Axiomatization of Frequent Sets

Toon Calders* and Jan Paredaens

Universiteit Antwerpen,
Departement Wiskunde-Informatica,
Universiteitsplein 1, B-2610 Wilrijk, Belgium.
{calders,pareda}@uia.ua.ac.be

Abstract. In data mining association rules are very popular. Most of the algorithms in the literature for finding association rules start by searching for frequent itemsets. The itemset mining algorithms typically interleave brute force counting of frequencies with a meta-phase for pruning parts of the search space. The knowledge acquired in the counting phases can be represented by frequent set expressions. A frequent set expression is a pair containing an itemset and a frequency indicating that the frequency of that itemset is greater than or equal to the given frequency. A system of frequent sets is a collection of such expressions. We give an axiomatization for these systems. This axiomatization characterizes *complete systems*. A system is complete when it explicitly contains all information that it logically implies. Every system of frequent sets has a unique completion. The completion of a system actually represents the knowledge that maximally can be derived in the meta-phase.

1 Introduction

Association rules are one of the most studied topics in data mining. They have many applications [1]. Since their introduction, many algorithms have been proposed to find association rules [1][2][8].

We start with a formal definition of the association rule mining problem as stated in [1]: Let $\mathcal{I} = \{I_1, I_2, \ldots, I_m\}$ be a set of symbols, called items. Let $\mathcal{D}$ be a set of transactions, where each transaction T is a set of items, $T \subseteq \mathcal{I}$, and a unique transaction ID. We say that a transaction T *contains* X, a set of some items in $\mathcal{I}$, if $X \subseteq T$. The fraction of transactions containing X is called the *frequency* of X. An *association rule* is an implication of the form $X \Rightarrow Y$, where $X \subseteq \mathcal{I}$, $Y \subseteq \mathcal{I}$, and $X \cap Y = \phi$. The rule holds in the transaction set $\mathcal{D}$ with *confidence* c if the fraction of the transactions containing X, that also contain Y is at least c. The rule $X \Rightarrow Y$ has *support* s in the transaction set $\mathcal{D}$ if the fraction of the transactions in $\mathcal{D}$ that contain $X \cup Y$ is at least s.

Most algorithms start with searching itemsets that are contained in at least a fraction s of the transactions. To optimize the search for frequent itemsets, the algorithms use the following monotonicity principle:

* Research Assistant of the Fund for Scientific Research - Flanders (Belgium)(F.W.O. - Vlaanderen).

> if $X \subseteq Y$, then the frequency of X will never be smaller than the frequency of Y.

This information is then used to *prune* parts of the search space *a priori*. To exploit this monotonicity as much as possible, the apriori-algorithm [2] starts by counting the single itemsets. In the second step, only itemsets $\{i_1, i_2\}$ are counted where $\{i_1\}$ and $\{i_2\}$ are frequent. All other 2-itemsets are discarded. In the third step, the algorithm proceeds with the 3-itemsets that only contain frequent 2-itemsets. This iteration continues until no itemsets that can be frequent are left. The search of frequent itemsets is thus basically an interleaving of a counting phase and a meta-phase. In the counting phase, the frequencies of some predetermined itemsets, the so-called *candidates* are counted. In the meta-phase the results of the counting phase are evaluated. Based on the monotonicity principle, some itemsets are a priori excluded.

Although the monotonicity of frequency is commonly used, there is to our knowledge no previous work that discusses whether in the general case this rule is *complete*, in the sense that it tells us everything we can derive from a given set of frequencies. In this paper we consider the notion of *a system of frequent sets*. A system of frequent sets contains, possibly incomplete, information about the frequency of every itemset. For example, $A :: 0.6, B :: 0.6, AB :: 0.1, \phi :: 0.5$ is a system of frequent sets. This system of frequent sets represents partial information (e.g. obtained in counting phases.) In this system, $A :: 0.6$ expresses the knowledge that itemset A has a frequency of at least 0.6. The system can be improved. Indeed; we can conclude that $AB :: 0.2$ holds, since $A :: 0.6$ and $B :: 0.6$ and there must be an overlap of at least a 0.2-fraction between the transactions containing A and the transactions containing B. We can also improve $\phi :: 0.5$, because $\phi :: 1$ always holds. Therefore, this system is called incomplete. When a system cannot be improved, it is complete. The completion of a system represents the maximal information that can be assumed in the meta-phase.

We give three rules **F1**, **F2**, and **F3** that characterize complete systems of frequent sets; e.g. a system is complete iff it satisfies **F1**, **F2**, **F3**. We show that, after a small modification to **F3**, this axiomatization is finite and every logical implication can be inferred using these axioms a finite number of times.

As an intermediate stage in the proofs, we introduce *rare sets*. A rare set expression $K : p_K$ expresses that at most a p_K-fraction of the transactions does not contain at least one item of K.

The structure of the paper is as follows: in Section 2 related work is discussed. In Section 3 we formally define a system of frequent sets. In Section 4, an axiomatization for complete systems of frequent sets is given. Section 5 discusses inference of complete systems using the axioms. Section 6 summarizes and concludes the paper.

Many proofs in this paper are only sketched. The full proofs can be found in [3].

ACKNOWLEDGMENT

We would like to thank Prof. Dirk Van Gucht and Prof. Ed Robertson from Indiana University, and Prof. Jef Wijsen from UMH (Mons, Belgium) for their preliminary thoughts and reflections on the topic of this paper. We also would like to thank the reviewers of our paper for their useful comments and for pointing out the links with probabilistic logic.

2 Related Work

In artificial intelligence literature, probabilistic logic is studied intensively. The link with this paper is that the frequency of an itemset I can be seen as the probability that a randomly chosen transaction from the transaction database satisfies I; i.e. we can consider the transaction database as an underlying probability structure.

Nilsson introduced in [12] the following *probabilistic logic problem*: given a finite set of m logical sentences $S_1, \ldots, S_m$ defined on a set $X = \{x_1, \ldots, x_n\}$ of n boolean variables with the usual boolean operators $\wedge, \vee$, and $\neg$, together with probabilities $p_1, \ldots, p_m$, does there exists a probability distribution on the possible truth assignments of X, such that the probability of S_i being true, is *exactly* p_i for all $1 \leq i \leq m$. *Georgakopoulos et al.* prove in [7] that this problem, they suggest the name *probabilistic satisfiability problem* (PSAT), is NP-complete. This problem, however, does not apply to our framework. In our framework, a system of frequent sets can *always* be satisfied. Indeed, since a system only gives *lower* bounds on the frequencies, the system is always satisfied by a transaction database where each transaction contains every item.

Another, more interesting problem, also stated by *Nilsson* in [12], is that of *probabilistic entailment*. Again a set of logical sentences $S_1, \ldots, S_m$, together with probabilities $p_1, \ldots, p_m$ is given, and one extra logical sentence S_{m+1}, the target. It is asked to find best possible upper and lower bounds on the probability that S_{m+1} is true, given $S_1, \ldots, S_m$ are satisfied with respective probabilities $p_1, \ldots, p_m$. The interval defined by these lower and upper bounds forms the so-called *tight entailment* of S_{m+1}. It is well known that both PSAT and probabilistic entailment can be solved nondeterministically in polynomial time using linear programming techniques. In our framework, a complete system of frequent sets is a system that only contains tight frequent expressions; i.e. the bounds of the frequent expressions in the complete system are the best possible in view of the system, and as such, this corresponds to the notion of tight entailment.

For a comprehensive overview of probabilistic logic, entailment and various extensions, we refer to [9][10]. Nilsson's probabilistic logic and entailment are extended in various ways, including assigning intervals to logical expressions instead of exact probability values and considering conditional probabilities [6].

In [4], *Fagin et al.* study the following extension. A *basic weight formula* is an expression $a_1 w(\phi_1) + \ldots + a_k w(\phi_k) \geq c$, where $a_1, \ldots, a_k$ and c are integers and $\phi_1, \ldots, \phi_k$ are propositional formulas, meaning that the sum of all a_i times the *weight* of ϕ_i is greater than or equal to c. A *weight formula* is a boolean combination of basic weight formulas. The semantics are introduced by an underlying probability space. The weight of a formula corresponds with the probability that it is true. The main contribution (from the viewpoint of our paper) of [4] is the description of a sound and complete axiomatization for this probabilistic logic. The logical framework in our paper is in some sense embedded into the logic in [4]. Indeed, if we introduce a propositional symbol P_i for each item i, the frequent set expression $K :: p_K$ can be translated as $w(\bigwedge_{i \in K} P_i) \geq p_K$. As such, by results obtained in [4], the implication problem in our framework is guaranteed to be decidable. Satisfiability, and thus also the implication problem, are NP-complete in Fagin's framework. Our approach differs from Fagin's approach in the sense that we only consider situations where for all expressions a probability is given.

Also in [6], axioms for a probabilistic logic are introduced. However, the authors are unable to proof whether the axioms are complete. For a sub-language (Type-A problems), they proof that their set of axioms is complete. However, this sub-language is not sufficiently powerful to express frequent itemset expressions.

On the other side of the spectrum, we have related work within the context of data mining. There have been attempts to proof some completeness results for itemsets in this area. One such attempt is described shortly in [11]. In the presence of constraints on the allowable itemsets, the authors introduce the notion of *ccc-optimality*[1]. ccc-optimality can intuitively be understood as "the algorithm only generates and tests itemsets that still can be frequent, using the current knowledge." Our approach however, is more general, since we do not restrict ourselves to a particular algorithm. No attempt is known to us in the context of data mining, that studies what we can derive from an arbitrary set of frequent itemsets.

Finally, we would like to add that in our paper the emphasis is on introducing a logical framework for frequent itemsets and not on introducing a new probabilistic logic, nor on algorithms.

3 Complete System of Frequent Sets

We formally define a system of frequent sets. We also define what it means for a system to be complete.

To represent a database with transactions, we use a matrix. The columns of the matrix represent the items and the rows represent the transactions. The matrix contains a one in the (i, j)-entry if transaction i contains item j, else this entry is zero. When R is a matrix where the columns represent the items in I, we say that R is a matrix over I. In our running example we regularly refer to the items with capital letters. With this notation, we get the following definition:

Definition 1. *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, and R be a matrix over I. The* frequency *of an itemset $K \subseteq I$ in R, denoted $freq(K, R)$ is the fraction of rows in R that have a one in every column of K.*

[1] ccc-optimality stands for <u>C</u>onstraint <u>C</u>hecking and <u>C</u>ounting-optimality

Example 1. In Fig. 1, a matrix is given, together with some frequencies. The frequency of DEF [2] is 0.2, because 2 rows out of 10 have a one in every column of DEF. Note that, because R is a matrix, R can have identical rows.

Matrix R

A	B	C	D	E	F
1	0	1	0	1	1
1	0	1	0	1	1
0	1	0	1	1	0
1	1	1	0	0	1
1	0	0	1	0	1
0	1	0	1	1	1
1	1	0	1	1	1
0	0	1	0	0	1
1	1	1	0	1	0
1	0	0	1	0	1

$freq(A, R) = 0.7$
$freq(B, R) = 0.5$
$freq(AB, R) = 0.3$
$freq(DEF, R) = 0.2$

R satisfies $A :: 0.5$, $AB :: 0.3$, $DEF :: 0.1$
R does not satisfy $A :: 0.8$, $ABC :: 0.4$, $DEF :: 0.3$

Fig. 1. A matrix together with some frequent set expressions

We now introduce logical implication and completeness of a system of frequent sets.

Definition 2. *Let $I = \{I_1, \ldots, I_n\}$ be a set of items.*

- *A* frequent set expression *over I is an expression $K :: p_K$ with $K \subseteq I$ and p_K rational with $0 \leq p_K \leq 1$.*
- *A matrix R over I* satisfies *$K :: p_K$ iff $freq(K, R) \geq p_K$. Hence itemset K has frequency at least p_K.*
- *A* system of frequent sets *over I is a collection*

$$\left\{_{K\subseteq I} K :: p_K\right.$$

of frequent set expressions, with exactly one expression for each $K \subseteq I$.
- *A matrix R over I* satisfies *the system $\left\{_{K\subseteq I} K :: p_K\right.$ iff R satisfies all $K :: p_K$.*

Example 2. In Fig. 1, the matrix R satisfies $A :: 0.6$, because the frequency of A in R is bigger than 0.6. The matrix does not satisfy $B :: 0.7$, because the frequency of B is lower than 0.7.

Definition 3. *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, and $K \subseteq I$.*

- *A system of frequent sets S over I* logically implies *$K :: p_K$, denoted $S \models K :: p_K$, iff every matrix that satisfies S, also satisfies $K :: p_K$. System S_1* logically implies *system S_2, denoted $S_1 \models S_2$, iff every $K :: p$ in S_2 is logically implied by S_1.*

[2] DEF denotes the set $\{D, E, F\}$

B, C, BC, ABC

A	B	C
1	1	0
1	1	1
1	0	1
1	1	1
0	1	1

A, AB, AC

A	B	C
1	1	0
1	1	1
1	1	1
0	1	1
0	1	1

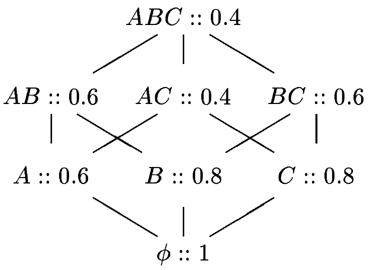


Fig. 2. Proof-matrices for a system of frequent sets

- *A system of frequent sets* $S = \{_{K \subseteq I}\ K :: p_K$ *is* complete *iff for each* $K :: p$ *logically implied by* S, $p \leq p_K$ *holds.*

Example 3. Let $I = \{A, B, C, D, E, F\}$. Consider the following system: $S = \{_{K \subseteq I}\ K :: p_K$, where $p_A = 0.7$, $p_B = 0.5$, $p_{AB} = 0.3$, $p_{DEF} = 0.2$, and $p_K = 0$ for all other itemsets K. The matrix in Fig. 1 satisfies S. S is not complete, because in every matrix satisfying $DEF :: 0.2$, the frequency of DE must be at least 0.2, and S contains $DE :: 0$. Furthermore, S *does not* logically imply $EF :: 0.5$, since R satisfies S, and R does not satisfy $EF :: 0.5$.

Consider the following system over $I = \{A, B, C\}$:
$\{\phi :: 1, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, BC :: 0.6, ABC :: 0.4\}$. This system is complete. We prove this by showing that for every subset K of I, there exists a matrix R_K that satisfies S, and $freq(K, R_K)$ is exactly p_K. These matrices then *prove* that for all K, we cannot further improve on K; i.e. make p_K larger. These proof-matrices are very important in the proof of the axiomatization that is given in the next section. In Fig. 2, the different proof-matrices are given.

When a system S is not complete, we can improve this system. Suppose a system $S = \{_{K \subseteq I}\ K :: p_K$ is not complete, then there is a frequent set expression $K :: p'_k$ that is logically implied by S, and $p'_K > p_K$. We can improve S by replacing $K :: p_K$ by $K :: p'_K$. The next proposition says that there exists a unique system $C(S)$, that is logically implied by S and that is complete.

Proposition 1. *Let* $I = \{I_1, \ldots, I_n\}$ *be a set of items, and* $S = \{_{K \subseteq I}\ K :: p_K$ *be a system of frequent sets. There exists a unique system* $C(S)$*, the* completion of S*, such that* $S \models C(S)$*, and* $C(S)$ *is a complete system.*

Proof. Let $M_K = \{p_K \mid S \models K :: p_K\}$. M_K always contains its supremum. This can easily be seen as follows: suppose a matrix M satisfies S. Let p be the frequency of K in M. Since M satisfies S, for all $p_K \in M_K$, $p \geq p_K$ holds, and hence $p \geq sup(M_K)$ holds. Hence, every matrix satisfying S, also satisfies $K :: sup(M_K)$, and thus $S \models K :: sup(M_K)$. It is straightforward that the system $\{_{K \subseteq I}\ K :: supp(M_K)$ is the unique completion of S.

Example 4. $I = \{A, B, C\}$. The system $\{\phi :: 1, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, \mathbf{BC} :: \mathbf{0.6}, ABC :: 0.4\}$ is the unique completion of the system $\{\phi :: 0.8, A :: 0.6, B :: 0.8, C :: 0.8, AB :: 0.6, AC :: 0.4, \mathbf{BC} :: \mathbf{0.4}, ABC :: 0.4\}$. $BC :: 0.6$ is implied by the second system, since there is an overlap of at least 0.6 between the rows having a one on B and the rows having a one on C.

Remark that when a system is complete, it is not necessary that there exists *one* matrix such that for all itemsets the frequency is exactly the frequency given in the system. Consider for example the following system: $\{\phi :: 1, A :: 0.5, B :: 0.5, C :: 0.1, AB :: 0, AC :: 0, BC :: 0, ABC :: 0\}$. This system is complete. However, we will never find a matrix in which the following six conditions are simultaneously true: $freq(A) = 0.5$, $freq(B) = 0.5$, $freq(C) = 0.1$, $freq(AB) = 0$, $freq(AC) = 0$, and $freq(BC) = 0$, because due to $freq(A) = 0.5$, $freq(B) = 0.5$, and $freq(AB) = 0$, every row has a one in A or in B. So, every row having a one in C has also a one in A or in B, and thus violates respectively $freq(AC) = 0$, or $freq(BC) = 0$.

4 Axiomatizations

We give an axiomatization for frequent sets. An axiomatization in this context is a set of rules that are satisfied by the system if and only if it is complete. In order to simplify the notation we first introduce rare sets. In Section 5 we will show how we can build finite proofs for all logical implications using the axioms as rules of inference.

4.1 Rare Sets

Definition 4. *Let $I = \{I_1, \ldots, I_n\}$ be a set of items, and $K \subseteq I$.*

- *Let R be a matrix over I. The* rareness *of an itemset $K \subseteq I$ in R, denoted $rare(K, R)$, is the fraction of rows in R that have a zero in at least one column of K.*
- *A* rare set expression *over I is an expression $K : p_K$ with $K \subseteq I$ and p_K rational with $0 \leq p_K \leq 1$.*
- *A matrix R over I* satisfies *$K : p_K$ iff $rare(K, R) \leq p_K$. Hence itemset K has rareness at most p_K.*
- *A* system of rare sets *over I is a collection $\{_{K \subseteq I} K : P_K$ of rare set expressions, with exactly one expression for each $K \subseteq I$.*
- *A matrix R over I* satisfies *the system $\{_{K \subseteq I} K : p_K$ iff R satisfies all $K : p_K$.*
- *A system of rare sets S over I* logically implies *$K : p$, denoted $S \models K : p$ iff every matrix that satisfies S also satisfies $K : p$. System S_1 logically implies system S_2, denoted $S_1 \models S_2$, iff every $K : p$ in S_2 is logically implied by S_1.*
- *A system of rare sets $S = \{_{K \in I} K : p_K$ is* complete *iff for each $K : p$ logically implied by S, $p_K \leq p$ holds.*

Example 5. In Fig. 1, the matrix R satisfies $A : 0.4$, because the rareness of A in R is smaller than 0.4. The matrix does not satisfy $B : 0.3$, because the rareness of B is greater than 0.3. Let $I = \{A, B\}$. The system $\{AB : 0.8, A : 0.3, B : 0.4, \phi : 0.4\}$ is not complete. The unique completion of this system is $\{AB : 0.7, A : 0.3, B : 0.4, \phi : 0\}$.

The next proposition connects rare sets with frequent sets. The connection between the two is straightforward. Indeed: the set of rows that have a zero in at least one column on K is exactly the complement of the set of rows having only ones in these columns. The second part of the proposition shows that an axiomatization for rare sets automatically yields an axiomatization for frequent sets.

Proposition 2. *Let $I = \{I_1 \dots I_n\}$ be a set of items. For every matrix R over I and every subset K of I holds that*

- $freq(K, R) + rare(K, R) = 1$.
- *R satisfies $K : p_K$ iff R satisfies $K :: 1 - p_K$.*

In the following subsection we prove an axiomatization for complete systems of rare sets. From this axiomatization, we can easily derive an axiomatization for frequent sets, using the last proposition.

4.2 Axiomatization of Rare Sets

Before we give the axiomatization, we first introduce our notation of bags.

Definition 5.

- *A* bag *over a set S is a total function from S into $\{0, 1, 2, \dots\}$.*
- *Let $\mathbf{K}$ be a bag over S and $s \in S$. We say that s* appears n times in $\mathbf{K}$ *iff $\mathbf{K}(s) = n$.*
- *If $\mathbf{K}$ and $\mathbf{L}$ are bags over S, then we define the* bag-union of $\mathbf{K}$ and $\mathbf{L}$, *notation $\mathbf{K} \bigcup \mathbf{L}$, as follows: for all $s \in S$, $(\mathbf{K} \bigcup \mathbf{L})(s) = \mathbf{K}(s) + \mathbf{L}(s)$.*
- *Let $S = \{s_1, s_2, \dots, s_n\}$. $\{\!\{ c_1{}'s_1, \dots, c_n{}'s_n \}\!\}$ denotes the bag over S in which s_i appears c_i times for $1 \leq i \leq n$.*
- *Let S be a set, $\mathbf{K}$ a bag over S. $\sum_{s \in S} \mathbf{K}(s)$ is the* cardinality of $\mathbf{K}$, *and is denoted by $|\mathbf{K}|$.*
- *Let $\mathbf{K}$ be a bag over the subsets of a set S. Then $\bigcup \mathbf{K}$ denotes the bag $\bigcup_{K \in \mathbf{K}} K$. The degree of an element $s \in S$ in $\mathbf{K}$, denoted $deg(s, \mathbf{K})$ is the number of times s appears in $\bigcup \mathbf{K}$.*

Example 6. $\mathbf{K} = \{\!\{ 1'\{a, b\}, 2'\{b, c\}, 2'\{b, d\} \}\!\}$ is a bag over the subsets of $\{a, b, c, d\}$. $\bigcup \mathbf{K} = \{\!\{ 1'a, 5'b, 2'c, 2'd \}\!\}$. $deg(b, \mathbf{K}) = 5$. $|\mathbf{K}| = 5$.

The next three rules form an axiomatization for complete systems of rare sets in the sense that the complete systems are exactly the ones that satisfy these three rules. The p_K's that appear in the rules, indicate the rareness-values given in the system for the set K; i.e. $K : p_K$ is in the system.

R1 $p_\phi = 0$
R2 If $K_2 \subseteq K_1$, then $p_{K_2} \leq p_{K_1}$
R3 Let $K \subseteq I$, $\mathbf{M}$ a bag of subsets of K. Then

$$p_K \leq \frac{\sum_{M \in \mathbf{M}} p_M}{k},$$

with $k = min_{a \in K}(deg(a, \mathbf{M}))$[3]

The next theorem is one of the most important results of this paper. The following lemma, proved in [3], will be used in the proof of the theorem.

Lemma 1. *Given a set of indices I and given rational numbers a_K, b_K for every non-empty $K \subseteq I$. Consider the following system of inequalities:*

$$\left\{_{K \subseteq I} \; a_K \leq \sum_{i \in K} X_i \leq b_K \right.$$

This system has a solution $(x_1, \ldots, x_{\#I})$, x_i rational, iff for all $\mathbf{K}$ *and* $\mathbf{L}$*, bags of subsets of I with $\bigcup \mathbf{K} = \bigcup \mathbf{L}$ holds that $\sum_{K \in \mathbf{K}} a_K \leq \sum_{L \in \mathbf{L}} b_L$.*

Theorem 1. *Let $S = \left\{_{K \subseteq I} K : p_K\right.$ be a system of rare sets over I. The following two statements are equivalent:*

- *S is a complete system.*
- *S satisfies* **R1***,* **R2***, and* **R3***.*

Proof. ($\Rightarrow$) **R1** and **R2** are trivial.
R3: Let $\mathbf{M}$ be a bag over the subsets of an itemset K, and $S = \left\{_{K \subseteq I} K : p_K\right.$ is a complete system. Let R be an arbitrary matrix that satisfies S. D_K^R is the bag that contains exactly those rows r for which there exists a k in K such that $r(k) = 0$. Then, for every L holds: $\frac{|D_L^R|}{|R|} \leq p_L$. If $r \in D_K^R$, then there exists a $a \in K$ such that $r(a) = 0$. a appears in at least $k = min_{a \in K} deg(a, \mathbf{M})$ of the sets of $\mathbf{M}$. Thus, $k|D_K^R| \leq \sum_{M \in \mathbf{M}} |D_M^R|$. We can conclude that in every matrix satisfying S, $rare(K, R) = \frac{|D_K^R|}{|R|} \leq \frac{\sum_{M \in \mathbf{M}} p_M}{k}$.
($\Leftarrow$) We show that if $S = \left\{_{K \subseteq I} K : p_K\right.$ satisfies **R1**, **R2**, and **R3**, we can for each itemset K find a proof-matrix $\widehat{R_K}$, such that $\widehat{R_K}$ satisfies S, and $rare(K, \widehat{R_K}) = p_K$ [4]. We specify $\widehat{R_K}$ by giving the frequency of every possible row r. β_Z denotes the fraction of rows that have a zero in every column of Z, and a one elsewhere. We will show that there exists such a matrix $\widehat{R_K}$ with only rows with at most one zero, and this zero, if present, must be in a column of K; i.e. whenever $|Z| > 1$ or $Z \not\subseteq K$, $\beta_Z = 0$.

[3] If $k = 0$, **R3** should be interpreted as "$p_K \leq 1$"

[4] Remark the similarities with the traditional Armstrong-relations in functional dependency theory [5]

This can be expressed by the following system of inequalities:

$$\begin{cases} \forall a \in K : 0 \leq \beta_a \leq 1 & \text{(1) all fractions are between 0 and 1} \\ 0 \leq \beta_0 \leq 1 & \text{(2) idem} \\ (\sum_{a \in K} \beta_a) + \beta_0 = 1 & \text{(3) the frequencies add up to one} \\ p_K = \sum_{a \in K} \beta_a & \text{(4) the rareness of } K \text{ is exactly } p_K \\ \forall L \subset K : p_L \geq \sum_{a \in L} \beta_a & \text{(5) for other sets } L,\ p_L \geq rare(L, \widehat{R_K}) \end{cases}$$

Every solution of this system describes a matrix that satisfies S. Only (5) needs a little more explanation. For an arbitrary itemset L, $rare(L, \widehat{R_K}) = rare(L \cap K, \widehat{R_K})$ due to the construction. Because S satisfies **R2**, $p_L \geq p_{K \cap L}$. Therefore, it suffices to demand that $rare(L, \widehat{R_K}) \leq p_L$, for all $L \subset K$.

The system has a solution if the following (simpler) system has a solution:

$$\left\{ \forall L \subseteq K : p_K - p_L \leq \sum_{a \in K} \beta_a - \sum_{a \in L} \beta_a \leq p_K \ (1') \right.$$

1 is ok: choose $L = K - \{a\}$, then $0 \leq^{(\mathbf{R2})} p_K - p_{K-\{a\}} \leq \beta_a \leq p_K \leq 1$

2+3 are ok: let $\beta_0 = 1 - \sum_{a \in K} \beta_a = 1 - p_K$

4 is ok: choose $L = \phi$, $p_L = 0$ (**R1**), and thus $p_K \leq \sum_{a \in K} \beta_K \leq p_K$

5 is ok: $p_L - p_K \geq \sum_{a \in L} \beta_a - \sum_{a \in K} \beta_a + 4$.

According to Lemma 1, this last system has a rational solution iff for all bags $\mathbf{M}$ and $\mathbf{N}$ over the subsets of K , such that $\bigcup \mathbf{M} = \bigcup \mathbf{N}$, $\sum_{M \in \mathbf{M}} (p_K - p_{K-M}) \leq \sum_{N \in \mathbf{N}} p_N$ holds.

Let $\mathbf{L} = \mathbf{N} \bigcup \{\!\{\ K - M \mid M \in \mathbf{M}\ \}\!\}$. Then, by **R3** we have that $\frac{\sum_{L \in \mathbf{L}} p_L}{k} \geq p_K$, with $k = min_{a \in K} \#(\{\!\{\ N \mid a \in N \wedge N \in \mathbf{N}\ \}\!\} \bigcup \{\!\{\ M \mid M \in \mathbf{M} \wedge a \notin M\ \}\!\})$. Because $\#\{\!\{\ M \mid M \in \mathbf{M} \wedge a \in M\ \}\!\} = \#\{\!\{\ N \mid N \in \mathbf{N} \wedge a \in n\ \}\!\}$, $k = \#\mathbf{M}$. We have: $\sum_{L \in \mathbf{L}} p_L \geq \#\mathbf{M} p_K$. Since $\sum_{L \in \mathbf{L}} p_L = \sum_{N \in \mathbf{N}} p_N + \sum_{M \in \mathbf{M}} p_{K-M}$ and $\#\mathbf{M} p_K = \sum_{M \in \mathbf{M}} p_K$, $\sum_{M \in \mathbf{M}} (p_K - p_{K-M}) \leq \sum_{N \in \mathbf{N}} p_N$ holds.

Example 7. The system $\{\phi : 0.5, A : 0.5, B : 0.25, C : 0.5, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since $\phi : 0.5$ violates **R1**.

The system $\{\phi : 0, A : 0.5, B : 0.25, C : 0.5, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since for example $AB : 0$ and $A : 0.5$ together violate **R2**.

The system $\{\phi : 0, A : 0, B : 0, C : 0, AB : 0, AC : 1, BC : 0, ABC : 1\}$ is not complete, since $A : 0$, $C : 0$, and $AC : 1$ together violate **R3**.

The system $\{\phi : 0, A : 0, B : 0, C : 0, AB : 0, AC : 0, BC : 0, ABC : 0\}$ is complete, since it satisfies **R1**, **R2**, and **R3**. This system is the unique completion of all systems in this example.

4.3 Axiomatization of Frequent Sets

From Proposition 2, we can now easily derive the following axiomatization for frequent sets.

F1 $p_\phi = 1$

F2 If $K_2 \subseteq K_1$, then $p_{K_2} \geq p_{K_1}$

F3 Let $K \subseteq I$, $\mathbf{M}$ a bag of subsets of K. Then

$$p_K \geq 1 - \frac{\#\mathbf{M} - \sum_{M \in \mathbf{M}} p_M}{k},$$

with $k = min_{a \in K}(deg(a, \mathbf{M}))$[5]

Theorem 2. *Let $S = \{_{K \subseteq I}\ K :: p_K$ be a system of frequent sets over I. The following two statements are equivalent:*

- *S is a complete system.*
- *S satisfies* **F1**, **F2**, *and* **F3**.

5 Inference

In the rest of the text we continue working with rare sets. The results obtained for rare sets can, just like the axiomatization, be carried over to frequent sets.

In the previous section we introduced and proved an axiomatization for complete systems of rare and frequent sets. There is however still one problem with this axiomatization. **R3** states a property that has to be checked for all bags over the subsets of K. This number of bags is infinite. In this section we show that it suffices to check only a finite number of bags: the minimal multi-covers. We show that the number of minimal multi-covers over a set is finite, and that they can be computed.

We also look at the following problem: when an incomplete system is given, can we compute its completion using the axioms? We show that this is indeed possible. We use **R1**, **R2**, and **R3** as inference rules to adjust rareness values in the system; whenever we detect an inconsistency with one of the rules, we improve the system. When the rules are applied in a systematic way, this method leads to a complete system within a finite number of steps.

Actually, the completion of a system of frequent sets can be computed in an obvious way by using linear programming. Indeed, when we look at the proof of theorem 1, we can compute the completion of the system of inequalities by applying linear programming. For all sets K, we can minimize p_K with respect to a system of inequalities expressing that the frequencies obey the system of rare sets. Since the system of inequalities has polynomial size in the number of frequent itemsets, this algorithm is even polynomial in the size of the system. However, in association rule mining, it is very common that the number of itemsets becomes very large and thus the system of inequalities will in practical situations certainly become prohibitive large. Therefore, solving the linear programming problem is a theoretical solution, but not a practical one. Also, as mentioned in [6], an axiomatization has as an advantage that it provides human-readable proofs, and that, when the inference is stopped before termination, still a partial solution is provided.

[5] If $k = 0$, **R3** should be interpreted as "$p_K \geq 0$"

5.1 Minimal Multi-covers

Definition 6.

- *A* k-cover *of a set* S *is a bag* $\mathbf{K}$ *over the subsets of* S *such that for all* $s \in S$, $deg(s, \mathbf{K}) = k$.
- *A bag* $\mathbf{K}$ *over the subsets of a set* S *is a* multi-cover *of* S *if there exists an integer* k *such that* $\mathbf{K}$ *is a* k*-cover of* S.
- *A* k*-cover* $\mathbf{K}$ *of* S *is* minimal *if it cannot be decomposed as* $\mathbf{K} = \mathbf{K}_1 \bigcup \mathbf{K}_2$, *with* $\mathbf{K}_1$ *and* $\mathbf{K}_2$ *respectively* k_1*- and* k_2*-covers of* S, $k_1 > 0$ *and* $k_2 > 0$.

Example 8. Let $K = \{A, B, C, D\}$. $\{\!\{\ 1'AB, 1'BC, 1'CD, 1'AD, 1'ABCD\ \}\!\}$ is a 3-cover of K. It is not minimal, because it can be decomposed into the following two minimal multi-covers of K: $\{\!\{\ 1'AB, 1'BC, 1'CD, 1'AD\ \}\!\}$ and $\{\!\{\ 1'ABCD\ \}\!\}$.

The new rule that replaces **R3** states that it is not necessary to check all bags; we only need to check the minimal multi-covers. This gives the following **R3**':

R3' Let $K \subseteq I$, $\mathbf{M}$ a minimal k-cover of K. Then

$$p_K \leq \frac{\sum_{M \in \mathbf{M}} p_M}{k} .$$

Theorem 3. *Let* S *be a system of rare sets over* I. *The following statements are equivalent:*

1. *S satisfies* **R1**, **R2**, *and* **R3**.
2. *S satisfies* **R1**, **R2**, *and* **R3**'.

Sketch of the proof. (1) The direction **R1**, **R2**, **R3** implies **R1**, **R2**, **R3**' is trivial, since every k-cover of K is also a bag over the subsets of K, where the minimal degree is k.
(2) Suppose the system S satisfies **R1** and **R2**, but violates **R3**. There exists a set K and a bag $\mathbf{K}$ over the subsets of K, such that $p_K > \frac{\sum_{L \in \mathbf{K}} p_L}{k}$, with $k = min_{a \in K} deg(a, \mathbf{K})$. Starting from this bag, one can construct a minimal multi-cover of K, that violates **R3**'. We show this construction with an example. Suppose $\mathbf{K} = \{\!\{\ AB, BC, ABC\ \}\!\}$. Every element appears at least 2 times in $\mathbf{K}$. We first construct a multi-cover from $\mathbf{K}$, by removing elements that appear more than others. In this example, B appears 3 times, and all other elements appear only 2 times. We remove B from one of the sets in $\mathbf{K}$, resulting in $\{\!\{\ A, BC, ABC\ \}\!\}$. The sum over $\mathbf{K}$ became smaller by this operation, since S satisfies **R2**. This multi-cover can be split into two different minimal multi-covers: $\mathbf{K}_1 = \{\!\{\ A, BC\ \}\!\}$, and $\mathbf{K}_2 = \{\!\{\ ABC\ \}\!\}$. Because now $\frac{\sum_{L \in \mathbf{K}} p_L}{2} = \frac{\sum_{L \in \mathbf{K}_1} p_L + \sum_{L \in \mathbf{K}_2} p_L}{1+1}$, for at least one i, $\frac{\sum_{L \in \mathbf{K}_i} p_L}{1}$ is smaller than $\frac{\sum_{L \in \mathbf{K}} p_L}{2}$.

Proposition 3. *Let* K *be a finite set. The number of minimal multi-covers of* K *is finite and computable.*

The proof can be found in [3].

5.2 Computing the Completion of a System with Inference Rules

We prove that by applying **R1**, **R2**, and **R3**' as rules, we can compute the completion of any given system of rare sets. Applying for example rule **R2** means that whenever we see a situation $K_1 \subseteq K_2$, and the system states $K_1 : p_{K_1}$ and $K_2 : p_{K_2}$, and $p_{K_2} < p_{K_1}$, we improve the system by replacing $K_1 : p_{K_1}$ by $K_1 : p_{K_2}$. It is clear that **R1** can only be applied once; **R2** and **R3** never create situations in which **R1** can be applied again.

R2 is a *top-down operation*, in the sense that the rareness values of smaller sets is adjusted using values of bigger sets. So, for a given system S we can easily reach a fixpoint for rule **R2**, by going top-down; we first try to improve the frequencies of the biggest itemsets, before continuing with the smaller ones.

R3 is a *bottom-up operation*; values of smaller sets are used to adjust the values of bigger sets. So, again, for a given system S, we can reach a fixpoint for rule **R3**, by applying the rule bottom-up.

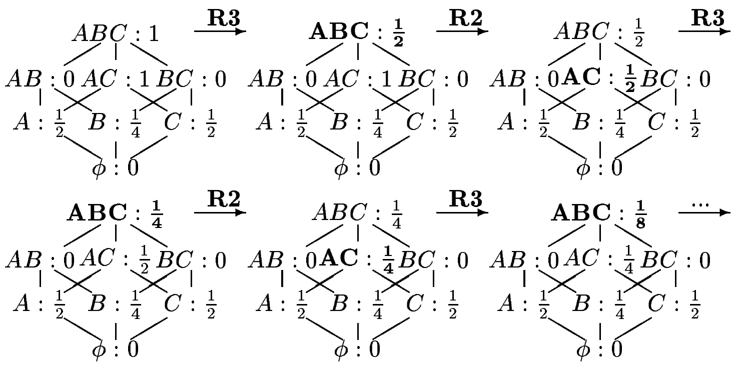
A trivial algorithm to compute the completion of a system is the following: apply **R1**, and then keep applying **R2** and **R3** until a fixpoint is reached. Clearly, the *limit* of this approach yields a complete system, but it is not clear that a fixpoint will be reached within a finite number of steps. Moreover, there are examples of situations in which infinite loops are possible. In Fig. 3, such an example is given. The completion of the first system, is clearly all rareness values equal to zero, because for every matrix satisfying the system, none of the rows have a zero in AB, and none have a zero in BC, so there are no zeros at all in the matrix. When we keep applying the rules as in Fig. 3, we never reach this fixpoint, since in step $2n$, the value for ABC is $\left(\frac{1}{2}\right)^n$. This is however not a problem; we show that when we apply the rules **R2** and **R3** in a systematic way, we always reach a fixpoint within a finite number of steps. This systematic approach is illustrated in Fig. 4. We first apply **R2** top-down until we reach a fixpoint for **R2**, and then we apply **R3** bottom-up until we reach a fixpoint for **R3**. The general systematic approach is written down in Fig. 5. We prove that for every system these two meta-steps are all there is needed to reach the completion.

Definition 7. *Let I be a set of items, $J \subseteq I$, and $S = \{_{K \subseteq I}\ K : p_K$ a system of rare sets over I. The* projection of S on J, *denoted $proj(S, J)$, is the system $S' = \{_{K \subseteq J}\ K : p_K$.*

Lemma 2. *Let I be a set of items, $J \subseteq I$, and $S = \{_{K \subseteq I}\ K : p_K$ a system of rare sets over I. If S satisfies* **R2**, *then $proj(C(S), J) = C(proj(S, J))$.*

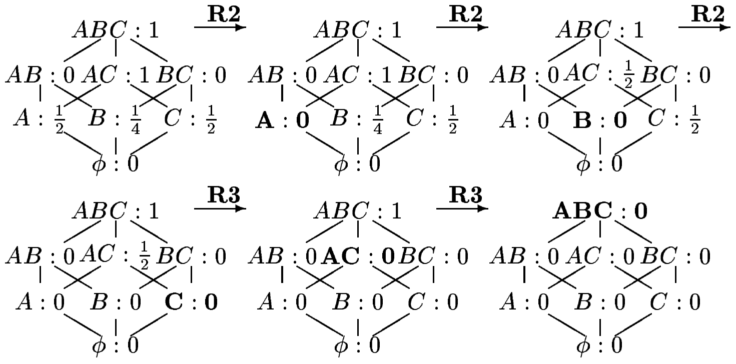
Theorem 4. *The algorithm in Fig. 5 computes the completion of the system of rare sets S.*

Sketch of the proof. Let $I = \{A, B, C\}$, and S be a system of rare sets over I. After the top-down step, the resulting system satisfies **R2**. First we apply **R3** to adjust the value of A. Because S satisfies **R2**, and after application of **R3** on A, the system $\{\phi : 0, A : p_A\}$ is complete, we cannot further improve



$ABC : 1$
$\xrightarrow{\mathbf{R3}}$
$AB : 0$ $AC : 1$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$\mathbf{ABC} : \frac{1}{2}$
$\xrightarrow{\mathbf{R2}}$
$AB : 0$ $AC : 1$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$ABC : \frac{1}{2}$
$\xrightarrow{\mathbf{R3}}$
$AB : 0$ $\mathbf{AC} : \frac{1}{2}$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$\mathbf{ABC} : \frac{1}{4}$
$\xrightarrow{\mathbf{R2}}$
$AB : 0$ $AC : \frac{1}{2}$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$ABC : \frac{1}{4}$
$\xrightarrow{\mathbf{R3}}$
$AB : 0$ $\mathbf{AC} : \frac{1}{4}$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$\mathbf{ABC} : \frac{1}{8}$
$\xrightarrow{\cdots}$
$AB : 0$ $AC : \frac{1}{4}$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$

Fig. 3. "Random" application of the rules can lead to infinite loops



$ABC : 1$
$\xrightarrow{\mathbf{R2}}$
$AB : 0$ $AC : 1$ $BC : 0$
$A : \frac{1}{2}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$ABC : 1$
$\xrightarrow{\mathbf{R2}}$
$AB : 0$ $AC : 1$ $BC : 0$
$\mathbf{A} : \mathbf{0}$ $B : \frac{1}{4}$ $C : \frac{1}{2}$
$\phi : 0$
$ABC : 1$
$\xrightarrow{\mathbf{R2}}$
$AB : 0$ $AC : \frac{1}{2}$ $BC : 0$
$A : 0$ $\mathbf{B} : \mathbf{0}$ $C : \frac{1}{2}$
$\phi : 0$
$ABC : 1$
$\xrightarrow{\mathbf{R3}}$
$AB : 0$ $AC : \frac{1}{2}$ $BC : 0$
$A : 0$ $B : 0$ $\mathbf{C} : \mathbf{0}$
$\phi : 0$
$ABC : 1$
$\xrightarrow{\mathbf{R3}}$
$AB : 0$ $\mathbf{AC} : \mathbf{0}$ $BC : 0$
$A : 0$ $B : 0$ $C : 0$
$\phi : 0$
$\mathbf{ABC} : \mathbf{0}$
$AB : 0$ $AC : 0$ $BC : 0$
$A : 0$ $B : 0$ $C : 0$
$\phi : 0$

Fig. 4. Systematic application of the rules avoids infinite computations

on A; $proj(C(S), \{A\}) = C(S, proj(S, \{A\}))$. We can use the same argument for B and C. Then we apply **R3** to adjust the value of AC. After this step, $\{\phi : 0, A : p_A, B : p_B, AC : p_{AC}\}$ satisfies **R3**. This system also satisfies **R2**, because otherwise we could improve on A or on B, and we just showed that we cannot further improve on A or B. Thus, the system $proj(S, \{A, C\})$ is closed, and thus we cannot further improve on AC. This way, we iteratively go up, and finally we can conclude that S must be complete after the full bottom-up step.

6 Summary and Further Work

We presented an axiomatization for complete systems of frequent sets. As an intermediate stage in the proofs, we introduced the notion of a system of rare sets. The axiomatization for rare sets contained three rules **R1**, **R2**, and **R3**. From these rules we could easily derive the axiomatization, **F1**, **F2**, and **F3** for frequent sets. Because rule **R3** yields a condition that needs to be checked for an infinite number of bags, we replaced **R3** by **R3'**. We showed that the completion

```
Close($S$)
    $p_\phi = 0$
    TopDown($S$)
    BottomUp($S$)

TopDown($S$)
for $i = n$ downto 1 do
    for all itemsets $K$ of cardinality $i$ do
        make $p_K = min_{K \subseteq L}(p_L)$

BottomUp($S$)
    for $i = 1$ to $n$ do
        for all itemsets $K$ of cardinality $i$ do
            make $p_K = min_{\mathbf{K}\text{, minimal } k\text{-cover of } K}\left(\frac{\sum_{K' \in \mathbf{K}} p_{K'}}{k}\right)$
```

Fig. 5. Algorithm Close for finding the completion of the system $S = \left\{_{K \subseteq I} K : p_K\right.$ over $I = \{I_1, \ldots, I_n\}$

can be computed by applying **R1**, **R2**, and **R3**' as inference rules. If these rules are applied first top-down, and then bottom-up, the completion is reached within a finite number of steps. In the future we want to study an axiomatization for systems in which not for every set a frequency is given. For some preliminary results on these *sparse systems*, we refer to [3]. Another interesting topic is expanding the axiomatization to include association rules and confidences.

References

[1] R. Agrawal, T. Imilienski, and A. Swami. Mining association rules between sets of items in large databases. In *Proc. ACM SIGMOD*, 1993

[2] R. Agrawal, R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. VLDB*, 1994

[3] T. Calders, and J. Paredaens. A Theoretical Framework for Reasoning about Frequent Itemsets. Technical Report 006, Universiteit Antwerpen, Belgium, http://win-www.uia.ac.be/u/calders/download/axiom.ps, June 2000.

[4] R. Fagin, J. Halpern, and N. Megiddo. A Logic for Reasoning about Probabilities. In *Information and Computation* 87(1,2): 78-128, 1990.

[5] R. Fagin, M. Y. Vardi. Armstrong Databases for Functional and Inclusion Dependencies. In *IPL 16(1): 13-19*, 1983.

[6] A. M. Frisch, P. Haddawy. Anytime Deduction for Probabilistic Logic. In *Artificial Intelligence* 69(1-2): 93-122, 1994.

[7] G. Georgakopoulos, D. Kavvadias, and C. H. Papadimitriou. Probabilistic Satisfiability. In *Journal of Complexity* 4:1-11, 1988.

[8] J. Han, J.Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. ACM SIGMOD*, 2000

[9] P. Hansen, B. Jaumard, G.-B. D. Nguets, M. P. de Aragão. Models and Algorithms for Probabilistic and Bayesian Logic. In *Proc. IJCAI*, 1995

[10] P. Hansen, B. Jaumard. Probabilistic Satisfiability. *Les Cahiers du GERAD* G-96-31, 1996

[11] L. V.S. Laksmanan, R.T. Ng, J. Han, and A. Pang. Optimization of Constrained Frequent Set Queries with 2-variable Constraints. *Proc. ACM SIGMOD*, 1999

[12] N. Nilsson. Probabilistic Logic. In *Artificial Intelligence* 28: 71-87, 1986