

Chapter 6

Summarizing Itemsets

6.1 Maximal and Closed Frequent Itemsets

Given a database over the tids \mathcal{T} , and items I , let \mathcal{F} denote the set of all frequent itemsets, i.e.,

$$\mathcal{F} = \{X \mid X \subseteq I \text{ and } \text{sup}(X) \geq \text{minsup}\} \tag{6.1}$$

A frequent itemset $X \in \mathcal{F}$ is called *maximal* iff it has no frequent supersets. Let \mathcal{M} be the set of all maximal frequent itemsets, which is given as

$$\mathcal{M} = \{X \mid X \in \mathcal{F} \text{ and } \nexists Y \supset X, \text{ such that } Y \in \mathcal{F}\} \tag{6.2}$$

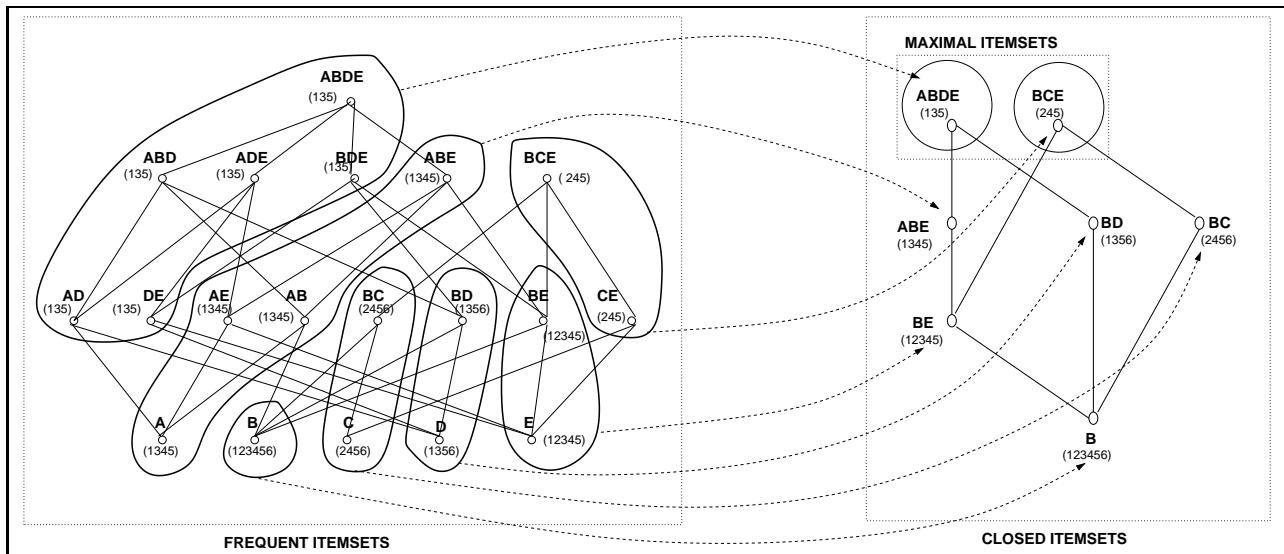


Figure 6.1: Frequent, Closed Frequent and Maximal Frequent Itemsets

A frequent set $X \in \mathcal{F}$ is called *closed* iff it has no frequent supersets with the same frequency. Let \mathcal{C} be the set of all closed frequent itemsets, which is given as

$$\mathcal{C} = \{X \mid X \in \mathcal{F} \text{ and } \nexists Y \supset X \text{ with } \text{sup}(X) = \text{sup}(Y)\} \tag{6.3}$$

The following relationship holds between these sets: $\mathcal{M} \subseteq \mathcal{C} \subseteq \mathcal{F}$, which is illustrated in Figure 6.1, based on the example dataset in Figure 5.1(b) and using $minsup = 3$. We can see clearly the *equivalence classes* of itemsets that have the same tidsets; the largest itemset in each class is a closed itemset. We can also see that the maximal itemsets are a subset of the closed itemsets.

6.2 Non-Derivable Itemsets

An itemset is set to be *non-derivable* if its support cannot be derived from some combination of the support of its subsets. The set of all frequent non-derivable itemsets ($NDI \subseteq \mathcal{F}$) are a summary or condensed representation of the set of all frequent itemsets. Furthermore, NDI is lossless with respect to support, that is, the exact support of all other frequent itemsets can be deduced from it.

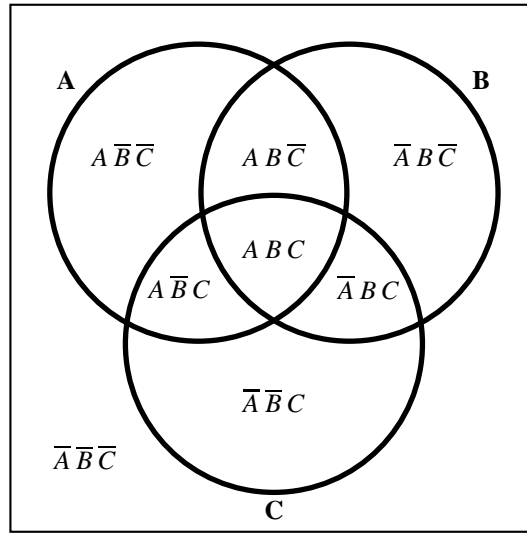


Figure 6.2: Tid Partition Induced by Three Items

Let \mathcal{T} be a set of tids, and let I be a set of items. Let A , B , and C be any three items in I . These three sets induce a partition on the space of all tids, as illustrated in the Venn Diagram of Figure 6.2. For example, the region labeled $A\bar{B}\bar{C}$ represents those tids that contain A but neither B nor C . In general any set of items X , with size $|X| = k$, induces a partition with 2^k distinct regions.

Notice that each region is in fact a *generalized itemset*, which is defined as an itemset that may contain either an item or its negation. A generalized itemset can be represented as $X\bar{Y}$, where X consists of regular items and Y consists of negated items. For example for $A\bar{B}\bar{C}$, we have $X = A$ and $Y = BC$. Define the support of a generalized itemset $X\bar{Y}$ as the number of transactions that contain all items in X , but no items in Y , given as

$$sup(X\bar{Y}) = |\{t \in \mathcal{T} \mid X \subseteq \mathbf{i}(t) \text{ and } Y \cap \mathbf{i}(t) = \emptyset\}| \quad (6.4)$$

Consider how we can compute the support of $A\bar{B}\bar{C}$ in Figure 6.2. We start with all the tids for A , and remove the tids contained in regions AB and AC . However, we realize that this removes the tids in ABC twice, so we need to add those back. In other words, the support of $A\bar{B}\bar{C}$ is given as

$$sup(A\bar{B}\bar{C}) = sup(A) - sup(AB) - sup(AC) + sup(ABC) \quad (6.5)$$

Notice that the support of \overline{ABC} is given by some combination of the supports of the (regular) supersets of A that do not have any negated items. In fact the supersets are obtained by combining A with any subset of the set of negated items $\{B, C\}$. Another way to say this is that if we let $I = \{A\} \cup \{B, C\} = ABC$, then the support of \overline{ABC} is given as a combination of itemsets in the set $\{J \mid A \subseteq J \subseteq ABC\}$.

Inclusion-Exclusion Principle

Let $X\bar{Y}$ be a generalized itemset, and let $I = X \cup Y$. Then the support of $X\bar{Y}$ can be expressed as some combination of the supports of supersets $J \supseteq X$, such that $J \subseteq I$. The precise formula is given by the well-known *inclusion-exclusion principle* in combinatorics:

$$\text{sup}(X\bar{Y}) = \sum_{X \subseteq J \subseteq I} (-1)^{|J \setminus X|} \text{sup}(J) \quad (6.6)$$

where $|J \setminus X| = J - X$. Let us verify this for $X\bar{Y} = \overline{ABC}$

$$\begin{aligned} \text{sup}(\overline{ABC}) &= (-1)^0 \text{sup}(A) + && J = A, |J \setminus X| = 0 \\ &(-1)^1 \text{sup}(AB) + && J = AB, |J \setminus X| = 1 \\ &(-1)^1 \text{sup}(AC) + && J = AC, |J \setminus X| = 1 \\ &(-1)^2 \text{sup}(ABC) && J = ABC, |J \setminus X| = 2 \\ &= \text{sup}(A) - \text{sup}(AB) - \text{sup}(AC) + \text{sup}(ABC) \end{aligned} \quad (6.7)$$

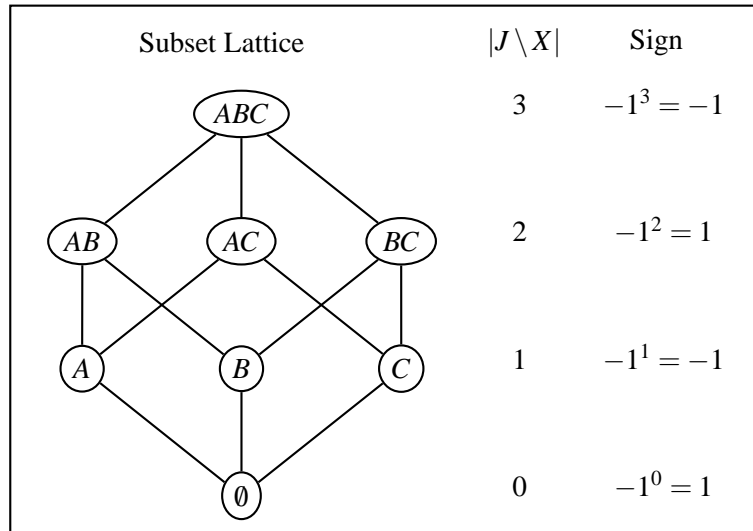


Figure 6.3: Subset Lattice and Signs for Different Levels

As another example, the support of $X\bar{Y} = \overline{ABC}$, with $X = \emptyset$ and $Y = ABC$. Figure 6.3 shows the cardinality of $J \setminus X$ and the corresponding signs to use in the inclusion-exclusion formula, which yields the following combination:

$$\begin{aligned} \text{sup}(\overline{ABC}) &= \text{sup}(\emptyset) \\ &\quad - \text{sup}(A) - \text{sup}(B) - \text{sup}(C) \\ &\quad + \text{sup}(AB) + \text{sup}(AC) + \text{sup}(BC) \\ &\quad - \text{sup}(ABC) \end{aligned} \quad (6.8)$$

Support Bounds for an Itemsets

Notice that the inclusion-exclusion formula for the support of $X\bar{Y}$, always starts at some itemset X and has terms for all subsets between X and $I = X \cup Y$. In other words, each of the eight regions in Figure 6.2 has a term for $sup(ABC)$. In general each of the $2^{|I|}$ regions in the partition induced by items in I has a term for I . Also note that the support of any generalized itemset must be at least zero (negative support would be a meaningless concept), i.e.,

$$sup(X\bar{Y}) \geq 0 \quad (6.9)$$

For example, setting $sup(\overline{ABC}) \geq 0$, gives us

$$\begin{aligned} sup(\overline{ABC}) &= sup(A) - sup(AB) - sup(AC) + sup(ABC) \geq 0, \text{ which implies} \\ sup(ABC) &\geq -sup(A) + sup(AB) + sup(AC) \end{aligned}$$

Notice that this rule gives a lower bound on the support of ABC .

As another example, setting $sup(\overline{ABC}) \geq 0$, yields

$$\begin{aligned} sup(\overline{ABC}) &= sup(\emptyset) - sup(A) - sup(B) - sup(C) + sup(AB) + sup(AC) + sup(BC) - sup(ABC) \geq 0 \\ \text{or } sup(ABC) &\leq sup(\emptyset) - sup(A) - sup(B) - sup(C) + sup(AB) + sup(AC) + sup(BC) \end{aligned}$$

Notice that this rule gives an upper bound on the support of ABC .

In fact, from each of the regions in Figure 6.2, we get one rule, and out of the eight possible rules, exactly four give upper bounds and the other four give lower bounds on the support of ABC , as shown below:

$$\begin{array}{ll} sup(ABC) \geq 0 & \text{when } X = ABC \\ \leq sup(AB) & \text{when } X = AB \\ \leq sup(AC) & \text{when } X = AC \\ \leq sup(BC) & \text{when } X = BC \\ \geq sup(AB) + sup(AC) - sup(A) & \text{when } X = A \\ \geq sup(AB) + sup(BC) - sup(B) & \text{when } X = B \\ \geq sup(AC) + sup(BC) - sup(C) & \text{when } X = C \\ \leq sup(AB) + sup(BC) + sup(AC) - sup(A) - sup(B) - sup(C) + sup(\emptyset) & \text{when } X = \emptyset \end{array}$$

The above derivation rules are schematically summarized in Figure 6.4.

In the general case, for any itemset of interest I , we can derive the rules for the upper and lower bound on its support, by using each set $X \subseteq I$ the starting point, and after rearranging the terms in the inclusion-exclusion formula (6.6), we get the following two general rules:

$$\textbf{Upper Bounds (}|I \setminus X| \text{ is odd): } sup(I) \leq \sum_{X \subseteq J \subseteq I} (-1)^{(|I \setminus J|+1)} sup(J) \quad (6.10)$$

$$\textbf{Lower Bounds (}|I \setminus X| \text{ is even): } sup(I) \geq \sum_{X \subseteq J \subseteq I} (-1)^{(|I \setminus J|+1)} sup(J) \quad (6.11)$$

Note that the only difference in the two equations is the inequality, which depends on the starting point X for the rule derivation.

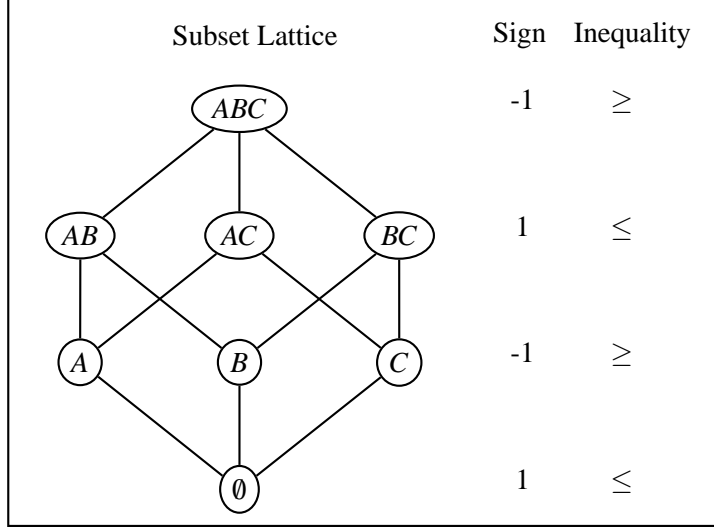


Figure 6.4: Derivation Rule Summary

Non-Derivable Itemsets

Since we have several derivation rules for the support lower and upper bounds for any itemset I , we can determine the *Least Upper Bound*, denoted $lub(I)$, among all the upper bounds, and the *Greatest Lower Bound*, denoted $glb(I)$, among all the lower bounds for $sup(I)$ from the different rules. It follows that $sup(I) \in [glb(I), lub(I)]$. In other words, the actual support of I must clearly be at least $glb(I)$, and at most $lub(I)$. If the $glb(I)$ and $lub(I)$ are identical, then $sup(I)$ can be derived exactly using the supports of its subsets, and I is called **derivable**. On the other hand, if $glb(I) \neq lub(I)$ then we cannot exactly determine the support of I from its subsets (we only know the range), and in this case I is called **non-derivable**. Thus the set of all **Non-Derivable Itemsets (NDI)** is given as follows:

$$NDI = \{I \in \mathcal{F} \mid glb(I) \neq lub(I)\} \quad (6.12)$$

where \mathcal{F} is the set of all frequent itemsets (for a given *minsup* threshold).

For example, consider the database shown in Figure 5.1(b), and the corresponding set of frequent itemsets \mathcal{F} shown in Figure 5.2. Here are a few examples of how to derive the support bounds for a given itemset. For example for $I = A$, we have

$$\begin{aligned} sup(A) &\leq sup(\emptyset) = 6 \\ &\geq 0 \end{aligned}$$

Thus $sup(A) \in [0, 6]$ and A is non-derivable.

For $I = AB$, we have

$$\begin{aligned} sup(AB) &\leq sup(A) = 4 \\ &\leq sup(B) = 6 \\ &\geq 0 \\ &\geq sup(A) + sup(B) - sup(\emptyset) = 4 + 6 - 6 = 4 \end{aligned}$$

Thus $sup(AB) \in [4, 4]$, or $sup(AB) = 4$, which means that AB is derivable.

As another example for $I = ABDE$, we get the following upper bounds:

$$\begin{aligned}
 \text{sup}(ABDE) &\leq \text{sup}(ABD) = 3 \\
 &\leq \text{sup}(ABE) = 4 \\
 &\leq \text{sup}(ADE) = 3 \\
 &\leq \text{sup}(BDE) = 3 \\
 \text{sup}(ABDE) &\leq \text{sup}(A) - \text{sup}(AB) - \text{sup}(AD) - \text{sup}(AE) + \text{sup}(ABD) + \text{sup}(ABE) + \text{sup}(ADE) \\
 &= 4 - (4 + 3 + 4) + (3 + 4 + 3) = 3
 \end{aligned}$$

From these upper bounds, we know that $\text{sup}(ABDE) \leq 3$. Let's consider the lower bound rule derived from AB , we get:

$$\begin{aligned}
 \text{sup}(ABDE) &\geq \text{sup}(ABD) + \text{sup}(ABE) - \text{sup}(AB) \\
 &= 3 + 4 - 4 = 3
 \end{aligned}$$

At this point we know that $\text{sup}(ABDE) \geq 3$, so without processing any further rules, we can immediately conclude that $\text{sup}(ABDE) \in [3, 3]$, which means that $ABDE$ is derivable.

For the example database in Figure 5.1(b), the set of non-derivable itemsets (with the support bounds) is given as:

$$NDI = \{A[0, 6], B[0, 6], C[0, 6], D[0, 6], E[0, 6], AD[2, 4], AE[3, 4], CE[3, 4], DE[3, 4]\}$$