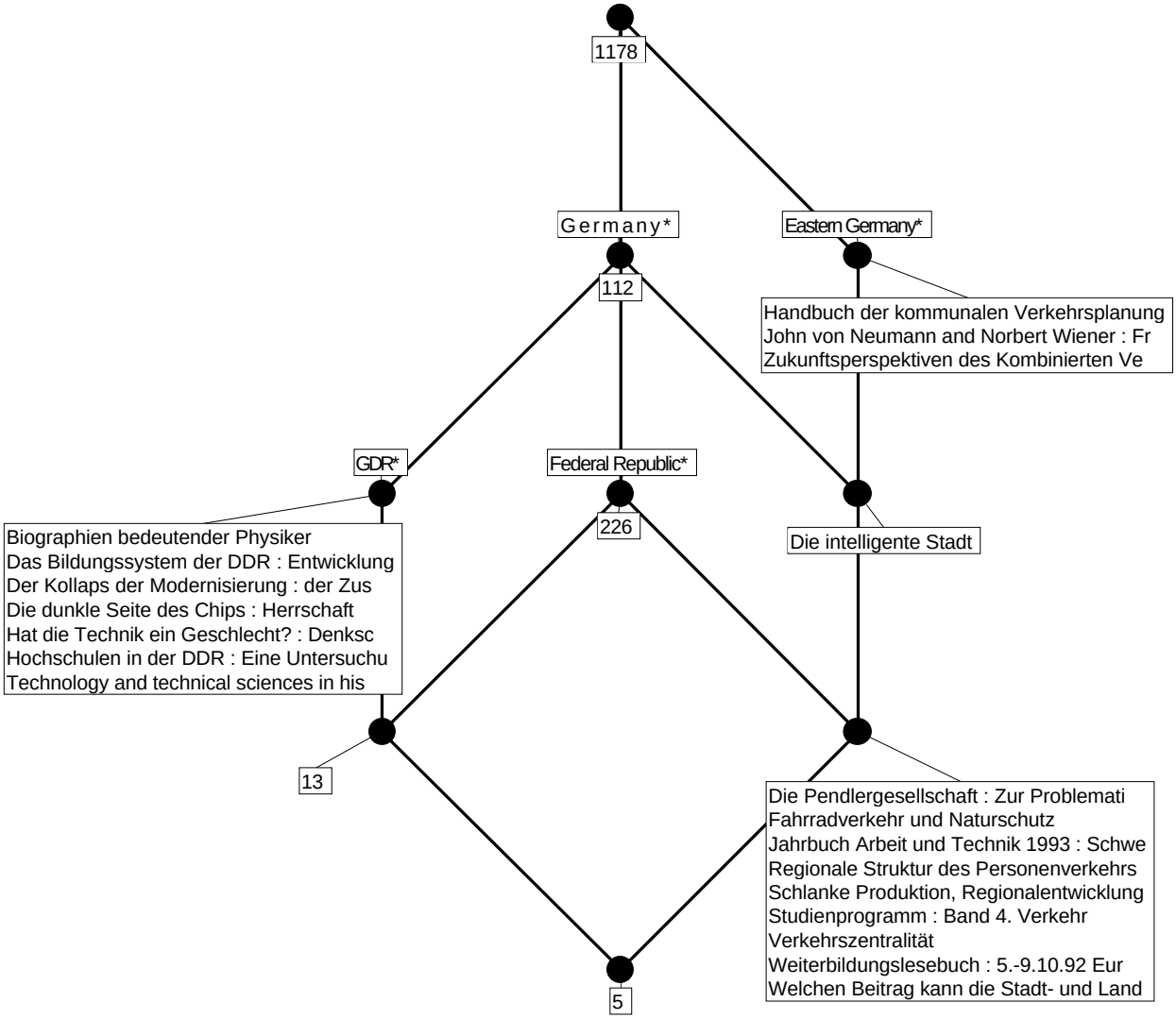


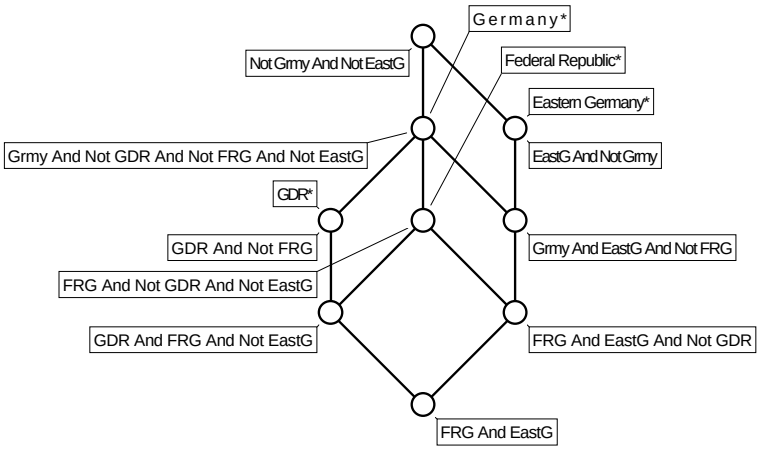
FIGURE 2: Concept lattice of the formal context in Figure 1

Definition: A *conceptual scale* for a subset $B \subseteq M$ of attributes is a (one-valued) formal context $\mathbb{S}_B := (G_B, B, \ni)$ with $G_B \subseteq \mathfrak{P}(B)$. The scale is called *consistent* with respect to $\mathbb{K} := (G, M, I)$ if $\{g\}' \cap B \subseteq G_B$ for each $g \in G$. For a consistent scale $\mathbb{S}_B$, the context $\mathbb{S}_B(\mathbb{K}) := (G, B, I \cap (G \times B))$ is called its *realized scale*.

Conceptual scales are used to group together related attributes. They are determined at the design phase, and the realized scales are derived from them at run-time.

In the ZIT library, there are in total 137 scales. One of them is the scale Germany[2] which is displayed in Figure 3. Its realized scale is the one we already saw in Figure 2. In the conceptual scale, the attributes are strings that are displayed later as attribute names in the realized scale. The objects are parts of WHERE-clauses of SQL queries which determine the

[2]We indicate all original catchwords with a *, while names of scales do not have a *. Thus Germany* is a *catchword* which appears in the *scale* Germany.



	Germany*	GDR*	Federal Republic*	Eastern Germany*
Not Grmy And Not EastG				
Grmy And Not GDR And Not FRG And Not EastG	×			
EastG And Not Grmy				×
GDR And Not FRG	×	×		
FRG And Not GDR And Not EastG	×		×	
Grmy And EastG And Not FRG	×			×
GDR And FRG And Not EastG	×	×	×	
FRG And EastG And Not GDR	×		×	×
FRG And EastG	×	×	×	×

FIGURE 3: The conceptual scale $\mathbb{S}_{\mathsf{Germany}}$

set of objects to be displayed in the realized scale.[3] Dt, OstD, ... are Boolean attributes in the database. The queries are optimized and G_B is chosen such that they fit exactly to the actual books. An update of the database without checking the scale for consistency may hence lead to displaying books at the wrong place.

In general, the difficult task of designing conceptual scales is to fix the set G_B. One can always choose the whole powerset $\mathfrak{P}(B)$, but then the scale may become unnecessarily large. On the other hand, if G_B is chosen too small, the line diagram gets smaller but may be inconsistent with the data. For the design of the scales for the library system, the tool DOKUANA has been used. It makes the scales just large enough to fit the actual data. This way of creating conceptual scales is called *data driven design*, in contrast to *theory driven design*. Theory driven design depends on an expert who decides which subsets of B are impossible according to his theory about the domain. In (Stumme 1999a) it is described how the knowledge acquisition for shifting from data driven to theory driven design can be supported by Attribute Exploration (Ganter 1987), an algorithm of B. Ganter. If there is no theory at all, then theory driven design leads to (large) Boolean scales, i. e., with $G_B = \mathfrak{P}(G)$. Scales generated by using data driven design have to be redesigned after major updates of the database (which is supported again by DOKUANA).

[3] In the current version, a bitmap encoding is used in the database. That makes queries faster, but the queries in the scale less understandable. Therefore Fig. 3 shows the old version.

The retrieval system for the ZIT library is implemented as a Conceptual Information System. *Conceptual Information Systems* consist of a (one- or many-valued) context together with a collection of conceptual scales. The context is implemented as a relational database. The collection of the scales is called *conceptual scheme* (Vogt, Wachter, Wille 1991; Scheich, Skorsky, Vogt, Wachter, Wille 1993). Beside the contexts of the conceptual scales, the conceptual scheme also contains the layout of their line diagrams. The layout has to be provided in advance, since, in general, well readable line diagrams cannot be generated fully automatically.

For Conceptual Information Systems, the management system TOSCANA (Kollewe, Skorsky, Vogt, Wille 1994; Vogt, Wille 1995) has been developed. Based on the paradigm of conceptual landscapes of knowledge (Wille 1997b), TOSCANA supports navigation through the data by using conceptual scales like maps designed for different purposes and in different granularities. It supports the ad hoc cmbination of arbitrry conceptual scales and their visualization by nested line diagrams as well as drill down into a concept of some scale by differentiating it with a second scale. In the next section, we see discuss how such a navigation process is facilitated when we add conceptual scales on a higher level.

3 Hierarchies of Conceptual Scales

In this section, we introduce *higher level scales* which group together scales on a higher level of abstraction. This has three advantages. Firstly it supports the user in retrieving the appropriate scales, and thus simplifies navigation. Secondly, higher level scales provide information about the distribution of the objects over the scales. Such an information cannot easily be discovered with the original scales. Thirdly, this approach allows to model thesauri and ontologies. Browsing an ontology or a thesaurus can then be supported by visualization techniques of Formal Concept Analysis.

We start with introducing a *taxonomy* $(\mathcal{M}, \leq)$ from which we derive *higher level scales.*

Definition: Let $\mathbb{K} := (G, M, I)$ be a formal context. We extend the set M of one-valued attributes to a partially ordered set $(\mathcal{M}, \leq)$ such that the set M contains exactly the minimal elements of $\mathcal{M}$. We say that $n \in \mathcal{M}$ is a lower neighbor $(n \prec m)$ of $m \in \mathcal{M}$, if $n < m$ and there is no $l \in \mathcal{M}$ with $n < l < m$. The set of all lower neighbors of $m \in \mathcal{M}$ is denoted by $m_\prec := \{n \in \mathcal{M} \mid n \prec m\}$.

We define the *extended context* $\widehat{\mathbb{K}} := (G, \mathcal{M}, \mathcal{I})$ by

$$(g, m) \in \mathcal{I} :\Longleftrightarrow \exists n \in M : n \leq m \wedge (g, n) \in I \ .$$

For each $m \in \mathcal{M} \setminus M$, let $G_{m_\prec} \subseteq \mathfrak{P}(m_\prec)$ such that the scale $\mathbb{S}_m := (G_{m_\prec}, m_\prec, \ni)$ is consistent with respect to $\widehat{\mathbb{K}}$. m is then called the *name* of the scale $\mathbb{S}_m$. For $m \in \mathcal{M}$ with $m_\prec \not\subseteq M$, we call $\mathbb{S}_m$ a *higher level scale.*

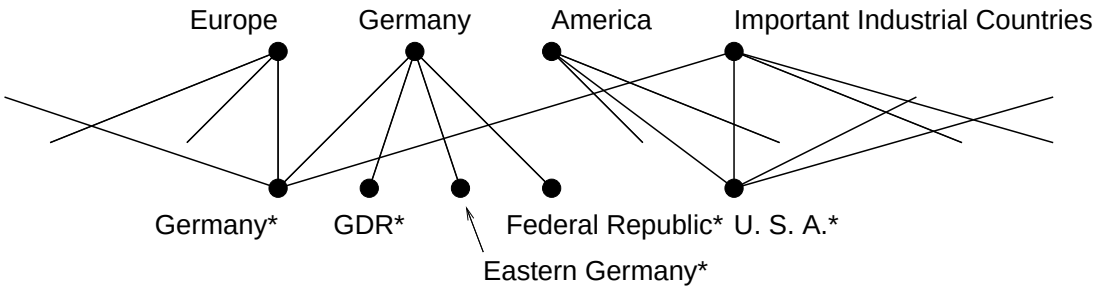


FIGURE 4: Part of the hierarchy of the original library retrieval system

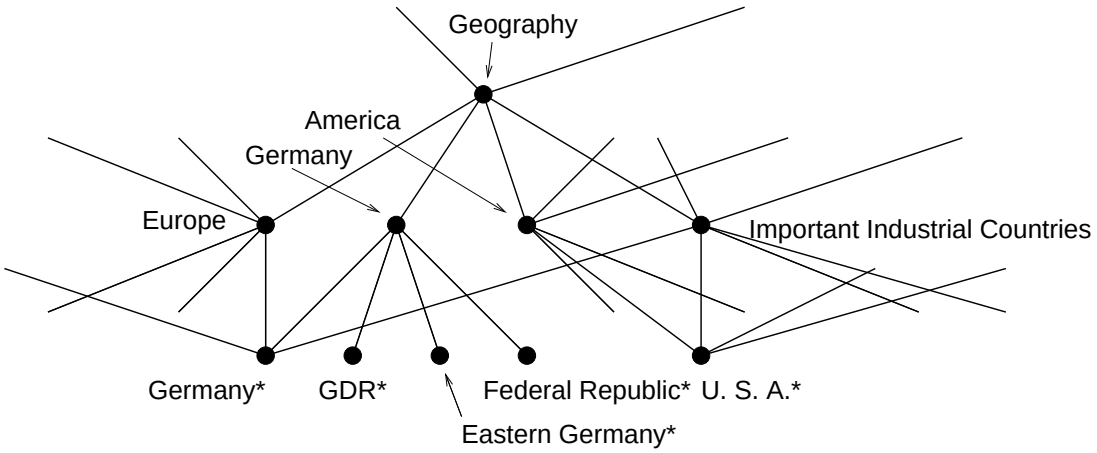


FIGURE 5: Part of the extended hierarchy

Hence, we consider all elements of the taxonomy (except the maximal elements) as attributes, and at the same time (except the minimal elements) as conceptual scales. This means that, for $l \prec m \prec n$, m appears in two different ways in our system. It is an *attribute* of scale $\mathbb{S}_n$ and at the same time the *name* of scale $\mathbb{S}_m$.

Standard Conceptual Information Systems (i. e., without higher level scales) are a special case of this definition: The partially ordered set $(\mathcal{M}, \leq)$ is then of height 2, i. e., each element is either a minimal or a maximal element. In other words, each $m \in \mathcal{M}$ is either an attribute of the original context (G, M, I) or the name of a scale. Figure 4 shows a part of this hierarchy for the original ZIT library system. In the diagram one can see that the partial order on $\mathcal{M}$ is not necessarily a tree. The catchword Germany*, for instance, appears in more than just one scale: beside in scale Germany also in scale Europe and scale Important Industrial Countries.

We can now extend the partial ordering by adding more general attributes (catchwords). Figure 5 shows how the hierarchy is changed when we add a new attribute Geography which we let be the upper neighbor of the scales Europe, Germany, America, and Important Industrial Countries. Data driven design generates then automatically the higher level scale $\mathbb{S}_{\mathsf{Geography}}$ which is shown in Figure 6.

Data driven designed higher level scales can be used for Knowledge Discovery in Conceptual Information Systems: 'Missing' concepts in line diagrams always indicate implications, hence provide (new) knowledge about the domain. In Figure 6, we recognize that, out of 16 possible

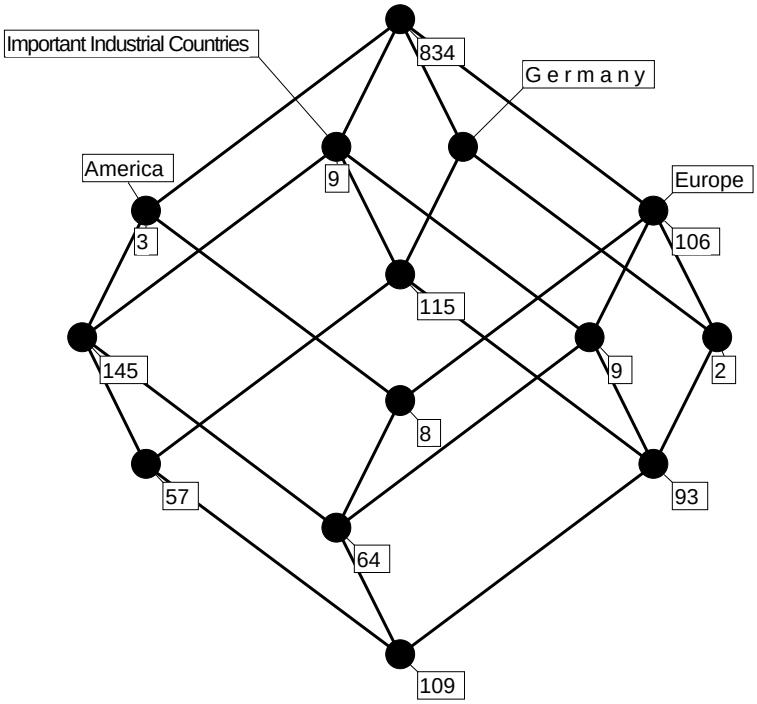


FIGURE 6: The realized higher level scale $\mathbb{S}_{\mathsf{Geography}}$

concepts, the scale consists of only 14 concepts. This leads us to the implication {America, Germany} $\rightarrow$ {Important Industrial Countries}. Hence each book having at least one catchword in each of the scales America and Germany also has at least one of the catchwords in the scale Important Industrial Countries.

Up to now, we have started from an existing Conceptual Information System with one level of scales. We have extended it to the partially ordered set $\mathcal{M}$. Now we consider the dual situation: We start from the subsumption hierarchy of an ontology or thesaurus and generate a Conceptual Information System for supporting navigation through the hierarchy. Hence the situation is as follows: Given is a partially ordered set $(\mathcal{M}, \leq)$ representing the subsumption hierarchy, and a context $(G, \mathcal{M}, \mathcal{I})$ which assigns the related thesaurus terms (the elements in $\mathcal{M}$) to the documents (which are collected in the set G). From these two, conceptual scales have to be derived for the Conceptual Information System.

For a subsumption hierarchy, we can assume that the following *compatibility condition* holds:

$$\forall g \in G,\ m, n \in \mathcal{M}\colon \quad (g, m) \in \mathcal{I},\ m \leq n \ \Rightarrow\ (g, n) \in \mathcal{I} \qquad (\ddagger)$$

(i.e., the assignment of thesaurus terms respects the transitivity of the partial order). In that case, conceptual scales can be derived automatically, one for each non-minimal element of $(\mathcal{M}, \leq)$, just as in the previous definition. Data-driven design of all these scales can be generated automatically by DOKUANA, a preparation tool for Conceptual Information Systems developed by NAVICON GMBH.

In thesauri however, the situation is more complex. Often the compatibility condition ($\ddagger$) does not hold. Even worse, many thesauri merge the subsumption relation (isa-relation, also called

generic hierarchical relationship) with the partitive relationship (part-whole relationship) and the instance relationship (Nikolai 1999). Hence the relation $\leq$ is no longer transitive, and $(\mathcal{M}, \leq)$ cannot longer be considered as partially ordered set. It remains a directed graph in which reasoning by transitivity is no longer meaningful.

However, since the definition of the conceptual scales only makes use of the direct neighbor relation $\prec$, the previous definition of the conceptual scales can still be applied. But then the user of the Conceptual Information System has to be aware that he can get additional objects into his scope when he drills down the hierarchy. If for each link it is known which kind of hierarchical relation it represents (an assumption which does not hold in many thesauri, cf. (Nikolai 1999)) one can, for each non-minimal element in $(\mathcal{M}, \leq)$, provide one scale for each kind of relationship. Depending on his choice, the user will then know if or if not he can expect transitivity.

4 Nested Scaling

Nested line diagrams have become an established way of displaying large concept lattices. They are based on the following theorem:

Theorem 1 (Ganter, Wille 1999) *Let $\mathbb{K} := (G, M, I)$ be a formal context, and $\mathbb{S}_{B_1}$ and $\mathbb{S}_{B_2}$ consistent scales. Then the concept lattice $\underline{\mathfrak{B}}(\mathbb{S}_{B_1}(\mathbb{K})|\mathbb{S}_{B_2}(\mathbb{K}))$ of the* apposition *$\mathbb{S}_{B_1}(\mathbb{K})|\mathbb{S}_{B_2}(\mathbb{K}) := (G, B_1 \cup B_2, I \cap (G \times (B_1 \cup B_2)))$ of the realized scales $\mathbb{S}_{B_1}(\mathbb{K})$ and $\mathbb{S}_{B_2}(\mathbb{K})$ can be embedded as a complete sub-$\bigvee$-semilattice in the direct product $\underline{\mathfrak{B}}(\mathbb{S}_{B_1}) \times \underline{\mathfrak{B}}(\mathbb{S}_{B_2})$.*

The theorem allows us to visualize the combination of two or more scales based only on the layouts of the line diagrams of the individual scales. Hence we need to prepare only a small number of line diagrams in advance, one for each conceptual scale. Whichever scales are chosen for conceptual scaling, the resulting concept lattice can always be visualized by using the precomputed diagrams for the concept lattices $\underline{\mathfrak{B}}(\mathbb{S}_{B_i})$. Figure 7 shows the *nested line diagram* for the scales $\mathbb{S}_{\mathsf{Geography}}$ and $\mathbb{S}_{\mathsf{Germany}}$. Each of the bold lines of the outer diagram has now to be read as a sheaf of nine parallel lines which link corresponding concepts of the inner scales. For instance, the concept at the lower left labeled with 20 is direct upper neighbor of the concept at the bottom labeled with 33.

The concept lattice $\underline{\mathfrak{B}}(\mathbb{S}_{\mathsf{Geography}}(\mathbb{K})|\mathbb{S}_{\mathsf{Germany}}(\mathbb{K}))$ is the one we are really interested in. Its embedding in the direct product is indicated by the bold circles. All other concepts are *not realized by the actual data*. This means that their attribute sets do not form intents with respect to the actual data, since all objects which have these attributes in common share at least one other additional attribute.

From the definition of the scale $\mathbb{S}_{\mathsf{Germany}}$ we know that, in all concepts of the outer scale which do not have the attribute Germany, none of the concepts of the inner scale beside the top element can be realized. Nested scaling is introduced for omitting these 'uninteresting' attributes.

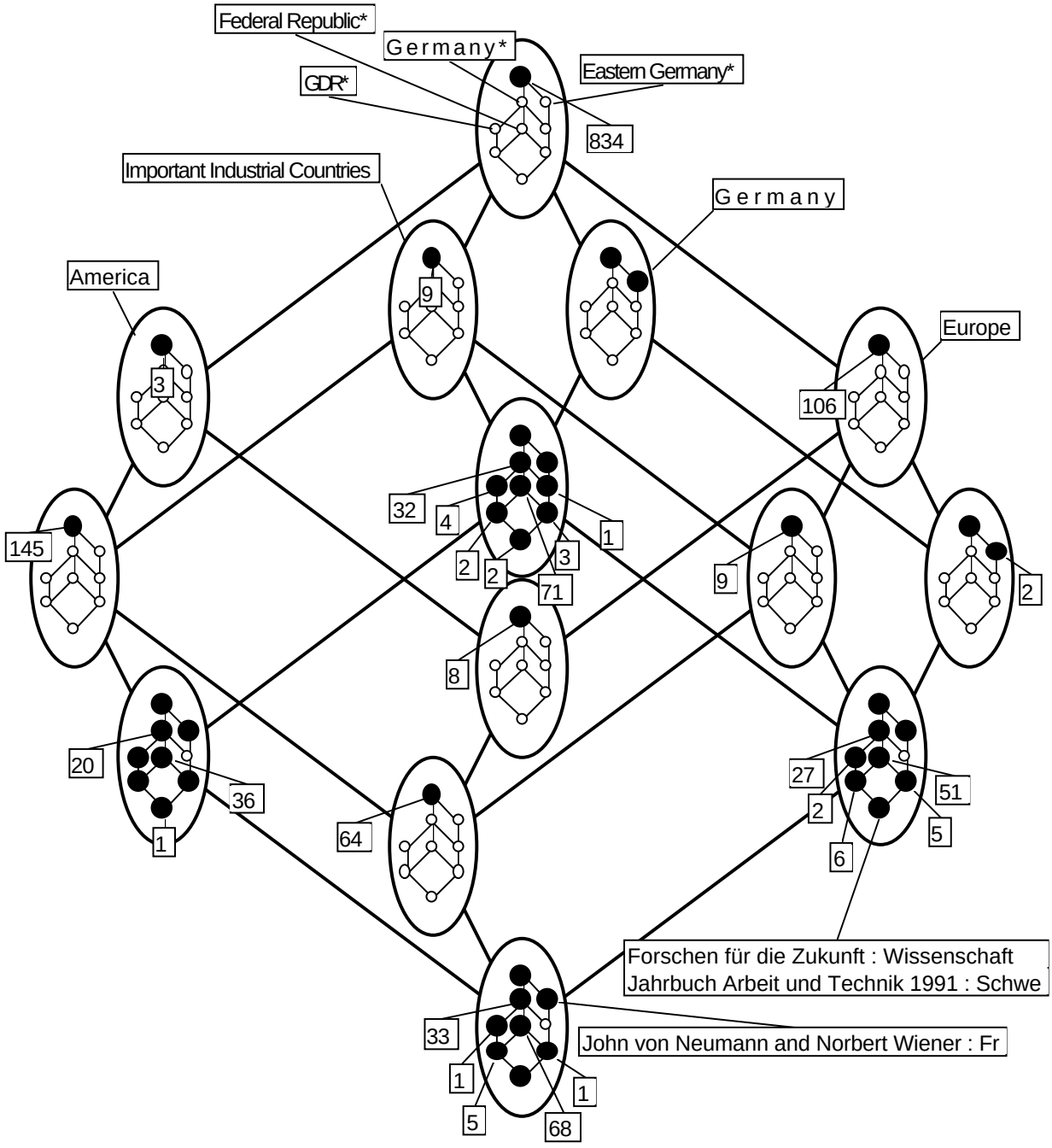


FIGURE 7: The concept lattice $\underline{\mathfrak{B}}(\mathbb{S}_{\mathsf{Geography}}(\mathbb{K})|\mathbb{S}_{\mathsf{Germany}}(\mathbb{K}))$ embedded in the direct product of the concept lattices $\underline{\mathfrak{B}}(\mathbb{S}_{\mathsf{Geography}})$ and $\underline{\mathfrak{B}}(\mathbb{S}_{\mathsf{Germany}})$

Nested scaling is the application of local scaling (Stumme 1996) to scales where one scale subsumes the other in the partially ordered set $(\mathcal{M}, \leq)$. *Local scaling* only 'blows up' those concepts of the outer, higher level scale where the inner scale provides some information, i. e., where there is at least one realized concept in the inner scale unequal to the top concept. From the definition of the attribute Germany, we know that, in the scale $\mathbb{S}_{\mathsf{Geography}}$, the

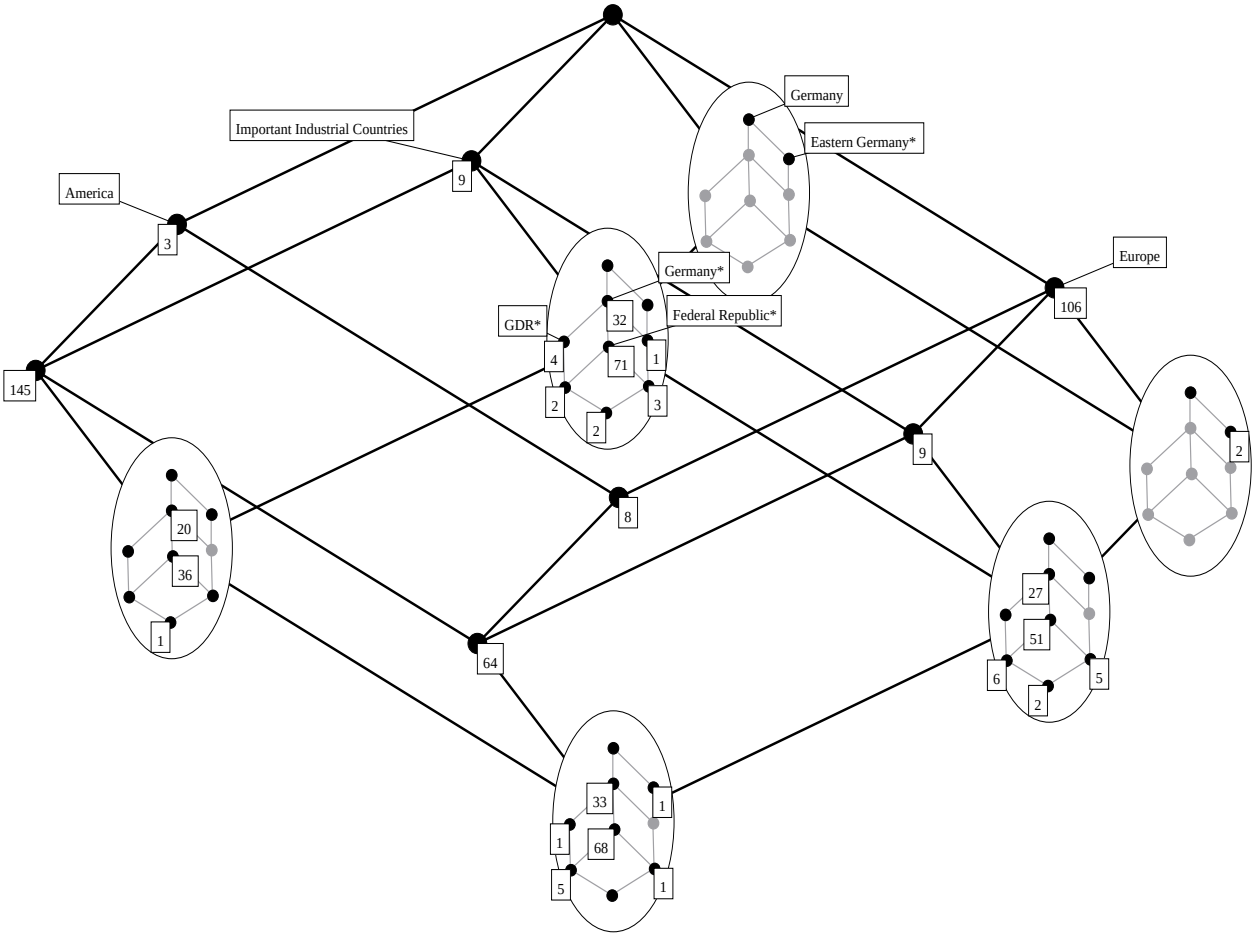


FIGURE 8: Nested scaling of $\mathbb{S}_{\mathsf{Geography}}$ and $\mathbb{S}_{\mathsf{Germany}}$

concepts to be blown up must all be below the concept labeled by Germany. The resulting diagram is shown in Figure 8. The diagram is better readable, as it omits all concepts for which it is clear from the construction of $(\mathcal{M}, \leq)$ that they are not realized. The reason for not omitting *all* non-realized concepts is that they indicate the underlying structure of the scale which is important for the readability of the diagram. Furthermore, in data analysis applications, unrealized concepts are important because they indicate implications (functional dependencies) between attributes. The concepts which are omitted by nested scaling are exactly those which are related to implications resulting from the construction of the taxonomy $(\mathcal{M}, \leq)$.

Nested scaling is based on the following theorem. It assures that the visualization is correct and conforms to the reading conventions introduced in Section 2. The theorem follows directly from the theorems stated and proved in (Stumme 1996). As it uses the compatibility condition (‡) for hiding the redundant information, it can only be applied when $\leq$ is known to be a partial order and when (‡) holds. This is especially the case for the subsumption relation.

Theorem 2 *Let (G, M, I) be a formal context and $(\mathcal{M}, \leq)$ a partially ordered set with $\min \mathcal{M} = M$ verifying the compatibility condition (‡). Let the scales $\mathbb{S}_l$, $l \in \mathcal{M}$, be defined as in Section 3. Let $m, n \in \mathcal{M}$ with $m \prec n$, and let $C := \{\mathfrak{b} \in \underline{\mathfrak{B}}(\mathbb{S}_n) \mid \mathfrak{b} \leq (\{m\}'', \{m\}')\}$.*

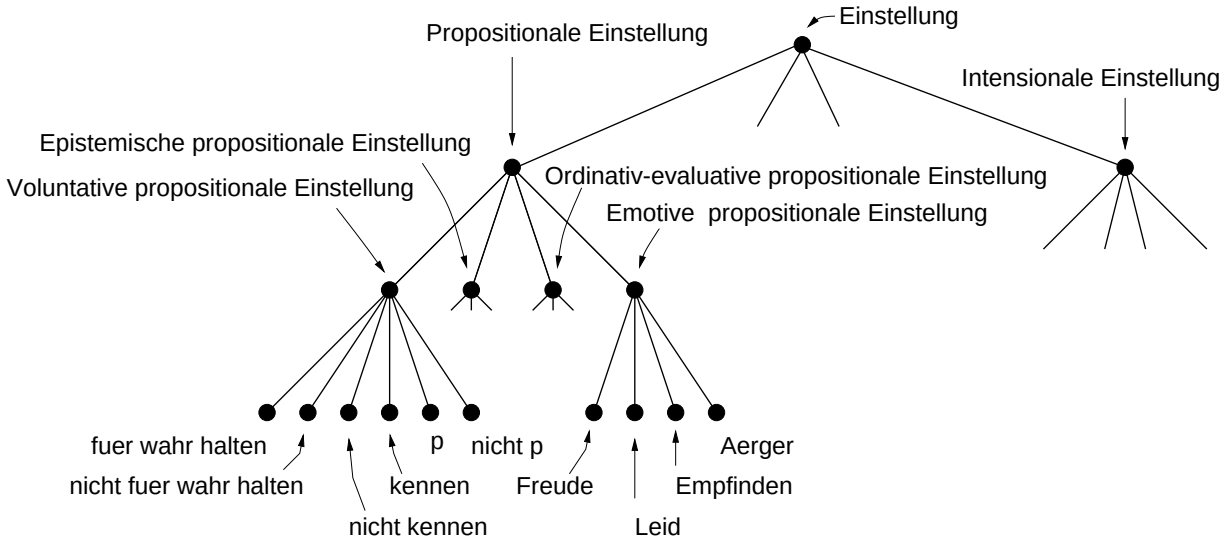


FIGURE 9: Part of the hierarchy for the Conceptual Information System about speech act verbs

Let $V := (\underline{\mathfrak{B}}(\mathbb{S}_n) \setminus C) \cup (C \times \underline{\mathfrak{B}}(\mathbb{S}_m))$ *with the partial order* $\leq$ *given by*

$$\begin{array}{rcl} \mathfrak{b}_n \leq \mathfrak{d}_n & :\Longleftrightarrow & \mathfrak{b}_n \leq_n \mathfrak{d}_n, \\ \mathfrak{b}_n \leq (\mathfrak{c}_n, \mathfrak{c}_m) & :\Longleftrightarrow & \mathfrak{b}_n \leq_n \mathfrak{c}_n \text{and } \mathfrak{c}_m = 1_{\underline{\mathfrak{B}}(\mathbb{S}_m)}, \\ (\mathfrak{c}_n, \mathfrak{c}_m) \leq \mathfrak{b}_n & :\Longleftrightarrow & \mathfrak{c}_n \leq_n \mathfrak{b}_n, \\ (\mathfrak{c}_n, \mathfrak{c}_m) \leq (\mathfrak{e}_n, \mathfrak{e}_m) & :\Longleftrightarrow & \mathfrak{c}_n \leq_n \mathfrak{e}_n \text{and } \mathfrak{c}_m \leq_m \mathfrak{e}_m \end{array}$$

for $\mathfrak{b}_n, \mathfrak{d}_n \in \underline{\mathfrak{B}}(\mathbb{S}_n) \setminus C$, $(\mathfrak{c}_n, \mathfrak{c}_m), (\mathfrak{e}_n, \mathfrak{e}_m) \in C \times \underline{\mathfrak{B}}(\mathbb{S}_m)$.

For $B \subseteq G_m \cup G_n$, *we define* $B_m := B \cap G_m$ *and* $B_n := B \cap G_n$.

Then, for $H \subseteq G$, *the mapping*

$$\varepsilon \colon \underline{\mathfrak{B}}(H, G_m \cup G_n, \mathcal{I} \cap (H \times (G_m \cup G_n))) \to V$$

with $(A, B) \mapsto (B'_n, B_n)$ *if* $m \notin B_n$, *and* $(A, B) \mapsto ((B'_n, B_n), (B'_m, B_m))$ *if* $m \in B_n$ *is a complete* $\bigvee$*–preserving embedding.*

Nested scaling is particularly interesting if the higher level scales are nominal. Then one can nest in each of their concepts another refining scale. The example in Figures 9 and 10 is taken from a Conceptual Information System about German speech act verbs (Großkopf, Harras 1999). The objects in the system are German speech act verbs, e. g., sagen (say), lügen (lie), etc. The attributes are characterizations of speech act verbs, like, e. g., Einstellung (Attitude) which indicates that the speech act described by the verb expresses some attitude.

Part of the partially ordered set $(\mathcal{M}, \leq)$ is shown in Figure 9. Figure 10 shows that it is possible to refine the conceptual scale Propositionale Einstellung (propositional attitude) simultaneously by four different scales and still obtain a diagram with a suitable size.

In this case, we have additionally made use of the fact that in nested scaling one can omit the bottom element of a scale if it has an empty extent. When in at least one scale the bot-

-

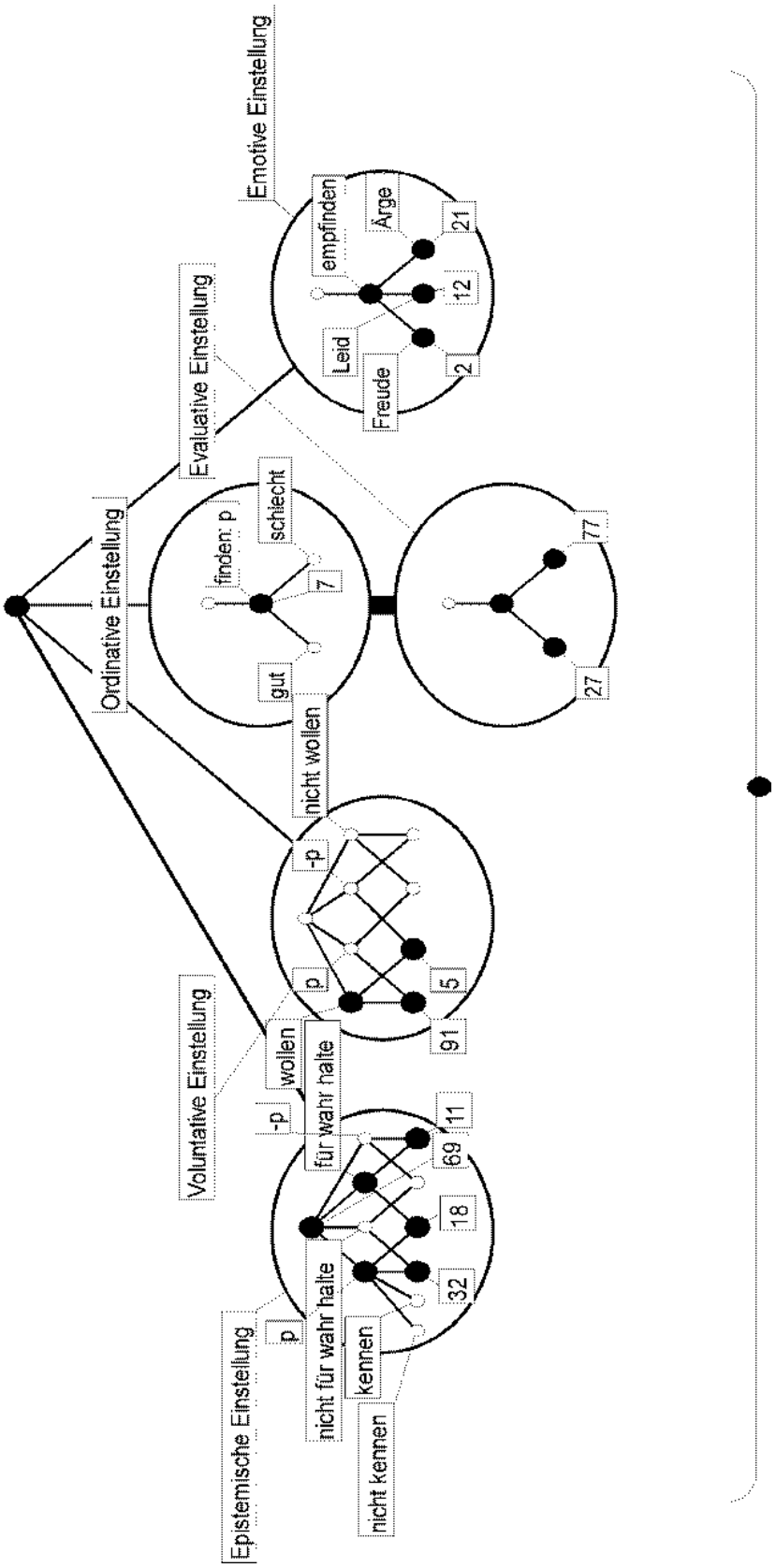


FIGURE 10: The conceptual scale Propositionale Einstellung refined by four different lower level scales

tom element is omitted then an additional global bottom element has to be added.Theorem 2 can be generalized to state this fact, but it makes the description quite technical. For more details, refer to (Stumme 1996). In Figure 10, this generalization allowed us to omit the bottom elements of the outer and all four inner scales and instead adding one new global bottom element which is indicated by the brace in the diagram.

5 Evaluation and Outlook

In this concluding section, we discuss strengths and limitations of our approach. The discussion will lead to topics of further research.

First of all, nested scaling, and more general the use of higher level scales, inherits the strengths (and most of the limitations) from Fromal Concept Analysis. It provides a representation of concepts and conceptual hierarchies which is based on an precise set-theoretical semantics. Furthermore it comes with a well-defined, unambiguous visualization. By structuring the level of conceptual scales, the presented approach enhances the navigation paradigm of Formal Concept Analysis by supporting navigation on this meta-level. This means that it can be applied in larger and more complex domains and still keep the conceptual and visual appeal of Formal Concept Analysis. In this paper, we have focussed mainly on Information Retrieval. In data analysis and knowledge discovery applications, higher level scales allow to discover relationships between scales (and not only between attributes as in the current state of Formal Concept Analysis). In these applications, nested scaling suppresses all information which results from the chosen hierarchy and hence is already known by the analyst.

We have started our considerations with the question how to support navigation in large Conceptual Information Systems. We have seen that nested scaling of conceptual scales with different granularities is a solution to this problem. However, our approach is bound to fail if each of the scales that we start with assigns at least one attribute to each of the objects. Indeed, this is often the case in Conceptual Information Systems based on many-valued contexts. We are currently studying *meta-scales* which have *scales as objects*. The system of meta-scales will provide a 'Meta Conceptual Information System' on top of the original system.

Another limitation of the approach is that it is not able to deal with relations between objects. In order to overcome this problem, there exist major research efforts to combine Formal Concept Analysis with the theory of Conceptual Graphs (Sowa 1984). (Wille 1997a), (Groh, Eklund 1999), and (Prediger, Wille 1999) indicate this development. At the moment, a combination of nested scaling and Conceptual Graphs is tested in an email exploration tool as part of a joint research project with Grifith University, Gold Coast, Australia. The project also deals with the question how the construction of interesting scales can be supported by using Conceptual Graphs. It is expected that results of this research will also be applicable to higher level scales. They could then be used as a knowledge acqusition technique for ontologies and thesauri.

In many applications it is not allowed to change the structure of the database. In particular it is often not possible to add additional attributes. In that case, the higher level scales have to be defined by using formulas of some logic. This is one of many reasons to establish *logical scaling* (Prediger 1997; Prediger, Stumme 1999), where conceptual scales are generated from logical formulas which are assigned to the attributes of the scales. From this assignment one can then derive *concrete scales* which are conceptual scales with the modification that the set of objects consists of logical formulas.

One can recognize that there are different approaches to extend the theory of Formal Concept Analysis in order to bring it together with other knowledge represntation techniques. One of these approaches is the work presented in this paper. Next research steps will include its combination with logical scaling and Conceptual Graphs.

References

A. Arnauld, P. Nicole (1668): La logique ou l'art de penser — contenant, outre les règles communes, plusieurs observations nouvelles, propres à former le jugement. Ch. Saveux, Paris

Deutsches Institut für Normung (1979): DIN 2330; Begriffe und Benennungen: Allgemeine Grundsätze. Beuth, Berlin-Köln

B. Ganter (1987): Algorithmen zur Begriffsanalyse. In: B. Ganter, R. Wille, K. E. Wolff (eds.): *Beiträge zur Begriffsanalyse*. B. I.-Wissenschaftsverlag, Mannheim, Wien, Zürich. 241–254

B. Ganter, R. Wille (1999): *Formal Concept Analysis: Mathematical Foundations*. Springer, Heidelberg (Translation of: Formale Begriffsanalyse: Mathematische Grundlagen. Springer, Heidelberg 1996)

B. Groh, P. Eklund (1999): Algorithms for creating relational power context families from conceptual graphs. In: W. Tepfenhart, W. Cyre (eds.): Conceptual structures: Standards and practices. LNAI **1640**, Springer, Heidelberg

A. Großkopf, G. Harras (1999): Eine TOSCANA-Anwendung für Sprechaktverben des Deutschen. In: G. Stumme and R. Wille (eds.), Begriffliche Wissensverarbeitung: Methoden und Anwendungen. Springer, Berlin-Heidelberg-New York

J. Hereth, G. Stumme, R. Wille, U. Wille (1999): *Conceptual Knowledge Discovery in Data Analysis*. FB4-Preprint, TU Darmstadt

W. Kollewe, M. Skorsky, F. Vogt, R. Wille (1994): TOSCANA — ein Werkzeug zur begrifflichen Analyse und Erkundung von Daten. In: R. Wille, M. Zickwolff (eds.): *Begriff liche Wissensverarbeitung — Grundfragen und Aufgaben*. B. I.–Wissenschaftsverlag, Mannheim

R. Nikolai (1999): *Semi-Automatic Thesaurus Integration: Does it work?*, FZI Karlsruhe, Preprint

N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal (1999): Efficient mining of association rules using closed itemset lattices. *Information systems*

S. Prediger (1997): Logical scaling in Formal Concept Analysis. In: D. Lukose, H. Delugach, M. Keeler, L. Searle, J. F. Sowa (eds.): *Conceptual Structures: Fulfilling Peirce's Dream.* LNAI **1257**, Springer, Berlin, 332–341

S. Prediger, G. Stumme (1999): Theory-driven Logical Scaling – Conceptual Information Systems Meet Description Logic. In: E. Franconi et al (eds.): *Proc. 6th Intl. Workshop Knowledge Representation Meets Databases.* CEUR Workshop Proc. Vol. 21. Also in: P. Lambrix et al (eds.): *Proc. Intl. Workshop on Description Logics.* CEUR Workshop Proc. Vol. 22. (http://SunSite.Informatik.RWTH-Aachen.de/Publications/CEUR-WS/)

S. Prediger, R. Wille (1999): The lattice of concept graphs of a relational scaled context. In: W. Tepfenhart, W. Cyre (eds.): Conceptual structures: Standards and practices. LNAI **1640**, Springer, Heidelberg

T. Rock, R. Wille (1999): Ein TOSCANA-System zur Literatursuche. In: G. Stumme and R. Wille (eds.): *Begriffliche Wissensverarbeitung: Methoden und Anwendungen.* Springer, Berlin-Heidelberg

P. Scheich, M. Skorsky, F. Vogt, C. Wachter, R. Wille (1993): Conceptual data systems. In: O. Opitz, B. Lausen, R. Klar (eds.): *Information and classification.* Springer, Heidelberg, 72–84

J. Sowa (1984): *Conceptual structures: Information Processing in mind and machine.* Addison-Wesley, Reading

G. Stumme (1996): Local Scaling in Conceptual Data Systems. In: P. W. Eklund, G. Ellis, G. Mann (Hrsg.): *Conceptual Structures: Knowledge Representation as Interlingua.* LNAI **1115**, Springer, Heidelberg, 308–320

G. Stumme (1998): On-Line Analytical Processing with Conceptual Information Systems. *Proc. 5th Intl. Conf. on Foundations of Data Organization*, 12.–13. November 1998, 117–126 (to be published by Kluwer)

G. Stumme (1999): Acquiring Expert Knowledge for the Design of Conceptual Scales. In: D. Fensel, R. Studer (eds.): *Knowledge Acquisition, Modeling, and Management.* Proc. 11th European Workshop on Knowledge Acqusition, Modeling, and Management. LNAI **1621**, Springer, Heidelberg, 271–286

G. Stumme (1999): *Conceptual Knowledge Discovery with Frequent Concept Lattices.* FB4-Preprint, TU Darmstadt

G. Stumme, R. Wille, U. Wille (1998): Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods. In: J. M. Żytkow, M. Quafofou (eds.): *Principles of Data Mining and Knowledge Discovery.* Proc. of the 2nd European Symposium on PKDD '98, LNAI **1510**, Springer, Heidelberg, 450–458

F. Vogt, C. Wachter, R. Wille (1991): Data analysis based on a conceptual file. In: H.-H. Bock, P. Ihm (eds.): *Classification, data analysis, and knowledge organization.* Springer, Heidelberg, 131–140

F. Vogt, R. Wille (1995): TOSCANA — A graphical tool for analyzing and exploring data. In: R. Tamassia, I. G. Tollis (eds.): *Graph Drawing '94*, Lecture Notes in Computer Sciences **894**, Springer, Heidelberg, 226–233

H. Wagner (1973): Begriff. In: H. M. Baumgartner, C. Wild (eds.): *Handbuch philosophischer Grundbegriffe.* Kösel Verlag, München, 191–209

R. Wille (1982): Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival (ed.): *Ordered sets.* Reidel, Dordrecht–Boston, 445–470

R. Wille (1997a): Conceptual Graphs and Forma Concept Analysis. In: D. Lukose, H. Delugach, M. Keeler, L. Searle, J. Sowa (eds.): *Conceptual structures: Fulfilling Peirce's Dream.* LNAI **1257**, Springer, Heidelberg, 290–303

R. Wille (1997b): Conceptual landscapes of knowledge: A pragmatic paradigm of knowledge processing. In: *Proceedings of the international conference on knowledge retrieval, use, and storage for efficiency*, Vancouver, Kanada, 11.–13. 8. 1997, 2–13