Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods

Gerd Stumme[1], Rudolf Wille[1], Uta Wille[2]

[1] Technische Universität Darmstadt, Fachbereich Mathematik, D–64289 Darmstadt, Germany, {stumme, wille}@mathematik.tu-darmstadt.de
[2] IBM Research Division, Zurich Research Laboratory, CH–8803 Rüschlikon, Switzerland, wille_u@jelmoli.ch

In this paper we discuss *Conceptual Knowledge Discovery in Databases* (*CKDD*) as it is developing in the field of *Conceptual Knowledge Processing* (cf. [29],[30]). 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 18 years. This approach relies on the pragmatic philosophy of Ch. S. Peirce [15] who claims that we can only analyze and argue within restricted contexts where we always rely on pre-knowledge and common sense. The development of Formal Concept Analysis led to the software system TOSCANA, which is presented as a CKDD tool in this paper. TOSCANA is a flexible navigation tool that allows dynamic browsing through and zooming into the data. It supports the exploration of large databases by visualizing conceptual aspects inherent to the data. We want to clarify that CKDD can be understood as a human-centered approach of Knowledge Discovery in Databases. The actual discussion about human-centered Knowledge Discovery is therefore briefly summarized in Section 1.

1 Human-Centered Knowledge Discovery

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"* [7].

According to R. S. Brachman and T. Anand [3], much attention and effort 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 processorientation of KDD tasks and argue in favor of a more human-centered approach for a successful development of knowledge-discovery support tools (see also [24], 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 [4]. Data Archaeology can be considered a highly human-centered process of asking, exploring, analyzing, interpreting, and learning in interaction with the underlying database.

Emphasizing the KDD process, comprehensive support of the analyst has to be provided that, according to [3], should be embedded into a knowledge-discovery *support environment*. A support environment should especially support the overall process of human-centered KDD, including Data Archaeology involved in many KDD applications. In this paper, we investigate and discuss how the process of human-centered KDD can be supported by *Formal Concept Analysis* methods. This is done with regard to the basic requirements formulated for human-centered KDD support tools.

In order to formulate requirements for knowledge discovery support tools, it is necessary to reflect the underlying understanding of knowledge. A human-centered approach to KDD that supports the overall KDD process should be based on a comprehensive notion of knowledge 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 [6], p. 8). The *landscape paradigm* of knowledge underlying *conceptual knowledge processing* as described in [30] 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. [30]).

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 supports 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 [3], 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 [4]. Requirements 1 to 5 of the subsequent list are explicitely stated in [4], p. 164, while the remaining requirements are implicit in [3] and [4].

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.

Before discussing Conceptual Knowledge Discovery in Databases with regard to these requirements in Section 3, we introduce some basic notions and ideas of *Formal Concept Analysis* and *conceptual data systems* in the next section.

2 Formal Concept Analysis

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* ([27],[28],[5]) 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. $I \subseteq G \times M$); in general, gIm ($\Leftrightarrow (g,m) \in I$) is read: "the object g *has* the attribute m". In Figure 1, a formal context is described by a table in which the crosses represent the binary relation I between the object set G (comprising the gates of Terminal 1 at Frankfurt Airport) and the attribute set M (consisting of certain gate types). 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 $\underline{\mathfrak{B}}(G, M, I)$.

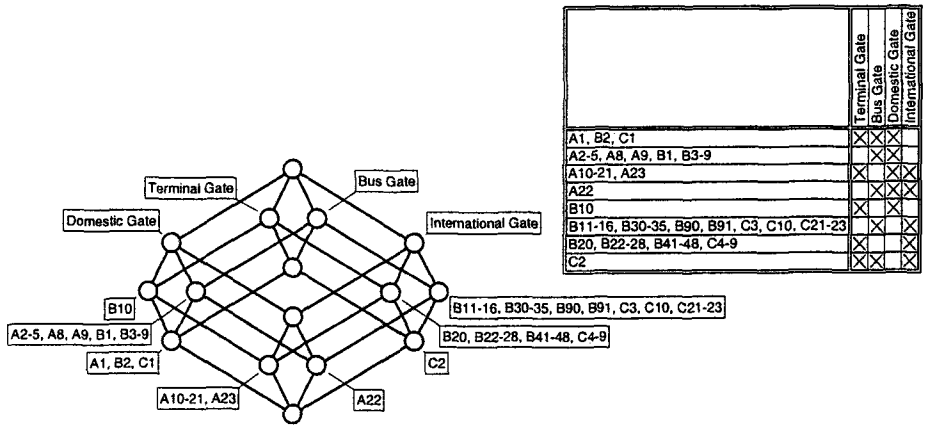

Fig. 1. A formal context concerning gates at Frankfurt Airport and its concept lattice

In this example, the intents of the formal context are exactly the subsets of its attribute set; hence its concept lattice is a 16-element Boolean lattice, as can be seen in Figure 1, which visualizes the concept lattice by a (*labeled*) *line diagram*. In a line diagram of a concept lattice, the name of an object g is always attached to the circle representing the smallest concept with g in its extent (denoted by γg); dually, the name of an attribute m is always attached to the circle representing the largest concept with m in its intent (denoted by μm). This labelling allows us to read the context relation from the diagram because $gIm \iff \gamma g \leq \mu m$, in words: *the object g has the attribute m if and only if there is an ascending path from the circle representing γg to the circle 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 circle in the line diagram of Figure 1 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 data systems" has been developed to activate concept lattices as query structures for databases. A *conceptual data 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 are implemented with the management system TOSCANA (see [20],[26]). 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 their runways, which are taken as attributes of the scale. The power of the 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

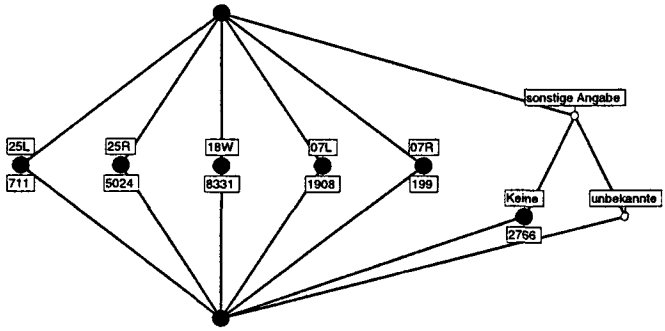

Fig. 2. The query structure *Runway*

both lattices or a line diagram of the second lattice refining a specific concept of the first; the latter alternative may be used for zooming further and further, which potentially allows us to navigate through the entire database.

3 Conceptual Knowledge Discovery in Databases

Conceptual data systems activated by the management system TOSCANA can be considered as *knowledge discovery support environments* that promote human-centered discovery processes and representations of their findings. In this section, we want to discuss how such processes of conceptual knowledge discovery fulfil the requirements listed in Section 1. As illustrating example, we use a TOSCANA system established by U. Kaufmann [10] 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 TOSCANA system, the *objects* of the underlying domain are stored structurally in a relational database so that they can be activated by SQL-statements for establishing updated *conceptual scales*. The objects are represented for the user in line diagrams of the concept lattices of conceptual scales as 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 circle 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 TOSCANA systems. The second requirement of extendibility of categorical structures is already realized by the great flexibility in forming conceptual scales; even during the process of discovery new insights may give rise to further conceptual scales. The third requirement of meaningful data segmentations

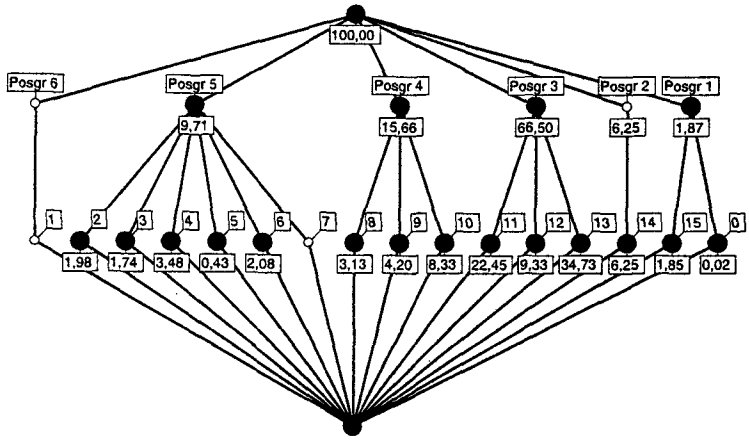

Fig. 3. The query structure *Wingspan Code and Position Size*

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 TOSCANA systems by continuing the investigation of Runway 18 West. In Figure 3 we zoom into the concept node labelled "18W" with the conceptual scale *Wingspan Code and Position Size*. Then we can study the size of the 8331 planes that took off from 18 West within the resulting line diagram shown in Figure 3. The 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 size of the extents is described by percentages instead of quantities. 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 overproportionally to the noise pollution. We test this hypothesis by zooming into the two concept nodes labelled *Posgr 5* and *Posgr 4* with the scale *Noise Class of the Plane by ICAO-Annex 16*. The two line diagrams in Figure 4 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 overproportionally large, but that more than 95% of them are categorized as silent.

TOSCANA 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 [14] how such processes may be understood as

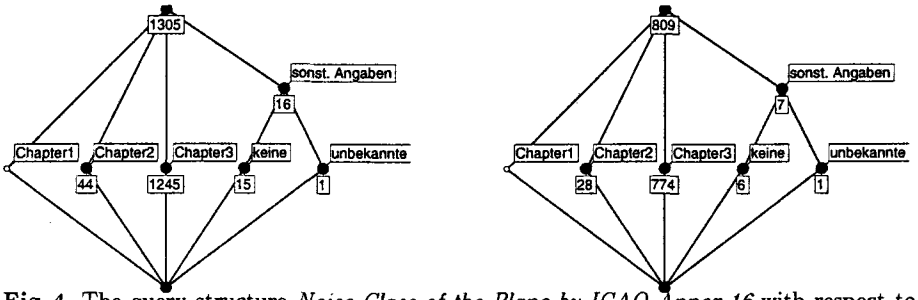

Fig. 4. The query structure *Noise Class of the Plane by ICAO-Annex 16* with respect to position sizes P4 and P5

procedures of qualitative theory building. 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, a TOSCANA system offers a conceptually shaped landscape 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. [30]). The graphical representation of interesting parts of the landscape, in particular, supports intersubjective communication and argumentation.

4 Applications

TOSCANA systems have been successfully elaborated for many purposes in different research areas, but also on the commercial level. Its range covers a variety of applications, that incorporate knowledge discovery. For instance, TOSCANA systems have been established for *analyzing* data of children with diabetes [20], for *investigating* international cooperations [11], for *exploring* laws and regulations in civil engineering [13], for *retrieving* books in a library [12], [17], for *assisting* engineers in designing pipings [25], for *developing* qualitative theories in music esthetics [14], for *studying* semantics of speech-act verbs [8], and for *examining* the medical nomenclature system SNOMED [18]. As a Conceptual Knowledge Discovery tool, TOSCANA was used to investigate deficiencies of the control system of the incineration plant Darmstadt [9]. One of the leading German mail-order companies is currently implementing a prototype of a TOSCANA system for its customer database, which shall be compared to statistical KDD tools.

Conceptual data systems can also be understood as On-Line Analytical Processing (OLAP) tools [22]. Roughly, the conceptual scales can be regarded as dimensions of a multi-dimensional data cube. The zooming-in on one of the concepts of a scale as described in the previous section corresponds to 'slicing' the data cube. 'Rotating' and 'Drill-Down' are also supported. Figure 6 shows how different scales can be combined and represented in a nested line diagram to visualize dependencies between different 'dimensions'. Here the positions of the aircraft are compared to the positions of the assigned baggage conveyors. In this application, it is not of interest to obtain general propositions, but to detect special cases. For instance, there are four aircraft that docked at Terminal 2, while their assigned baggage conveyors are in Terminal 1. Vice versa, 180 aircrafts at Terminal 1 were assigned conveyors in Terminal 2. The 7+17 cases in which the aircraft docks at one of the two terminals,

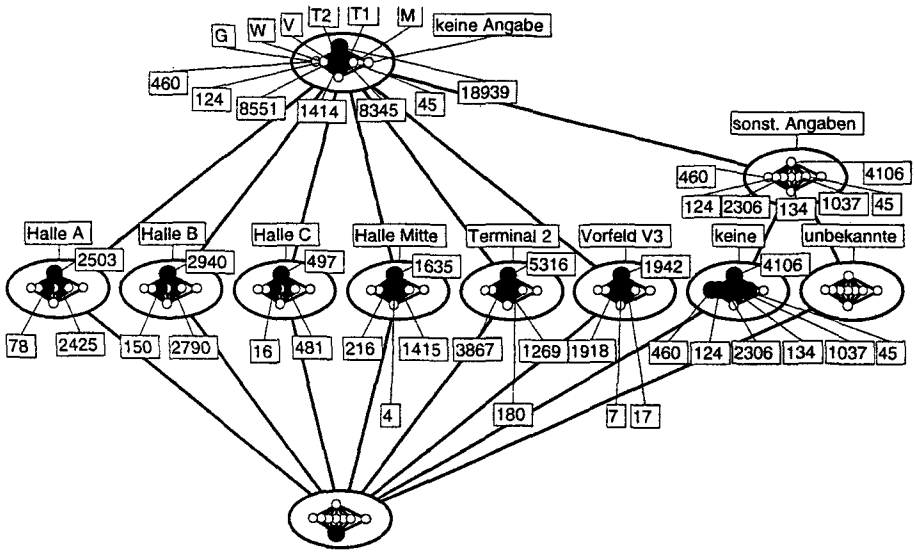

Fig. 5. Nested line diagram of the scales *Position of baggage conveyor* and *Positions*

while the assigned conveyor is on the apron, should also be considered. In all these cases, one can drill down to the original data by clicking on the number to obtain the flight movement numbers, which in turn lead to the data set stored in the INFO-80 system.

Further research in Conceptual Knowledge Processing aims at developing *conceptual knowledge systems* by extending the functionalities of conceptual data systems, especially by logic-based components. As Formal Concept Analysis and Description Logics are closely related and have similar purposes (see, e. g., [4],[19]), first steps in integrating both theories have been made ([1], [2], [16], [21]). For hybrid knowledge processing, an extension of conceptual data systems is foreseen by incorporating statistical and computational components [23]. This indicates a promising development in terms of extending TOSCANA systems toward a wider range of CKDD applications.

References

1. Baader, F., Computing a Minimal Representation of the Subsumption Lattice of all Conjunctions of Concepts Defined in a Terminology. In: *Proc. of KRUSE '95.* August 11–13, 1995, 168–178
2. Berg, H., *Terminologische Begriffslogik.* Diplomarbeit, FB4, TU Darmstadt, 1997.
3. Brachman, R.J., Anand, T., The Process of Knowledge Discovery in Databases. In [7]
4. Brachman, R.J. et. al., Integrated Support for Data Archaeology. *Int. J. of Intelligent and Cooperative Information Systems*, **2**(2), 1993, 159–185.
5. Ganter, B., Wille, R., *Formale Begriffsanalyse: Mathematische Grundlagen.* Berlin-Heidelberg: Springer-Verlag, 1996 (English translation to appear).
6. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., From Data Mining to Knowledge Discovery: An Overview. In [7]
7. Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds. *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, Cambridge 1996.
8. Grosskopf, A., Harras, G.,*Eine TOSCANA-Anwendung für Sprechaktverben des Deutschen.* FB4-Preprint, TU Darmstadt 1998.

9. Kalix, E., *Entwicklung von Regelungskonzepten für thermische Abfallbehandlungsanlagen.* Diplomarbeit, FB13, TU Darmstadt, 1997.
10. Kaufmann, U., *Begriffliche Analyse von Daten über Flugereignisse — Implementierung eines Erkundungs- und Analysesystems mit TOSCANA.* TU Darmstadt, 1996.
11. Kohler-Koch, B., Vogt F., *Normen und regelgeleitete internationale Kooperationen.* FB4-Preprint 1632, TU Darmstadt, 1994.
12. Kollewe, W., Sander, C., Schmiede, R., Wille, R., TOSCANA als Instrument der bibliothekarischen Sacherschließung. In: H. Havekost and H.J. Wätjen (eds.), *Aufbau und Erschließung begrifflicher Datenbanken.* (BIS)-Verlag, Oldenburg, 1995, 95-114.
13. Kollewe, W., Skorsky, M., Vogt, F., Wille, R., TOSCANA — ein Werkzeug zur begrifflichen Analyse und Erkundung von Daten. In: R. Wille and M. Zickwolff (eds.), *Begriffliche Wissensverarbeitung - Grundfragen und Aufgaben.* B.I.-Wissenschaftsverlag, Mannheim, 1994, 267-288.
14. Mackensen, K., Wille, U., *Qualitative Text Analysis Supported by Conceptual Data Systems.* Preprint, ZUMA, Mannheim, 1997.
15. Peirce, Ch. S., *Collected Papers.* Havard University Press, Cambridge, 1931–35.
16. Prediger, S., 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.
17. Rock, T., Wille, R., Ein TOSCANA-System zur Literatursuche. In: G. Stumme and R. Wille (eds.): *Begriffliche Wissensverarbeitung: Methoden und Anwendungen.* Springer, Berlin-Heidelberg (to appear)
18. Roth-Hintz, M., Mieth, M, Wetter, T., Strahringer, S., Groh, B., Wille, R., Investigating SNOMED by Formal Concept Analysis. Submitted to: *Artificial Intelligence in Medicine.*
19. Selfridge, P. D., Srivastava, D., Wilson, L. O., IDEA: Interactive Data Exploration and Analysis. SIGMOD '96, Montreal, Canada 1996
20. Scheich, P., Skorsky, M., Vogt, F., Wachter, C., Wille, R., Conceptual Data Systems. In: O. Opitz, B. Lausen, R. Klar (eds.): *Information and Classification.* Springer, Berlin-Heidelberg, 1993, 72–84.
21. Stumme, G., 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
22. Stumme, G., Conceptual Information Systems and Conceptual On-Line Analytical Processing. *Proc. of FODO '98.* Springer, Heidelberg 1998 (submitted)
23. Stumme, G., Wolff, K.E., Computing in Conceptual Data Systems with Relational Structures. In In: *Proc. of KRUSE '97.* Vancouver, August 11–13, 1997, 206–219
24. Uthurusamy, R., From Data Mining to Knowledge Discovery: Current Challenges and Future Directions. In [7]
25. Vogel, N., *Ein begriffliches Erkundungssystem für Rohrleitungen.* TU Darmstadt, 1995.
26. Vogt, F., Wille, R., 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.
27. Wille, R., Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts. In I. Rival (ed.): *Ordered Sets.* Boston-Dordrecht: Reidel, 1982, 445–470.
28. Wille, R., Concept Lattices and Conceptual Knowledge Systems. *Computers & Mathematics with Applications, 23,* 1992, 493-515.
29. Wille, R., 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.
30. Wille, R., Conceptual Landscapes of Knowledge: A Pragmatic Paradigm for Knowledge Processing. In: *Proc. of KRUSE '97.* Vancouver, August 11–13, 1997, 2–13