Conceptual Knowledge Discovery and Data Analysis

Joachim Hereth[1], Gerd Stumme[1], Rudolf Wille[1], and Uta Wille[2]

[1] Technische Universität Darmstadt, Fachbereich Mathematik, Schloßgartenstr. 7, D–64289 Darmstadt, Germany,
{hereth,stumme,wille}@mathematik.tu-darmstadt.de
[2] Jelmoli AG, Data Management, Postfach 3020, Ch–8021 Zürich, Switzerland;
wille_u@jelmoli.ch

Abstract. In this paper, we discuss Conceptual Knowledge Discovery in Databases (CKDD) in its connection with Data Analysis. Our approach is based on Formal Concept Analysis, a mathematical theory which has been developed and proven useful during the last 20 years. Formal Concept Analysis has led to a theory of conceptual information systems which has been applied by using the management system TOSCANA in a wide range of domains. In this paper, we use such an application in database marketing to demonstrate how methods and procedures of CKDD can be applied in Data Analysis. In particular, we show the interplay and integration of data mining and data analysis techniques based on Formal Concept Analysis. The main concern of this paper is to explain how the transition from data to knowledge can be supported by a TOSCANA system. To clarify the transition steps we discuss their correspondence to the five levels of knowledge representation established by R. Brachman and to the steps of empirically grounded theory building proposed by A. Strauss and J. Corbin.

Contents

1 Conceptual Knowledge Discovery in Databases

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-centered process of discovering knowledge from data by visualizing and analyzing the formal 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. [35]). The general approach of CKDD and the qualities of TOSCANA as a KDD support tool have previously been discussed in [27] with respect to Brachman and Anand's fundamental requirements for knowledge discovery support environments (cf. [4]). Therefore, the basic notions and the philosophical background of CKDD are only briefly summarized in this paper. For a comprehensive presentation of the mathematical foundations of Formal Concept Analysis see [10]; basics of Conceptual Knowledge Processing are explained in [31],[32],[33],[35].

The overall theme and contribution of the volume *"Advances in Knowledge Discovery and Data Mining"* [7] is a process-centered view of KDD considering KDD as an interactive and iterative process between a human and a database that may strongly involve background knowledge of the analyzing domain expert. In particular, R. S. Brachman and T. Anand [4] argue in favor of a more human-centered approach to knowledge discovery support referring to the constitutive character of human interpretation for the discovery of knowledge and stressing the complex, interactive process of KDD as being led by human thought.

Following Brachman and Anand, CKDD pursues a human-centered approach to KDD based on a comprehensive notion of knowledge as a part of human thought and argumentation. The *landscape paradigm* of knowledge underlying CKDD is based on the pragmatic philosophy of Ch. S. Peirce [16] where knowledge is understood as always being incomplete, formed and continuously assured by human discourse within an intersubjective community of communication (cf. [35]). Emphasizing the intersubjective character of knowledge, CKDD considers knowledge communication as an important part of the overall discovery process with respect to both the dialog between user and system, and also as a part of human communication and argumentation. Therefore, a major focus of CKDD is to provide knowledge discovery support that guarantees a high transparency of the discovery process and a representation of its (interim) findings to support human argumentation and establishment of intersubjectively assured knowledge. CKDD especially supports a wide-ranging and unpredictable interactive exploration of the data ("data archaeology", cf. [5]) where the software tools TOSCANA and CHIANTI serve as a knowledge discovery support environment in which CKDD applications can be efficiently implemented (see [27]).

2 Conceptual Data Analysis

CKDD is based on methods and procedures of *Conceptual Data Analysis* that allow the analysis of given data by examination and visualization of their conceptual structure. The derived graphical representations have proven to be useful for making the data communicable in addition to identifying conceptual relationships in the data. Knowledge is discovered in interaction with the data during an iterative process which activates techniques of Conceptual Data Analysis and is guided by theoretical preconceptions and declared purposes of the domain expert. In the following paragraphs, we briefly introduce the basic notions and procedures of Conceptual Data Analysis using an application in database marketing.

Based on a philosophically grounded formalization of concept (see [34]), Conceptual Data Analysis allows data to be mathematically treated and processed. Formal Concept Analysis, the mathematical theory underlying Conceptual Data Analysis, formalizes concept and conceptual hierarchy to reflect the philosophical understanding of a concept as a unit of thought constituted by its extension and its intension. The extension comprises all objects belonging to the concept while the intension consists of all attributes valid for those objects. To allow a mathematical description of extension and intension, Formal Concept Analysis always starts with a formal context:

Definition 1. *A* formal context *is a set structure* $\mathbb{K} := (G, M, I)$ *where* G *and* M *are sets and* I *is a binary relation between* G *and* M *(i. e.* $I \subseteq G \times M$*). The elements of* G *and* M *are called (*formal*)* objects *and* attributes*, respectively, and* gIm *(*$\Leftrightarrow (g, m) \in I$*) is read: "the object* g has *the attribute* m*".* Derivations *are defined by* $X' := \{m \in M \mid \forall g \in X : gIm\}$ *for* $X \subseteq G$ *and* $Y' := \{g \in G \mid \forall m \in Y : gIm\}$ *for* $Y \subseteq M$*. A* formal concept *of the formal context* $\mathbb{K}$ *is a pair* (A, B) *with* $A \subseteq G$*,* $B \subseteq M$*,* $A = B'$*, and* $B = A'$*; 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 formal concepts of $\mathbb{K}$ *together with the order relation* $\leq$ *is always a complete lattice, called the* concept lattice *of* $\mathbb{K}$ *and denoted by* $\underline{\mathfrak{B}}(\mathbb{K})$*.*

The concept lattices can be graphically represented by line diagrams which have been proven to be useful representations for the understanding of conceptual relationships in data. Before we illustrate this by examples, we introduce the notion of a *many-valued context* as a formalization of data tables that reports, for objects under consideration, specific values with respect to given attributes. In order to obtain a concept lattice of a many-valued context, the context has to be formally transformed to a formal context (also called a *one-valued context*). This transformation is performed by using *conceptual scales* which reflect specific interpretations of the data.

Definition 2. *A many-valued context is a set structure* $\mathbb{K} := (G, M, W, I)$ *where* G*,* M*, and* W *are sets and* I *is a ternary relation between* G*,* M*, and* W *(i.e.* $I \subseteq G \times M \times W$*) such that* $(g, m, w_1) \in I$ *and* $(g, m, w_2) \in I$ *always imply* $w_1 = w_2$*. The elements of* G*,* M*, and* W *are called objects, attributes, and attribute values, respectively, and* $(g, m, w) \in I$ *is read: "the object* g *has the attribute value* w *for the attribute* m*". An attribute* m *may be considered as a (partial) mapping from* G *to* W*; therefore,* $m(g) = w$ *is often written instead of* $(g, m, w) \in I$*. A conceptual scale for an attribute* $m \in M$ *is a one-valued context* $\mathbb{S}_m := (G_m, M_m, I_m)$ *with* $m(G) \subseteq G_m$*. The context* $\mathbb{R}_m := (G, M_m, J_m)$ *with* $gJ_mn :\Longleftrightarrow m(g)I_mn$ *is called the realized scale for the attribute* $m \in M$*. The*

Fig. 1. Line diagrams showing the cross-selling between travel accessories, perfumery, and ladies' accessories

derived context *of* $\mathbb{K}$ *with respect to the conceptual scales* $\mathbb{S}_m := (G_m, M_m, I_m)$ $(m \in M)$*is the formal context* $(G, \bigcup_{m\in M}\{m\} \times M_m, J)$ *with* $gJ(m,n) :\iff m(g)I_m n$*; its concept lattice is considered as the* concept lattice *of the many-valued context* $\mathbb{K}$ scaled *by the conceptual scales* $\mathbb{S}_m := (G_m, M_m, I_m)$ $(m \in M)$*. A many-valued context together with a collection of appertaining conceptual scales with line diagrams of their concept lattices is called a* conceptual data system*.*

Conceptual data systems can be implemented with the management system TOSCANA (see [29]). 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, thus allowing users to navigate through the data and to analyze specific sets of objects by activating scales that interpret relevant aspects of the given data. Conceptual data systems stored in a database and implemented with a management system such as TOSCANA are called *conceptual information systems*.

In the following paragraphs, we illustrate how conceptual data analysis may be performed with a TOSCANA information system implemented to support the database marketing of a Swiss department store. The conceptual scales together with line diagrams of their concept lattices are derived from a database recording the activity of individual customers with respect to the various departments of the store. The analysis was undertaken to reveal potentials for cross-selling activities. For instance, to select the target group of a direct mail for promoting the ladies' wear department, one may start with unfolding the cross-selling behavior between departments where women typically buy.

The line diagram on the left side in Figure 1 shows the cross-selling behavior between travel accessories, perfumery, and ladies' accessories. The line diagram represents the concept lattice of the realized scale having as formal objects all customers with purchases in at least one of the three departments and having the three formal attributes 'purchased in travel accessories', 'purchased in perfumery', and 'purchased in ladies' accessories' while the binary relation records who bought in which department. The formal concepts of the realized scale are

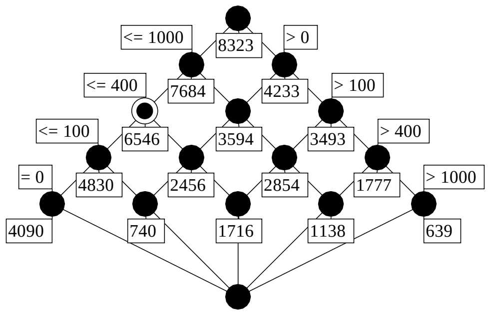

Fig. 2. Line diagram showing sales in women's clothing accrued by perfume and ladies' accessories customers

represented in the diagram by the little circles. The name of a formal object g is always attached to the circle representing the smallest concept with g in its extent (denoted by γg); dually, the name of a formal attribute m is always attached to the little circle representing the largest concept with m in its intent (denoted by μm). This labelling allows to read the context relation from the diagram because of $gIm \iff \gamma g \leq \mu m$, in words:

> *The object g has the attribute m if and only if there is an ascending path of line segments from the circle labelled with the name of g to the circle labelled with the name of m.*

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 $B = \{m \in M \mid (A, B) \leq \mu m\}$. The line diagrams in this paper show instead of the object names only the number of those names attached to the appertaining circle. Therefore, the diagram shows that there were 1075 customers who bought travel accessories only, 8182 perfumes only, and 3964 ladies' accessories only, but nothing in either of the other two departments. Furthermore, there were 967 customers who purchased travel accessories and something from perfumery but no ladies' accessories, and 1849 customers who were active in all three departments. From the diagram questions naturally arise, for example, why do 8182 customers buy perfumery goods but no travel or ladies' accessories even though both departments are right next to each other?

For the forementioned mailing select to promote sales in ladies' clothing, interesting are the 6474 + 1849 = 8323 customers because, in general, it is easier to develop active customers into better customers. The diagram on the right hand side in Figure 1 represents the same facts as the left one, but the number of customers are summed from the bottom up. To study the group of perfume and ladies' accessory buyers in further detail, TOSCANA allows users to "zoom into" the circle in the right diagram representing the 8323 customers who bought perfumery goods, ladies' accessories and, in some cases, travel accessories. Figure 2 shows a segmentation of those customers with respect to their previous

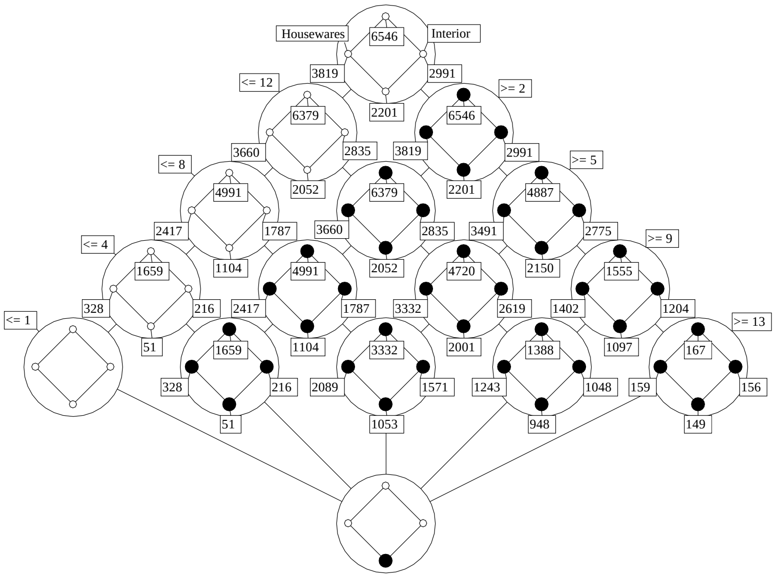

Fig. 3. Nested line diagram combining numbers of visited departments with the cross-selling between Housewares and Interior

activity in the ladies' wear department (formal, business, and casual wear). In this diagram, the number of customers are again summed from the bottom up; for instance, there are 1777 customers in the group of 8323 who spent more than 400 SFr for women's clothing, 639 who spent more than 1000 SFr, and 1138 who spent between 400 and 1000 SFr. The customers with low or no activity in ladies' wear were chosen as the targets of the mailing select, as the rest of the customers were identified as already being good ladies' wear customers.

In Figure 3 the activity of the 6546 customers with 400 or less sFr sales of women's clothing is shown. The *nested line diagram* presents two aspects of the activity of the 6546 customers: the line diagram representing the number of departments in which customers shopped (outer part) is combined with the cross-selling line diagram between housewares and interior (inner part). The circles of the first line diagram have been enlarged so that a copy of the second line diagram could be drawn in each enlarged circle. The nested line diagram can be read like an ordinary one if we replace the lines beween the large circles by parallel lines between the correspondeng circles of the inner diagrams. For instance, we can read from the diagram that there are 4720 customers who shopped in 5 or more but less than 13 departments of the store, and that 2001 of those bought housewares as well as interiors which seems to be a good target group for a direct mailing.

The examples should have made it clear that a TOSCANA information system enables an interactive and iterative process of conceptual data analysis leading to useful knowledge. The experiences with many TOSCANA systems have shown that domain experts are mostly stimulated by navigating through the graphical representations because they have a rich background knowledge about the appertaining domain and special interests for activating substantial questions. The process of knowledge discovery with TOSCANA systems is always accompanied by a learning process which increases the ability of the user to better understand the goals and possibilities of the specific exploration procedure. All these are reasons for viewing TOSCANA information systems as human-centered support of knowledge discovery, as Brachman and Anand advocated in [4].

3 From Data to Knowledge

In the previous section it is demonstrated through examples of conceptual data analysis how a conceptual information system may function as a *knowledge discovery support environment* that promotes human-centered discovery processes. In this section we want to explain in general the transition from data to knowledge for the discovery processes supported by a TOSCANA system. To clarify the transition steps from data (understood as symbolic representation of realities) to human knowledge, we call upon an analysis of knowledge representations in semantic networks performed by R. Brachman [3] who identified the following five representation levels (cf. [14]):

– *Implementational Level*: The primitives are nodes and links where links are merely pointers and nodes are simply destinations for links. On this level, there are only data structures from which logical forms can be build.
– *Logical Level*: The primitives are logical predicates, operators, and propositions together with a structured index over those primitives. On this level, logical adequacy is responsible for meaningfully prestructuring knowledge.
– *Epistemological Level*: The primitives are conceptual units, conceptual subpieces, inheritance and structuring relations. On this level, conceptual units are determined by their inherent structure and their interrelationships.
– *Conceptual Level*: The primitives are word senses and case relations, object- and action-types. On this level, small sets of language-independent conceptual elements and relationships are fixed and from which all expressible concepts can be constructed.
– *Linguistic Level*: The primitives are arbitrary concepts, words, and expressions. On this level, the primitives are language-dependent, and are expected to change in meaning as the network grows.

The grading of the levels, from implementational to linguistic, orders the representations from simple and abstract to complex and concrete; hence the grading should not misunderstood as a chronological ordering, although there are connections between the grading and the course of the transition from data to knowledge. In the following, the representation levels shall be characterized according to their functionalities for supporting the process from data to knowledge as performed by a TOSCANA information system.

On the *implementational level*, the basic data structures are defined as one- and many-valued contexts. Already on this elementary level, there are instances for establishing connections to human knowledge, namely the formal objects, attributes, and attribute values of the contexts and the incidence relations between those elements. On this level, data contexts are merely considered as formal set structures without any content. Implementational issues for TOSCANA systems are discussed in [28] in detail.

On the *logical level*, names for the formal objects, attributes, attribute values, and incidence relations are formally taken as logical predicates which allow the composition of further predicates by logical connectives and quantifiers. Syntax and formal contextual semantics of those predicates have been elaborated to the so-called *Terminological Attribute Logic* (see [18],[11]) and *Terminological Concept Logic* (see [2]) which are both related to description logics. Both terminological logics may assist the formation of abstract scales for the methods of conceptual, relational, and logical scaling (see [17],[19]). The management system TOSCANA allows the activation of used logical expressions by representing them as SQL-queries. The combination of abstract scales to larger contexts is also performed on the logical level, namely by various context constructions; the mostly used context construction is the semiproduct which is basic for 'plain conceptual scaling' (see [9]), and the apposition which underlies the nested line diagrams used by TOSCANA as exemplified in Section 2 (see [29]).

The *epistemological level* addresses "the possibility of organizations of conceptual knowledge into units more structured than simple nodes and links or predicates and propositions" [3]. Formal concepts are indeed more internally structured than just a node or a predicate: they unify an object set (the extent) and an attribute set (the intent) so that each of these parts determines the other. Furthermore, the internal structure of the formal concepts gives rise to a conceptual hierarchy which mathematically forms a complete lattice if the formal concepts are those of a given formal context. Thus, the rich mathematical theory of Formal Concept Analysis (see [10]) yields a substantial contribution to Brachman's epistemological level. As Formal Concept Analysis is founded on lattice theory, lattice constructions and decompositions can be activated for establishing more complex concept hierarchies out of simpler ones, and, vice versa, for reducing complex concept hierarchies to simpler ones. Constructions like (sub-) direct products and tensor products of concept lattices and decompositions like subdirect and atlas decompositions have been successfully applied in data analysis and knowledge processing. For supporting the process of knowledge discovery, the visualization of concept lattices and their constructions and decompositions by specific line diagrams are of great importance. Those visualizations (also belonging to the epistomological level) are able to stabilize knowledge acquisition and communication (cf. [32]).

On the *conceptual level*, word senses are represented by the context attributes which lead to a contextual representation of concept intensions. As primitive case relations, there are defined four basic relations: an object has an attribute, an object belongs to a concept, a concept abstracts to an attribute, and a concept is a subconcept of another concept (cf. [12]). These four relations are basic for the knowledge representation in conceptual information systems because, together with the word senses, they can represent a large amount of language-independent knowledge structures. Such structures are the *concrete scales* of TOSCANA systems which are used to capture the intensional content of an application domain (the extensional side of those scales are still abstract).

On the *linguistic level*, TOSCANA systems work with realized scales which are obtained by actualizing the abstract objects of their concrete scales according to real data. This realization particularly allows to deduce concept graphs representing verbal texts (see [20]). On this level, the knowledge representation is language-dependent so that users of the conceptual information system can best activate their background knowledge and common sense. The navigation through the conceptual landscape of the system, visualized by labelled line diagrams, can be performed successfully because the interplay between formal and material thinking stimulated by the diagrams gives purposeful orientations (cf. [35]).

The given characterization of the five representation levels for TOSCANA information systems shall now be used for explaining the discovery process from data to knowledge. This process can be seen in correspondence with the process of empirically grounded theory building proposed by A. Strauss and J. Corbin in [22] (see also [21]). According to Strauss and Corbin (p.57), empirically grounded theory building starts from data which are broken down, conceptualized, and put back together in new ways to generate a rich, tightly woven, explanatory theory that closely approximates the reality it represents. Although Strauss and Corbin are concentrating on theory building as the most systematic way of forming, synthesizing, and integrating scientific knowledge, their methodology may also apply to structuring and explaining the discovery process from data to knowledge in the more general case. This shall be outlined by means of the TOSCANA system discussed in the previous section.

The first step of breaking down the data is performed to establish the implementational level: the raw data are shaped to obtain elementary data structures which allow further formal treatments. In the case of our example, the raw data are coded in a relational database as a list of purchase transactions, each described by the ID number of the customer, the date, the department, and the purchase amount. From these data, suitable many-valued contexts are derived and represented in a data-warehouse as, for example, a many-valued context with the customers as formal objects structured by the many-valued attributes ‘department’, ‘date’, and ‘purchase amount’. Establishing one- and many-valued contexts is a first move toward a conceptualization of the data.

The next step of conceptualization is, according to Strauss and Corbin, concerned with categorization. For TOSCANA systems, categorization is performed by methods of conceptual, relational, and logical scaling which, on the logical level, are only understood formally. In Figure 2, an example of a conceptual scale is shown having formal attributes described by formal expressions which can be represented by SQL-queries in the management system TOSCANA. The apposition construction yielding the nested line diagram in Figure 3, which enlarges the attribute categorization, also belongs to the logical level.

The formal conceptualization is fully elaborated on the epistemological level. The concept lattices and the line diagrams as abstract structures are located on this level such as the formal procedures which make those lattices and diagrams to a successful support of knowledge acquisition and communication. The categorization leading to attributes of an abstract conceptual scale are now embedded into the significantly richer structure of the concept lattice of the scale which becomes human readable by a suitable line diagram. The richness of information given by such graphical representation may be seen in Figure 3; the nested structure shown in this figure reflects a subdirect product construction of the two combined concept lattices.

On the conceptual level the formal structures of the first three levels receive intensional meaning. For instance, the attribute names in Figure 1 are (on this level) understood by their literal meaning; thereby, the intensions of a represented concept can be described by combining all those meanings which belong to the attribute names attached to its superconcepts. Since the numbers in Figure 1 come from actual customers, they obtain their full meaning, discussed in Section 2, only on the linguistic level. On the conceptual level the concept lattices in Figure 1 represent a concrete scale which, according to Strauss and Corbin, may be understood as a intensionally determined dimension for the data to be analysed.

The full support for knowledge discovery is given on the linguistic level where the formal objects also carry meaning and, therefore, the formal concepts can unify intensional and extensional meaning. Of course, if further customers are considered in the presented example then the extensional meaning may change (although the intensional meaning of the concrete scales keeps the same). On this level, we can produce substantial interpretations of the data by suitable comparisions using nested line diagrams as in Figure 3; these diagrams correspond to the axial coding of Strauss and Corbin. Clearly, the rich, tightly woven, suggestive landscape of concept lattices that closely approximates the reality it represents, can serve through its representation by a TOSCANA information system, as a stimulating knowledge discovery support environment.

4 Procedures of Conceptual Knowledge Discovery

In most applications, classical *data analysis* and *decision support* facilities (for instance Online Analytical Processing (OLAP) or statistical packages) are already present when data mining tools are added to the knowledge discovery support environment. For supporting the analyst in the overall process of human-centered knowledge discovery, both decision support and data mining tools should provide a homogeneous environment. In particular, this shows the need of a *unified knowledge representation*. In conceptual information systems, concept lattices are used as such a unified knowledge representation. TOSCANA information systems have shown their use for data analysis in over 30 implementations. The relationship between conceptual information systems and Online Analytical Processing is discussed in [23].

In the first part of this section, we show how data analysis and data mining techniques based on Formal Concept Analysis may support each other. In the second part, we go one step further: there, we present CHIANTI, a new tool that integrates data mining and data analysis in the framework of Conceptual Knowledge Discovery (CKDD).

4.1 Interplay of Data Analysis and Knowledge Discovery: Association Rules and Frequent Concept Lattices

In this subsection, we discuss how Formal Concept Analysis may support the mining of association rules, and how, vice versa, results of association rules mining may be used for decreasing the complexity of the visualization of traditional data analysis within conceptual information systems. *Association rules* are statements of the type '37 % of the customers buying coffee also buy milk'. The task of mining association rules is to determine all rules that have a certain *confidence* (37 % in the example) and a certain *support* (the percentage of customers buying coffee and milk). Mining association rules can nowadays be considered as one of the core tasks of KDD. Algorithmic aspects of mining association rules within the framework of Formal Concept Analysis are discussed in more detail in [15] and [30].

Improving the mining of association rules by using Formal Concept Analysis techniques. In terms of Formal Concept Analysis, the problem is the following: Let $\mathbb{K} := (G, M, I)$ be a formal context (for instance, G could be the set of transactions registered during a certain time period in the department store, M the set of products (or items) sold by the store, and $(g, m) \in I$ means that item m was purchased in transaction g). Each subset X of M is called an *itemset*. The *support* of X is defined by $\mathrm{supp}(X) := \frac{|X'|}{|G|}$. An *association rule* $X \to Y$ consists of two subsets X and Y of M. We say that the rule $X \to Y$ holds with *support* $\mathrm{supp}(X \to Y) := \frac{|(X \cup Y)'|}{|G|}$ and with *confidence* $\mathrm{conf}(X \to Y) := \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X)}$ (in short: $X \xrightarrow{s,c} Y$ with $s := \mathrm{supp}(X \to Y)$ and $c := \mathrm{conf}(X \to Y)$). The task is now to compute, for given *minsupp*, *minconf* $\in [0, 1]$, all association rules $X \xrightarrow{s,c} Y$ with $s \geq$ *minsupp* and $c \geq$ *minconf*.

The notion of association rules and their application to large databases was introduced by R. Agrawal, T. Imielinski, and A. Swami in [1]. They stated the problem and provided a first algorithm. Now there are several algorithms for mining association rules in the literature, see for instance [15] for details.

Rules that hold only with a certain confidence have been investigated before by many researchers. For instance, in the framework of Formal Concept Analysis, M. Luxenburger [13] has called them *partial implications*. They are a generalization of *implications* which play an important role in Conceptual Data Analysis based on Formal Concept Analysis. Implications are association rules which hold for all objects but have no restriction on the support, i. e., they are exactly the association rules with $minconf = 1$ and $minsupp = 0$.

One problem in presenting the mined association rules to the user is that they usually form a long list, from which only very few are of interest to the domain expert. Using the following theorem ([15, 26]) one can reduce the list without losing any information:

Theorem. Let $X, Y \subseteq M$. Then $X \xrightarrow{s_1, c_1} Y$ and $X'' \xrightarrow{s_2, c_2} Y''$ have the same support and the same confidence.

It is based on the fact that, for any frequent itemset Y, the smallest concept intent which contains Y (i. e., Y'') has the same support and hence is also frequent. For the development of algorithms, this property permits the consideration of only concept intents (instead of all itemsets) for determining the set $\mathcal{F}$ of frequent itemsets [15, 30]. Especially in strongly correlated data, the algorithm can thereby skip many itemsets.

Using the theorem, one can present a significantly shorter list of association rules without loosing any information. The list is composed of the so-called *Duquenne-Guigues basis for exact association rules* and the *Luxenburger basis for approximate association rules*. Both bases are introduced in [30], together with algorithms for their computation.

Reducing the complexity of data visualization in conceptual information systems by using results from association rule mining. For examining cross-selling (cf. Section 2), the concepts having many attributes – and hence only relatively few objects! – are of special importance. In those cases, one needs the whole line diagram for an analysis of how well cross-selling works. But there are many applications where concepts which differentiate the population too much are not interesting – at least not for a first overview. In that situation, frequent concepts, as defined above, can be utilized. By fixing a threshold *minsupp*, all infrequent concepts of the conceptual scale can be pruned. 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 4. Two facts can be easily seen a) the birthyear of more than half the credit card customers is unknown, and b) 4690 of all credit card customers were born before 1973. Hence, there are very few customers with a known birthyear who are younger than 25 and have paid with a credit card.

4.2 Integration of Data Analysis and Knowledge Discovery: Guided Learning

In the expression *supervised learning* (as a task of *Machine Learning*), 'learning' is used in a metaphorical way. One expects the software to find an intensional

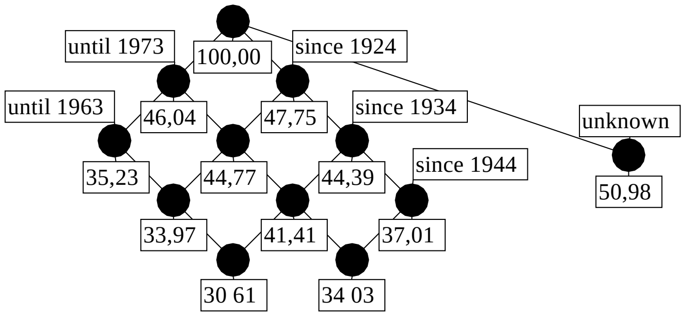

Fig. 4. Conceptual Scale 'Year of Birth' restricted to frequent concepts with $minsupp = 25\%$

description of some subpopulation, based on a training set. As CKDD is seen as a human-centered knowledge discovery process, our aim is to support the learning process (in its literal meaning) of a human expert. Human knowledge always relies on background knowledge which is formed by intersubjective argumentation, and only part of this knowledge can be expressed explicitly. Knowledge which can be made explicit may be treated by procedures of Machine Learning. But if one considers *all* aspects of knowledge, then it becomes clear that learning can only be supported by a knowledge discovery environment, but can never be completely automated.

In this setting, we understand *guided learning* as a technical support for the learning process of the human expert.[1] Guided learning shall automatically lead the user to conceptual scales (or combinations of conceptual scales) which are expected to provide interesting information, combined with the freedom of navigating around. As in supervised learning, the problem we tackle is to gain more knowledge about a given subpopulation. The difference is that we do not necessarily require an *explicit* description of the behavior. For instance, we might want to learn (in its literal meaning) more about the differences in buying behavior between high- and low-spending credit card customers.

For this purpose, we have developed the new tool Chianti, based on [24] and [25]. Chianti takes as input two subpopulations which are defined by SQL queries. In the following example, we have divided the population in two parts: those customers who spent more than 1000 SFr and those who spent less. This tool compares the distribution of the two subpopulations in all scales of the conceptual information system and returns a ranking of all scales. In the ranking, the scales which appear at the top are those where the distribution differs the most. The current implementation of Chianti provides two measures for the distribution: 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

[1] The expression 'guided learning' is also used for education and training software, but here we use it to show the analogy to supervised learning.

Skala	Wert
Xselling Housewares/Interior	0,0684
Xselling Food/Wine	0,0570
Xselling Travel Access./Perfumery/Ladies' Accessories	0,0325
Xselling Perfumery/Housewares/Food	0,0324
Xselling Perfumery/Ladies' Fashion	0,0305
Xselling Wine/Men's Fashion/Perfumery	0,0275
Xselling Ladies' Fashion/Men's Fashion/Sports	0,0229
Xselling Sports/Children/Travel Accessories	0,0160
Xselling Men's Clothing (incl. Underwear)	0,0160
Xselling Ladies' Wear	0,0123
Xselling Men's City	0,0009

Fig. 5. Ranking of conceptual scales related to cross-selling

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. At the moment, CHIANTI only works on the contingents (this means that, for the measure, the cardinality of the concept extents is not used, only the number of objects which generate the concept). As the difference of the distributions of the two populations may be more significant in more general concepts (which are not necessarily generated by single objects), the next version of CHIANTI will also analyze concept extents.

Figure 5 shows the ranking of all scales related to cross-selling for the two subpopulations mentioned above with 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 3. This means that among all cross-selling scales, this scale differentiates the two groups the most. The scale 'Cross-selling Housewares/Interior' also appears as topmost scale in the ranking according to the maximum norm.

By combining the topmost scales with the scale 'Money spent $\leq$ / $>$ 1000 SFr' we can analyze the distribution of the two groups in more detail. The combination of this scale together with the scale 'Cross-selling Housewares/Interior' is shown in Figure 6. In the diagram, we have set the top element of each inner scale to 100% in order to facilitate comparison. We see that the high-spending customers buy over-proportionally in the departments Housewares (265% more often) and Interior (322% more often). Furthermore, for this customer group, the cross-selling between both departments is much higher than for the rest: The percentage of high-spending customers who were active in both interior and housewares (36.98%) is much greater than that of low-spending customers (5.56%).

We emphasize that — unlike many other statistical techniques — the ranking of the scales is not the final result, but a suggestion to the analyst of certain combination of scales for analyzing the situation in more detail. The ranking alone does not indicate that the buying behavior in the housewares 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

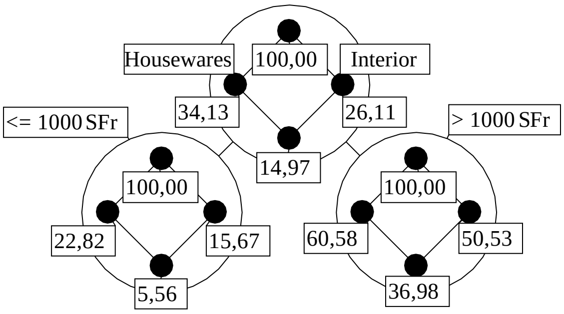

Fig. 6. Customers of the Housewares department differentiated by the amount of money spent

consequence of the different distribution, as is clearly demonstrated by studying the ranking of all the 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 spends less than 1000 SFr in a single transaction; hence to spend more money, he has to visit the department store more than once. The next scale then is the scale 'Cross-selling Housewares/Interior'.

The insight that the scale about the active time is not useful for this kind of analysis can only be gained by referring to the implicit background knowledge 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 conceived of in advance. However, it is promising for further research to consider such a repository which 'learns' (in the metaphorical meaning) from the behavior of the analyst which combinations are of interest and which are not.

References

1. R. Agrawal, T. Imielinski, A. Swami: Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD*, 1993.
2. H. Berg: *Terminologische Begriffslogik.* Diplomarbeit. FB Mathematik, TU Darmstadt 1997.
3. R. J. Brachman: On the epistemological status of semantic networks. In: N. V. Findler (ed.): *Associative networks: representation and use of knowledge by computers.* Academic Press, New York 1979, 3–50.
4. R. J. Brachman, T. Anand: The process of knowledge discovery in databases. In [7]
5. 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.

6. J.-L. Guigues, V. Duquenne: Familles minimales d' implications informatives resultant d' un tableau de données binaires. *Math. Sci. Humaines* **95**, 1986, 5–18.
7. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.): *Advances in Knowledge Discovery and Data Mining.* AAAI/MIT Press, Cambridge 1996.
8. 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.
9. B. Ganter, R. Wille: Conceptual scaling. In: F. Roberts (ed.): *Applications of combinatorics and graph theory to the biological and social sciences.* Springer, Berlin-Heidelberg-New York 1989, 139–167.
10. B. Ganter, R. Wille: *Formal Concept Analysis: Mathematical Foundations.* Springer, Berlin-Heidelberg 1999 (Translation of: *Formale Begriffsanalyse: Mathematische Grundlagen.* Springer, Berlin-Heidelberg, 1996).
11. B. Ganter, R. Wille: Contextual Attribute Logic. *Proc. ICCS '99*, LNAI **1640**, Springer, Heidelberg 1999, 377–388
12. P. Luksch, R. Wille: A mathematical model for conceptual knowledge systems. In: H.-H. Bock, P. Ihm (eds.): *Classification, data analysis, and knowledge organization.* Springer, Berlin-Heidelberg 1991, 156–162.
13. M. Luxenburger: Implications partielles dans un contexte. *Mathématiques, informatique et sciences humaines* **113**, 1991, 35–55.
14. G. Mineau, G. Stumme, R. Wille: Conceptual Structures Represented by Conceptual Graphs and Formal Concept Analysis. *Proc. ICCS '99*, LNAI **1640**. Springer, Heidelberg 1999, 423–441
15. N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal: Efficient mining of association rules using closed itemset lattices. *Journal of Information systems*, **24** (1999), 25–46.
16. Ch. S. Peirce: *Collected Papers.* Harvard University Press, Cambridge 1931-35.
17. 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-New York 1997, 332–341.
18. S. Prediger: Terminologische Merkmalslogik in der Formalen Begriffsanalyse. In: G. Stumme, R. Wille (eds.): *Begriffliche Wissensverarbeitung: Methoden und Anwendungen.* Springer, Berlin-Heidelberg 2000, 99–124.
19. S. Prediger, G. Stumme: Theory-Driven Logical Scaling. *Proc. KRDB '99.* (Also in *Proc. DL '99*). CEUR Workshop Proc. **21+22**, 1999 (http://SunSITE.Informatik.RWTH-Aachen.DE/Publications/CEUR-WS/)
20. S. Prediger, R. Wille: The lattice of concept graphs of a relationally scaled context. FB4-Preprint, TU Darmstadt 1999.
21. S. Strahringer, R. Wille, U. Wille: Mathematical support for empirical theory building. FB4-Preprint, TU Darmstadt 1999.
22. A. Strauss, J. Corbin: *Basics of qualitative research: grounded theory procedures and techniques.* Sage Publ., Newbury Park 1990.
23. G. Stumme: 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)
24. 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.
25. G. Stumme: Dual Retrieval in Conceptual Information Systems. in: A. P. Buchmann (ed.): *Datenbanksysteme in Büro, Technik und Wissenschaft.* Springer, Heidelberg 1999, 328–342

26. G. Stumme: Conceptual Knowledge Discovery with Frequent Concept Lattices. FB4-Preprint, TU Darmstadt 1999
27. 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, Heidelberg 1998, 450–458.
28. F. Vogt: *Formale Begriffsanalyse mit C++: Datenstrukturen und Algorithmen.* Springer, Berlin–Heidelberg–New York 1996.
29. F. Vogt, R. Wille: TOSCANA – a graphical tool for analyzing and exploring data. In: R. Tamassia, I. G. Tollis (eds.): *Graph Drawing '94.* Lecture Notes in Computer Science 894. Springer, Berlin-Heidelberg-New York 1995, 226-233.
30. R. Taouil, Y. Bastide, N. Pasquier, G. Stumme, L. Lakhal: Mining bases for association rules based on Formal Concept Analysis. *Proc. ECAI 2000* (submitted)
31. R. Wille: Concept Lattices and Conceptual Knowledge Systems. *Computers & Mathematics with Applications*, *23*, 1992, 493-515.
32. 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.
33. R. Wille: Plädoyer für eine philosophische Grundlegung der Begrifflichen Wissensverarbeitung. In: R. Wille, M. Zickwolff (eds.): *Begriffliche Wissensverarbeitung: Grundfragen und Aufgaben.* B.I.-Wissenschaftsverlag, Mannheim 1994, 11–25.
34. R. Wille: Begriffsdenken: Von der griechischen Philosophie bis zur Künstlichen Intelligenz heute. Dilthey-Kastanie, Ludwig-Georgs-Gymnasium Darmstadt 1995, 77–109.
35. R. Wille: Conceptual Landscapes of Knowledge: A Pragmatic Paradigm for Knowledge Processing. In: *Proc. of KRUSE '97.* Vancouver, August 11–13, 1997, 2–13.