Boolean Formulas and Frequent Sets

Jouni K. Seppänen and Heikki Mannila

HIIT Basic Research Unit, Lab. Computer and Information Science,
FI-02015 Helsinki University of Technology, Finland
{Jouni.Seppanen, Heikki.Mannila}@hut.fi

Abstract. We consider the problem of how one can estimate the support of Boolean queries given a collection of frequent itemsets. We describe an algorithm that truncates the inclusion-exclusion sum to include only the frequencies of known itemsets, give a bound for its performance on disjunctions of attributes that is smaller than the previously known bound, and show that this bound is in fact achievable. We also show how to generalize the algorithm to approximate arbitrary Boolean queries.

1 Introduction

Algorithms for mining frequent itemsets continue to be a subject of recent data mining research [GZ03] long after the original publications [AIS93, AMS+96]. Less attention has been received by the question of how one can utilize the frequent itemsets one has mined. The original motivation was provided by association rules, but we claim that the collection of frequent itemsets is good for much more than rule mining: they give us a picture of the joint distribution of the data, and can therefore be used to approximately evaluate Boolean queries over the original data.

The simple idea of approximating exponentially long inclusion-exclusion sums using a small collection of frequent itemsets was considered in [MT96]. Thus the frequent sets can be seen as a condensed representation of the data. In this paper we give a more thorough presentation of the issues involved. We sharpen Theorem 5 in [MT96], which shows that the approximation error for a disjunctive query is bounded by $2^{b-2}/b$, where b is the size of the negative border. Here we prove a bound of $\binom{b}{\lceil b/2 \rceil}/b$, and give a family of examples for which this bound is reached. Our main contribution is the generalization of the discussion to arbitrary Boolean formulas.

Related work includes using maximum entropy to approximate the joint distribution [PMS00, PS01] and linear programming to find upper and lower bounds for queries [BSH04]. These approaches share the problem that they require exponential space in the number of attributes involved. There has also been much work on reducing the size of the itemset collection, such as free-sets [BBR00] and non-derivable sets [CG02]. However, most such work concentrates on algorithms for discovering itemsets, not on using the itemsets obtained to evaluate queries.

The rest of this paper is structured as follows. We start in Section 2 from the almost trivial case of estimating conjunctive queries, introducing notation and

J.-F. Boulicaut et al. (Eds.): Constraint-Based Mining, LNAI 3848, pp. 348–361, 2005.

showing the general idea of the results that follow. In Section 3 we discuss the much more interesting case of disjunctive queries, and in Section 4 we generalize our results to arbitrary Boolean queries.

2 Conjunction of Attributes

The goal of this section is to introduce our notation and the basic task in a simple setting. Throughout the paper, we will denote by U a set of *attributes* or *items*, and by r a *binary relation* over U: that is, r is a multiset of *tuples* $T \subset U$. *Itemsets* are arbitrary subsets of U, denoted by X or Y. We will denote by $g(X)$ the fraction of tuples in r that are equal to X, and call this quantity the *exact frequency* of X. By contrast, the *frequency* $f(X)$ is the fraction of tuples that contain all attributes in X. Thus

$$f(X) = \sum_{Y \supset X} g(X).$$

An itemset is called *frequent* if its frequency is at least a predefined threshold σ. The collection of σ-frequent itemsets is denoted $\mathcal{F}_\sigma$. As is well known, this collection is *downward closed*: given itemsets $X \subset Y \in \mathcal{F}_\sigma$, we have $X \in \mathcal{F}_\sigma$. We will in general denote by $\mathcal{F}$ an arbitrary downward closed collection of itemsets.

The task that we are interested in is estimating the result of a *Boolean query* φ. For now, we let φ be simply a conjunction of attributes: let $C \subset U$ be any itemset, and let

$$\varphi = \bigwedge C = \bigwedge_{A \in C} A.$$

We say that a tuple $T \in r$ *supports* φ, denoted $T \models \varphi$, if T includes every attribute in the query. The *frequency* of the query $f(\varphi)$ is the fraction of tuples in r that support φ. In the case of conjunctions, one sees immediately that a tuple T supports φ if and only if $C \subset T$, and thus the frequency of the query φ is equivalent to the frequency of the itemset C.

We can now present the task Approximate Query$(\mathcal{F}_\sigma, \varphi)$: given the collection $\mathcal{F}_\sigma$ of frequent itemsets, and a Boolean query φ, find an estimate $\hat{f}(\varphi)$ that should be close to $f(\varphi)$. Any solution to this task will be evaluated on its worst-case accuracy, i.e., how far from the actual frequency it can be. The measure of accuracy that we will use is the maximum absolute value of the error $e(\varphi) = f(\varphi) - \hat{f}(\varphi)$.

It is of course easy to come up with a simple and fairly accurate solution, which we name Truncate Sum (for reasons that will become clear later): $\hat{f}(\varphi) = f(C)$ if $C \in \mathcal{F}_\sigma$, otherwise $\hat{f}(\varphi) = 0$. Again, C is the set of attributes in the conjunction φ. If $C \in \mathcal{F}_\sigma$, we have $\hat{f}(\varphi) = f(\varphi)$ and therefore $e(\varphi) = 0$. Otherwise, we know that $0 \leq f(C) < \sigma$, which implies that $|e(\varphi)| \leq \sigma$. Thus our bound for the maximum absolute error is σ. We have shown the following result:

Proposition 1. *For a conjunction of attributes* φ, Truncate Sum *yields results to* Approximate Query *that have maximal absolute error* σ.

3 Disjunction of Attributes

In this section we consider the case where φ is a disjunction of attributes. The main results are that the worst-case error of Truncate Sum can be bounded by an expression that depends exponentially on the size of the negative border (which will be defined later), and that this worst-case bound cannot be decreased at all: that is, there are collections of frequent sets where the bound holds with equality.

We now investigate queries of the form

$$\varphi = \bigvee D = \bigvee_{A \in D} A.$$

A tuple $T \in r$ supports the query if at least one of the attributes in D appears in T: in logical notation, $T \models \varphi$ if $D \cap T \neq \emptyset$. The frequency of φ is, again, the fraction of tuples that support φ. A basic result in combinatorics is that this frequency can be obtained by the inclusion-exclusion principle:

$$f(\varphi) = \sum_{X \subseteq D} [X \neq \emptyset] \, (-1)^{|X|+1} f(X). \tag{1}$$

Here, and in the sequel, we avoid long sum conditions in subscripts by using the "Iverson notation"

$$[P] = \begin{cases} 1, & P \text{ is true,} \\ 0, & P \text{ is false,} \end{cases}$$

popularized by Knuth [Knu92].

The inclusion-exclusion principle is fine if we know the frequency of every itemset X that is a subset of D. If we do not, our approach is to compute the sum over all itemsets whose frequencies we do know:

$$\hat{f}(\varphi) = \sum_{X \subseteq D} [\emptyset \neq X \in \mathcal{F}_\sigma] \, (-1)^{|X|+1} f(X). \tag{2}$$

The exponentially long sum (1) is truncated to the terms (2) that we know; thus the name Truncate Sum. The error made in the approximation is

$$e(\varphi) = \sum_{X \subseteq D} [X \in \mathcal{G}] \, (-1)^{|X|+1} f(X),$$

where by $\mathcal{G}$ we denote the family of non-frequent sets, i.e., the complement of $\mathcal{F}_\sigma$.

An intuition for this estimate is provided by the well-known Bonferroni inequalities, which state that if our collection $\mathcal{F}_\sigma$ happens to contain exactly the itemsets of size at most k, then the error is bounded by the sum of frequencies of itemsets of size $k+1$. [GS96] The proof is simple, although not entirely trivial: the error consists of exponentially many terms, which happen to mostly cancel out. We would like to prove analogues of the Bonferroni inequalities for our more general case.

We start from a simple upper bound that is not very interesting in itself but will be used in the proof of Theorem 1. Recall that $\mathcal{G}$ is the complement of $\mathcal{F}_\sigma$.

Lemma 1. *For a disjunction of attributes* $\varphi = \bigvee D$,

$$|e(\varphi)| \leq \sum_{X \in \mathcal{G}} \binom{|X|}{\lceil |X|/2 \rceil} g(X).$$

Proof. We write the frequency as a sum over tuples: $f(X) = |r|^{-1} \sum_{T \in r} [T \supset X]$, and therefore

$$e(\varphi) = |r|^{-1} \sum_{T \in r} \sum_{X \subset T} [X \in \mathcal{G}] \, (-1)^{|X|+1}.$$

One way of proving the Bonferroni inequalities is based on pairing up most of the tuples in $\mathcal{G}$; we proceed similarly.

We first introduce some notation: $t = |T|$, and $t' = \lceil t/2 \rceil$. It is well known that the power set $\mathcal{P}(T)$ can be written as a union of $\binom{t}{t'}$ disjoint *chains*, where a chain means a collection C of sets where given any two sets $X, Y \in C$, either $X \subset Y$ or $Y \subset X$. [Bol88, Theorem 1 of Section 4] The construction of Bollobás yields chains that are symmetric and consist of consecutive sets: if we write $C = \{ X_1, X_2, \ldots, X_k \}$ with $X_1 \subset X_2 \subset \ldots X_k$, then $|X_1| + |X_k| = d$ and $|X_{j+1}| = |X_j| + 1$ for all $1 \leq j < k$. Thus, if d is *odd*, each chain C is of even length, and the alternating sum $\sum_{X \in C} (-1)^{|X|+1}$ is zero. If d is *even*, the chains from the construction are of odd length. However, we can remove one attribute A from T, perform the construction on $T \setminus \{ A \}$ to obtain a collection of $\binom{d-1}{d'-1}$ chains, and then add to the collection a duplicate of each chain with A added to every set: the result is a partition of $\mathcal{P}(T)$ into $2\binom{t-1}{t'-1} = \binom{t}{t'}$ chains, each of which consists of an even number of consecutive sets.

We can thus assume that there is a partition $T = C_1 \cup C_2 \cup \cdots \cup C_m$ of T into $m = \binom{t}{t'}$ disjoint chains, such that $\sum_{X \in C_j} (-1)^{|X|+1} = 0$ for each chain C_j. Now

$$\sum_{X \subset T} [X \in \mathcal{G}] \, (-1)^{|X|+1} = \sum_{j=1}^{m} \sum_{X \in C_j} [X \in \mathcal{G}] \, (-1)^{|X|+1}.$$

Every chain C_j that is wholly contained in either $\mathcal{F}_\sigma$ or $\mathcal{G}$ contributes 0 to this sum. Every other chain contributes either 0 or ± 1. Therefore,

$$\left| \sum_{X \subset T} [X \in \mathcal{G}] \, (-1)^{|X|+1} \right| \leq m = \binom{t}{t'}.$$

The claim follows by observing that $g(X) = |r|^{-1} \sum_{T \in r} [T = X]$. □

Recall that the Bonferroni inequalities, which apply to the case where $\mathcal{F}_\sigma$ consists of all itemsets of size at most k, give an error bound related to the itemsets of size $k+1$. An analogue of the size $k+1$ itemsets that is both intuitively appealing and practically useful is the *negative border* Bd^- [MT96], defined as the family of minimal non-frequent sets:

$$Bd^- = \{ X \in \mathcal{G} \mid Y \in \mathcal{F}_\sigma \, \forall Y \subsetneq X \}.$$

Note that if $\mathcal{F}_\sigma$ consists of sets of size at most k, then Bd^- is exactly the family of sets of size $k+1$. The practical usefulness stems from the famous Apriori algorithm, which computes the family $\mathcal{F}_\sigma$ and finds Bd^- as a byproduct of its stopping condition [AMS+96].

We will prove a lemma connecting the negative border to the error $e(\varphi)$. First, we need to define some more notation: $\mathcal{G}_D$ is the set of non-frequent subsets of D,

$$\mathcal{G}_D = \{\, X \mid X \in \mathcal{G}, X \subset D \,\},$$

and the *negative border relative to* D, denoted Bd^-_D, consists of the minimal sets in $\mathcal{G}_D$. Note that $Bd^-_D \subset Bd^-$, since if X is minimal in $\mathcal{G}_D$, all subsets of X are in $\mathcal{F}_\sigma$, and therefore X is minimal also in $\mathcal{G}$. We will also need the concept of *exact frequency of* X *relative to* D, defined for $X \subset D$ as the fraction of tuples whose intersection with D is X, which we can write as

$$g_D(X) = f\Big(\bigwedge_{A \in X} A \;\wedge\; \bigwedge_{A \in D \setminus X} \neg A \Big).$$

Lemma 2. *Consider the query* $\varphi = \bigvee D$. *If* $Bd^-_D \neq \{\, \emptyset \,\}$, *the algorithm* Truncate Sum *has an error of*

$$e(\varphi) = \sum_{\emptyset \neq \mathcal{E} \subset Bd^-_D} (-1)^{|\mathcal{E}|+|\bigcup \mathcal{E}|} g_D\big(\bigcup \mathcal{E}\big). \tag{3}$$

To illustrate the lemma, consider some simple examples. If Bd^-_D consists of a single set $B \neq \emptyset$, the error is an inclusion-exclusion sum

$$e(\varphi) = \sum_X [B \subset X \subset D]\, (-1)^{|X|+1} f(X),$$

which is of course exactly the expression for $(-1)^{|B|+1} g_D(B)$.

Likewise, if Bd^- is the two-set family $\{\, B_1, B_2 \,\}$ with $B_j \cap D \neq \emptyset$ for $j = 1, 2$, we obtain

$$e(\varphi) = \sum_X [B_1 \subset X \subset D \text{ or } B_2 \subset X \subset D]\, (-1)^{|X|+1} f(X).$$

We can use inclusion-exclusion to decompose the condition on X:

$$\begin{aligned}[B_1 \subset X \subset D \text{ or } B_2 \subset X \subset D] \\ \qquad = [B_1 \subset X \subset D] + [B_2 \subset X \subset D] - [B_1 \cup B_2 \subset X \subset D]\end{aligned}$$

Thus we can break the formula for $e(\varphi)$ into three components, which sum up to $g_D(B_1)$, $g_D(B_2)$ and $-g_D(B_1 \cup B_2)$. The proof of the lemma is a straightforward extension of this idea.

Proof of Lemma 2. The error is given by

$$e(\varphi) = \sum_X [X \in \mathcal{G}_D]\,(-1)^{|X|+1} f(X). \tag{4}$$

We can rewrite the condition $X \in \mathcal{G}_D$ in terms of the minimal sets in $\mathcal{G}_D$ as follows: We have $X \in \mathcal{G}_D$ if and only if $X \supset B$ for some $B \in Bd_D^-$. We apply inclusion-exclusion on the Iverson function:

$$\begin{aligned}[X \in \mathcal{G}_D] &= \sum_{B \in Bd_D^-} [B \subset X \subset D] - \sum_{B_1, B_2 \in Bd_D^-} [B_1 \cup B_2 \subset X \subset D] + \cdots \\ &= \sum_{\emptyset \neq \mathcal{E} \subset Bd_D^-} (-1)^{|\mathcal{E}|+1} \left[\bigcup \mathcal{E} \subset X \subset D\right].\end{aligned}$$

Plugging this in the error sum (4) and changing the order of summation, we obtain

$$e(\varphi) = \sum_{\emptyset \neq \mathcal{E} \subset Bd_D^-} (-1)^{|\mathcal{E}|+1} \sum_X \left[\bigcup \mathcal{E} \subset X \subset D\right] (-1)^{|X|+1} f(X).$$

It now suffices to show that for $Y \subset D$,

$$(-1)^{|Y|} g_D(Y) = \sum_X [Y \subset X \subset D]\,(-1)^{|X|} f(X),$$

for then letting $Y = \bigcup \mathcal{E}$ yields (3). This is an easy exercise in inclusion-exclusion: given a tuple $T \in r$, write $T = R \cup S$ with $R \subset D$, $S \subset U \setminus D$. The tuple will contribute $(-1)^{|R|}$ to all terms corresponding to $X \subset R \cup S$. In the case $R = Y$, the contribution is exactly $(-1)^{|Y|}$; otherwise, the contributions cancel out. □

Based on the lemma, we can prove an analogue to the Bonferroni inequalities that gives, however, rather larger bounds than the Bonferroni case.

Theorem 1. *For a disjunction of attributes $\varphi = \bigvee D$, the absolute error $|e(\varphi)|$ of* Truncate Sum *is bounded by*

$$\binom{|Bd_D^-|}{\lceil |Bd_D^-|/2 \rceil} |Bd_D^-|^{-1} \sum_{X \in Bd_D^-} f(X).$$

Proof. Arrange the sum (3) in the form

$$e(\varphi) = \sum_{X \in \mathcal{G}_D} \nu(X) g_D(X).$$

For the coefficients $\nu(X)$ we have

$$\nu(X) = (-1)^{|X|} \sum_{\mathcal{E} \subset Bd_X^-} \left[\bigcup \mathcal{E} = X\right] (-1)^{|\mathcal{E}|}.$$

In this sum, the condition $\left[\bigcup \mathcal{E} = X\right]$ defines an upwards-closed subfamily of the powerset of Bd_X^-. We know from Lemma 1 that the absolute value of this alternating sum is bounded by $\binom{m}{m'}$ with $m = |Bd_X^-|$ and $m' = \lceil m/2 \rceil$.

Arrange also the sum $\sum_{X \in Bd_D^-} f(X)$ in the form

$$\sum_{X \in \mathcal{G}_D} \mu(X) g_D(X).$$

We have for the coefficients $\mu(X)$

$$\mu(X) = \sum_{Y \in Bd_D^-} [Y \subset X] = |Bd_X^-| > 0$$

for all $X \in \mathcal{G}_D$. The ratio $|\nu(X)|/\mu(X)$ is bounded by $\binom{m}{m'}/m$, and this bound is largest for $X = D$. Thus

$$\begin{aligned} |e(\varphi)| &\leq \sum_{X \in \mathcal{G}_D} |\nu(X)| g_D(X) \leq \sum_{X \in \mathcal{G}_D} \left(\max \frac{|\nu(X)|}{\mu(X)}\right) \mu(X) g_D(X) \\ &\leq \binom{|Bd_D^-|}{\lceil |Bd_D^-|/2 \rceil} |Bd_D^-|^{-1} \sum_{X \in Bd_D^-} f(X). \end{aligned} \tag{5}$$

□

Using the inequality $f(X) < \sigma$ for $X \in Bd^-$, we can obtain a form of the bound that is independent of the actual frequencies of sets in the border.

Corollary 1. *For a disjunction of attributes $\varphi = \bigvee D$,*

$$e(\varphi) \leq \binom{|Bd^-|}{\lceil |Bd^-|/2 \rceil} \sigma.$$

Thus, the bound depends superpolynomially on the size of the border. A natural question is whether the bound can be decreased. We will next show that the answer is negative: the bound is in the worst case tight. The example will have a small negative border. The key part in the proof is constructing the family $\mathcal{F}_\sigma$ so that when Lemma 1 is used in the proof of Theorem 1, equality holds. This is the case when the minimal families $\mathcal{E} \subset Bd_X^-$ that satisfy the condition $\bigcup \mathcal{E} = X$ are exactly of size $\lceil |Bd_X^-|/2 \rceil$.

Theorem 2. *There exists a set U, a relation r over U, and a downward-closed collection of itemsets $\mathcal{F}$ such that for the disjunctive query $\varphi = \bigvee U$ the absolute error of* Truncate Sum *is*

$$|e(\varphi)| = \binom{|Bd^-|}{\lceil |Bd^-|/2 \rceil} |Bd^-|^{-1} \sum_{X \in Bd^-} f(X).$$

Proof. Choose integer parameters $p > k > 1$; p will be the number of sets in the negative border, and we will see later that choosing $p = 2k + 1$ suits our purposes well. We will need $n = \binom{p}{k}$ attributes: let $U = [n] = \{1, \ldots, n\}$. We will set up Bd^- so that for all families $\mathcal{E} \subset Bd^-$, $|\mathcal{E}| \leq k$ implies $\bigcup \mathcal{E} \neq U$, and $|\mathcal{E}| > k$ implies $\bigcup \mathcal{E} = U$. To achieve this, we first enumerate all the k-element subsets of $[p]$; there are n of them, and we will name them $K_1, K_2, \ldots, K_n$ in any arbitrary order. Then for all $q \in [p]$, we define W_q as the set of those i such that $q \notin K_i$. Let $Bd^- = \{W_q \mid q \in [p]\}$. Note that Bd^- is an antichain, since all sets W_q have the same number of elements; thus we can define $\mathcal{F}$ as the downward-closed collection of sets that are not supersets of any sets in Bd^-, and Bd^- will automatically be the negative border corresponding to $\mathcal{F}$.

We must now prove the assertion that for $\mathcal{E} \subset Bd^-$, $\bigcup \mathcal{E} = U$ if and only if $|\mathcal{E}| > k$. Given any collection $\mathcal{E}$ of border sets, we can write $\mathcal{E} = \{W_q \mid q \in Q\}$ for some index set $Q \subset [p]$. If $|\mathcal{E}| = |Q| \leq k$, some set K_i must be a superset of the index set Q, since we have enumerated all k-element subsets of $[p]$. But then we have that $i \notin \bigcup \mathcal{E}$, and thus $\bigcup \mathcal{E} \neq U$. Conversely, if $\bigcup \mathcal{E} \neq U$, there must be some $i \notin \bigcup \mathcal{E}$, and therefore for all $q \in Q$ we must have $q \in K_i$, because $i \notin W_q$. But this means that $Q \subset K_i$, and therefore $|\mathcal{E}| = |Q| \leq |K_i| = k$. We have thus shown that $\bigcup \mathcal{E} = U$ if and only if $|\mathcal{E}| > k$.

We will let $\varphi = \bigvee U$ over all the attributes. Thus, the terms $g_D(X)$ will be the usual exact frequencies $g(X)$ and the family Bd_D^- will be the usual negative border Bd^-. We will also let $g(U) = 1$ and $g(X) = 0$ for all $X \notin \mathcal{F}$, $X \neq U$. We can let $g(X)$ be some sufficiently high number for all $X \in \mathcal{F}$ so that $\mathcal{F} = \mathcal{F}_\sigma$ for some σ.

Now we are in a position to apply Lemma 2. The sum over $\mathcal{E} \subset Bd^-$ becomes a sum over those $\mathcal{E}$ for which $\bigcup \mathcal{E} = U$, since $g(\bigcup \mathcal{E}) = 0$ otherwise. By the construction, these are exactly those $\mathcal{E}$ such that $|\mathcal{E}| > k$. Thus

$$e(\varphi) = \sum_{\mathcal{E} \subset Bd^-} (-1)^{|\mathcal{E}|+|r|} [|\mathcal{E}| > k] = (-1)^{|r|} \sum_{j=k+1}^{n} (-1)^j \binom{p}{j}.$$

It is an easy proof by induction that

$$\sum_{j=k+1}^{n} (-1)^j \binom{p}{j} = (-1)^{k+1} \binom{p-1}{k}.$$

If we now let $p = 2k + 1$, we have

$$|e(\varphi)| = \binom{2k}{k} = \binom{|Bd^-|}{|Bd^-|/2}.$$

Since we have $f(X) = 1$ for all $X \in Bd^-$, the frequency sum of sets in the border is $\sum_X [X \in Bd^-] f(X) = |Bd^-| = |Bd_D^-|$. This completes the proof. □

While the construction creates a small number of sets in the border, there are of course many sets that are "almost" in the border, which is not true in the usual Bonferroni situation. The following theorem is another analogue of the Bonferroni inequalities.

Theorem 3. *Define the* thick negative border Bd^-_* *as the family of itemsets that are not frequent but that have at least one frequent subset. Then*

$$e(\varphi) \leq \sum_{X \in Bd^-_*} f(X).$$

Proof. Again, we will write $e(\varphi)$ as a sum over all tuples $T \in r$. We will show that the contribution made by T toward $e(\varphi)$ is bounded by the number of sets in Bd^-_* that include T, which implies the claimed upper bound.

First of all, if $T \in \mathcal{F}_\sigma$, the contribution is zero. Otherwise, the contribution is

$$\sum_{X \subset T} [X \notin \mathcal{F}_\sigma] \, (-1)^{|X|+1}. \tag{6}$$

Select any attribute $A \in T$, and delete from the sum (6) all pairs $X, Y \notin \mathcal{F}_\sigma$ such that $Y = X \cup \{ A \}$. What we have left is

$$\sum_{X \subset T} [X \notin \mathcal{F}_\sigma] \, [X \setminus \{ A \} \in \mathcal{F}_\sigma] \, (-1)^{|X|+1}.$$

All sets fulfilling both conditions of the sum are in $Bd^-_* \cap \mathcal{P}(T)$, and thus the absolute value of the contribution is bounded by $|Bd^-_* \cap \mathcal{P}(T)|$. Summing these inequalities for all contributions yields

$$|e(\varphi)| \leq \sum_{T \in R} |Bd^-_* \cap \mathcal{P}(T)| \leq \sum_{T \in R} |Bd^-_*| = \sum_{X \in Bd^-_*} f(X). \qquad \square$$

We have proved two theorems for upper-bounding the absolute error: Theorems 1 and 3. Both theorems are problematic in practice: the bound of Theorem 1 grows exponentially, and the thick border of Theorem 3 can be very large. It would be useful to find a bound for Truncate Sum in-between these two theorems. Note that the construction of Theorem 2 creates a large number of maximal frequent sets. By analogy with the negative border, one can define the *positive border* Bd^+ as the collection of these sets. For the construction, Bd^+ is large and Bd^- is small; in many practical cases, Bd^+ is smaller and Bd^- larger. The set $Bd^+ \cup Bd^-$ is worth investigating, and we conjecture (again [Man02]) that

$$e(\varphi) \leq \sum_X [X \in Bd^- \cup Bd^+] \, f(X).$$

4 General Queries

In this section we will generalize the preceding discussion: we will define Truncate Sum for arbitrary Boolean formulas φ, and prove counterparts of Lemma 2 and Theorem 1. The bounds provided by these results can be even larger than the disjunction-specific bounds of the previous section.

Let now φ be an arbitrary Boolean formula, i.e., an expression consisting of negation $\neg$, conjunction $\wedge$, disjunction $\vee$ and attributes $A \in U$. We define the semantics of such formulas in the usual way: $T \models \varphi$ if φ is true when the attributes are substituted by their values in T. The goal remains the same: to approximate $f(\varphi)$, the fraction of tuples supporting φ, given the collection $\mathcal{F}_\sigma$ of σ-frequent itemsets.

The support of the query formula can obviously be written as

$$f(\varphi) = \sum_X [X \models \varphi]\, g(X).$$

We denote the coefficients $\zeta(X) = [X \models \varphi]$. What we want to do is write

$$f(\varphi) = \sum_X \xi(X) f(X)$$

with suitable new coefficients $\xi(X)$, and then truncate the sum, obtaining

$$\hat{f}(\varphi) = \sum_X [X \in \mathcal{F}_\sigma]\, \xi(X) f(X).$$

To compute the new coefficients, we can use inclusion-exclusion: since

$$g(X) = \sum_Y [Y \supset X]\, (-1)^{|Y \setminus X|} f(Y),$$

we have

$$\begin{aligned} f(\varphi) &= \sum_X \zeta(X) g(X) = \sum_X \sum_Y \zeta(X)\, [Y \supset X]\, (-1)^{|Y \setminus X|} f(Y) \\ &= \sum_Y f(Y) \sum_X \zeta(X)\, [Y \supset X]\, (-1)^{|Y \setminus X|}. \end{aligned}$$

The required coefficients are thus given by

$$\xi(Y) = \sum_X [X \subset Y]\, (-1)^{|Y \setminus X|} \zeta(X).$$

Next we prove a generalization of Lemma 2.

Lemma 3. *When φ is an arbitrary Boolean formula with exact-frequency coefficients $\zeta(X) = [X \models \varphi]$ and the border Bd^- does not contain the empty set,*

$$e(\varphi) = \sum_X \nu(X) g(X),$$

where

$$\nu(X) = (-1)^{|X|} \sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} (-1)^{|\mathcal{E}|+1} \sum_Y \left[X \setminus \bigcup \mathcal{E} \subset Y \subset X\right] (-1)^{|Y|} \zeta(Y).$$

Proof. The error is

$$\begin{aligned} e(\varphi) &= \sum_X [X \in \mathcal{G}]\, \xi(X) f(X) \\ &= \sum_{X,Y} [X \in \mathcal{G}]\, f(X)\, [Y \subset X]\, (-1)^{|X \setminus Y|} \zeta(Y). \end{aligned}$$

Again we apply inclusion-exclusion on the condition $X \in \mathcal{G}$:

$$[X \in \mathcal{G}] = \sum_{\emptyset \neq \mathcal{E} \subset Bd^-} (-1)^{|\mathcal{E}|+1} \left[X \supset \bigcup \mathcal{E}\right],$$

obtaining

$$\begin{aligned} e(\varphi) &= \sum_{\emptyset \neq \mathcal{E} \subset Bd^-} (-1)^{|\mathcal{E}|+1} \sum_{X,Y} \left[X \supset \bigcup \mathcal{E}\right] f(X)\, [Y \subset X]\, (-1)^{|X \setminus Y|} \zeta(Y) \\ &= \sum_{\emptyset \neq \mathcal{E} \subset Bd^-} (-1)^{|\mathcal{E}|+1} \sum_Y (-1)^{|Y|} \zeta(Y) \sum_X \left[X \supset \bigcup \mathcal{E} \cup Y\right] (-1)^{|X|} f(X) \\ &= \sum_{\emptyset \neq \mathcal{E} \subset Bd^-} (-1)^{|\mathcal{E}|+1} \sum_Y (-1)^{|Y| + |\bigcup \mathcal{E} \cup Y|} \zeta(Y) g\big(\bigcup \mathcal{E} \cup Y\big). \end{aligned}$$

Regrouping the terms yields

$$e(\varphi) = \sum_X g(X) (-1)^{|X|} \sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} (-1)^{|\mathcal{E}|+1} \sum_Y \left[X \setminus \bigcup \mathcal{E} \subset Y \subset X\right] (-1)^{|Y|} \zeta(Y),$$

which is the claim. □

To see that this generalizes Lemma 2, let $\varphi = \bigvee D$. Then $\zeta(X) = [X \cap D \neq \emptyset]$. Consider the sum over Y:

$$\sum_Y \left[X \setminus \bigcup \mathcal{E} \subset Y \subset X\right] (-1)^{|Y|} [Y \cap D \neq \emptyset]. \tag{7}$$

We may assume that $\bigcup \mathcal{E} \subset X$, since the outer sum is taken over $\mathcal{E} \subset Bd_X^-$. Furthermore, if $\bigcup \mathcal{E}$ contains any attribute A that is not in D, we can pair up terms corresponding to $Y \ni A$ and $Y \setminus \{A\}$, and thus show that the sum (7) is 0. On the other hand, if X contains any attributes that are in D but not in $\bigcup \mathcal{E}$, the Iverson function $[Y \cap D \neq \emptyset]$ is always 1, and since $\bigcup \mathcal{E} \neq \emptyset$, and the sum (7) is seen to compute the difference in number of even and odd subsets of $\bigcup \mathcal{E}$, which is of course 0.

Assume now that $X = \bigcup \mathcal{E} \cup Z$ with $\bigcup \mathcal{E} \subset D$ and $Z \cap D = \emptyset$. We thus have for $Z \subset Y \subset X$ that $[Y \cap D \neq \emptyset] = [Y \neq Z]$, and the sum (7) becomes $-(-1)^{|Z|} = (-1)^{|X \setminus \bigcup \mathcal{E}|+1} = (-1)^{|X|+|\bigcup \mathcal{E}|+1}$, since $\bigcup \mathcal{E} \subset X$.

We have shown for all X that

$$\sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} (-1)^{|\mathcal{E}|+1} \sum_Y \left[X \setminus \bigcup \mathcal{E} \subset Y \subset X\right] (-1)^{|Y|} \zeta(Y)$$
$$= \sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} \left[\bigcup \mathcal{E} \subset D\right] \left[(X \setminus \bigcup \mathcal{E}) \cap D = \emptyset\right] (-1)^{|\mathcal{E}|+|\bigcup \mathcal{E}|}.$$

The result of Lemma 2 follows by noting that for $X \subset D$

$$g_D(X) = \sum_Y [Y \cap D = \emptyset]\, g(X \cup Y)$$

and rearranging terms.

The coefficients $\nu(X)$ used in the statement of the lemma have already played a role in proving Theorem 1: the key part was showing that $|\nu(X)| \leq 2^{|Bd_X^-|}$ for disjunctions φ. A natural question then is, how large can $|\nu(X)|$ be for general queries? To answer this question, we rearrange the sum as

$$\nu(X) = (-1)^{|X|} \sum_Y [Y \subset X]\, (-1)^{|Y|} \zeta(Y) \sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} (-1)^{|\mathcal{E}|+1} \left[X \setminus \bigcup \mathcal{E} \subset Y\right]. \quad (8)$$

Denote by S the innermost sum. We can rewrite it in the form

$$S = \sum_{\emptyset \neq \mathcal{E} \subset Bd_X^-} \left[X \setminus Y \subset \bigcup \mathcal{E}\right] (-1)^{|\mathcal{E}|+1},$$

which is seen to be an inclusion-exclusion sum over the upwards-closed subfamily

$$\left\{ \mathcal{E} \subset Bd_X^- \;\middle|\; \bigcup \mathcal{E} \supset X \setminus Y \right\} \quad (9)$$

of the powerset of Bd_X^-. Applying Lemma 1 to this sum, we have for $|S|$ an upper bound of $\binom{p}{\lceil p/2 \rceil}$, where $p = |Bd_X^-|$. Combining this with the fact that $\zeta(Y)$ is always 0 or 1, we obtain

$$|\nu(X)| \leq 2^{|X|-1} \binom{|Bd_X^-|}{\lceil |Bd_X^-|/2 \rceil}.$$

We thus have the following analogue of Theorem 1.

Theorem 4. *For an arbitrary query φ, the absolute error $|e(\varphi)|$ of* Truncate Sum *is bounded by*

$$2^{|U|-1} \binom{|Bd^-|}{\lceil |Bd^-|/2 \rceil} |Bd^-|^{-1} \sum_{X \in Bd^-} f(X).$$

The bound in the general case is even larger than the one in the disjunction case. How close to the bound can we come? Consider the sum (8). The form of the alternating sum over Y suggests that a parity-like function would be a difficult case: if $\zeta(Y) = 1$ if and only if $|Y|$ is even, the sum becomes

$$\nu(X) = (-1)^{|X|} \sum_{Y \subset X} \big[|Y| \text{ even}\big]\, S,$$

where S is the inclusion-exclusion sum mentioned in the proof of Theorem 4. The bound for $|S|$ used Lemma 1, where it is easy to see that equality holds if the upwards-closed family (9) consists of those sets $\mathcal{E} \subset Bd_X^-$ that have $|\mathcal{E}| = \lceil |Bd_X^-|/2 \rceil$. But for $Y = \emptyset$ exactly this is achieved by the construction in the proof of Theorem 2. For larger sets $Y \subset X$, S is smaller; however, this suffices to show that if the statement of Theorem 4 is to be strengthened, one cannot simply decrease the general bound for $|S|$, but more careful analysis of the double sum (8) is required.

5 Conclusion and Future Work

We have described the Approximate Query problem and analyzed the Truncate Sum algorithm, expanding upon the foundations in [MT96]. The results are disappointing in a sense: for the simple-looking query class of disjunctions of attributes, the behavior is not even polynomial in the size of the border. However, this is a worst-case situation that may not be very realistic in practical, sparse datasets. In the proof of Theorem 1, the key inequality (5) is based upon bounding the ratio $|\nu(X)|/\mu(X)$. However, the ratio is multiplied by the quantity $g_D(X)$, the exact frequency of X when the data is projected to the attributes in D, and in sparse data it is reasonable that this quantity should vanish for most large itemsets X. This observation suggests a modified algorithm: when mining the frequent itemsets, remove from the data those tuples where the ratio would be large, and store them separately; if the data is sparse, there should not be too many of these tuples. Queries can be computed exactly for the difficult, dense tuples, and approximated for the easy part of the data condensed into the frequent itemset representation.

More generally, assume that there is space for storing some extra information along with the frequent itemsets. The question then is, what is a good class of information to store in order to approximate a wide variety of queries?

Another avenue for future research is to use the information inherent in frequent itemsets in some way other than truncating the inclusion-exclusion sum. In the Bonferroni case, Linial and Nisan have shown that if the frequencies are known for sets X with $|X| \geq \Omega(\sqrt{|D|})$, there are good approximations to $f(\bigvee D)$ using multipliers other than ± 1, and if the frequencies are known only for sets X with $|X| \leq O(\sqrt{|D|})$, no approximation can be very good [LN90]. It would be interesting to extend this approach to the general case of frequent itemsets that do not form such a level family, and to queries more general than disjunctions.

References

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *SIGMOD '93*, pages 207–216, 1993.

[AMS+96] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307–328. AAAI Press, 1996.

[BBR00] Jean-François Boulicaut, Artur Bykowski, and Christophe Rigotti. Approximation of frequency queries by means of free-sets. In *PKDD '00*, volume 1910 of *LNCS*, pages 75–85. Springer, 2000.

[Bol88] Béla Bollobás. *Combinatorics: set systems, hypergraphs, families of vectors and combinatorial probability*. U Cambridge, 1988.

[BSH04] Artur Bykowski, Jouni K. Seppänen, and Jaakko Hollmén. Model-independent bounding of the supports of Boolean formulae in binary data. In Rosa Meo, Pier Luca Lanzi, and Mika Klemettinen, editors, *Database Support for Data Mining Applications: Discovering Knowledge with Inductive Queries*, volume 2682 of *LNAI*, pages 234–249. Springer, 2004.

[CG02] Toon Calders and Bart Goethals. Mining all non-derivable frequent itemsets. In *PKDD '02*, volume 2431 of *LNAI*, pages 74–85. Springer, 2002.

[GS96] Janos Galambos and Italo Simonelli. *Bonferroni-type Inequalities with Applications*. Probability and its Applications. Springer, 1996.

[GZ03] Bart Goethals and Mohammed J. Zaki, editors. *Proceedings of the Workshop on Frequent Itemset Mining Implementations (FIMI–03)*, volume 90 of *CEUR-WS*, Melbourne, Florida, 2003. http://CEUR-WS.org/Vol-90/.

[Knu92] Donald E. Knuth. Two notes on notation. *Am. Math. Monthly*, 99(5):403–422, 1992.

[LN90] Nathan Linial and Noam Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.

[Man02] Heikki Mannila. Local and global methods in data mining: Basic techniques and open problems. In P. Widmayer, F. Triguero, R. Morales, M. Hennessy, S. Eidenbenz, and R. Conejo, editors, *Automata, Languages and Programming*, volume 2380 of *LNCS*, pages 57–68. Springer, 2002.

[MT96] Heikki Mannila and Hannu Toivonen. Multiple uses of frequent sets and condensed representations. In *KDD '96*, pages 189–194, Portland, Oregon, August 1996. AAAI Press.

[PMS00] Dmitry Pavlov, Heikki Mannila, and Padhraic Smyth. Probabilistic models for query approximation with large sparse binary datasets. In *UAI*, 2000.

[PS01] Dmitry Pavlov and Padhraic Smyth. Probabilistic query models for transaction data. In *KDD '01*, 2001.