A Note on Depth-First Mining of Frequent Itemsets

Jef Wijsen

February 22, 2008

Abstract

This note may help you in understanding Section 6.6 in the textbook [1].

1 Vertical Database Layout

Let $(\mathcal{I}, \prec)$ be a linearly ordered, finite set of *items*. Let $\mathcal{T}$ be a set of transaction identifiers. Under the *vertical database layout*, a *transaction database* is a total function $\mathsf{cover} : \mathcal{I} \rightarrow 2^{\mathcal{T}}$. All the following definitions are relative to such a fixed transaction database cover.

Example 1 Assume $a \prec b \prec c \prec d \prec e$. The following transaction database is shown in Fig. 6.24 on page 364 of [1].

$$
\begin{array}{lllllllllll}
\mathsf{cover}(a) & = \{ & 1, & & 3, & 4, & 5, & 6, & 7, & 8, & 9 & & \} \\
\mathsf{cover}(b) & = \{ & 1, & 2, & & & 5, & 6, & & 8, & 9, & 10 & \} \\
\mathsf{cover}(c) & = \{ & & 2, & 3, & & 5, & 6, & & 8, & & 10 & \} \\
\mathsf{cover}(d) & = \{ & & 2, & 3, & 4, & & 6, & & & 9 & & \} \\
\mathsf{cover}(e) & = \{ & & & 3, & 4, & & & & & & 10 & \}
\end{array}
$$

An *itemset* is a sequence $a_1 a_2 \ldots a_n$ where $n \geq 1$ and $a_1 \prec a_2 \prec \ldots \prec a_n$ (the technical treatment can be easily extended to deal with the empty itemset). The function cover naturally extends to itemsets:

$$\mathsf{cover}(a_1 a_2 \ldots a_n) = \bigcap_{i=1}^{n} \mathsf{cover}(a_i)$$

Example 2

$$\mathsf{cover}(abc) = \{5, 6, 8\}$$

The (left) concatenation operator $\cdot$ is defined as usual: if $s = a_1 a_2 \ldots a_n$ is an itemset and $a_0 \prec a_1$, then $a_0 \cdot s$ denotes the itemset $a_0 a_1 \ldots a_n$.

2 Conditional Cover

Let s be an itemset. We define the "conditional cover" as follows:

$$
\begin{array}{lll}
\mathsf{cocov}[s] = \{(a, T) \mid & a \in \mathcal{I} \text{ and} & \\
& a \text{ precedes (w.r.t. } \prec) \text{ the leftmost symbol of } s \text{ and} & \\
& T = \mathsf{cover}(a \cdot s) & \} .
\end{array}
$$

If b be the leftmost symbol of s, then $\mathsf{cocov}[s]$ is a transaction database with domain $\{a \in \mathcal{I} \mid a \prec b\}$.

Example 3

$$
\begin{array}{lll}
\mathsf{cocov}[de](a) & = & \{3, 4\} \\
\mathsf{cocov}[de](b) & = & \{\} \\
\mathsf{cocov}[de](c) & = & \{3\}
\end{array}
$$

Property 1 *If $b \cdot s$ is an itemset and $a \prec b$, then*

$$\mathsf{cocov}[b \cdot s](a) = \mathsf{cocov}[s](b) \cap \mathsf{cocov}[s](a) \ .$$

Proof.

$$\begin{aligned}
\mathsf{cocov}[b \cdot s](a) &= \mathsf{cover}(a \cdot b \cdot s) \\
\mathsf{cover}(a \cdot b \cdot s) &= \mathsf{cover}(a \cdot s) \cap \mathsf{cover}(b \cdot s) \\
\mathsf{cover}(a \cdot s) &= \mathsf{cocov}[s](a) \\
\mathsf{cover}(b \cdot s) &= \mathsf{cocov}[s](b)
\end{aligned}$$

□

This property suggests a recursive procedure for computing frequent itemsets with a support count greater than σ. Compute recursively the following sequence, while outputting frequent itemsets and ignoring infrequent items.

- $\mathsf{cocov}[e]$
  - $\mathsf{cocov}[de]$
    - $\mathsf{cocov}[cde]$
      - $\mathsf{cocov}[bcde]$
    - $\mathsf{cocov}[bde]$
  - $\mathsf{cocov}[ce]$
    - $\mathsf{cocov}[bce]$
  - $\mathsf{cocov}[be]$
- $\mathsf{cocov}[d]$
  - $\mathsf{cocov}[cd]$
    - $\mathsf{cocov}[bcd]$
  - $\mathsf{cocov}[bd]$
- $\mathsf{cocov}[c]$
  - $\mathsf{cocov}[bc]$
- $\mathsf{cocov}[b]$
- $\mathsf{cocov}[a]$

Example 4 Assume $\sigma = 2$.

- $\mathsf{cocov}[e]$ $\begin{array}{|ll|} \hline a & \{3,4\} \\ b & \{10\} \\ c & \{3,10\} \\ d & \{3,4\} \\ \hline \end{array}$ $\begin{array}{ll} \rightsquigarrow & \{a,e\} \\ \times & \\ \rightsquigarrow & \{c,e\} \\ \rightsquigarrow & \{d,e\} \end{array}$

  In the *horizontal database layout*, this yields (omitting the infrequent item b): $\{(3, acd), (4, ad), (10, c)\}$. The FP-tree is shown in Fig. 6.27 (b) on page 368 of [1].

  - $\mathsf{cocov}[de]$ $\begin{array}{|ll|} \hline a & \{3,4\} \\ c & \{3\} \\ \hline \end{array}$ $\begin{array}{ll} \rightsquigarrow & \{a,d,e\} \\ \times & \end{array}$

    In the *horizontal database layout*, this yields (omitting the infrequent item c): $\{(3, a), (4, a)\}$. The FP-tree is shown in Fig. 6.27 (d) on page 368 of [1].

  - $\mathsf{cocov}[ce]$ $\begin{array}{|ll|} \hline a & \{3\} \\ \hline \end{array}$ $\times$

- $\mathsf{cocov}[d]$ $\begin{array}{|ll|} \hline a & \{3,4,6,9\} \\ b & \{2,6,9\} \\ c & \{2,3,6\} \\ \hline \end{array}$ $\begin{array}{ll} \rightsquigarrow & \{a,d\} \\ \rightsquigarrow & \{b,d\} \\ \rightsquigarrow & \{c,d\} \end{array}$

  - $\mathsf{cocov}[cd]$ $\begin{array}{|ll|} \hline a & \{3,6\} \\ b & \{2,6\} \\ \hline \end{array}$ $\begin{array}{ll} \rightsquigarrow & \{a,c,d\} \\ \rightsquigarrow & \{b,c,d\} \end{array}$
    - $\mathsf{cocov}[bcd]$ $\begin{array}{|ll|} \hline a & \{6\} \\ \hline \end{array}$ $\times$
  - $\mathsf{cocov}[bd]$ $\begin{array}{|ll|} \hline a & \{6,9\} \\ \hline \end{array}$ $\rightsquigarrow \{a,b,d\}$

- $\mathsf{cocov}[c]$ $\begin{array}{|ll|} \hline a & \{5,6,8\} \\ b & \{2,5,6,8,10\} \\ \hline \end{array}$ $\begin{array}{ll} \rightsquigarrow & \{a,c\} \\ \rightsquigarrow & \{b,c\} \end{array}$

  - $\mathsf{cocov}[bc]$ $\begin{array}{|ll|} \hline a & \{5,6,8\} \\ \hline \end{array}$ $\rightsquigarrow \{a,b,c\}$

- $\mathsf{cocov}[b]$ $\begin{array}{|ll|} \hline a & \{1,5,6,8,9\} \\ \hline \end{array}$ $\rightsquigarrow \{a,b\}$

References

[1] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining.* Addison Wesley, 2006.