

The background features a complex network of thin, intersecting lines in shades of red, orange, and grey, creating a web-like structure. Scattered throughout are small, colored dots in green, blue, and orange. On the left side, there is a vertical strip with a grid of small, light-colored squares. In the upper left corner, there is a small inset image showing a cluster of orange and red dots. The overall aesthetic is technical and data-oriented.

Mining Frequent Patterns by Exploring Vertical Data Format

Exploring Vertical Data Format: ECLAT

- ❑ ECLAT (Equivalence Class Transformation): A depth-first search algorithm using set intersection [Zaki et al. @KDD'97]
- ❑ Tid-List: List of transaction-ids containing an itemset
- ❑ Vertical format: $t(e) = \{T_{10}, T_{20}, T_{30}\}$; $t(a) = \{T_{10}, T_{20}\}$; $t(ae) = \{T_{10}, T_{20}\}$
 $a, e \rightarrow \text{Tid-lists}$
- ❑ Properties of Tid-Lists
 - ★ $t(X) = t(Y)$: X and Y always happen together (e.g., $t(ac) = t(d)$)
 - ★ $t(X) \subset t(Y)$: transaction having X always has Y (e.g., $t(ac) \subset t(ce)$)
- ❑ Deriving frequent patterns based on vertical intersections
- ❑ Using diffset to accelerate mining
 - ❑ Only keep track of differences of tids
 - ❑ $t(e) = \{T_{10}, T_{20}, T_{30}\}$, $t(ce) = \{T_{10}, T_{30}\} \rightarrow \text{Diffset}(ce, e) = \{T_{20}\}$ 节省空间

A transaction DB in Horizontal Data Format

Tid	Itemset
10	a, c, d, e
20	a, b, e
30	b, c, e

The transaction DB in Vertical Data Format

Item	TidList
a	10, 20
b	20, 30
c	10, 30
d	10
e	10, 20, 30