Technische Universität Dresden
Herausgeber: Der Rektor

Expected Numbers of Proper Premises and Concept Intents

Felix Distel[1] und Daniel Borchmann[1]

MATH–AL–03-2011–2011 July 2011

[1]TU Dresden

1 Introduction

For many years computing the stem base has been the default method for extracting a small but complete set of implications from a formal context. There exist mainly two algorithms to achieve this [4, 8] and both of them compute not only the implications from the stem base, but also concept intents. This is problematic as a context may have exponentially many concept intents. Recent theoretical results suggest that existing approaches at computing the stem base may not lead to algorithms with better worst-case complexity [3, 1].

Proper premises provide another approach for obtaining an implicational base of a formal context. Because this set of implications does not have minimal cardinality, proper premises have been outside the focus of the FCA community for many years. However, there are substantial arguments to reconsider using them. Existing methods for computing proper premises avoid computing concept intents. Thus, in contexts with many concept intents they may have a clear advantage in runtime over the stem base algorithms.

We want to be able to make a prognosis whether we can expect stem base algorithms or proper premises algorithms to perform better on context of a given size. Therefore, it is interesting to know what number in intents and what number of proper premises one can expect in a formal context. We make no further assumptions about this formal context, instead we assume that it is chosen uniformly at random among all formal contexts of a given size.

Knowing the expected behaviour is also useful when conducting experiments such as in [2]. They allow us to compare the theoretically expected behaviour to experimental results.

2 Preliminaries

We provide a short summary of the most common definitions in formal concept analysis. A *formal context* is a triple $\mathbb{K} = (G, M, I)$ where G is a set of objects, M a set of attributes, and $I \subseteq G \times M$ is a relation that expresses whether an object $g \in G$ has an attribute $m \in M$. If $A \subseteq G$ is a set of objects then A' denotes the set of all attributes that are shared among all objects in A, i.e. $A' = \{\, m \in M \mid \forall g \in G \colon gIm \,\}$.

Likewise, for some set $B \subseteq M$ we define $B' = \{\, g \in G \mid \forall m \in B\colon gIm \,\}$. Pairs of the form (A, B) where $A' = B$ and $B' = A$ are called formal concepts. Formal concepts of the form $(\{\, m \,\}', \{\, m \,\}'')$ for some attribute $m \in M$ are called *attribute concept* and are denoted by μm. We define the partial order $\leq$ on the set of all formal concepts of a context to be the subset order on the first component. The first component of a formal concept is called the *concept extent* while the second component is called the *concept intent*.

Formal Concept Analysis provides methods to mine implicational knowledge from formal contexts. An *implication* is a pair (B_1, B_2) where $B_1, B_2 \subseteq M$, usually denoted by $B_1 \to B_2$. We say that *the implication* $B_1 \to B_2$ *holds* in a context $\mathbb{K}$ if $B_1' \subseteq B_2'$. An implication $B_1 \to B_2$ *follows from a set of implications* $\mathcal{L}$ if for every context $\mathbb{K}$ in which all implications from $\mathcal{L}$ hold $B_1 \to B_2$ also holds. We say that $\mathcal{L}$ is *sound* for $\mathbb{K}$ if all implications from $\mathcal{L}$ hold in $\mathbb{K}$, and we say that $\mathcal{L}$ is *complete* for $\mathbb{K}$ if all implications that hold in $\mathbb{K}$ follow from $\mathcal{L}$. There exists a sound and complete set of implications for each context which has minimal cardinality [6]. This is called the stem base. The exact definition of the stem base is outside the scope of this work.

A sound and complete set of implications can also be obtained using *proper premises*. For a given set of attributes $B \subseteq M$ we define $B^\bullet$ to be the set of those attributes in $M \setminus B$ that follow from B but not from a strict subset of B, i.e.

$$B^\bullet = B'' \setminus \left(B \cup \bigcup_{S \subsetneq B} S'' \right).$$

B is called a *proper premise* if $B^\bullet$ is not empty. It is called a *proper premise for* $m \in M$ if $m \in B^\bullet$. It can be shown that $\mathcal{L} = \{\, B \to B^\bullet \mid B \text{ proper premise} \,\}$ is sound and complete [5].

We write $g \not\asymp m$ if g' is maximal with respect to the subset order among all object intents which do not contain m.

3 Expected Number of Concept Intents

We provide formulae for the statistical expectation of the number of intents and proper premises in a formal context that is chosen uniformly at

random among all $n \times m$-contexts for fixed natural numbers n and m.[1]

First we consider a fixed set $Q \subseteq M$ with $|Q| = q \leq m$ and a fixed set $R \subseteq G$ with $|R| = r \leq n$. We compute the number of $n \times m$-contexts that satisfy $Q' = R$ and $R' = Q$. Table **??** illustrates this computation. From $R \subseteq Q'$ we obtain that the relation $I \cap (Q \times R)$ must be the full relation, and therefore there is only one choice for $I \cap (Q \times R)$. $Q' \subseteq R$ requires that no object g in $G \setminus R$ is in relation gIq with all $q \in Q$. Within the quadrant $Q \times (G \setminus R)$ we thus obtain $2^q - 1$ choices per object, i.e. in total $(2^q - 1)^{n-r}$ choices for $I \cap (Q \times (G \setminus R))$. Analogously, $R' = Q$ yields $(2^r - 1)^{m-q}$ choices for $I \cap ((M \setminus Q) \times R)$. No restrictions apply to $I \cap ((M \setminus Q) \times (G \setminus R))$, yielding $2^{(m-q)(n-r)}$ possibilities. Hence the total number of contexts with $Q' = R$ and $R' = Q$ is

$$2^{(m-q)(n-r)}(2^q - 1)^{n-r}(2^r - 1)^{m-q}.$$

Summing over all possible choices for R gives us the number N_Q of all contexts that have Q as their intent:

$$N_Q = \sum_{r=0}^{n} \binom{n}{r} 2^{(m-q)(n-r)}(2^q - 1)^{n-r}(2^r - 1)^{m-q}.$$

It must hold that $\sum_{\mathbb{K} \text{ context}} N_{\mathbb{K}} = \sum_{Q \subseteq M} N_Q$ where $N_{\mathbb{K}}$ is the number of intents of $\mathbb{K}$. The expected number of intents is then obtained as

$$\begin{aligned}
\mathbb{E}_{\text{intent}} &= 2^{-nm} \sum_{\mathbb{K} \text{ context}} N_{\mathbb{K}} \\
&= 2^{-nm} \sum_{Q \subseteq M} N_Q \\
&= \sum_{q=0}^{m} \binom{m}{q} \sum_{r=0}^{n} \binom{n}{r} 2^{-rq}(1 - 2^{-r})^{m-q}(1 - 2^{-q})^{n-r}.
\end{aligned}$$

4 Expected Number of Hypergraph-Transversals

Before we can compute the expected number of proper premises we need to consider hypergraph transversals, as they are closely related to proper

[1] We ignore renaming of attributes and objects.

Table 1: Computing the Expected Number of Concept Intents

	... Q ...	...
$\vdots$ R $\vdots$	1	$(2^r - 1)^{m-q}$
$\vdots$	$(2^q - 1)^{n-r}$	$2^{(m-q)(n-r)}$

premises. This relationship will be examined more closely in Section **??**. Let V be a finite set of vertices. A *hypergraph* H is simply a subset of the power set of V. Intuitively, each set $E \in H$ represents an edge of the hypergraph, which, in contrast to classical graph theory, may be incident to more or less than two vertices. A set $S \subseteq V$ is called a *hypergraph transversal* of H if it intersects every edge $E \in H$, i.e.

$$\forall E \in H \colon S \cap E \neq \emptyset.$$

S is called a *minimal hypergraph transversal* The *transversal hypergraph* of H is the set of all minimal hypergraph transversals of H. It is denoted by $Tr(H)$.

Notice that there is a correspondence between hypergraphs and formal contexts, where the attributes of the formal context correspond to the vertices and the object intents correspond to the edges. Table **??** is to be understood in this sense.

We present a formula for the expected number of hypergraph transversals for a hypergraph that is chosen uniformly at random among all hypergraphs with m vertices and n edges.

We start by computing the number of hypergraphs that have a given set $Q \subseteq V$ as a minimal hypergraph transversal. We make use of the following proposition.

Proposition 1. *In a hypergraph H a set of vertices $Q \subseteq V$ is a minimal hypergraph transversal if and only if*

1. $Q \cap E \neq \emptyset$ *for all* $E \in H$, *and*

2. *for all* $v \in Q$ *there exists some* $E \in H$ *such that* $Q \cap E = \{v\}$.

Proof. The first property is simply the transversal property. The second property is equivalent to minimality, since for all $v \in Q$ the set $Q \setminus \{v\}$ is a hypergraph transversal iff there is no E such that $Q \cap E = \{v\}$. □

Let the edges of H be numbered $E_1, \ldots, E_n$. By Propostion **??** if Q is a minimal hypergraph transversal then we can find values $p_1 < \cdots < p_m$ such that $Q \cap E_{p_i} = \{v_i\}$ for some $v_i \in V$ and $Q \cap E_{p_i} \neq Q \cap E_j$ for all $i \in \{1, \ldots, m\}$ and all $j < p_i$. By choosing the p_i in this way we can avoid counting the same context multiple times.

Table **??** illustrates the number of choices for the n edges of H. For each edge E the entries in $E \setminus Q$ are irrelevant and therefore allow $2^{(m-q)n}$ choices. For $E_{p_i} \cap Q$ there is only one choice, namely $E_{p_i} \cap Q = \{v_i\}$. For each edge E_j, $j < p_1$ it must hold that $E_j \cap Q \neq \emptyset$ since Q is a hypergraph transversal and $E_j \cap Q \neq \{v\}$ for all $v \in Q$ because of our choice of the p_i. Hence there are $2^q - 1 - q$ choices for $Q \cap g'_j$. Similarly, we obtain $2^q - 1 - (q-1)$ choices for $E_j \cap Q$ where $p_1 < j < p_2$ since now $Q \cap E_j$ can also take the value $Q \cap E_j = \{v_1\}$. The total number of contexts $\mathbb{K}$ that have Q as a minimal hypergraph transversal is then obtained by summing over all possible values for the p_i and multiplying with $k!$ to account for permutations of the v_i.

$$N_Q(n, m, q) = q! \cdot 2^{(m-q)n} \sum_{\substack{(p_1,\ldots,p_q) \in \mathbb{N}^q \\ 1 \leq p_1 < \cdots < p_q \leq n}} \prod_{i=0}^{q} (2^q - 1 - i)^{p_{i+1} - p_i - 1}$$

where we define $p_0 = 0$ and $p_{q+1} = n + 1$. Using similar arguments as for

Table 2: Computing the Expected Number of Hypergraph Transversals

	... Q ...	...
$\vdots$	$(2^q - 1 - q)^{p_1 - 1}$	
E_{p_1}	1	
$\vdots$	$(2^q - 1 - (q-1))^{p_2 - p_1 - 1}$	
E_{p_2}	1	$2^{(m-q)n}$
$\vdots$	$\vdots$	
E_{p_q}	1	
$\vdots$	$(2^q - 1)^{n - p_q}$	

$\mathbb{E}_{\text{intent}}$ we obtain

$$\begin{aligned}
\mathbb{E}_{\text{HG}-\text{trans}} &= 2^{-mn} \sum_{\mathbb{K} \text{ context}} N_{\mathbb{K}} \\
&= 2^{-mn} \sum_{Q \subseteq M} N_Q(n, m, q) \\
&= 2^{-mn} \sum_{q=0}^{m} \binom{m}{q} N_Q(n, m, q) \\
&= \sum_{q=0}^{m} \binom{m}{q} q!\, 2^{-q^2} \sum_{\substack{(p_1, \ldots, p_q) \in \mathbb{N}^q \\ 1 \le p_1 < \cdots < p_q \le n}} \prod_{i=0}^{q} \left(1 - 2^{-q}(1+i)\right)^{p_{i+1} - p_i - 1}.
\end{aligned}$$

5 Expected Number of Proper Premises

We present a connection between proper premises and minimal hypergraph transversals. This connection can help us to obtain a formula for the expected value of proper premises from the formula for $\mathbb{E}_{\mathrm{HG-trans}}$. This connection has been exploited in database theory to the purpose of mining functional dependencies from a database relation [7]. Implicitly, it has also been known for a long time within the FCA community. However, the term *hypergraph* has not been used in this context (cf. Proposition 23 from [5]). The following proposition can be found in [5] among others.

Proposition 2. *$P \subseteq M$ is a premise of m iff*

$$(M \setminus g') \cap P \neq \emptyset$$

holds for all $g \in G$ with $g \not\downarrow m$. P is a proper premise for m iff P is minimal (with respect to $\subseteq$) with this property.

We immediately obtain the following corollary.

Corollary 1. *P is a premise of m iff P is a hypergraph transversal of*

$$\{M \setminus g' \mid g \in G, g \not\downarrow m\}.$$

The set of all proper premises of m is exactly the transversal hypergraph

$$Tr(\{M \setminus g' \mid g \in G, g \not\downarrow m\}).$$

Notice that the sets g' with $g \not\downarrow m$ are by definition exactly the maximal elements of $\{g' \mid g \in G, m \notin g'\}$. Hence, $\{M \setminus g' \mid g \in G, g \not\downarrow m\}$ contains exactly the minimal elements of $\{M \setminus g' \mid g \in G, m \notin g'\}$. When searching for hypergraph transversals it suffices to look at the minimal edges in a hypergraph and therefore a set S is a hypergraph transversal of $\{M \setminus g' \mid g \in G, g \not\downarrow m\}$ if and only if S is a hypergraph transversal of $\{M \setminus g' \mid g \in G, m \notin g'\}$. This yields the following corollary.

Corollary 2. *The set of all proper premises of m is exactly the transversal hypergraph*

$$Tr(\{M \setminus g' \mid g \in G, m \notin g'\}).$$

From Corollary **??** we obtain for the expected value of proper premises of an attribute v_0

$$\begin{aligned}\mathbb{E}_{\mathrm{pp}}(n,m) &= \sum_{R\subseteq G} 2^{-n}\mathbb{E}_{\mathrm{HG-trans}}(|R|,m)\\ &= \sum_{r=0}^{n}\binom{n}{r}2^{-n}\mathbb{E}_{\mathrm{HG-trans}}(r,m)\\ &= 2^{-n}\sum_{r=0}^{n}\binom{n}{r}\sum_{q=0}^{m-1}\binom{m}{q}q!\,2^{-q^2}\sum_{\substack{(p_1,\ldots,p_q)\in\mathbb{N}^q\\ 1\le p_1<\cdots<p_q\le r}}\prod_{i=0}^{q}\left(1-2^{-q}(1+i)\right)^{p_{i+1}-p_i-1}.\end{aligned}$$

where the sum is over all possible ways to choose the attribute intent v_0', where each choice has probability 2^{-n}.

6 Conclusion

In this work formulae have been obtained for the expected values of intents and proper premises in a context that has been chosen uniformly at random. A third formula gives the expected number of hypergraph transversals in a hypergraph that has been chosen uniformly at random. Our hope is that these formulae can help to examine and compare the behaviour of existing algorithms for finding an implicational base for a formal context. The plots in Figure 1 show that when the number of objects is large compared to the number of attributes proper premise algorithms appear to have an advantage over algorithms that compute both pseudo-intents and intents.

References

[1] Mikhail Babin and Sergei Kuznetsov. Recognizing pseudo-intents is coNP-complete. In Marzena Kryszkiewicz and Sergei Obiedkov, editors, *Proc. of the 7th Int. Conf. on Concept Lattices and Their Applications (CLA 2010)*, volume 672. CEUR Workshop Proceedings, 2010.

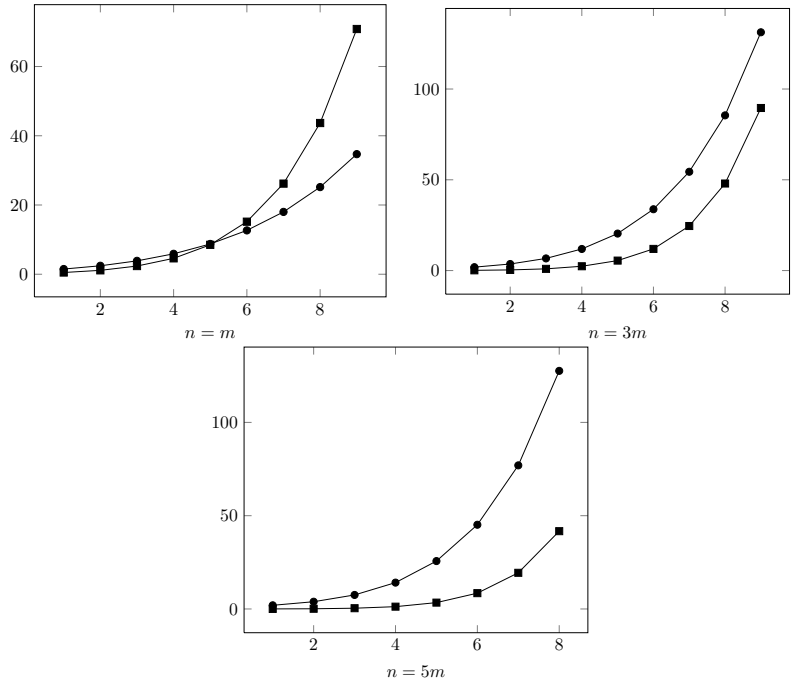

Figure 1: Expected Number of Intents and Proper Premises for Certain Families of Formal Contexts

[2] Daniel Borchmann. Decomposing finite closure operators by attribute exploration. In *Supplementary Proc. of ICFCA '11*, 2011.

[3] Felix Distel. Hardness of enumerating pseudo-intents in the lectic order. In Barış Sertkaya and Léonard Kwuida, editors, *Proc. of the 8th Int. Conf. on Formal Concept Analysis (ICFCA 2010)*, volume 5986 of *Lecture Notes in Artificial Intelligence*, pages 124–137. Springer, 2010.

[4] Bernhard Ganter. Two basic algorithms in concept analysis. Preprint 831, Fachbereich Mathematik, TU Darmstadt, Darmstadt, Germany, 1984.

[5] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, New York, 1997.

[6] J.-L. Guigues and V. Duquenne. Familles minimales d'implications informatives résultant d'un tableau de données binaires. *Math. Sci. Humaines*, 95:5–18, 1986.

[7] Heikki Mannila and Kari-Jouko Räihä. Algorithms for inferring functional dependencies from relations. *Data & Knowledge Engineering*, 12(1):83 – 99, 1994.

[8] S. Obiedkov and V. Duquenne. Attribute-incremental construction of the canonical implication basis. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):77–99, 2007.