Intelligent Structuring and Reducing of Association Rules with Formal Concept Analysis

Gerd Stumme,[1] Rafik Taouil,[2] Yves Bastide,[3] Nicolas Pasquier,[4] Lotfi Lakhal[5]

[1] Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB), Universität Karlsruhe (TH), D–76128 Karlsruhe, Germany; stumme@aifb.uni-karlsruhe.de
[2] INRIA Lorraine, LORIA, BP 239, F–54506 Vandœuvre–lès–Nancy, France; rafik.taouil@loria.fr
[3] Laboratoire d'Informatique (LIMOS), Université Blaise Pascal, Complexe Scientifique des Cézeaux, 24 Av. des Landais, F–63177 Aubière Cedex, France; bastide@libd2.univ-bpclermont.fr
[4] I3S, CNRS UMR–6070, Université de Nice – Sophia Antipolis, Les Algorithmes, 2000 route des Lucioles, F–06903 Sophia Antipolis, France; nicolas.pasquier@unice.fr
[5] LIM, CNRS FRE-2246, Université de la Méditerranée, Case 90, 163 Avenue de Luminy, F–13288 Marseille Cedex 9, France; lotfi.lakhal@lim.univ-mrs.fr

Abstract. Association rules are used to investigate large databases. The analyst is usually confronted with large lists of such rules and has to find the most relevant ones for his purpose. Based on results about knowledge representation within the theoretical framework of Formal Concept Analysis, we present relatively small bases for association rules from which all rules can be deduced. We also provide algorithms for their calculation.[1]

1 Introduction and Motivation

One of the core tasks of *Knowledge Discovery in Databases* (*KDD*) is the mining of association rules (conditional implications). *Association rules* are statements of the type '67 % of the customers buying cereals and sugar also buy milk (where 7% of all customers buy all three items)'. The task of mining association rules is to determine all rules whose *confidences* (67 % in the example) and *supports* (7 % in the example) are above user-defined thresholds. Since the problem was stated [1], various approaches have been proposed for an increased efficiency of rule discovery in very large databases [2, 7, 11, 30, 31]. However, fully taking advantage of exhibited rules means providing capabilities to handle them. The problem is especially critical when collected data is highly correlated or dense, like in statistical databases [11]. For instance, when applied to a census dataset of 10,000 objects, each of which characterized by values of 73 attributes, experiments result in more then 2,000,000 rules with support and confidence greater

[1] This paper is a revised and extended version of a presentation given at the workshop "Bases de Données Avancées", Bordeaux, France, 1999 [29], and of the technical report [37].

than or equal 90%. Thus the question arises: How can long lists of association rules be reduced in size?

Approaches addressing the described issue provide users with mechanisms for filtering rules, for instance by user defined templates [4, 21], Boolean [26, 35] or SQL-like [25] operators or by introducing further measures of "usefulness" [8]; or they attempt to minimize the number of extracted rules a priori by using information about taxonomies [17, 19, 34] or by applying statistical measures like Pearson's correlation or the χ^2-test [10]. All these approaches have in common that they lose some information.

Our approach, on the other hand, allows us to significantly reduce the number of rules without losing any information. We extract only a subset of all association rules, called *basis*, from which all other rules can be derived. This approach is orthogonal to the ones mentioned above and can be combined with them.

We make use of techniques of *Formal Concept Analysis* (FCA). Formal Concept Analysis [41, 15] arose as a mathematical theory for the formalization of the concept of 'concept' in the early 80ies and is nowadays considered as an AI theory. It has since then grown to a technique for data analysis, information retrieval, and knowledge representation with over 200 applications, for analyzing flight movements at Frankfurt Airport [20], for studying semantics of German speech-act verbs [16], for examining the medical nomenclature system SNOMED [33], for IT-security management [9], and for database marketing [18]. FCA provides a framework for KDD, especially for conceptual clustering and association rules. A broad discussion of the role of Formal Concept Analysis in data analysis, decision support, and KDD is provided in [18] and [36].

We use results of Duquenne and Guigues ([12], cf. also [15]) and Luxenburger [22, 23]. The former have studied bases (i. e., minimal non-redundant sets of rules from which all other rules can be derived) for association rules with 100 % confidence, and the latter association rules with less than 100 % confidence, but neither of them considered the support of the rules. We adopt their results to association rules (where both the support and the confidence are considered) and provide algorithms for computing the new bases by using *iceberg concept lattices* [39]. We follow an approach in two steps. In the first step, we compute the iceberg concept lattice for the given parameters. It consists of all FCA concepts whose extents exceed the user-defined minimum support. In the second step, we derive the bases for the association rules. In this paper, we focus on the second step. For the first step, we refer to the Pascal [6] and Titanic [38] algorithms.

This two-step approach has two advantages compared to the classical two-step approach [2] (which computes all frequent itemsets as intermediate result, and not only those which are intents of frequent FCA concepts):

1. It allows to determine bases for non-redundant association rules and thus to prune redundancy.
2. It speeds up the computation, especially for strongly correlated data or when the minimum support is low.

In [5], we have presented another pair of bases, which provide rules with minimal antecedents and maximal consequents. Compared to the results presented here, they have the disadvantage of a higher total number of rules. For the approximate rules, M. Zaki has presented similar results in [44]. However, he does not provide inference rules for support and confidence derivation, does not discuss minimality of his results, and does not provide algorithms for the computation of the bases.

The remainder of this paper is as follows. After having recalled some basic definitions in Section 2, we introduce two bases for association rules in Section 3: the *Duquenne-Guigues basis for exact association rules* (i. e., for all rules with a 100% confidence), and the *Luxenburger basis for approximate association rules* (i. e., with a confidence < 100%). In Section 4, algorithms are given which compute the two bases. We conclude the paper with the presentation of experimental results (Section 5) and a discussion of future work (Section 6).

2 Formal Concept Analysis and the Association Rule Framework

In this section, we briefly recall the basic notions of Formal Concept Analysis [41, 15] and the association rule problem [1]. For a more extensive introduction into Formal Concept Analysis refer to [15].

Definition 1. *A* formal context *is a triple* $\mathbb{K} := (G, M, R)$ *where* G *and* M *are sets and* $R \subseteq G \times M$ *is a binary relation. A* data mining context *(or* dataset*) is a formal context where* G *and* M *are finite sets. Its elements are called* objects *and* items*, respectively.* $(o, i) \in R$ *is read as "object* o *is related to item* i*".*

For $O \subseteq G$*, we define* $f(O) := \{i \in M \mid \forall o \in O\colon (o, i) \in R\}$*; and for* $I \subseteq M$*, we define dually* $g(I) := \{o \in G \mid \forall i \in I\colon (o, i) \in R\}$*. A* formal concept *is a pair* $(O, I) \in \mathfrak{P}(G) \times \mathfrak{P}(M)$ *with* $f(O) = I$ *and* $g(I) = O$*.* O *is called* extent *and* I *is called* intent *of the concept. The set of all concepts of a formal context* $\mathbb{K}$ *together with the partial order* $(O_1, I_1) \leq (O_2, I_2) :\Longleftrightarrow O_1 \subseteq O_2$ $(\Longleftrightarrow I_2 \subseteq I_1)$ *is a complete lattice, called* concept lattice *of* $\mathbb{K}$*.*

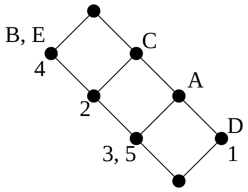
In this setting, we call each subset of M *also* itemset*, and each intent* I *also* closed itemset *(since it satisfies the equation* $I = f(g(I))$*). For two closed itemsets* I_1 *and* I_2*, we note* $I_1 \prec I_2$ *if* $I_1 \subset I_2$ *and if there does not exist a closed itemset* I_3 *with* $I_1 \subset I_3 \subset I_2$*.*[2]

In the following, we will use the composed function $h := f \circ g\colon \mathfrak{P}(M) \to \mathfrak{P}(M)$ which is a closure operator on M (i. e., it is extensive, monotonous, and idempotent). The related closure system (i. e., the set of all $I \subseteq M$ with $h(I) = I$) is exactly the set of the intents of all concepts of the context.

Definition 2. *Let* $I \subseteq M$*, and let minsupp, minconf* $\in [0, 1]$*. The* support count *of the itemset* I *in* $\mathbb{K}$ *is supp*$(I) := \frac{|g(I)|}{|G|}$*.* I *is said to be* frequent *if supp*$(I) \geq$*minsupp. The set of all frequent itemsets of a context is denoted* FI*.*

[2] We write $X \subset Y$ if and only if $X \subseteq Y$ and $X \neq Y$.

	A	B	C	D	E
1	×		×	×	
2		×	×		×
3	×	×	×		×
4		×			×
5	×	×	×		×



Exact rule	Supp	Approximate rule	Supp	Conf	Approximate rule	Supp	Conf	Approximate rule	Supp	Conf
ABC → E	0.4	BCE → A	0.4	2/3	C → ABE	0.4	2/4	C → BE	0.4	3/4
ABE → C	0.4	AC → BE	0.4	2/3	E → ABC	0.4	2/4	E → BC	0.6	3/4
ACE → B	0.4	BE → AC	0.4	2/4	A → BC	0.4	2/3	A → B	0.4	2/3
AB → CE	0.4	CE → AB	0.4	2/3	B → AC	0.4	2/4	B → A	0.4	2/4
AE → BC	0.4	AC → B	0.4	2/3	C → AB	0.4	2/4	C → A	0.6	3/4
AB → C	0.4	BC → A	0.4	2/3	A → BE	0.4	2/3	A → E	0.4	2/3
AB → E	0.4	BE → A	0.4	2/4	B → AE	0.4	2/4	E → A	0.4	2/4
AE → B	0.4	AC → E	0.4	2/3	E → AB	0.4	2/4	C → B	0.6	3/4
AE → C	0.4	CE → A	0.4	2/3	A → CE	0.4	2/3	C → E	0.6	3/4
BC → E	0.6	BE → C	0.6	3/4	C → AE	0.4	2/4	E → C	0.6	3/4
CE → B	0.6	A → BCE	0.4	2/3	E → AC	0.4	2/4	∅ → E	0.8	4/5
A → C	0.6	B → ACE	0.4	2/4	B → CE	0.6	3/4	∅ → BE	0.8	4/5
B → E	0.8	∅ → C	0.8	4/5	∅ → AC	0.6	3/5	∅ → BC	0.6	3/5
E → B	0.8	∅ → A	0.6	3/5	∅ → B	0.8	4/5	∅ → CE	0.6	3/5
								∅ → BCE	0.6	3/5

Fig. 1. The example data mining context $\mathbb{K}$ and its concept lattice. The table shows all association rules that hold in $\mathbb{K}$ for *minsupp* $= 0.4$ and *minconf* $= 1/2$.

An association rule *is a pair of itemsets I_1 and I_2, denoted $I_1 \to I_2$, where $I_2 \neq \emptyset$. I_1 and I_2 are called* antecedent *and* consequent *of the rule, respectively. The* support *and* confidence *of an association rule $r := I_1 \to I_2$ are defined as follows: $supp(r) := \frac{|g(I_1 \cup I_2)|}{|G|}$, $conf(r) := \frac{supp(I_1 \cup I_2)}{supp(I_1)}$. If conf(r)=1, then r is called* exact association rule *(or* implication*), otherwise r is called* approximate association rule.

An association rule r holds *in the context if $supp(r) \geq$ minsupp and $conf(r) \geq$ minconf. The set of all association rules holding in $\mathbb{K}$ for given minsupp and minconf is denoted AR.*

Remark 1. The definition of association rules often includes the additional condition $I_1 \cap I_2 = \emptyset$. This condition helps pruning rules which are obviously redundant, as $I_1 \to I_2$ and $I_1 \to I_2 \setminus I_1$ have same support and same confidence. In this paper, we omit the condition, in order to simplify definitions. When discussing the algorithms, however, we will use the condition since it saves memory.

The association rule framework has first been formulated in terms of Formal Concept Analysis independently in [28], [37], and [42]. [28] provided also the first algorithm (named Close) based on this approach.

Example 1. An example data mining context $\mathbb{K}$ consisting of five objects (identified by their OID) and five items is given in Figure 1 together with its concept lattice. The association rules holding for *minsupp* $= 0.4$ and *minconf* $= 1/2$ are shown in the lower table.

In the *line diagram*, the name of an object g is always attached to the node representing the smallest concept with g in its extent; dually, the name of an attribute m is always attached to the node representing the largest concept with

m in its intent. This allows us to read the context relation from the diagram because an object g has an attribute m if and only if there is an ascending path from the node labeled by g to the node labeled by m. The extent of a concept consists of all objects whose labels are below in the diagram, and the intent consists of all attributes attached to concepts above in the hierarchy. For example, the concept labeled by 'A' has $\{1,3,5\}$ as extent, and $\{\mathrm{A},\mathrm{C}\}$ as intent.

An example for an exact rule (implication) which holds in the context is $\{\mathrm{A},\mathrm{B}\} \rightarrow \{\mathrm{C},\mathrm{E}\}$. It can also be read directly in the line diagram: the largest concept having both A and B in its intent is the one labeled by 3 and 5, and it is below or equal to (here the latter is the case) the largest concept having both C and E in its intent. This implication can be derived from two simpler implications, namely $\{\mathrm{A}\} \rightarrow \{\mathrm{C}\}$ and $\{\mathrm{B}\} \rightarrow \{\mathrm{E}\}$. The aim of the frequent Duquenne-Guigues-basis which we introduce in the next section is to provide only a minimal, non-redundant set of implications to the user. That basis will include the two simpler implications.

At the end of this section, we give some simple facts about association rules. We will refer to them later as derivation rules.

Lemma 1. *Rules 1 and 2 hold for* $\phi \in \{\mathrm{conf},\mathrm{supp}\}$.

1. $\phi(X \rightarrow Y) = \phi(X \rightarrow Y \setminus Z)$, *for all* $Z \subseteq X \subseteq M$, $Y \subseteq M$.

2. $\phi(h(X) \rightarrow h(Y)) = \phi(X \rightarrow Y)$, *for all* $X, Y \subseteq M$.

3. $\mathrm{conf}(X \rightarrow Y) = p \wedge \mathrm{conf}(Y \rightarrow Z) = q \Longrightarrow \mathrm{conf}(X \rightarrow Z) = p \cdot q$, *for all frequent concept intents* $X \subset Y \subset Z$.

3'. $\mathrm{supp}(X \rightarrow Z) = \mathrm{supp}(Y \rightarrow Z)$, *for all* $X, Y \subseteq Z$.

4. $\mathrm{conf}(X \rightarrow X) = 1$, *for all* $X \subseteq M$.

Proof. The proofs for the confidence are given in [23].

1. $\mathrm{supp}(X \rightarrow Y) = \mathrm{supp}(X \rightarrow Y \setminus Z)$ follows from $X \cup Y = X \cup (Y \setminus Z)$ and the definition of the support count.
2. $\mathrm{supp}(h(X) \rightarrow h(Y)) = \mathrm{supp}(X \rightarrow Y)$ follows from $g(h(X) \cup h(Y)) = g(h(X)) \cap g(h(Y)) = g(f(g(X))) \cap g(f(g(Y))) = g(X) \cap g(Y) = g(X \cup Y)$ by using the facts $g(f(g(X))) = g(X)$ and $g(X \cup Y) = g(X) \cap g(Y)$ provided in [15].

3'. $\mathrm{supp}(X \rightarrow Z) = \frac{|g(X \cup Z)|}{|G|} = \frac{|g(Z)|}{|G|} = \frac{|g(Y \cup Z)|}{|G|} = \mathrm{supp}(Y \rightarrow Z)$ □

3 Bases for Association Rules

In this section, we recall the definition of *iceberg concept lattices* and show that one can derive all frequent itemsets and association rules from them. Then we characterize the *Duquenne-Guigues basis for exact association rules* and the *Luxenburger basis for approximate association rules* and show that all other association rules can be derived from these two bases.

Frequent closed itemset	support
$\emptyset$	1.0
{C}	0.8
{AC}	0.6
{BE}	0.8
{BCE}	0.6
{ABCE}	0.4

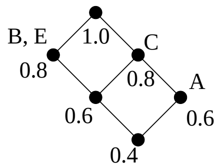


Fig. 2. Frequent closed itemsets extracted from $\mathbb{K}$ for *minsupp* $= 0.4$.

Definition 3. *A concept* (O, I) *is called* frequent concept *if* $supp(I)\,(= \frac{|O|}{|G|}) \geq$ *minsupp. The set of all frequent concepts is called* iceberg concept lattice. *An itemset* I *is called* frequent intent *(or* frequent closed itemset*) if it is intent of a frequent concept (i. e., its support is at least minsupp). The set of all frequent closed itemsets in* $\mathbb{K}$ *is denoted* FC.

Example 2. The frequent closed itemsets in the context $\mathbb{K}$ for *minsupp*=0.4 are presented in Figure 2 together with the semi-lattice of all frequent concepts. Both the table and the diagram provide the same information. Note that, in general, the set of frequent concepts is not a lattice, but only a semi-lattice (consider e. g. *minsupp*= 0.5 in the example).

Lemma 2 ([31]). *i) The support of an itemset* I *is equal to the support of the smallest closed itemset containing* I*, i. e.,* $supp(I) = supp(h(I))$.

ii) The set of maximal frequent itemsets $\{I \in FI \mid \nexists I' \in FI\colon I \subset I'\}$ *is identical to the set of maximal frequent closed itemsets* $\{I \in FC \mid \nexists I' \in FC\colon I \subset I'\}$.

The next theorem shows that the set of frequent closed itemsets with their support is a small collection of frequent itemsets from which all frequent itemsets with their support and all association rules can be derived. I. e., it is a condensed representation in the sense of Mannila and Toivonen [24]. This theorem follows from Lemma 2.

Theorem 1. *All frequent itemsets and their support, as well as all association rules holding in the dataset, their support, and their confidence can be derived from the set* FC *of frequent closed itemsets with their support.*

3.1 Duquenne-Guigues Basis for Exact Association Rules

Next we present the Duquenne-Guigues basis for exact association rules. It is based on the following closure operator.

Theorem 2. *The set* $FI \cup \{M\}$ *is a closure system on* M*, and its related closure operator* $\overline{\cdot}$ *is given by* $\overline{I} := h(I)$ *if* $supp(I) \geq$*minsupp and* $\overline{I} := M$ *else.*

Proof. The set of all frequent itemsets together with M is a closure system, as well as the set of all concept intents. Hence $FI \cup \{M\}$ is, as intersection of those two closure systems, also a closure system. The proof of the fact that $\overline{\cdot}$ is the corresponding closure operator is straightforward. $\square$

Our basis adopts the results of [12] to the association rule framework, where additionally the support of the rules has to be considered.

Definition 4. *An itemset $I \subseteq M$ in $\mathbb{K}$ is a* $\overline{\cdot}$–pseudo-closed itemset *(or* pseudo-closed itemset *for short)* [3] *if $\overline{I} \neq I$ and for all pseudo-closed itemsets J with $J \subset I$, we have $\overline{J} \subset I$. The set of all frequent pseudo-closed itemsets in $\mathbb{K}$ is denoted FP, the set of all infrequent pseudo-closed itemsets is denoted IP. In the (unlikely) case that all itemsets are frequent except the whole set M, we let $IP := \{M\}$ (in order to distinguish this situation from the one where all itemsets are frequent).*

The Duquenne-Guigues basis for exact association rules *(or* frequent Duquenne-Guigues basis*) is defined as the tuple $FDG := (\mathcal{L}, IP)$ with $\mathcal{L} := \{I_1 \to h(I_1) \mid I_1 \in FP\}$ and IP as defined above.*

Theorem 3. *From the Duquenne-Guigues basis for exact association rules one can derive all exact association rules holding in the dataset by applying the following rules. Rules ii) to iv) can be applied to $\mathcal{L}$ as long as they do not contradict (i).*

i) If there exists $I \in IP$ with $I \subseteq I_1 \cup I_2$, then $I_1 \to I_2$ does not hold (because its support is too low).
ii) $X \to X$ holds.
iii) If $X \to Z$ holds, then also $X \cup Y \to Z$.
iv) If $X \to Y$ and $Y \cup Z \to W$ hold, then also $X \cup Z \to W$.

Proof. We only sketch the proof here, which applies results of [12] (see also [15]). One has to check that $\mathcal{L} \cup \{I \to M \mid I \in IP\}$ is the Duquenne-Guigues-basis (in the traditional sense, cf. to [12, 15]) of the closure system $FC \cup \{M\}$. Rule (i) reflects the implications of the form $I \to M$. $\square$

The Duquenne-Guigues basis for exact association rules is not only minimal with respect to set inclusion, but also minimal with respect to the number of rules in $\mathcal{L}$ plus the number of elements in IP, since there can be no complete set with fewer rules than there are frequent pseudo-closed itemsets [12, 15]. Observe that, although it is possible to derive all exact association rules from the Duquenne-Guigues basis, it is not possible in general to determine their support.[4]

Example 3. The set of frequent pseudo-closed itemsets of $\mathbb{K}$ for *minsupp*=0.4 and *minconf*=1/2 is $FP = \{\{A\}, \{B\}, \{E\}\}$, the set of infrequent pseudo-closed itemsets is $IP = \{\{D\}\}$. The Duquenne-Guigues basis is presented in Figure 3.

[3] We do not consider pseudo-closed itemsets with respect to other closure operators than $\overline{\cdot}$ (especially not with respect to h) in this paper.

[4] Even if the support for all rules in the basis is known. With the knowledge about all frequent closed itemsets and their support however, this is possible (see Theorem 1).

3.2 Luxenburger Basis for Approximate Association Rules

In [22, 23], M. Luxenburger discusses bases for partial implications. A *partial implication* is an association rule where the support is not considered. He observed that it is sufficient to consider rules between concept intents only, since $\mathrm{conf}(X \to Y) = \mathrm{conf}(h(X) \to h(Y))$. However, his derivation process does not only consist of deduction rules which can be applied in a straightforward manner, but it requires to solve a system of linear equations.

In the KDD process, however, we have to consider the trade-off between the amount of information presented to the user, and the degree of its explicitness. The appearance of the system of linear equations indicates that Luxenburger's results are in favor for a minimal amount of information presented, and against a higher degree of explicitness. As one of the requirements to KDD is that the results should be "ultimately understandable" [13], we want to emphasize more on the explicitness of the results. Therefore we restrict now the expressiveness of the derivation process. This forces the association rules presented to the user to be more explicit.[5]

In the sequel, we consider the derivation rules given in Lemma 1. We present a basis for the approximate association rules for these derivation rules.

Definition 5. *The* Luxenburger basis for approximate association rules *is given by* $LB := \{(r, supp(r), conf(r)) \mid r = I_1 \to I_2,\ I_1, I_2 \in FC,\ I_1 \prec I_2,\ conf(r) \geq minconf,\ supp(I_2) \geq minsupp\}$.

Theorem 4. *From the Luxenburger basis LB for approximate association rules one can derive all association rules holding in the dataset together with their support and their confidence by using the rules given in Lemma 1. Furthermore, LB is minimal (with respect to set inclusion) with this property.*

Proof. In order to determine if an association rule $r := I \to J$ holds in a context (and for determining its support and its confidence) one can consider the rule $I' \to J'$ with $I' := h(I)$ and $J' := h(I \cup J)$ which has (by Rules 1 & 2) the same support and the same confidence. If $I' = J'$, then $conf(r) = 1$ and $supp(r) = \mathrm{supp}(I')$. If $I' \neq J'$, then exists a path of approximate rules, i. e., there are frequent closed itemsets $I_1, \ldots, I_n$ with $I_i \to I_{i+1} \in LB$ and $I' = I_1$ and $I_n = J'$. Support and confidence of r can now be determined by $supp(r) = supp(I_n)$ (Rule 3') and $conf(r) = \Pi_{i=1}^{n-1} conf(I_i \to I_{i+1})$ (Rule 3).

Now we show the minimality of LB. Let $r := I \to J \in LB$. We show that the confidence of r cannot be derived from $LB \setminus \{r\}$ by applying the rules of Lemma 2. Rule 1 cannot be applied forward since J already contains I. It cannot be applied backward because of $I \prec J$. Rule 2 cannot be applied forward since $I = h(I)$ and $J = h(J)$. It cannot be applied backward as LB contains only rules with closed antecedent and closed consequent. Rule 3 cannot be applied since there is no $K \subset M$ with $I \to K \in LB \setminus \{r\}$ and $K \to J \in LB \setminus \{r\}$ (because of $I \prec J$). Rule 4 cannot be applied since $I \neq J$. □

[5] Note that in the KDD setting the user will never actually perform longer series of inference steps.

Luxenburger basis		
Approximate rule	Support	Confidence
BCE → A	0.4	2/3
AC → BE	0.4	2/3
BE → C	0.6	3/4
C → BE	0.6	3/4
C → A	0.6	3/4
∅ → BE	0.8	4/5
∅ → C	0.8	4/5

Duquenne-Guigues basis	
$\mathcal{L}$	(Support)
A → C	0.6
B → E	0.8
E → B	0.8
$IP = \{\{D\}\}$	

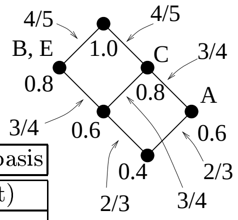


Fig. 3. Duquenne-Guigues and Luxenburger bases for *minsupp*=0.4 and *minconf*=1/2.

Remark 2. A basis in the sense of [23] is a maximal spanning tree of our basis (when considered as *undirected* graph) containing at most one rule with M as conclusion.[6]

Example 4. The Luxenburger basis for approximate association rules of $\mathbb{K}$ for *minsupp*=0.4 and *minconf*=1/2 is also presented in Figure 3. It provides the same information as the list in Figure 1, but in a more condensed form. The Luxenburger basis is visualized in the line diagram in Figure 3: From its definition it is clear, that each approximate rule in the basis corresponds to (at most)[7] one edge in the diagram. The edge is labeled by the confidence of the rule (as a fraction), and its lower vertice is labeled by its support (as a rational). Implications (exact rules) can be read in the diagram in the standard way described in Section 2.

As example for the proof of Theorem 4, let us check if $\{B\} \to \{A\}$ holds in the context for *minsupp*=0.4 and *minconf*=1/2. We have $I := \{B\}$ and $J := \{A\}$. The smallest frequent closed itemset containing B is $I' := \{B, E\}$ and the smallest one containing A and B is $J' := \{A, B, C, E\}$. In the diagram, I' and J' are always represented by the largest concepts which are below all attributes in I and $I \cup J$, resp. Between the two concepts we find the path $I_1 := I'$, $I_2 := \{B, C, E\}$, and $I_3 := J'$. Hence $supp(B \to A) = supp(J') = 0.4 \geq minsupp$ and $conf(B \to A) = conf(I_1 \to I_2) \cdot conf(I_2 \to I_3) = 3/4 \cdot 2/3 = 2/4 \geq minconf$, which means that the rule holds.

4 Algorithms for Computing the Bases

The algorithms presented in this paper assume that the iceberg concept lattice is already computed. There are several algorithms for computing iceberg concept

[6] The second condition is negligible in KDD, as it follows directly from minsupp > 0 %.

[7] In general, there may be edges which do not represent any rule in the Luxenburger basis. Consider for instance *minconf*= 7/10. In this case, the two lowest edges would not stand for a valid approximate rule, and would hence not be labelled.

Algorithm 1 Generating the Duquenne-Guigues basis with Gen-FDG.

1) $\mathcal{L} \leftarrow \{\}$;
2) **if** $(FC_0 = \{\})$ **then** $FP_0 \leftarrow \emptyset$;
3) **else** $FP_0 \leftarrow \{\}$;
4) **for** $(i \leftarrow 1;\ i \leq k;\ i{+}{+})$ **do begin**
5) $FP_i \leftarrow FI_i \setminus FC_i$;
6) **forall** $L \in FP_i$ **do begin**
7) $pseudo \leftarrow true$;
8)   **forall** $P \in FP_j$ with $j < i$ **do begin**
9)   **if** $(P \subset L)$ **and** $(P.\text{closure} \not\subseteq L)$
10)   **then do begin**
11) $pseudo \leftarrow false$;
12) $FP_i \leftarrow FP_i \setminus \{L\}$;
13)   **endif**
14)   **end**
15)   **if** $(pseudo = true)$ **then** $L.\text{closure} \leftarrow \min_{\subseteq}(\{C \in FC_{j>i} \mid L \subseteq C\})$;
16) **end**
17) **end**
18) **forall** $P \in \bigcup_{i=1}^{n} FP_i$ **do** $\mathcal{L} \leftarrow \mathcal{L} \cup \{P \rightarrow (P.\text{closure} \setminus P)\}$;
19) $IP \leftarrow \emptyset$;
20) **forall** $L \in MI$ **do** $IP \leftarrow IP \cup \{\mathcal{L}^*\text{-closure}(I)\}$;
21) $IP \leftarrow \min_{\subseteq} IP$;

lattices: the algorithm Close for strongly correlated data [31], the algorithm A-Close for weakly correlated data [30], the algorithms CLOSET [32], ChARM [43], and TITANIC [38,39]. The algorithm PASCAL [6] computes all (closed and non-closed) frequent itemsets, but can be upgraded to determine also their closures with almost no additional computation time by using the fact that, for $I \subseteq M$,

$$h(I) = I \cup \{m \in M \setminus I \mid supp(I) = supp(I \cup \{m\})\} \ .$$

When the iceberg concept lattice is computed, then the Duquenne–Guigues basis and finally the Luxenburger basis are computed.

4.1 Generating the Duquenne-Guigues basis for Exact Association Rules with Gen-FDG

In this section, we present an algorithm that determines the Duquenne–Guigues basis using the iceberg concept lattice. This algorithm (which has not been presented before) implements Definition 4. As it needs to know the closure of frequent itemsets, it is best applied after an algorithm like PASCAL with the modification mentioned above, ChARM, or CLOSET.

The pseudo-code is given in Algorithm 1. The algorithm takes as input the sets FI_i, $1 \leq i \leq k$, containing the frequent itemsets and their support, and the sets FC_i, $0 \leq i \leq k$, containing the frequent closed itemsets and their support. It first computes the frequent pseudo-closed itemsets iteratively (steps 2 to 17). In steps 2 and 3, the empty set is examined. (It must be either a closed or a

Algorithm 2 Function $\mathcal{L}^*$-closure reads X and returns its $\mathcal{L}^*$-closure $\mathcal{L}^*(X)$.

$Y \leftarrow X$;
for $(i \leftarrow 1; i = n; i++)$ **do** i.used$\leftarrow$ *false*;
repeat
    changed $\leftarrow$ *false*;
    If Subsets$(IP, Y) \neq \emptyset$ **then begin** $Y \leftarrow M$; changed$\leftarrow$ *true* **end**
    else for $(i \leftarrow 1; i \leq n; i++)$ **do**
            if $X_i \subset Y$ **then begin** $Y \leftarrow Y \cup Y_i$; changed$\leftarrow$ *true* **end**
until not changed;
return Y

pseudo-closed itemset by definition.) The loop from step 4 to 17 is a direct implementation of Definition 4 for the frequent pseudo-closed itemsets. The frequent pseudo-closed i-itemsets, their closure and their support are stored in FP_i. They are used to generate the set $\mathcal{L}$ of implications of the Duquenne-Guigues basis for exact association rules DG (step 18).

The set of infrequent pseudo-closed itemsets is determined in steps 19 to 21 using the function $\mathcal{L}^*$-closure (Algorithm 2). This function uses the fact that, for a given closure system, the set of all closed or pseudo-closed sets forms again a closure system [14]. Hence one can generate all closed sets and pseudo-closed sets iteratively by using the corresponding closure operator $\mathcal{L}^*\text{-closure}(Z) := \bigcup_{i=0}^{\infty} Z_i$ with $Z_0 := Z$ and $Z_{i+1} := Z_i \cup \bigcup\{Y | X \rightarrow Y \in \mathcal{L}, X \subset Z_i\}$ [14]. The set $\mathcal{L}$ of implications has the form $\mathcal{L} = \{X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n\}$.

4.2 Generating the Luxenburger Basis for Approximate Association Rules with Gen-LB

The pseudo-code generating the Luxenburger basis for approximate association rules is presented in Algorithm 3. The algorithm takes as input the sets FC_i, $0 \leq i \leq k$, containing the frequent closed itemsets and their support. The output of the algorithm is the Luxenburger basis for approximate association rules LB.

The algorithm iteratively considers all frequent closed itemsets $L \in FC_i$ for $2 \leq i \leq k$. It determines which frequent closed itemsets $L' \in \bigcup_{j<i} FC_j$ are covered by L and generates association rules of the form $L' \rightarrow L \setminus L'$ that have sufficient confidence. During the i^{th} iteration, each itemset L in FC_i is considered (steps 3 to 13). For each set FC_j, $1 \leq j < i$, a set S_j containing all frequent closed j-itemsets in FC_j that are subsets of L is created (step 4). Then, all these subsets of L are considered in decreasing order of their sizes (steps 5 to 12). For each of these subsets $L' \in S_j$, the confidence of the approximate association rule $r := L' \rightarrow L \setminus L'$ is computed (step 7). If the confidence of r is sufficient, r is inserted into LB (step 9) and all subsets L'' of L' are removed from S_l, for $l < j$ (step 10). At the end of the algorithm, the set LB contains all rules of the Luxenburger basis for approximate association rules. The proof of the correctness of the algorithm is given in [27].

Algorithm 3 Generating the Luxenburger basis with Gen-LB.

```
LB ← {};
for (i ← 2; i ≤ k; i++) do begin
    forall L ∈ FC_i do begin
        for (j ← 0; J < i; j + +) do S_j ← Subsets(FC_j, L);
        for (j ← i − 1; J ≥ 1; j − −) do begin
            forall L' ∈ S_j do begin
                conf ← L.support / L'.support;
                if (conf ≥ minconf)
                    then LB ← LB ∪ {(L' → (L \ L'), L.support, conf)};
                for (l ← j; l ≥ 1; l − −) do S_l ← S_l \ Subsets(S_l, L');
            end
        end
    end
end
```

5 Experimental Results

We have preformed several experiments on synthetic and real data. The characteristics of the datasets used in the experiments are given in Table 1. These datasets are the T10I4D100K synthetic dataset that mimics market basket data,[8] the C20D10K and the C73D10K census datasets from the PUMS sample file,[9] and the MUSHROOMS dataset describing mushroom characteristics.[10] In all experiments, we attempted to choose significant minimum support and confidence threshold values. We varied these thresholds and, for each couple of values, we analyzed rules extracted in the bases.

Table 1. Datasets.

Name	Number of objects	Average size of objects	Number of items
T10I4D100K	100,000	10	1,000
MUSHROOMS	8,416	23	127
C20D10K	10,000	20	386
C73D10K	10,000	73	2,177

Number of Rules. Table 2 compares the size of the Duquenne-Guigues basis for exact rules with the number of all exact association rules, and the size of the Luxenburger basis for approximate rules with the number of all approximate rules. In the case of weakly correlated data (T10I4D100K), no exact rule is generated. The reason is that in such data all frequent itemsets are frequent

[8] http://www.almaden.ibm.com/cs/quest/syndata.html
[9] ftp://ftp2.cc.ukans.edu/pub/ippr/census/pums/pums90ks.zip
[10] ftp://ftp.ics.uci.edu/~cmerz/mldb.tar.Z

Table 2. Number of exact and approximate association rules compared with the number of rules in the Duquenne-Guigues and Luxenburger bases.

Dataset (Minsupp)	Exact rules	D.-G. basis	Minconf	Approximate rules	Luxenburger basis
T10I4D100K (0.5%)	0	0	90%	16,269	3,511
			70%	20,419	4,004
			50%	21,686	4,191
			30%	22,952	4,519
MUSHROOMS (30%)	7,476	69	90%	12,911	563
			70%	37,671	968
			50%	56,703	1,169
			30%	71,412	1,260
C20D10K (50%)	2,277	11	90%	36,012	1,379
			70%	89,601	1,948
			50%	116,791	1,948
			30%	116,791	1,948
C73D10K (90%)	52,035	15	95%	1,606,726	4,052
			90%	2,053,896	4,089
			85%	2,053,936	4,089
			80%	2,053,936	4,089

closed itemsets. However, the Luxenburger basis is relatively small compared to the number of all rules, since only immediate neighbors with respect to the subset order (and not arbitrary pairs of sets) are considered. In the case of strongly correlated data (MUSHROOMS, C20D10K and C73D10K), the ratio between the size of the bases to the number of all rules which hold is much smaller than in the weekly correlated case, because here only few of the frequent itemsets are closed and have to be considered.

Relative Performance. Our experiments also show that in all cases the execution time of Gen-FDG and Gen-LB are insignificantly small compared to those of the computation of the iceberg concept lattice, since both algorithms need not access the database. We can conclude that without additional computation time (compared to other approaches, like e. g. Apriori) our approach not only computes all frequent closed itemsets but also the two bases described in Section 2.

6 Outlook

In this paper, we introduced bases which significantly reduce the number of association rules presented to the user without losing any information; and provided algorithms for computing them. This work is currently extended in different directions:

Integrating reduction methods. Templates, as defined in [4, 21], can directly be used for extracting all association rules matching some user specified patterns from the bases. Information in taxonomies and ontologies associated with the dataset can also be integrated in the process as proposed in [17, 34] for extracting bases for generalized (multi-level) association rules. Integrating item constraints [8, 26, 35] and statistical measures [10] in the generation of bases requires further work.

Integration of association rule visualization in Conceptual Information Systems. Using the technique of conceptual scaling, Conceptual Information Systems present the information contained in large databases to the user in conceptual hierarchies of a manageable size [40, 36, 18]. We work on exploiting this visualization techniques for presenting also association rules to the user.

Supporting the creation of new concepts in Description Logics. In Description Logics, currently approaches are discussed to support the domain expert in creating new concepts which regroup more specific similar concepts [3]. Those approaches extend the partial order of the concepts in the terminology to a lattice and suggest new concepts to the user. Since the more specific concepts are often defined incoherently, the user is often interested in only approximate relationships between those concepts, and on a general level only. It is planned to adapt the bases and the algorithms presented in this paper to that task.

References

1. R. Agrawal, T. Imielinski, A. Swami. Mining association rules between sets of items in large databases. *Proc. SIGMOD Conf.*, 1993, 207–216
2. R. Agrawal and R. Srikant. Fast algorithms for mining association rules. *Proc. VLDB Conf.*, 1994, 478–499 (Expanded version in IBM Report RJ9839)
3. F. Baader, R. Molitor: Building and structuring Description Logic knowledge bases using least common subsumers and Concept Analysis. In: B. Ganter, G. W. Mineau (eds.): *Conceptual Structures: Logical, Linguistic, and Computational Issues.* Proc. ICCS 2000. LNAI **1867**, Springer, Heidelberg 2000, 292–305
4. E. Baralis and G. Psaila. Designing templates for mining association rules. *Journal of Intelligent Information Systems* 9(1), 1997, 7–32
5. 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. 1st Intl. Conf. on CL (6th Intl. Conf. on Database Systems). LNAI **1861**, Springer, Heidelberg 2000, 972–986
6. 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, 66–75
7. R. J. Bayardo. Efficiently mining long patterns from databases. *Proc. SIGMOD Conf.*, 1998, 85–93
8. R. J. Bayardo, R. Agrawal, D. Gunopulos. Constraint-based rule mining in large, dense databases. *Proc. ICDE Conf.*, 1999, 188–197
9. K. Becker, G. Stumme, R. Wille, U. Wille, M. Zickwolff: Conceptual Information Systems discussed through an IT-security tool. In: R. Dieng, O. Corby (Eds.): *Knowledge Engineering and Knowledge Management. Methods, Models, and Tools.* Proc. EKAW '00. LNAI **1937**, Springer, Heidelberg 2000, 352–365
10. S. Brin, R. Motwani, C. Silverstein: Beyond market baskets: Generalizing association rules to correlation. *Proc. SIGMOD Conf.*, 1997, 265–276
11. S. Brin, R. Motwani, J. D. Ullman, S. Tsur: Dynamic itemset counting and implication rules for market basket data. *Proc. SIGMOD Conf.*, 1997, 255–264
12. V. Duquenne, J.-L. Guigues: Famille minimale d'implication informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines* 24(95), 1986, 5–18
13. U. M. Fayyad, G. Piatetsky–Shapiro, P. Smyth, R. Uthurusamy (eds.): Advances in knowledge discovery and data mining. AAAI Press, Cambridge 1996
14. B. Ganter, K. Reuter: Finding all closed sets: A general approach. *Order.* Kluwer Academic Publishers, 1991, 283–290

15. B. Ganter, R. Wille: *Formal Concept Analysis: Mathematical Foundations.* Springer, Heidelberg 1999
16. A. Grosskopf and G. Harras: Eine TOSCANA-Anwendung für Sprechaktverben des Deutschen. In: G. Stumme and R. Wille (eds.), Begriffliche Wissensverarbeitung: Methoden und Anwendungen. Springer, Berlin-Heidelberg-New York 2000.
17. J. Han, Y. Fu: Discovery of multiple-level association rules from large databases. *Proc. VLDB Conf.*, 1995, 420–431 1995.
18. J. Hereth, G. Stumme, U. Wille, R. Wille: Conceptual Knowledge Discovery and Data Analysis. In: B. Ganter, G. W. Mineau (eds.): *Conceptual Structures: Logical, Linguistic, and Computational Issues.* Proc. ICCS 2000. LNAI **1867**, Springer, Heidelberg 2000, 421–437
19. J. Hipp, A. Myka, R. Wirth, U. Güntzer: A new algorithm for faster mining of generalized association rules. LNAI **1510**, Springer, Heidelberg 1998
20. U. Kaufmann: Begriffliche Analyse über Flugereignisse – Implementierung eines Erkundungs- und Analysesystems mit TOSCANA. Diplomarbeit, FB4, TU Darmstadt 1996.
21. M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, A. I. Verkamo: Finding interesting rules from large sets of discovered association rules. *Proc. CIKM Conf.*, 1994, 401–407
22. M. Luxenburger: Implications partielles dans un contexte. *Mathématiques, Informatique et Sciences Humaines*, 29(113), 1991, 35–55
23. M. Luxenburger: Partial implications. Part I of *Implikationen, Abhängigkeiten und Galois Abbildungen.* PhD thesis, TU Darmstadt. Shaker, Aachen 1993
24. H. Mannila, H. Toivonen: Multiple uses of frequent sets and condensed representations (Extended abstract). *Proc. KDD 1996*, 189–194
25. R. Meo, G. Psaila, S. Ceri: A new SQL-like operator for mining association rules. *Proc. VLDB Conf.*, 1996, 122–133
26. R. T. Ng, V. S. Lakshmanan, J. Han, A. Pang: Exploratory mining and pruning optimizations of constrained association rules. *Proc. SIGMOD Conf.*, 1998, 13–24
27. N. Pasquier: *Data Mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données.* PhD thesis. Université Blaise Pascal, Clermont-Ferrand II, 2000
28. N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal: Pruning closed itemset lattices for association rules. *Proc. 14ièmes Journées Bases de Données Avancées (BDA '98)*, Hammamet, Tunisie, 177–196
29. N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal: Closed set based discovery of small covers for association rules. *Proc. 15èmes Journées Bases de Données Avancées*, Bordeaux, France, 25–27 October 1999, 361–381
30. N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal: Discovering frequent closed itemsets for association rules. *Proc. ICDT Conf.*, 1999, 398–416
31. 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
32. J. Pei, J. Han, R. Mao: CLOSET: An efficient algorithm for mining frequent closed itemsets. *Proc. ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 2000*, 21–30
33. M. Roth-Hintz, M. Mieth, T. Wetter, S. Strahringer, B. Groh, R. Wille: Investgating SNOMED by Formal Concept Analysis. Preprint, FB4, TU Darmstadt 1997.
34. R. Srikant, R. Agrawal: Mining generalized association rules. *Proc. VLDB Conf.*, 1995, 407–419
35. R. Srikant, Q. Vu, R. Agrawal: Mining association rules with item constraints. *Proc. KDD Conf.*, 1997, 67–73
36. 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. 2nd European Symposium on PKDD '98, LNAI **1510**, Springer, Heidelberg 1998, 450–458
37. G. Stumme: *Conceptual Knowledge Discovery with Frequent Concept Lattices.* FB4-Preprint **2043**, TU Darmstadt 1999
38. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, L. Lakhal: Fast Computation of Concept Lattices Using Data Mining Techniques. *Proc. 7th Intl. Workshop on Knowledge Representation Meets Databases*, Berlin, 21–22. August 2000. CEUR-Workshop Proceeding. http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/
39. G. Stumme, R. Taouil, Y. Bastide, N. Pasquier, L. Lakhal: Computing Iceberg Concept Lattices with Titanic. *J. on Knowledge and Data Engineering.* (submitted)
40. F. Vogt, R. Wille: TOSCANA – A graphical tool for analyzing and exploring data. LNCS **894**, Springer, Heidelberg 1995, 226–233
41. R. Wille: Restructuring lattice theory: an approach based on hierarchies of concepts. In: I. Rival (ed.). *Ordered sets.* Reidel, Dordrecht–Boston 1982, 445–470
42. M. J. Zaki, M. Ogihara: Theoretical Foundations of Association Rules, *3rd SIGMOD'98 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD)*, Seattle, WA, June 1998, 7:1–7:8
43. M. J. Zaki, C.–J. Hsiao: ChARM: An efficient algorithm for closed association rule mining. Technical Report 99–10, Computer Science Dept., Rensselaer Polytechnic Institute, October 1999
44. M. J. Zaki: Generating non-redundant association rules. *Proc. KDD 2000.* 34–43