A Formal Concept Analysis Approach to Rough Data Tables

Bernhard Ganter and Christian Meschke*

Institut für Algebra, TU Dresden, Germany
{Bernhard.Ganter,Christian.Meschke}@tu-dresden.de

Abstract. In order to handle very large data bases efficiently, the data warehousing system ICE [5] builds so-called *rough tables* containing information that is abstracted from certain blocks of the original table. In this article we propose a formal description of such rough tables. We also investigate possibilities of mining them for implicational knowledge.

1 Introduction

Consider a large data table. It has rows, describing certain *objects*, and columns for *attributes* which these objects may have. The entry in row g and column m gives the *attribute value* that attribute m has for object g. By "large" we mean that the table has many rows, perhaps 10^9, or more. Even for a moderate number of attributes the size of such a table may be in the terabytes.

Data analysis on such a table faces complexity problems and requires a good choice of strategy. In the present paper we investigate an approach by Infobright using rough objects and granular data, and combine it with methods from Formal Concept Analysis.

Infobright Community Edition (ICE) [5,7] is an open source data warehousing system which is optimized to obtain high compression rates and to process analytic queries very quickly. ICE chops the stream of rows into so-called *rough rows*, each subsuming 65536 rows. The rough rows divide the columns into so-called *data packs*. Each data pack gets stored in a compressed form. For processing a query one does not want to decompress all data packs. Therefore, ICE creates a so-called *data pack node* to every data pack. A data pack node contains meta–information about the corresponding data pack. If for instance the column contains numeric values, the data pack nodes could consist, e.g., of minimum, maximum and the sum of the data pack values. The *rough table* is the data table that has the rough rows as rows, the same attributes as the original large data table, and the data pack nodes as values.

In order to sound the possibilities of getting interesting information about the original data table from the rough table, Infobright offered a contest [6] for which they provided a rough table with 15259 rough rows (the original table has one billion rows) and 32 attributes. Furthermore, Infobright invited to propose ways to do data mining in such rough tables.

Our approach is a systematic one. Our focus is on "what can be done" rather than on "how to get quick results". Although it is likely that a large data table will contain erroneous and imprecise data, we first concentrate on the case of precise data. Approximative and fault-tolerant methods shall later be build on this basis.

2 Partial formal contexts

We assume that the reader is familiar with the basic notions of Formal Concept Analysis [4]. This theory will be used here to provide the basic data model. To

* Supported by the DFG research grant no. GA 216/10-1.

encode the above mentioned granulation process we use the notion of a partial formal context. The information we are mining is in the form of implications or, more loosely, of association rules. Our aim is to infer such rules in the full data set from rules in the granulated data.

Definition 1 A **partial formal context** (G, M, i) consists of two sets G and M together with a mapping $i : G \times M \to \{\times, \bullet, \mathbf{?}\}$. $\Diamond$

We call the elements of G the *objects* of the partial formal context, those of M the *attributes*. We read $i(g, m)$ as follows:

$$i(g,m) = \begin{cases} \times & \text{the object } g \text{ has the attribute } m, \\ \bullet & \text{the object } g \text{ does not have the attribute } m, \\ \mathbf{?} & \text{ it is unknown if object } g \text{ has attribute } m. \end{cases}$$

A single row of such a partial context is called a *partial object description (POD)*. Partial formal contexts have been considered under different aspects by several authors [1,2,3]. A partial formal context (G, M, j) is said to **extend** (G, M, i) if one can build it from (G, M, i) by replacing question marks "**?**", i.e., it holds that

$$i^{-1}(\{\times\}) \subseteq j^{-1}(\{\times\}) \quad \text{and} \quad i^{-1}(\{\bullet\}) \subseteq j^{-1}(\{\bullet\}).$$

Partial formal contexts which are maximal w.r.t. to this *extension order* are called **complete**. A formal context (G, M, I) in the usual sense, where $I \subseteq G \times M$ is a relation, is called a **completion** of a partial formal context (G, M, i) iff

$$i^{-1}(\{\times\}) \subseteq I \subseteq i^{-1}(\{\times, \mathbf{?}\}).$$

We say that an implication $A \to B$, where $A, B \subseteq M$, **holds** in a partial formal context (G, M, i) iff it holds in every completion. An equivalent condition is that the following holds for every object $g \in G$:

$$\text{if } i(g,m) \in \{\times, \mathbf{?}\} \text{ for all } m \in A \text{ then } i(g,n) = \times \text{ for all } n \in B.$$

An implication $A \to B$ **is refuted** by the partial formal context (G, M, i) if it holds in no completion. This is equivalent to the existence of an object g with

$$i(g,m) = \times \text{ for all } m \in A \text{ and } i(g,n) = \bullet \text{ for some } n \in B.$$

In order to better handle canonical formal contexts related to the partial context (G, M, i) we define for $S \subseteq \{\times, \bullet, \mathbf{?}\}$

$$i_S := \{(g,m) \in G \times M \mid i(g,m) \in S\} = i^{-1}(S).$$

We leave away the set brackets of S. For instance we write $i_{\times,\mathbf{?}}$ instead of $i_{\{\times,\mathbf{?}\}}$.

Proposition 1 *Let* (G, M, i) *and* (G, M, j) *be partial formal contexts such that* (G, M, j) *extends* (G, M, i)*. Then*

- *every implication that holds in* (G, M, i) *also holds in* (G, M, j)*, and*
- *every implication that is refuted by* (G, M, i) *is also refuted by* (G, M, j)*.*

Proof For every implication $A \to B$ that holds in (G, M, i) it follows for $g \in G$ that[1]

$$A \subseteq g^{j_{\times,\mathbf{?}}} \implies A^{\complement} \supseteq g^{j_\bullet} \supseteq g^{i_\bullet} \implies A \subseteq g^{i_{\times,\mathbf{?}}} \implies B \subseteq g^{i_\times} \subseteq g^{j_\times}.$$

The second item follows immediately from the observation that every object that refutes an implication in (G, M, i) also refutes this implication in (G, M, j). □

[1] Let (G, M, R) be a formal context, i.e., $R \subseteq G \times M$. For $g \in G$ we define $g^R := \{m \in M \mid gRm\}$. For $A \subseteq G$ we define $A^R := \{m \in M \mid aRm \text{ for all } a \in A\}$. Dually one defines m^R and B^R for $m \in M$ and $B \subseteq M$, see [4].

3 Partial contexts obtained from streams

There is a natural way how partial formal contexts arise from complete ones. Let (G, M, I) be a formal context and let $\mathcal{F}$ be a family of nonempty subsets of the object set G, i.e. $\mathcal{F} \subseteq \mathfrak{P}_{>0}(G) := \mathfrak{P}(G) \setminus \{\emptyset\}$. We obtain a partial formal context $(\mathcal{F}, M, i)$ by defining for every **block** $F \in \mathcal{F}$

$$i(F, m) := \begin{cases} \times & \text{if } F \subseteq m^I, \\ \bullet & \text{if } F \cap m^I = \emptyset, \\ ? & \text{else.} \end{cases}$$

We refer to $(\mathcal{F}, M, i)$ as the **$\mathcal{F}$-granulated** partial context to (G, M, I). Note that this reflects the situation of Infobright's rough tables from the contest [6] and is only formulated in a different language. For further details we refer the reader to the following Section 4.

Proposition 2 *Let $(\mathfrak{P}_{>0}(G), M, i)$ be constructed from (G, M, I) as defined above, for the special case that $\mathcal{F} := \mathfrak{P}_{>0}(G)$. Then*

- *an implication that holds in $(\mathfrak{P}_{>0}(G), M, i)$ also holds in (G, M, I) and*
- *an implication is refuted by $(\mathfrak{P}_{>0}(G), M, i)$ iff it does not hold in (G, M, I).*

Proof The rows of (G, M, I) correspond to the rows belonging to singleton blocks $\{g\}$ in $(\mathfrak{P}_{>0}(G), M, i)$ in the following way: $\{g\}^{i_\times} = g^I$, $\{g\}^{i_?} = \emptyset$ and $\{g\}^{i_\bullet} = g^{\not{I}}$. Hence, the first item and "$\Leftarrow$" of the second item are trivial. The remaining direction of the second item is a special case of the following Proposition 3. □

Proposition 3 *For every $\mathcal{F} \subseteq \mathfrak{P}_{>0}(G)$ it is true that*

- *no implication refuted by $(\mathcal{F}, M, i)$ holds in (G, M, I).*

If $\mathcal{F}$ is a covering of G then it is true that

- *every implication that holds in $(\mathcal{F}, M, i)$ also holds in (G, M, I).*

Proof Let $F \in \mathcal{F}$ be block that refutes $A \to B$ in $(\mathcal{F}, M, i)$. Then it holds that $A \subseteq F^{i_\times} = F^I$ and $B \not\subseteq F^{i_{\times,?}}$. Hence, it follows $B \not\subseteq F^{i_\times} = F^I$ which implies that $A \to B$ cannot hold in (G, M, I), since F^I is an intent containing the premise A, but not containing B.

Let $A \to B$ be an implication that holds in $(\mathcal{F}, M, i)$ and let $g \in G$. Since $\mathcal{F}$ is a covering there is a block $F \in \mathcal{F}$ containing g. Hence, it holds that $g^I \subseteq F^{i_{\times,?}}$ which implies

$$A \subseteq g^I \implies A \subseteq F^{i_{\times,?}} \implies B \subseteq F^{i_\times} \implies B \subseteq g^I.$$

□

Now suppose that (G, M, I) is given as a stream of rows, and is chopped into data packs as described in the introduction. For each pack we take notes only if each object in the pack does have the attribute, in which case we note an "×" for the pack, or if no object in the pack has that attribute. We then note down "•". If some have and some do not, we note a question mark. This is a very strict rule, and we refer to it as **hard granulation**. Its disadvantage is that its outcome can drastically be changed by a single value in the pack. It shares this property with logical analysis: If a given logical formula does or does not hold in the original data, may be decided by a single counterexample. Proposition 3 above shows our possibilities to argue about implicational information of (G, M, I) based only on the granulated context $(\mathcal{F}, M, i)$. It is therefore necessary to investigate the circumstances under which an implication holds in or is refuted by $(\mathcal{F}, M, i)$. For both concerns it is sufficient to just take a look on implications of the form $A \to b$, where $A \subseteq M$ and $b \in M$.

Proposition 4 *For $F \in \mathcal{F}$ the following three statements are equivalent:*

(a) F refutes $A \to b$ in $(\mathcal{F}, M, i)$,
(b) $F \subseteq A^I \setminus b^I$,
(c) every single object $g \in F$ refutes $A \to b$ in (G, M, I).

Proof F refutes $A \to b$ iff it holds that $A \subseteq F^{i_\times} = F^I$ and $b \in F^{i_\bullet} = F^{\not I}$, which again is equivalent to $F \subseteq A^I$ and $F \subseteq b^{I^\complement}$. The rest follows immediately. □

The preceding propositions clarify under which conditions an implication $A \to b$ is refuted by the granulated context $(\mathcal{F}, M, i)$. If we insist on a definite answer, an answer that proves a refutation in the full data set on basis of the granulated data, these seem to be the natural conditions. But how likely is it that these conditions are satisfied? We attempt to give a first estimation. Obviously, the number $r := |A^I \setminus b^I|$ of all objects from the original data table (G, M, I) that share all attributes from A but do not have attribute b has to be large enough. Let k be a fixed number. For the probability that a block F of cardinality k refutes $A \to b$ the following holds:

$$P(F \text{ refutes } A \to b) \;=\; \frac{\binom{r}{k}}{\binom{n}{k}} \;=\; \frac{r \cdot (r-1) \cdot \ldots \cdot (r-k+1)}{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)} \;\leq\; \left(\frac{r}{n}\right)^k .$$

We now assume that all $F \in \mathcal{F}$ have the same cardinality k. With the inequality from above we can conclude the following upper approximation of the probability that a partial context $(\mathcal{F}, M, i)$ refutes $A \to b$:

$$P((\mathcal{F}, M, i) \text{ refutes } A \to b) \;\leq\; \sum_{F \in \mathcal{F}} P(F \text{ refutes } A \to b) \;\leq\; |\mathcal{F}| \cdot \left(\frac{r}{n}\right)^k .$$

If we for instance assume that 95% of all objects in (G, M, I) refute $A \to b$ and that $\mathcal{F}$ contains one million blocks, i.e., $\frac{r}{n} = 95\%$ and $|\mathcal{F}| = 1.000.000$, we get that already for relatively small block sizes of $k \geq 539$ the probability that $(\mathcal{F}, M, i)$ refutes $A \to b$ is smaller than one part of a million.

Proposition 5 *For $A, B \subseteq M$ the following three statements are equivalent:*

(a) $A \to B$ holds in $(\mathcal{F}, M, i)$,
(b) for all $F \in \mathcal{F}$ the implication $A \subseteq F^{i_{\times, ?}} \implies B \subseteq F^{i_\times}$ holds,
(c) for all $F \in \mathcal{F}$ the implication $A \subseteq \bigcup_{g \in F} g^I \implies B \subseteq \bigcap_{g \in F} g^I$ holds.

Proof Omitted. □

If one takes a look at the third condition it becomes obvious that the bigger the block sizes $|F|$ are, the more likely it becomes that the premisses are valid, and the less likely it becomes that the conclusions hold. Hence, if the number of the blocks and the sizes of the blocks are relatively large, we do not expect a lot of implications to hold in $(\mathcal{F}, M, i)$. In order to underline this argumentation more formally we make the (very debateable) assumption that the crosses in (G, M, I)

are distributed independently with a fixed probability $q := \frac{|I|}{|G \times M|}$. For $A, B \subseteq M$ and $F \subseteq G$ we get that:

$$\begin{aligned} P(A \subseteq \bigcup_{g \in F} g^I) &= (1-(1-q)^{|F|})^{|A|}, \\ P(B \subseteq \bigcap_{g \in F} g^I) &= (1-q)^{|F| \cdot |B|}. \end{aligned}$$

Hence, it follows that

$$P(A \subseteq F^{i_{\times,?}} \Rightarrow B \subseteq F^{i_\times}) \leq \Big(1-(1-(1-q)^{|F|})^{|A|}\Big) + (1-q)^{|F| \cdot |B|}.$$

But the expression on the right side tends to be very small for a large block size $|F|$ and $|A| \ll |F|$. Hence, for large block sizes it is very unlikely that a given implication $A \rightarrow B$ holds in $(\mathcal{F}, M, i)$. Summing up our thoughts about implications in $(\mathcal{F}, M, i)$ we have to conclude that for large block sizes it is not very likely that one can decide (using Proposition 3) if a given implication does or does not hold in (G, M, I) if one just knows the granulated data $(\mathcal{F}, M, i)$.

4 The Contest Data Set

The Infobright data set does not come as a formal context right away, but needs some (uncritical) transformation. The formalisation of a data table which we use is that of a *many-valued context* (G, M, W, J), where G is a set of objects, M a set of many-valued attributes, W a set of attribute values and J is a ternary incidence relation satisfying

$$(g, m, v) \in J \text{ and } (g, m, w) \in J \text{ implies } v = w.$$

The standard interpretation of $(g, m, v) \in J$ is that the value of attribute m for object g is v. The value the object g has regarding to attribute m is commonly denoted with $m(g)$. To better distinguish such many-valued contexts from the formal contexts introduced first we shall refer to these sometimes as *one-valued.*

One of the standard techniques in Formal Concept Analysis expresses many-valued contexts as one-valued ones by means of *conceptual scales.* With conceptual scaling, every many-valued attribute is represented by several one-valued attributes, and the incidence to these depends on the respective attribute value. Details can be found in [4], but for the moment it suffices to know that with this technique, a data table can be transformed to a (one-valued) formal context, and this transformation can be done object-wise, one after another. As a consequence, we may transform a stream of objects with many-valued attributes into a stream of objects in a formal context. To keep things simple, we summarize: Conceptual scaling associates to each column m of the data table a set of attributes (the "scale attributes for the many-valued attribute m").

In the case of Infobright's contest data set we may think of the underlying, very large data table as a many-valued context (G, M_0, W, J) in which for every attribute $m \in M_0$ the set

$$W_m := m[G] := \{w \in W \mid (g, m, w) \in J \text{ for some } g \in G\}$$

of all values occurring in the column of m are ordered linearly in a canonical way. Depending on the data type of the attribute m this canonical order $\leq_m$ can for instance be the natural order of numbers or the alphabetical order of character strings. If one transforms this data table (G, M_0, W, J) into the formal context

(G, M, I) via scaling every attribute from M_0 *interordinally*, this formal context (G, M, I) directly yields to the granulated partial formal context $(\mathcal{F}, M, i)$ which contains exactly the same information as the contests rough table from [6].

This implicit interordinal scaling works as follows. For every $m \in M_0$ and every $w \in W_m$ one creates the two scale attributes "$\leq_m w$" and "$\geq_m w$". An object $g \in G$ now *has* the scale attribute "$\geq_m w$" in the formal context (G, M, I) iff $m(g) \leq_m w$ holds. The incidences of "$\geq_m w$" are defined in the dual way. Formally, the attribute set of the scaled context (G, M, I) is defined as follows:

$$M \;:=\; \bigcup_{m \in M_0} \{\text{“}\leq_m w\text{”}, \text{“}\geq_m w\text{”} \mid w \in W_m\}.$$

The incidence relation I is defined in the *natural way* as described above. If we now take again an arbitrary collection $\mathcal{F}$ of nonempty object sets, we receive the granulated partial context $(\mathcal{F}, M, i)$ as described in the previous section. In the contest data $\mathcal{F}$ is a partition with classes having (almost ever) size 2^{16}. For $F \in \mathcal{F}$ and "$\leq_m w$"$\in M$ it holds that $i(F, \text{“}\leq_m w\text{”}) = \times$ iff w is an upper bound of the values the objects from F have in the m-column, i.e., $m(g) \leq_m w$ for every $g \in F$. Hence, it holds that

$$\max_{g \in F} m(g) \;=\; \min\{w \in W_m \mid i(F, \text{“}\leq_m w\text{”}) = \times\}.$$

Dually, one can also read the minimal m-value of F from $(\mathcal{F}, M, i)$. One could now calculate the implications that hold in $(\mathcal{F}, M, i)$ and the ones that are refuted by $(\mathcal{F}, M, i)$. But as we have seen in the previous section we expect a very small amount of such implications simply because $(\mathcal{F}, M, i)$ will contain far too many question marks **?**. Of course one can already see this in the original data: For many attributes $m \in M_0$ it holds that for almost every rough row $F \in \mathcal{F}$ the minimal and maximal m-values in F are exactly the overall minimal and maximal m-values, i.e.,

$$\min_{f \in F} m(f) \;=\; \min_{g \in G} m(g) \quad \text{and} \quad \max_{f \in F} m(f) \;=\; \max_{g \in G} m(g).$$

And if this equalities do not hold, they *almost* hold in the sense that the intervals

$$[\min_{f \in F} m(f), \max_{f \in F} m(f)]$$

are usually very, very large subintervals of W_m, which is reflected by a superiority of question marks in the granulated context.

5 Soft Granulation

There is a reason why the approach of the previous section led to rather disappointing results: Our definition of the granulation process was too rigid. We defined that a block has a certain object if *all* members of a pack have the attribute, etc. As an example from the Infobright data, we mention the *minimum* parameter: It expresses that all members of the pack have a value greater or equal this one.

For a rough estimation, such parameters that can drastically be changed by a single member of the block seem inappropriate. It seems more promising to work with parameters which reflect the "tendency" of the data packs. The simplest suggestion is counting: Let us record for each data pack (F, m) the relative frequency of objects having the attribute. Formally:

$$\texttt{freq}(F, m) \;:=\; \frac{|m^I \cap F|}{|F|}.$$

The absolute frequency of an data pack (F, m) is called its **support** and is defined as

$$\texttt{supp}(F, m) \;:=\; |m^I \cap F|.$$

The number of objects of an block F that do not have attribute m is called its **negative support** and is defined as

$$\texttt{nsupp}(F, m) \;:=\; |F \setminus m^I|.$$

Our granulation will now work as follows: The formal context (G, M, I) leads us to the $\mathbb{N}_0$-valued context $(\mathcal{F}, M, i)$, i.e., $i : \mathcal{F} \times M \to \mathbb{N}_0$, with

$$i(F, m) \;:=\; \texttt{supp}(F, m).$$

What we are trying to do is mining in $(\mathcal{F}, M, i)$ for association rules that hold in (G, M, I). An **association rule** $A \to B$ consists of two attribute sets: the *premises* A and the *conclusion* B. We call

$$\begin{array}{rcl}
\texttt{supp}(A) := & |A^I| & \text{the } \textbf{support}^2 \text{of } A, \\
\texttt{supp}(A \to B) := & \texttt{supp}(A \cup B) & \text{the } \textbf{support} \text{ of the rule } A \to B, \text{ and} \\
\texttt{conf}(A \to B) := & \frac{\texttt{supp}(A \cup B)}{\texttt{supp}(A)} & \text{the } \textbf{confidence} \text{ of the rule } A \to B.
\end{array}$$

Furthermore, for given thresholds $\texttt{minsupp} \in \mathbb{N}_0$ and $\texttt{minconf} \in [0, 1]$ we say an association rule **holds** in (G, M, I) if its support exceeds $\texttt{minsupp}$ and its confidence exceeds $\texttt{minconf}$. Hence, association rules are a generalization of the attribute implications: The implications that hold in a formal context are exactly the association rules that hold with $\texttt{minsupp} = 0$ and $\texttt{minconf} = 1$. We say an attribute set (or a rule) is **frequent** if its support is greater or equal $\texttt{minsupp}$.

Given an association rule $A \to B$ we ask ourself what we can say about support and confidence in (G, M, I) just knowing the granulated context $(\mathcal{F}, M, i)$. By the way we also assume that the block sizes $|F|$ are known for every $F \in \mathcal{F}$. Hence, for every data pack (F, m) we know its support and its negative support.

From now on we assume that $\mathcal{F}$ is a partition of the object set G. We define approximations of the above mentioned measures which just use knowledge given by the supports and the negative supports of the data packs:

$$\begin{aligned}
\underline{\texttt{supp}}(A) &:= \sum_{F \in \mathcal{F}} \max\Big\{0, |F| - \sum_{a \in A} \texttt{nsupp}(F, a)\Big\}, \\
\overline{\texttt{supp}}(A) &:= \sum_{F \in \mathcal{F}} \min_{a \in A} \texttt{supp}(F, a), \\
\underline{\texttt{conf}}(A \to B) &:= \frac{\sum_{F \in \mathcal{F}} \max\Big\{0, \min_{a \in A} \texttt{supp}(F, a) \;-\; \sum_{b \in B \setminus A} \texttt{nsupp}(F, b)\Big\}}{\overline{\texttt{supp}}(A)}.
\end{aligned}$$

Proposition 6 *For $A, B \subseteq M$ it holds that:*

$$\underline{\textit{supp}}(A) \;\leq\; \textit{supp}(A) \;\leq\; \overline{\textit{supp}}(A).$$

Furthermore, the inequality $\underline{\textit{conf}}(A \to B) \leq \textit{conf}(A \to B)$ holds.

Proof Omitted. □

[2] It is more common to define the *support* of A as the quotient $\frac{|A^I|}{|G|}$. We choose to define it the *absolute* way since it makes the following formulas more readable.

Definition 2 We say an association rule **holds** in the granulated partial context $(\mathcal{F}, M, i)$ if

$$\texttt{minsupp} \;\le\; \underline{\texttt{supp}}(A \cup B) \quad\text{and}\quad \texttt{minconf} \;\le\; \underline{\texttt{conf}}(A \to B).$$

$\Diamond$

Corollary 1 *Every association rule that holds in $(\mathcal{F}, M, i)$ also holds in (G, M, I).*

But how likely is it that an association rule with a very high support and a high confidence can be read from the granulated context? For a singleton conclusion $B = \{b\}$ we can further approximate the lower approximation $\underline{\texttt{conf}}(A \to b)$ of the confidence of rule $A \to b$ (with $b \notin A$) in the following way[3]:

$$\underline{\texttt{conf}}(A \to b) \;\ge\; 1 - \frac{|b^{\not I}|}{\overline{\texttt{supp}}(A)}.$$

The right side of this inequality exceeds `minconf` iff the following holds:

$$\frac{|b^{I}|}{|G|} \;\ge\; 1 - (1 - \texttt{minconf}) \cdot \frac{\overline{\texttt{supp}}(A)}{|G|}.$$

Let us take for instance $\texttt{minconf} = 70\%$ and $\overline{\texttt{supp}}(A) = 0.6 \cdot |G|$. If in this case 82% $(= 1 - 0.3 \cdot 0.6)$ of all objects have attribute b, we can for sure read from the granulated context $(\mathcal{F}, M, i)$ that the rule is frequent. By the way, the support

$$|b^{I}| \;=\; \sum_{F \in \mathcal{F}} \texttt{supp}(F, b)$$

of the attribute b can be read from the granulated context $(\mathcal{F}, M, i)$. Hence, we get that at least for association rules with very high support and with a conclusion containing very frequent attributes, the chances that its $\underline{\texttt{conf}}$ value exceeds the threshold `minconf` are not too bad. But when is the rule a frequent rule? Let C be an attribute set (for instance $C = A \cup \{b\}$). It holds that

$$\begin{aligned}\underline{\texttt{supp}}(C) \;&\ge\; |G| - \sum_{m \in C} |m^{\not I}| \\ &\ge\; |G| - |C| \cdot (|G| - \texttt{supp}(C))\end{aligned}$$

The right side exceeds `minsupp` iff

$$\texttt{supp}(C) \;\ge\; \frac{(|C| - 1) \cdot |G| + \texttt{minsupp}}{|C|}.$$

If we take for instance $|C| = 4$ and $\texttt{minsupp} = 0.2 \cdot |G|$, we get that C can be detected as frequent by just using the granulated context $(\mathcal{F}, M, i)$ if its actual support (in (G, M, I)) is at least 80% of $|G|$. Note that the approximations of $\underline{\texttt{supp}}(C)$ we made above were quite rigid. Hence, in practice we expect that $\underline{\texttt{supp}}$ gives a much better lower approximation of the actual support `supp` in (G, M, I) than our example may suggest.

Due to a lack of space we have to leave out the details on how to calculate a *basis* of the association rules that hold in $(\mathcal{F}, M, i)$. We will do this in a future paper. In summary our procedure will use the fact that $\underline{\texttt{supp}}$ yields to a closure system on M. The frequent closed attribute sets will be used to to build a Luxenburger-type basis [8]. Furthermore, the following paper should investigate how to improve the approximations $\underline{\texttt{conf}}$ and $\underline{\texttt{supp}}$ if one considers background knowledge that can for instance be given by the scales used in the scaling process.

[3] by applying the inequality $\max\{0, x\} \ge x$ to every summand in the numerator in the definition of $\underline{\texttt{conf}}$.

6 Conclusion

We proposed a way to describe the rough tables occurring at the data warehousing system ICE. We did that from the standpoint of Formal Concept Analysis and tried to mine these rough tables for implicational knowledge. We argued that its very unlikely that the very rigid *minimum* and *maximum* parameters as for instance used in the contest data set [6] will yield to satisfying results. We constituted that – having in mind the data mining in rough tables – in the process of building the data pack nodes it is worth to create more sophisticated parameters that allow to give a better estimation of the distribution of the values in the data packs (like counting the number of incidences in the data packs of the scaled data table).

Ongoing work has to include the following issues: How can one efficiently calculate a basis of the association rules in Section 5? One has to explain how background knowledge can be used to improve data mining in the granulated contexts. Furthermore, experimental results are needed to find out whether the soft granulation described in Section 5 will lead to satisfying results in practice.

References

1. Franz Baader, Bernhard Ganter, Ulrike Sattler, and Barış Sertkaya : *Completing Description Logic Knowledge Bases using Formal Concept Analysis.* , Vol. 258 CEUR-WS.org (2007) .
2. Radim Bělohlávek and Vilém Vychodil: *What is a fuzzy concept lattice?* CLA 2005, pp. 34–45, ISBN 80–248–0863–3.
3. Peter Burmeister and Richard Holzer: *On the Treatment of Incomplete Knowledge in Formal Concept Analysis.*, Vol. 1867 Springer (2000), pp. 385-398.
4. Bernhard Ganter and Rudolf Wille: *Formal Concept Analysis – Mathematical Foundations.* Springer, Heidelberg (1999).
5. http://www.infobright.org.
6. http://web.iitd.ac.in/%7Epremi09/infobright.pdf.
7. Infobright Community Edition, *Technology White Paper*, http://www.infobright.org/wiki/662270f87c77e37e879ba8f7ac2ea258/.
8. Lotfi Lakhal and Gerd Stumme, *Efficient Mining of Association Rules Based on Formal Concept Analysis*, Vol. 3626 Springer (2005), p. 180-195.