Mining Frequent Itemsets using the vertical data format

Vertical Data Format Approach

- Mining in Horizontal data format done using TID-itemset
- Mining in vertical data format using item-TID_set
  - TID_set is the set of transaction identifiers containing the item

Vertical Data Format Approach

- Convert the transaction db into vertical data format using item and transaction-set
- Mining can be performed by intersecting the TID_sets of every pair of frequent single items.
- All non-empty subsets with minimum support count belong to the set of frequent 2-itemsets
- So apart from the pair {I1,I4} and {I3,I5} all other Itemset forms the frequent 2-itemsets

Table 6.3 The Vertical Data Format of the Transaction Data Set *D* of Table 6.1

itemset	*TID_set*
I1	{T100, T400, T500, T700, T800, T900}
I2	{T100, T200, T300, T400, T600, T800, T900}
I3	{T300, T500, T600, T700, T800, T900}
I4	{T200, T400}
I5	{T100, T800}

Table 6.4 2-Itemsets in Vertical Data Format

itemset	*TID_set*
{I1, I2}	{T100, T400, T800, T900}
{I1, I3}	{T500, T700, T800, T900}
{I1, I4}	{T400}
{I1, I5}	{T100, T800}
{I2, I3}	{T300, T600, T800, T900}
{I2, I4}	{T200, T400}
{I2, I5}	{T100, T800}
{I3, I5}	{T800}

Vertical Data Format Approach

- Based on the Apriori property the given 3-itemset is a candidate 3-itemset only if every 2-item subsets are frequent
- Candidate generation process will generate only two 3-itemsets by intersecting the TID_sets of corresponding frequent 2-itemsets

Table 6.5 3-Itemsets in Vertical Data Format

itemset	TID_set
{I1, I2, I3}	{T800, T900}
{I1, I2, I5}	{T100, T800}

SSN

Vertical Data Format General Approach

- Transform horizontal formatted data into vertical format data.
- Support count is length of the TID_set of the (k+1) itemsets based on the Apriori property.
- Computation done by intersection of TID_sets of k itemsets to compute TID_sets of (k+1) itemsets.
- Process repeats by incrementing k until no frequent itemsets or candidate itemsets can be found.
- No need to scan the database for (k>=1)

Vertical Data Format Approach

- To reduce the cost of space and computation time for long sets.
- Use the technique called diffset, keeps track of only difference of the TID-sets of (k+1) itemset and corresponding k-itemset.
- {I1}={T100,T400,T500,T700,T800,T900}

  {I1,I2}={T100,T400,T800,T900}

  Diffset:{{I1},{I1,I2}}={T500,T700}

When dataset contains long patterns the technique reduce the total cost of vertical format mining of frequent itemsets.