

The background of the slide is a complex, abstract composition. It features a network of thin, intersecting lines in shades of red, orange, and brown, creating a web-like structure. Scattered throughout this network are numerous small, colored dots in green, blue, and orange. In the upper left corner, there is a rectangular inset showing a different data visualization: a grid of small squares with varying shades of orange and red, overlaid with a network of lines. The title text is centered in a white, semi-transparent rectangular area.

Extensions or Improvements of Apriori

Apriori: Improvements and Alternatives

- ❑ Reduce passes of transaction database scans
 - ❑ Partitioning (e.g., Savasere, et al., 1995)  To be discussed in subsequent slides
 - ❑ Dynamic itemset counting (Brin, et al., 1997)
- ❑ Shrink the number of candidates
 - ❑ Hashing (e.g., DHP: Park, et al., 1995)  To be discussed in subsequent slides
 - ❑ Pruning by support lower bounding (e.g., Bayardo 1998)
 - ❑ Sampling (e.g., Toivonen, 1996)
- ❑ Exploring special data structures
 - ❑ Tree projection (Agarwal, et al., 2001) *FP-Tree*
 - ❑ H-miner (Pei, et al., 2001)
 - ❑ Hypercube decomposition (e.g., LCM: Uno, et al., 2004)

Partitioning: Scan Database Only Twice

- Theorem: Any itemset that is potentially frequent in TDB must be frequent in at least one of the partitions of TDB 分成 k 个部分.

Here is the proof!

$$\begin{array}{ccccccc}
 \boxed{} & & \boxed{} & & \dots & & \boxed{} \\
 \text{TDB}_1 & + & \text{TDB}_2 & + & \dots & + & \text{TDB}_k = \text{TDB} \\
 \text{sup}_1(X) < \sigma |\text{TDB}_1| & & \text{sup}_2(X) < \sigma |\text{TDB}_2| & & \dots & & \text{sup}_k(X) < \sigma |\text{TDB}_k| \quad \text{sup}(X) < \sigma |\text{TDB}| \\
 \text{sup}(X) = \sum_{i=1}^k \text{sup}_i(X) < \sigma \sum_{i=1}^k |\text{TDB}_i| = \sigma |\text{TDB}| \Rightarrow \text{sup}(X) < \sigma |\text{TDB}|
 \end{array}$$

- Method: (A. Savasere, E. Omiecinski and S. Navathe, VLDB'95)
 - Scan 1: Partition database (how?) and find local frequent patterns
 - Scan 2: Consolidate global frequent patterns (how to?)
- Why does this method guarantee to scan TDB only twice?

Direct Hashing and Pruning (DHP)

- DHP (Direct Hashing and Pruning): Reduce the number of candidates (J. Park, M. Chen, and P. Yu, SIGMOD'95)
- Observation: A k -itemset whose corresponding hashing bucket count is below the threshold cannot be frequent

□ Candidates: a, b, c, d, e

□ Hash entries

□ {ab, ad, ae}

□ {bd, be, de}

□ ...

| Itemsets | Count |
|--------------|-------|
| {ab, ad, ae} | 35 |
| {bd, be, de} | 298 |
| | ... |
| {yz, qs, wt} | 58 |

Hash Table

□ Frequent 1-itemset: a, b, d, e

□ ab is not a candidate 2-itemset if the sum of count of {ab, ad, ae} is below support threshold

→ count 是所有 itemsets 支持总数。

if count < threshold :
itemsets not frequent.

else: maybe frequent