Conceptual Knowledge Discovery – a Human-Centered Approach

Joachim Hereth[*], Gerd Stumme[†], Rudolf Wille[*], Uta Wille[‡]

Abstract

In this paper we discuss *Conceptual Knowledge Discovery in Databases* (*CKDD*) as it is developing in the field of *Conceptual Knowledge Processing*. Conceptual Knowledge Processing is based on the mathematical theory of *Formal Concept Analysis* which has become a successful theory for data analysis during the last two decades. CKDD aims to support a human-centered process of discovering knowledge from data by visualizing and analyzing the conceptual structure of the data. We dicuss how the management system TOSCANA for conceptual information systems supports CKDD, and illustrate it by two applications in database marketing and flight movement analysis. Finally, we present a new tool for conceptual deviation discovery, CHIANTI.

1 Introduction

Knowledge Discovery in Databases (*KDD*) is aimed at the development of methods, techniques, and tools that support human analysts in the overall process of discovering useful information and knowledge in databases. Many real-world knowledge discovery tasks are both too complex to be accessible by simply applying a single learning or data mining algorithm and too knowledge-intensive to be performed without repeated participation of the domain expert. Therefore, knowledge discovery in databases is considered an interactive and iterative process between a human and a database that may strongly involve background knowledge of the analyzing domain expert. This process-centered view of KDD is the overall theme and contribution of the volume *"Advances in Knowledge Discovery and Data Mining"* [FPSU96].

Conceptual Knowledge Discovery in Databases (*CKDD*) has been developed in the field of *Conceptual Knowledge Processing*. Based on the mathematical theory of *Formal Concept Analysis*, CKDD aims to support a human-

[*]Technische Universität Darmstadt, Fachbereich Mathematik, Schloßgartenstr. 7, D–64289 Darmstadt, Germany, {hereth, wille}@mathematik.tu-darmstadt.de

[†]Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB), Universität Karlsruhe (TH), D–76128 Karlsruhe, Germany, stumme@aifb.uni-karlsruhe.de

[‡]Jelmoli AG, Data Management, Postfach 3020, Ch–8021 Zürich, Switzerland; wille_u@jelmoli.ch

centered process of discovering knowledge from data by visualizing and analyzing the conceptual structure of the data. Implementing the basic methods of Formal Concept Analysis, the management system TOSCANA has been used as a knowledge discovery tool in various research and commercial projects (cf. [Wi97, SW00]). To show its capabilities for CKDD, we address the fundamental requirements for a knowledge discovery support environment stated by R. Brachman in [BS+93, BA96]. They will be discussed using as example an information system implemented for Frankfurt Airport. Then a typical analysis process will be presented by a data warehousing system established for a Swiss department store. This project is also used to discuss a new tool for conceptual deviation analysis, called CHIANTI.

This article consolidates elements of work presented previously in the conference papers [SWW98] and [HSWW00].

Organization of the Paper

After the introduction of some basic notions and ideas of *Formal Concept Analysis* and *conceptual information systems* in the next section, we discuss human-centered knowledge discovery and provide Brachman's list of requirements for knowledge discovery support environments in Section 3. In Section 4, we illustrate CKDD by a flight movement analysis, and discuss along this example how CKDD matches Brachman's requirements. An inter-active knowledge discovery process performed in the context of a data warehouse is presented in Section 5, followed by a presentation of CHIANTI in Section 6. After a discussion of related work in Section 7, the paper concludes with Section 8.

2 Formal Concept Analysis and Conceptual Information Systems

Concepts are necessary for expressing human knowledge. Therefore, the process of discovering knowledge in databases benefits from a comprehensive formalization of concepts which can be activated to communicatively represent knowledge coded in databases. *Formal Concept Analysis* ([Wi82, Wi92a, GW99]) offers such a formalization by mathematizing concepts that are understood as units of thought constituted by their extension and intension.

To allow a mathematical description of extensions and intensions, Formal Concept Analysis always starts with a *formal context* defined as a triple (G, M, I), where G is a set of (*formal*) *objects*, M is a set of (*formal*) *attributes*, and I is a binary relation between G and M (i. e., $I \subseteq G \times M$). $(g, m) \in I$ is read "the object g *has* the attribute m".

In Figure 1, a formal context is given. The object set G contains all gates of Terminal 1 at Frankfurt Airport, and the attribute set M contains certain gate types. The cross table represents the binary relation I between the objects and the attributes.

	Terminal Gate	Bus Gate	Domestic Gate	International Gate
A1, B2, C1	×	×	×	
A2-5, A8, A9, B1, B3-9		×	×	
A10-21, A23	×		×	×
A22		×	×	×
B10	×		×	
B11-16, B30-35, B90, B91, C3, C10, C21-23		×		×
B20, B22-28, B41-48, C4-9	×			×
C2	×	×		×

Figure 1: A formal context about terminal gates at Frankfurt Airport

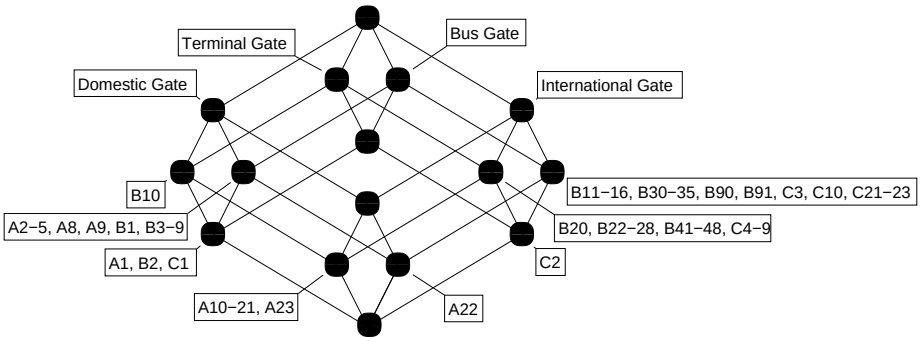

Figure 2: The concept lattice of the context in Figure 1

A *formal concept* of a formal context (G, M, I) is defined as a pair (A, B) with $A \subseteq G$ and $B \subseteq M$ such that (A, B) is maximal with the property $A \times B \subseteq I$; the sets A and B are called the *extent* and the *intent* of the formal concept (A, B). The *subconcept–superconcept relation* is formalized by $(A_1, B_1) \leq (A_2, B_2) :\Longleftrightarrow A_1 \subseteq A_2 \quad (\Longleftrightarrow B_1 \supseteq B_2)$. The set of all concepts of a context (G, M, I) together with the order relation $\leq$ is always a complete lattice, called the *concept lattice* of (G, M, I) and denoted by $\underline{\mathfrak{B}}(G, M, I)$.

In this example, the intents of the formal context are exactly all the subsets of its attribute set; hence its concept lattice is a 16-element Boolean lattice. The concept lattice is visualized in Figure 2 by a *(labeled) line diagram*.

In a line diagram of a concept lattice, the name of an object g is always attached to the node representing the smallest concept with g in its extent (denoted by γg); dually, the name of an attribute m is always attached to the node representing the largest concept with m in its intent (denoted by μm). This labeling allows us to read the context relation from the diagram because

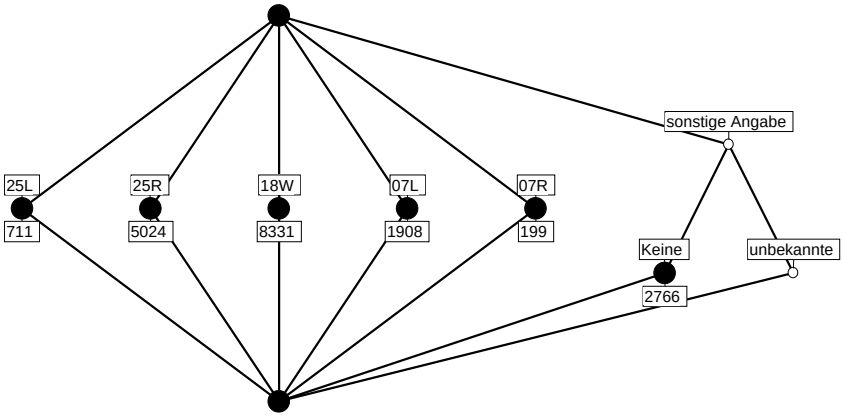

Figure 3: The query structure *Runway*

$(g, m) \in I \Longleftrightarrow \gamma g \leq \mu m$, in words: *the object g has the attribute m if and only if there is an ascending path from the node representing γg to the node representing γm.* The extent and intent of each concept (A, B) can also be recognized because $A = \{g \in G \mid \gamma g \leq (A, B)\}$ and $B = \{m \in M \mid (A, B) \leq \mu m\}$.

For example, the node in the line diagram of Figure 2 labeled "A2–5, ..." represents the concept with the extent {A1, A2, A3, A4, A5, A8, A9, A22, B1, B2, B3, B4, B5, B6, B7, B8, B9, C1} and the intent {Domestic Gate, Bus Gate}. A typical information one can obtain from such a diagram is the fact that gates A10 to A23 provide the flexibility of being used either as Domestic or International Gate, but that with the exception of bus gate A22 they all are terminal gates only.

Graphically represented concept lattices have proven to be extremely useful in discovering and understanding conceptual relationships in given data. Therefore a theory of 'conceptual information systems' has been developed to activate concept lattices as query structures for databases. A *conceptual information system* consists of a (relational) database and a collection of formal contexts, called *conceptual scales*, together with line diagrams of their concept lattices; such systems can be implemented with the management system TOSCANA (see [SS+93, VW95]). For a chosen conceptual scale, TOSCANA presents a line diagram of the corresponding concept lattice indicating all objects stored in the database in their relationships to the attributes of the scale. For instance, as result of a TOSCANA query, Figure 3 shows the concept lattice of the conceptual scale *Runway* indicating as objects 18939 takeoffs at Frankfurt Airport (during one specific month). These objects are classified according to the runways used for take-off, which are taken as attributes of the scale.

3 Human-Centered Knowledge Discovery

According to R. S. Brachman and T. Anand [BA96], much attention and effort in the field of KDD has been focused on the development of data-mining techniques, but only a minor effort has been devoted to the development of tools that support the analyst in the overall discovery task. They see a clear need to emphasize the process-orientation of KDD tasks and argue in favor of a more human-centered approach for a successful development of knowledge-discovery support tools (see also [Ut96], p. 564). All in all, human-centered KDD refers to the constitutive character of human interpretation for the discovery of knowledge, and stresses the complex, interactive process of KDD as being lead by human thought.

Real-world knowledge-discovery applications obviously vary in terms of underlying data, complexity, the amount of human involvement required, and their degree of possible automation of parts of the discovery process. In most applications, however, an indispensable part of the discovery process is that the analyst explores the data and sifts through the raw data to become familiar with it and to get a feel for what the data may cover. Often an explicit specification of what one is looking for only arises during an interactive process of data exploration, analysis, and segmentation. R. S. Brachman et al. introduced the notion of *Data Archaeology* for KDD tasks in which a precise specification of the discovery strategy, the crucial questions, and the basic goals of the task have to be elaborated during such an unpredictable interactive exploration of the data [BS+93]. Data Archaeology can be considered a highly human-centered process of asking, exploring, analyzing, interpreting, and learning in interaction with the underlying database.

A human-centered approach to KDD that supports the overall KDD process should be based on a comprehensive notion of knowledge as a part of human thought rather than on a restrictive formalization as it is used for the evaluation of automated knowledge-discovery or data-mining findings (for example [FPS96], p. 8). The *landscape paradigm* of knowledge underlying *Conceptual Knowledge Processing* as described in [Wi97] provides such a comprehensive and human-centered notion of knowledge. Although there is some similarity with the archaeology metaphor, the landscape paradigm places more emphasizes on the intersubjective character of knowledge. Following Peirce's pragmatic philosophy, knowledge is understood as always being incomplete, formed and continuously assured by human argumentation within an intersubjective community of communication (cf. [Wi97, Wi01]).

Knowledge discovery based on such an understanding of knowledge should support knowledge communication as a part of the KDD process, both with respect to the dialog between user and system and also as a part of human communication and argumentation. This presupposes a high transparency of the discovery process and a representation of its (interim) findings that support human argumentation to establish intersubjectively assured knowledge. Further fundamental requirements for human-centered KDD support tools have been stated by R. S. Brachman and T. Anand (see [BA96], p. 53). In addition to tools

that support the individual phases of the KDD process, they basically demand support for the coupling of the overall process, for exploratory Data Archaeology, and some help in deciding what discovery techniques to choose. Most of the content of these claims is covered by the more explicit and detailed requirements formulated already in [BS+93]. Requirements 1 to 5 of the subsequent list are explicitely stated in [BS+93, p. 164], while the remaining requirements are implicit in [BA96] and [BS+93].

1. The system should represent and present to the user the underlying domain in a natural and appropriate fashion. Objects from the domain should be easily incorporated into queries.

2. The domain representation should be extendible by the addition of new categories formed from queries. These categories (and their representative individuals) must be usable in subsequent queries.

3. It should be easy to form tentative segmentations of data, to investigate the segments, and to re-segment quickly and easily. There should be a powerful repertoire of viewing and analysis methods, and these methods should be applicable to segments.

4. Analysts should be supported in recognizing and abstracting common analysis (segmenting and viewing) patterns. These patterns must be easy to apply and modify.

5. There should be facilities for monitoring changes in classes or categories over time.

6. The system should increase the transparency of the KDD process, and document its different stages.

7. Analysis tools should take advantage of explicitly represented background knowledge of domain experts, but should also activate the implicit knowledge of experts.

8. The system should allow highly flexible processes of knowledge discovery respecting the open and procedural nature of productive human thinking. This means in particular the support of intersubjective communication and argumentation.

Conceptual Knowledge Discovery provides means to satisfy these requirements, as we will argue in the next section along a system for analyzing flight movements.

4 Conceptual Knowledge Discovery in Databases

Conceptual information systems using the management system TOSCANA can be considered as *knowledge discovery support environments* that promote human-

centered discovery processes and representations of their findings. As illustrating example, we use a conceptual information system established by U. Kaufmann [Ka96] for exploring data of the information system INFO-80 of the 'Flughafen Frankfurt Main AG'. This information system supports planning, realization, and control of business transactions related to flight movements at Frankfurt Airport.

In a conceptual information system, the *objects* of the underlying domain are stored in a relational database so that they can be activated by SQL-statements for establishing *conceptual scales*. The objects are presented to the user in line diagrams of the concept lattices of conceptual scales as already demonstrated in Figure 3. In general, the objects are first listed in quantities describing the size of the extents of the represented concepts. For instance, in Figure 3 the number 8331 attached to the node labelled "18W" informs that there were 8331 takeoffs on Runway 18 West. If one wants more specific information about objects, one can obtain the object names for an extent by clicking on the attached number, or even more information about a single object by clicking on its name. Of course, larger numbers as in Figure 3 first have to be differentiated by further scales before considering single objects. But the distribution of the quantities may be already informative: in our example the number 8331 indicates that more than 40% of all takeoffs are from Runway 18 West; this high proportion is interesting because there was a strong controversy about the construction of this runway regarding noise pollution.

Our discussion shows that the first requirement of appropriate object representations is fulfilled in conceptual information systems. The second requirement of extendibility of categorical structures is already realized by the large flexibility in forming conceptual scales; even during the discovery process new insights may give rise to further conceptual scales. The third requirement of meaningful data segmentations is also fulfilled because the conceptual scales and their combinations yield an almost unlimited multitude of conceptual segmentations and with that a powerful repertoire of different views for exploring and analyzing data. This flexible repertoire supports analysts in recognizing and abstracting the interpretable patterns for which the fourth requirement asks.

Let us demonstrate some of the discussed abilities of conceptual information systems by continuing the investigation of Runway 18 West. In Figure 3 we zoom into the concept node labeled '18W' with the conceptual scale *Wingspan Code and Position Size*. Then we can study the sizes of the 8331 planes that took off from 18 West in the line diagram shown in Figure 4. Position sizes indicate, in increasing order, the size of the docking position of the plane prior to takeoff, while the wingspan codes decrease by increasing wingspan (Code 0 stands for 'helicopter'). The cardinality of the extents are now described by percentages rather than by absolute cardinalities. From the diagram we obtain that most of the machines that took off from Runway 18 West had position size 4 or 5, hence are rather large. This might lead to the hypothesis that those machines contribute over-proportionally to the noise pollution. We test this hypothesis by zooming into the two concept nodes labeled *Posgr 4* and *Posgr 5* with the scale *Noise Class of the Plane by ICAO-Annex 16*. The two line

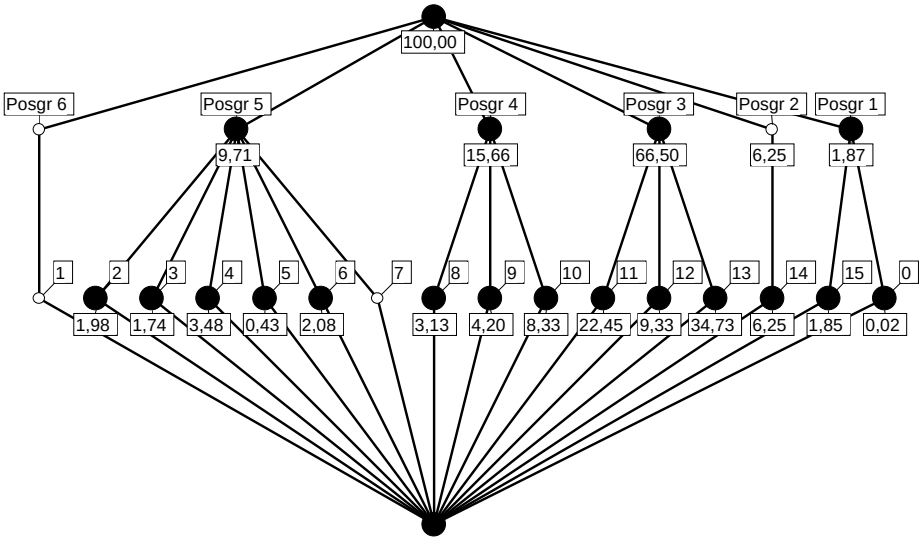

Figure 4: The query structure *Wingspan Code and Position Size*

diagrams in Figure 5 indicate that for both position sizes more than 95% of the planes that took off from Runway 18 West are quite silent (as classified by Chapter 3 of the Chicago Treaty). Hence the hypothesis is not supported by the data. Summarizing our investigation, we can conclude that the planes taking off from Runway 18 West are over-proportionally large, but that more than 95% of them are categorized as silent.

Conceptual information systems offer also facilities for fulfilling the other requirements listed. Changes in classes or categories over time may be documented in specific scales so that they can be easily monitored. Processes of knowledge discovery are developing in a network of conceptual scales that yields increasing transparency of the process and can be used for documenting the different phases of the process. K. Mackensen and U. Wille have even shown in [MW97] how such processes may be understood as procedures of qualitative theory building (see also [SWW01]). Background knowledge of domain experts enters the process of knowledge discovery via conceptual scales in which experts have explicitly coded formal aspects of their knowledge in structurally representing a certain theme, thereby also making connections to their implicit knowledge. Overall, conceptual information systems offer conceptually shaped landscapes of structurally coded knowledge allowing diverse excursions, during which a learning process yields an increasingly better understanding of what to collect and where to continue (cf. [Wi97]). The graphical representation of interesting parts of the landscape, in particular, supports intersubjective communication and argumentation.

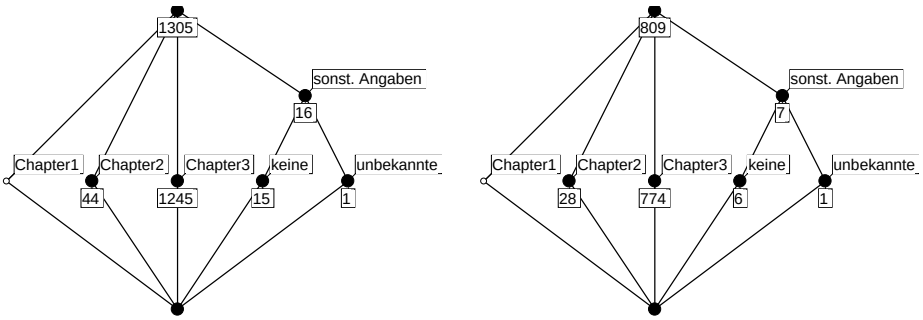

Figure 5: The query structure *Noise Class of the Plane by ICAO-Annex 16* with respect to position sizes P4 (left) and P5

5 Conceptual Knowledge Discovery and Data Warehousing

In this section, we present a second application of Conceptual Knowledge Discovery. In a joint research project of Darmstadt University of Technology and a Swiss department store, the integration of procedures of Formal Concept Analysis and Data Warehousing has been investigated. The resulting conceptual information system is in daily use in the database marketing of the department store. The implementation of this system led to new research topics, which resulted in a new analysis tool, CHIANTI, which will be presented in the following section.

The department store and its business partners provide a special credit card to their customers. If customers purchase products at the department store or one of its partners, they collect bonus points, which allow for reductions on special occasions. The credit card transactions are stored in a central database and allow the Database Marketing department to analyze the behaviour of their customers and to improve the personalization of marketing mailings, e. g., offering specific products the individual customers are more likely to be interested in. The database used at time of implementation contained data from more than nine million transactions and more than forty thousand customers. About 160 conceptual scales have been established within the conceptual information system for analyzing the data.

The line diagram on the left side in Figure 6 shows the cross-selling behavior between travel accessories, perfumery, and women's accessories. In the concept lattice, the objects are all customers with purchases in at least one of the three departments, and the attributes are 'purchased in the department for travel accessories', 'purchased in the perfumery', and 'purchased in the department for women's accessories'. The diagram shows that there are 1075 customers who bought in the travel accessories only, 8182 in the perfumery only, and 3964 in the women's accessories department only, and in non of the other two departments. Furthermore, there are 967 customers who bought travel accessories and some-

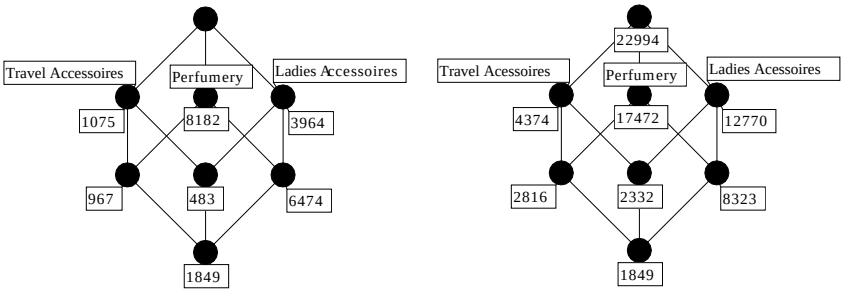

Figure 6: Line diagrams showing the cross-selling between travel accessories, perfumery, and women's accessories

thing from the perfumery but no women's accessories, and 1849 customers who where active in all three departments. Here we are already pointed to a marketing problem, namely, why are there 8182 customers buying perfumery goods but no accessories although both departments are right next to each other?

Interesting for the mailing select are the 6474 + 1849 = 8323 customers, because, in general, it is more promising to make active customers to better customers. The diagram on the right hand side in Figure 6 represents the same facts as the left one, but the amounts of customers are summed up from the bottom (i. e., to each concept is attached the cardinality of its extent). To study the group of perfume and women's accessories buyers in further detail, TOSCANA allows us to 'zoom into' the node in the right diagram representing the 8323 customers that bought perfumery goods and accessories and eventually travel accessories. Figure 7 shows a segmentation of those customers with respect to their previous activity in the ladies wear department (formal, business, and casual wear). In this diagram, the amounts of customers are again summed up from the bottom; for instance, there are 1777 customers in that group of 8323 who spent more than 400 SFr (Swiss Francs) for women's clothing (639 spent more than 1000 SFr and 1138 between 400 and 1000 SFr). Since there is no need to attract already good customers of the ladies wear department by incentives advertised in direct mailings, one has to look for high potential customers with low or no activity in the ladies wear department.

The power of TOSCANA systems lies in the possibility to refine a presented concept lattice by another one so that one obtains either a nested line diagram of a combination of both lattices, or a line diagram of the second lattice refining a single specific concept of the first; the latter may be used for zooming iteratively, which potentially allows us to navigate through the entire database. Figure 8 shows how different scales can be combined and represented in a nested line diagram to visualize dependencies between different 'dimensions'. It shows the activity of the 6546 customers with 400 or less SFr expenses for women's clothing. The *nested line diagram* presents two aspects of the activity of those 6546 customers: the line diagram representing the number of departments in

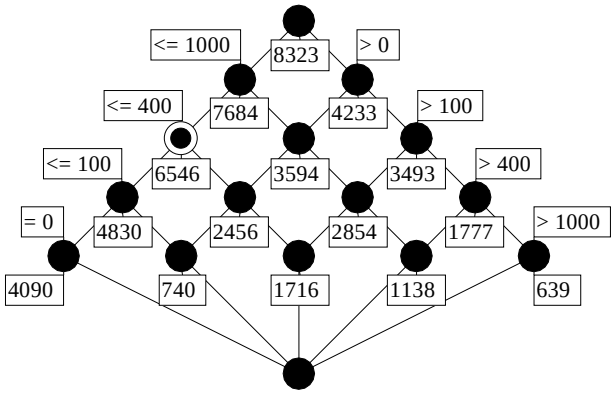

Figure 7: Line diagram showing the amount customers of perfume and women's accessories spent for women's clothing

which customers shopped (outer part) is combined with the cross-selling line diagram between houseware and interieur (inner part). The nodes of the first line diagram have been enlarged so that a copy of the second line diagram could be drawn in each enlarged node. The nested line diagram can be read like an non-nested one if we replace the lines between the large circles by parallel lines between the corresponding nodes of the inner diagrams. For instance, we can read from the diagram that there are 4720 customers who shopped in five or more but in at most twelve departments of the store, and that 2001 of those bought houseware as well as interieur, which seems to be a good target group for direct mailing.

For examining cross-selling, concepts having many attributes – and hence only relatively few objects – are of special importance. In those cases, one needs the whole line diagram for the analysis of how well cross-selling works. But there are many applications where one is not interested in concepts which differentiate the population too much – at least not for a first overview. In that situation one benefits from *iceberg concept lattices*: by fixing a support threshold *minsupp*, one can prune all infrequent concepts of the conceptual scale (cf. [St99b, ST+01b]). Then only the frequent concepts are displayed. For instance, if we want to have a first glance at the distribution of the age of the customers, then the conceptual scale 'Age' may be too detailed. By fixing $minsupp := 25\%$, we prune 18 of the 30 concepts of the scale 'Year of Birth'. The remainder is shown in Figure 9. One can for instance see that for more than half of the customers the year of birth is unknown, and that 46.04% of all customers (paying with credit card) are born before 1973. Hence there are very few customers (of which the year of birth is known) who are younger than 25 years and having paid with credit card.

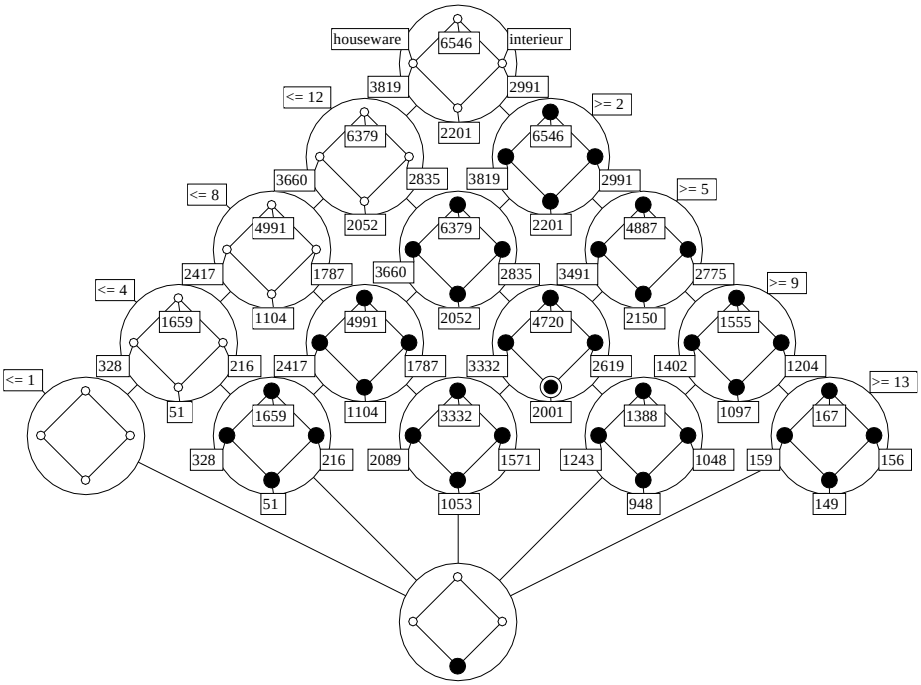

Figure 8: Nested line diagram combining numbers of visited departments with the cross-selling between houseware and interieur

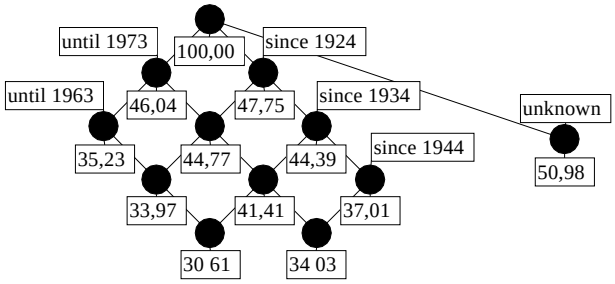

Figure 9: Conceptual Scale 'Year of Birth' restricted to frequent concepts with $minsupp = 25\%$

6 Conceptual Deviation Discovery with Chianti

Up to now, the search and analysis was guided by already present knowledge. On other occasions, the analyst is looking for new ideas, is trying to gain new knowledge based on the collected data. To help the human expert in those

more complex cases, a tool should automatically indicate conceptual scales (or combinations of conceptual scales) which are expected to provide interesting information, combined with the freedom of navigating around. The problem we tackle here is to gain more knowledge about a given sub-population. We do not necessarily require an *explicit* description of the behavior. For instance, in the context of Database Marketing, the user might want to learn (in its literal meaning) more about how customers spending much money differ from the other customers in their buying behavior. For this purpose, we have developed the tool CHIANTI, based on [St98] and [St99a]. While being an external tool at first, it is currently becoming an integral feature of the ToscanaJ Project [To01].

CHIANTI takes as input two sub-populations which are defined by SQL queries. It returns a list of conceptual scales ranked according to the difference in the distribution of the two sub-populations. We explain the use of CHIANTI by an example.

In the following example, we have divided the customer population in two parts: those customers who spent more than 1000 SFr and those who spent less. CHIANTI then compares the distributions of the two sub-populations in all scales of the conceptual information system and returns a ranking of all scales. In the ranking, those scales appear at the top where the distributions differ the most. The current implementation provides two measures for the similarity of the distributions: The χ^2-measure (hence the name of the program) and the maximum norm. While the first measure takes the differences in all concepts into account (the larger ones over proportionally), the second measure only regards the concept with the largest difference. This approach is useful when an easy interpretation of the ranking is desired. CHIANTI does not only work on contingents (i. e., the set of all objects belonging to the extent of the concept but not to the extent of any sub-concept), but also on extents, and is thus providing more possibilities for defining deviation measures on the conceptual scales.

Figure 10 shows the ranking of all scales related to cross-selling for the two sub-populations mentioned above according to the χ^2-measure. The scale at the top is the scale 'Cross-selling houseware/interieur' (which we have already seen as inner scale in Figure 8). This means that among all cross-selling scales, this scale differentiates the two groups the most. This scale also appears as topmost scale in the ranking according to the maximum norm.

By combining the topmost scales of the ranking with the scale 'Money spent $\leq$ / $>$ 1000 SFr', we can then analyze the distribution of the two groups in more detail. The combination of this scale together with the scale 'Cross-selling houseware/interieur' is shown in Figure 11. In the diagram, we have set the top element of each inner scale to 100% in order to facilitate comparison. We see that the customers spending much money buy over-proportionally in the Houseware (265% more often) and Interieur (322% more often) departments. Furthermore, for this customer group, the cross-selling between both departments is much higher than for the rest: The proportion from 36.98% to 60.58% and 50.53% is much higher than the one from 5.56% to 22.82% and 15.67%.

We emphasize that — unlike many statistical techniques — the final result is not the ranking of the scales. It provides just a suggestion to the analyst to con-

Scale	Val.
Xselling houseware/interieur	0,0684
Xselling food/wine	0,0570
Xselling travel accessoires/perfumery/ladies accessoir	0,0325
Xselling perfumery/houseware/food	0,0324
Xselling perfumery/ladies fashion	0,0305
Xselling wine/mens fashion/perfumery	0,0275
Xselling ladies fashion/mens fashion/sports	0,0229
Xselling sports/children/travel accessoires	0,0160
Xselling mens clothing (incl_ underwear)	0,0160
Xselling ladies wear	0,0123
Xselling mens city	0,0009

Figure 10: Ranking of conceptual scales related to cross-selling

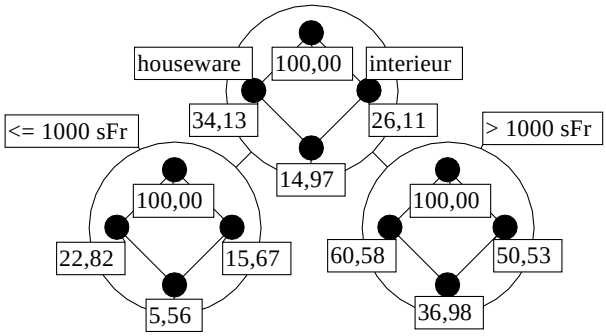

Figure 11: Customers of the houseware department differentiated by the amount of money spent

sider a certain combination of scales for analyzing the situation in more detail. It is *not* justified to deduce from the ranking alone that the buying behavior in the houseware department determines the value of the customer. In particular, it is not possible to decide automatically if a prominent position in the ranking indicates a cause for or a consequence of the different distribution. This becomes clear when one studies the ranking of all scales. The topmost scales are then all scales related to the amount of money spent. In those scales one will hardly discover new insights. The next scale is then 'Active Time (in days)'. This scale does not provide an interesting insight either, since it is intuitively clear that a typical customer usually pays less than 1000 SFr in a single transaction; hence for spending more money, he has to visit the department store more than once. The next scale then is the scale 'Cross-selling houseware/interieur'.

The insight that the scale about the active time is not useful for this kind of analysis can only be gained by refering to the implicit background knowledge

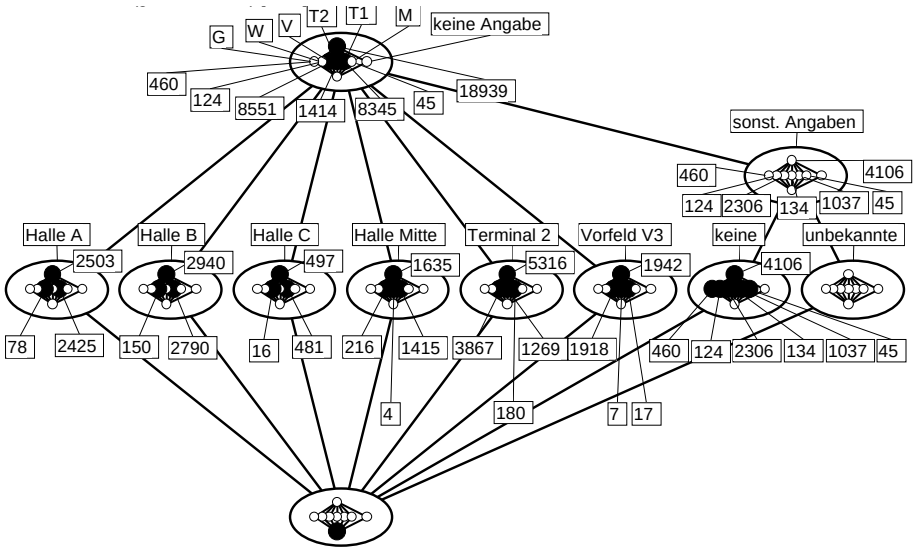

Figure 12: Nested line diagram of the scales *Position of baggage conveyor* and *Position of aircraft*

of the domain expert. A repository which stores such information explicitly cannot overcome the general problem. There is an almost boundless number of possible combinations of conceptual scales in a conceptual information system which cannot be overviewed in advance. However it is promising for further research to consider such a repository which 'learns' from the behavior of the analyst which combinations are of interest and which are not.

In this context, a more detailed study of different deviation measures has been started. We discuss it along the flight movement analysis example. Figure 12 shows the nested line diagram of the scales 'Position of baggage conveyor' and 'Position of the aircraft'. By a short investigation, some problematic cases can be found. For instance, there are 180 aircraft at Terminal 1 with baggage conveyors assigned in Terminal 2. The four aircraft that docked at Terminal 2, while their assigned baggage conveyors are in Terminal 1, as well as the 7+17 cases in which the aircraft docks at one of the two terminals, while the assigned conveyor is on the apron, are also considered problematic.

We applied different deviation measures for analyzing in which conceptual scales (out of a representative subset of scales) the deviation of the subpopulation of the 180 aircraft compared to the total population is the largest (cf. [KSK01]). To formally present those measures, we use the following notations: Given the multi-valued context (G, M, W, I), the two sub-populations used for the comparision are described by the subsets $G_1, G_2 \subseteq G$ and induce the sub-contexts $\mathbb{K}_i := (G_i, M, W, I \cap (G_i \times M \times W))$ with $i \in \{1, 2\}$.

Let $\mathcal{C}_\mathbb{S}$ be the set of all concepts of a conceptual scale $\mathbb{S}$. For every $\mathfrak{c} \in \mathcal{C}_\mathbb{S}$, $\mathfrak{c}_i$ shall be the corresponding concept of the scale $\mathbb{S}$ realized over $\mathbb{K}_i$. For a concept

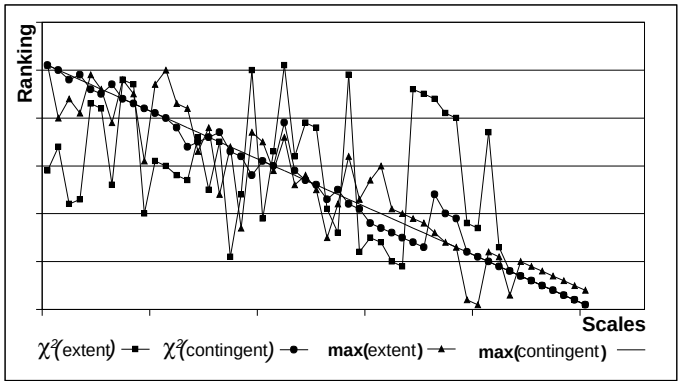

Figure 13: A comparison of four possible deviation measures

$\mathfrak{c}$, the functions extent and contingent return the number of objects in its extent and contingent respectively.

Using these notations, we can now describe the four deviation measures:

$$\begin{aligned}
\chi^2(\text{extent})(\mathbb{S}) &:= \frac{1}{|\mathcal{C}_\mathbb{S}|} \sum_{\mathfrak{c}\in\mathcal{C}_\mathbb{S}} \left(\frac{\text{extent}(\mathfrak{c}_1)}{|G_1|} - \frac{\text{extent}(\mathfrak{c}_2)}{|G_2|}\right)^2 \\
\chi^2(\text{contingent})(\mathbb{S}) &:= \frac{1}{|\mathcal{C}_\mathbb{S}|} \sum_{\mathfrak{c}\in\mathcal{C}_\mathbb{S}} \left(\frac{\text{contingent}(\mathfrak{c}_1)}{|G_1|} - \frac{\text{contingent}(\mathfrak{c}_2)}{|G_2|}\right)^2 \\
\max(\text{extent})(\mathbb{S}) &:= \frac{1}{|\mathcal{C}_\mathbb{S}|} \max_{\mathfrak{c}\in\mathcal{C}_\mathbb{S}} \left|\frac{\text{extent}(\mathfrak{c}_1)}{|G_1|} - \frac{\text{extent}(\mathfrak{c}_2)}{|G_2|}\right| \\
\max(\text{contingent})(\mathbb{S}) &:= \frac{1}{|\mathcal{C}_\mathbb{S}|} \max_{\mathfrak{c}\in\mathcal{C}_\mathbb{S}} \left|\frac{\text{contingent}(\mathfrak{c}_1)}{|G_1|} - \frac{\text{contingent}(\mathfrak{c}_2)}{|G_2|}\right|
\end{aligned}$$

Our analysis indicates that the deviation measures provide different rankings. According to two of them, the conceptual scale 'Airline' was ranked highest, while according to the two other, the scale 'Partner of the airlines' was ranked highest. Figure 13 shows the differences between the different deviation measures. Along the x-axis the scales are listed (ordered by their ranking with respect to the maximum-measure on the contingents). The y-axis provides the rank given to the scale by each of the four measures. If all of them would measure the same information, they would be all close to each other, i. e., all would be very close to the diagonal. As can be seen in the figure, this is not the case. Actually, there are conceptual scales that are esteemed important following the maximum-on-contingents measure but unimportant by the χ^2-on-extents measure – and vice-versa. This indicates the need for more experience in real-world applications based on judgments of domain experts. This would help to define ranking functions adapted to the domain of analysis.

7 Related Work

In this paper, we have presented an overview over some applications of Conceptual Knowledge Discovery. Now we briefly discuss some other ways (taken by ourselves as well as by other authors) of using Formal Concept Analysis to support the knowledge discovery process. This collection is by no means complete, but shall provide an insight in the variety of applications of Formal Concept Analysis for Knowledge Discovery.

A prominent research area in KDD are association rules. Formal Concept Analysis is used in two ways for computing association rules. Firstly, the lattice structure is exploited for improving the fiiciency of the algorithms. The first algorithm of this kind was Close [PBTL99]. [PHM00, BTP+00] present two more algorithms for efficient computation. Secondly, Formal Concept Analysis is used to reduce the number of rules presented to the user. The approaches described in [BPT+00, Za00, ST+01a] allow for rule reduction without loss of information. Another unsupervised KDD technique where Formal Concept Analysis was applied is conceptual clustering: [CR93, StW93, MG95, ST+01b] describe different approaches. B. Ganter and S. Kuznetsov [GK00, GK01] and E. M. Nguifo and P. Njiwoua [NN01] have applied Formal Concept Analysis for supervised learning in

In [CES00], an application of CKDD in email collections is discussed.

Conceptual information systems can also be understood as On-Line Analytical Processing (OLAP) tools. Roughly, the conceptual scales can be regarded as dimensions of a multi-dimensional data cube. Zooming in one of the concepts of a scale as described in the previous sections corresponds to 'slicing' the data cube. 'Rotating' and 'drill-down' are also supported [St00]. The combination of the conceptual structures of Formal Concept Analysis and the numerical capabilities presented by today's databases is incorporated in the ToscanaJ project (cf. [To01]) and will be the base for future applications in this domain.

While the ToscanaJ project is preparing for very large databases, there is also a need for small and handy analysis tools. For small and ad-hoc qualitative data analysis, the software tool CERNATO from Navicon GmbH[1] has been developed. CERNATO is a highly integrated system for data management and interactive analysis. It includes facilities for building conceptual information systems. The line diagrams are produced on the fly and are animated.

Further research in Conceptual Knowledge Processing aims at developing *conceptual knowledge systems* by extending the functionalities of conceptual information systems, especially by logic-based components. As Formal Concept Analysis and Description Logics are closely related and have similar purposes (see, e. g., [BS+93, SSW96]), first steps in integrating both theories have been made ([Ba95, Be97, Pr97, St96]). For hybrid knowledge processing, an extension of conceptual information systems is foreseen by incorporating statistical and computational components [SWo97]. This indicates a promising development in terms of extending conceptual information systems toward a wider range of

[1] http://www.navicon.de

CKDD applications.

A second line of extending Formal Concept Analysis to Conceptual Knowledge Processing beside CKDD is *Contextual Logic* [Wi96, Pr00]. It is based on the elementary doctrines of concepts, judgments, and conclusions as discussed in traditional philosophical logic. In this framework, Formal Concept Analysis is considered as a theory for concepts, while Conceptual Graphs [So84] are used as building blocks for a theory for judgments and conclusions. Details of Contextual Logic and its philosophical foundations are discussed in [Wi96, Wi97, Pr98, Pr00, Wi00].

8 Conclusion

In this paper, we have discussed the need for a human-centered approach to Knowledge Discovery. We presented Conceptual Knowledge Discovery as such an approach, and argued how it fulfills the criteria for a human-oriented knowledge discovery provided by R. Brachman. We illustrated CKDD by two applications at Frankfurt Airport and at a Swiss warehouse, and presented the new CHIANTI tool for conceptual deviation discovery.

References

[AIS93] R. Agrawal, T. Imielinski, A. Swami: Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD*, 1993

[Ba95] F. Baader: Computing a Minimal Representation of the Subsumption Lattice of all Conjunctions of Concepts Defined in a Terminology. *Proc. of KRUSE '95.* August 11–13, 1995, 168–178

[BPT+00] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, L. Lakhal: Mining Minimal Non-Redundant Association Rules Using Frequent Closed Itemsets. In: J. Lloyd, V. Dahl, U. Furbach, M. Kerber, K.–K. Lau, C. Palamidessi, L. M. Pereira, Y. Sagiv, P. J. Stuckey (eds.): *Computational Logic — CL 2000.* Proc. CL '00. LNAI **1861**, Springer, Berlin–Heidelberg 2000, 972–986

[BTP+00] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, L. Lakhal: Mining Frequent Patterns with Counting Inference. *SIGKDD Explorations* **2**(2), Special Issue on Scalable Algorithms, 2000, 71–80

[Be97] H. Berg: *Terminologische Begriffslogik.* Diplomarbeit, FB4, TU Darmstadt, 1997

[BA96] R. J. Brachman, T. Anand: The Process of Knowledge Discovery in Databases. In [FPSU96], 37–57

[BS+93] R. J. Brachman, P. G. Selfridge, L. G. Terveen, B. Altman, A. Borgida, F. Halper, T. Kirk, A. Lazar, D. L. McGuinnes,

L. A. Resnick: Integrated Support for Data Archaeology. *International Journal of Intelligent and Cooperative Information Systems* 2 (1993), 159–185

[CR93] C. Carpineto, G. Romano: GALOIS: An Order-Theoretic Approach to Conceptual Clustering. *Machine Learning*. Proc. ICML 1993, Morgan Kaufmann Prublishers 1993, 33–40

[CES00] R. Cole, P. Eklund, G. Stumme: CEM — A Program for Visualization and Discovery in Email. In: D. A. Zighed, J. Komorowski, J. Zytkow (eds.): *Principles of Data Mining and Knowledge Discovery*. Proc. PKDD 2000. LNAI **1910**, Springer, Berlin–Heidelberg 2000, 367–374

[DS01] H. S. Delugach, G. Stumme (eds.): *Conceptual Structures: Broadening the Base*. Proc. ICCS 2001. LNAI **2120**. Springer, Berlin–Heidelberg 2001

[DG86] J.-L. Guigues, V. Duquenne: Familles minimales d' implications informatives resultant d' un tableau de données binaires. *Math. Sci. Humaines* **95**, 1986, 5–18

[FPS96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth: From Data Mining to Knowledge Discovery: An Overview. In [FPSU96], 1–34

[FPSU96] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.): *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, Cambridge 1996

[Ga87] B. Ganter: Algorithmen zur Formalen Begriffsanalyse. In: B. Ganter, R. Wille, K. E. Wolff (eds.): *Beiträge zur Begriffsanalyse*. B.I.-Wissenschaftsverlag, Mannheim 1987, 241–254

[GK00] B. Ganter, S. Kuznetsov: Formalizing hypotheses with concepts. In: [GM00], 342–356

[GK01] B. Ganter, S. Kuznetsov: Pattern structures and their projections. In: [DS01], 129–142

[GM00] B. Ganter, G. W. Mineau (eds.): *Conceptual Structures: Logical, Linguistic, and Computational Issues*. Proc. ICCS 2000. LNAI **1867**. Springer, Berlin–Heidelberg 2000

[GW96] B. Ganter, R. Wille: *Formale Begriffsanalyse: Mathematische Grundlagen*. Springer, Berlin–Heidelberg 1996

[GW99] B. Ganter, R. Wille: *Formal Concept Analysis : Mathematical Foundations*. Springer, Berlin–Heidelberg 1999. (Translation of [GW96])

[HSWW00] J. Hereth, G. Stumme, U. Wille, R. Wille: Conceptual Knowledge Discovery and Data Analysis. In: [GM00], 421–437

[Ka96] U. Kaufmann, *Begriffliche Analyse von Daten über Flugereignisse — Implementierung eines Erkundungs- und Analysesystems mit TOSCANA*. TU Darmstadt, 1996

[KSK01] R. Kasperzec, F. Seidemann, J. Kircher: *Begriffliche Analyse von Teilpopulationen in Datenbanken.* Seminararbeit, Universität Karlsruhe, 2001

[Lu91] M. Luxenburger: Implications partielles dans un contexte. *Mathématiques, informatique et sciences humaines* **113**, 1991, 35–55

[MW97] K. Mackensen, U. Wille: Qualitative text analysis supported by conceptual data systems. *Quality and Quantity: International journal of methodology* 2/33 (1999), 135–156

[MG95] G. Mineau, R. Godin: Automatic Structuring of Knowledge Bases by Conceptual Clustering. *IEEE Transactions on Knowledge and Data Engineering* **7**(5),1995, 824–829

[NN01] E. M. Nguifo, P. Njiwoua: IGLUE: A Lattice-based Constructive Induction System. In: *Intl. Journal of Intelligent Data Analysis (IDA)* 4(4), 2001, 1–49

[PBTL99] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal: Efficient mining of association rules using closed itemset lattices. *Journal of Information Systems* **24**(1), 1999, 25–46

[PHM00] J. Pei, J. Han, R. Mao: Closet: An efficient algorithm for mining frequent closed itemsets. *Proc. Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD)*, May 2000, 21–30

[Pe35] Ch. S. Peirce: *Collected Papers.* Harvard University Press, Cambridge 1931-35

[Pr97] S. Prediger: 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–Heidelberg 1997, 332–341

[Pr98] S. Prediger: *Kontextuelle Urteilslogik mit Begriffsgraphen. Ein Beitrag zur Restrukturierung der mathematischen Logik.* Dissertation, TU Darmstadt. Shaker Verlag, Aachen 1998

[Pr00] S. Prediger: Mathematische Logik in der Wissensverarbeitung. Historisch–philosophische Gründe für eine Kontextuelle Logik. *Math. Semesterberichte* **47**(2), 2000, 165–191

[SSW96] P. D. Selfridge, D. Srivastava, L. O. Wilson, IDEA: Interactive Data Exploration and Analysis. In: *Proc. SIGMOD '96*, Montreal, Canada 1996

[SS+93] P. Scheich, M. Skorsky, F. Vogt, C. Wachter, R. Wille: Conceptual Data Systems. In: O. Opitz, B. Lausen, R. Klar (eds.): *Information and Classification*. Springer, Berlin–Heidelberg 1993, 72–84

[So84] J. F. Sowa: *Conceptual structures: Information processing in mind and machine*. Adison-Wesley, Reading 1984

[StW93] S. Strahringer, R. Wille: Conceptual clustering via convex-ordinal structures. In: O. Opitz, B. Lausen, R. Klar (eds.): *Information and Classification*. Springer, Berlin–Heidelberg 1993, 85–98

[SWW01] S. Strahringer, R. Wille, U. Wille: Mathematical support of empirical theory building. In: [DS01], 169–186

[St96] G. Stumme: The Concept Classification of a Terminology Extended by Conjunction and Disjunction. In: N. Foo, R. Goebel (eds.): *PRICAI'96: Topics in Artificial Intelligence*. LNAI **1114**, Springer, Berlin–Heidelberg 1996, 121–131

[St98] G. Stumme: Exploring Conceptual Similarities of Objects for Analyzing Inconsistencies in Relational Databases. *Proc. Workshop on Knowledge Discovery and Data Mining, 5th Pacific Rim Intl. Conf. on Artificial Intelligence*. Singapore, Nov. 22–27, 1998, 41–50

[St99a] G. Stumme: Dual Retrieval in Conceptual Information Systems. In: A. P. Buchmann (ed.): *Datenbanksysteme in Büro, Technik und Wissenschaft*. Springer, Berlin–Heidelberg 1999, 328–342

[St99b] G. Stumme: Conceptual Knowledge Discovery with Frequent Concept Lattices. FB 4–Preprint **2043**, TU Darmstadt 1999

[St00] G. Stumme: Conceptual On-Line Analytical Processing. In: K. Tanaka, S. Ghandeharizadeh, Y. Kambayashi (eds.): *Information Organization and Databases.* Chpt. 14. Kluwer, Boston–Dordrecht–London 2000

[ST+01a] G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, L. Lakhal: Intelligent Structuring and Reducing of Association Rules with Formal Concept Analysis. In: F. Baader. G. Brewker, T. Eiter (eds.): *KI 2001: Advances in Artificial Intelligence*. Proc. KI 2001. LNAI **2174**, Springer, Berlin–Heidelberg 2001, 335–350

[ST+01b] G. Stumme, R. Taouil, Y. Bastide, N. Pasqier, L. Lakhal: Computing Iceberg Concept Lattices with Titanic. *J. on Knowledge and Data Engineering* (to appear)

[SW00] G. Stumme, R. Wille (eds.): *Begriffliche Wissensverarbeitung: Methoden und Anwendungen*. Springer, Berlin–Heidelberg 2000

[SWW98] G. Stumme, R. Wille, U. Wille: 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, Lecture Notes in Artificial Intelligence **1510**, Springer, Berlin–Heidelberg 1998, 450–458

[SWo97] G. Stumme, K. E. Wolff: Computing in Conceptual Data Systems with Relational Structures. In: *Proc. of KRUSE '97*. Vancouver, August 11–13, 1997, 206–219

[To01] *The ToscanaJ-Project: An Open-Source Reimplementation of TOSCANA* . http://toscanaj.sourceforge.net

[Ut96] R. Uthurusamy: From Data Mining to Knowledge Discovery: Current Challenges and Future Directions. In [FPSU96], 561–569

[VW95] F. Vogt, R. Wille: TOSCANA — A Graphical Tool for Analyzing and Exploring Data. In R. Tamassia, I. G. Tollis (eds.): *Graph Drawing '94*. LNCS **894**. Springer, Berlin–Heidelberg 1995, 226–233

[Wi82] R. Wille: Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In I. Rival (ed.): *Ordered Sets*. Boston-Dordrecht: Reidel, 1982, 445–470

[Wi92a] R. Wille: Concept Lattices and Conceptual Knowledge Systems. *Computers & Mathematics with Applications* **23**, 1992, 493–515

[Wi92b] R. Wille: Begriffliche Datensysteme als Werkzeug der Wissenskommunikation. In: H. H. Zimmermann, H.-D. Luckhardt, A. Schulz (eds.): *Mensch und Maschine — Informationelle Schnittstellen der Kommunikation*. Univ.-Verl. Konstanz, 1992, 63–73

[Wi96] R. Wille: Restructuring mathematical logic: an approach based on Peirce's pragmatism. In: A. Ursini und P. Agliano (eds.): *Logic and Algebra*. Marcel Dekker, New York 1996, 267-281

[Wi97] R. Wille: Conceptual Landscapes of Knowledge: A Pragmatic Paradigm for Knowledge Processing. In: *Proc. of KRUSE '97*. Vancouver, August 11–13, 1997, 2–13

[Wi00] R. Wille: Contextual Logic summary. In: G. Stumme (ed.): *Working with Conceptual Structures: Contributions to ICCS 2000*. Shaker-Verlag, Aachen 2000, 265–276

[Wi01] R. Wille: Why can concept lattices support knowledge discovery in databases? In: E. M. Nguifo, V. Duquenne, M. Liquiere (eds.): *Concept Lattice-based theory, methods and tools for Knowledge Discovery in Databases.* Proc. of Workshop of the 9th Intl. Conf. on Conceptual Structures (ICCS '01). http://CEUR-WS.org/Vol-42/

[Za00] M. J. Zaki: Generating non-redundant association rules. *Proc. KDD 2000.* 34–43