Bài tập Lý thuyết môn Khai phá Dữ liệu

Họ và tên: Huỳnh Nguyễn Thế Dân

MSSV: 21110256

Lớp: 21TTH1

Questions

1. Consider the transaction database in the table below:

```
| tid |   Items   |
|-----|-----------|
| 1   | a, b, c, d|
| 2   | b, c, e, f|
| 3   | a, d, e, f|
| 4   |    a, e, f|
| 5   |    b, d, f|
```

Determine the absolute support of itemsets {a, e, f}, and {d, f}. Convert the absolute support to the relative support.

2. For the database in Exercise 1, compute all frequent patterns at absolute minimum support values of 2, 3, and 4.

3. For the database in Exercise 1, determine all the maximal frequent patterns at absolute minimum support values of 2, 3, and 4.

4. Represent the database of Exercise 1 in vertical format.

5. Consider the transaction database in the table below:

```
| tid |    Items    |
|-----|-------------|
| 1   | a, c, d, e  |
| 2   | a, d, e, f  |
| 3   | b, c, d, e, f|
| 4   | b, d, e, f  |
| 5   | b, e, f     |
| 6   | c, d, e     |
| 7   | c, e, f     |
| 8   | d, e, f     |
```

Determine all frequent patterns and maximal patterns at support levels of 3, 4, and 5.

Answer

*For exercise 1:

Absolute Support

Absolute support of an itemset is defined as the number of transactions in which the itemset appears.

  1. For the itemset {a, e, f}:

- Transaction 3: contains {a, d, e, f}
- Transaction 4: contains {a, e, f} Hence, the absolute support for {a, e, f} is 2 (appears in transactions 3 and 4).

  2. For the itemset {d, f}:

- Transaction 3: contains {a, d, e, f}
- Transaction 5: contains {b, d, f} Hence, the absolute support for {d, f} is 2 (appears in transactions 3 and 5).

Relative Support

Relative support is calculated by dividing the absolute support by the total number of transactions in the database.

Total number of transactions = 5.

  1. For the itemset {a, e, f}:

- Relative support = Absolute support of {a, e, f} / Total number of transactions
- Relative support = 2 / 5 = 0.4

  2. For the itemset {d, f}:

- Relative support = Absolute support of {d, f} / Total number of transactions
- Relative support = 2 / 5 = 0.4

Thus, the relative support for both {a, e, f} and {d, f} is 0.4.

---

*For exercise 2: Which requires computing all frequent patterns at absolute minimum support values of 2, 3, and 4.

Support Threshold of 2:

Absolute Support Count for Key Itemsets:

- {a, e, f}: Appears in T3, T4 → Support = 2

- {d, f}: Appears in T3, T5 → Support = 2
- {b, c}: Appears in T1, T2 → Support = 2
- {e, f}: Appears in T2, T3, T4 → Support = 3 (also frequent at higher thresholds)
- {b, d}: Appears in T1, T5 → Support = 2
- {a, f}: Appears in T3, T4 → Support = 2

Frequent Itemsets at Support ≥ 2:

- Singletons: All items {a}, {b}, {c}, {d}, {e}, {f} have support ≥ 2 since all appear in at least two transactions.
- Pairs and Triples:
- {a, e, f}
- {d, f}
- {b, c}
- {b, d}
- {a, f}
- {e, f} (also frequent at higher thresholds)

Support Threshold of 3:

Since the item {f} appears in 4 transactions, it's the most frequent single item.

- {e, f}: Appears in T2, T3, T4 → Support = 3
- {a}: Appears in T1, T3, T4 → Support = 3
- {b}: Appears in T1, T2, T5 → Support = 3
- {d}: Appears in T1, T3, T5 → Support = 3

Frequent Itemsets at Support ≥ 3:

- {e, f}
- {a}, {b}, {d}, {f}

Support Threshold of 4:

- {f}: Appears in T2, T3, T4, T5 → Support = 4

Frequent Itemsets at Support ≥ 4:

- {f}

Corrected Frequent Itemsets Summary:

- Support ≥ 2: {a, e, f}, {d, f}, {b, c}, {b, d}, {a, f}, {e, f}, {a}, {b}, {c}, {d}, {e}, {f}
- Support ≥ 3: {e, f}, {a}, {b}, {d}, {f}
- Support ≥ 4: {f}

This revised and corrected list accurately represents the frequent itemsets across the different minimum support thresholds.

*For exercise 3:

We'll use the frequent patterns already identified in a previous response and further identify which of these are maximal by checking if they can be extended without their support dropping below the respective thresholds.

Recap of Frequent Itemsets:

From the previous analysis:

- Support ≥ 2: {a, e, f}, {d, f}, {b, c}, {b, d}, {a, f}, {e, f}, {a}, {b}, {c}, {d}, {e}, {f}
- Support ≥ 3: {e, f}, {a}, {b}, {d}, {f}
- Support ≥ 4: {f}

Identifying Maximal Frequent Itemsets:

We need to check each frequent itemset and ensure there is no superset with the same or higher frequency.

- Maximal Frequent Itemsets for Support ≥ 2:
  - {a, e, f}: No superset has the same or greater support.
  - {b, c}: No superset has the same or greater support.
  - {b, d}: No superset has the same or greater support.

Other frequent itemsets such as {d, f} and {e, f} are not maximal because {a, d, e, f} and {a, e, f} include these items with equal or greater support.

- Maximal Frequent Itemsets for Support ≥ 3:
  - {e, f}: No superset of {e, f} is frequent at this level of support.

Note that the individual items {a}, {b}, {d}, {f} are not maximal as they are parts of larger frequent itemsets at this level.

- Maximal Frequent Itemsets for Support ≥ 4:
  - {f}: As {f} is the only itemset frequent at this level, and no superset of {f} meets this threshold, it is maximal by default.

Conclusion:

Here are the maximal frequent itemsets for each support threshold:

- Support ≥ 2: {a, e, f}, {b, c}, {b, d}
- Support ≥ 3: {e, f}
- Support ≥ 4: {f}

These itemsets represent the largest sets of items that frequently appear together in the database without any extension that remains frequent under the respective thresholds.

---

*For exercise 4:

To represent the transaction database from Exercise 1 in a vertical format, we list each item along with the set of transaction IDs (tids) in which the item appears. This format is particularly useful for efficient computation of itemset supports in frequent pattern mining, allowing easier intersection of sets to find common transactions.

Here is the vertical representation of the database:

```
| Items |     Appears      |
|-------|------------------|
|   a   |     {1, 3, 4}    |
|   b   |     {1, 2, 5}    |
|   c   |     {1, 2}       |
|   d   |     {1, 3, 5}    |
|   e   |     {2, 3, 4}    |
|   f   |     {2, 3, 4, 5} |
```

Explanation:

- a: Appears in transactions 1, 3, and 4.
- b: Appears in transactions 1, 2, and 5.
- c: Appears in transactions 1 and 2.
- d: Appears in transactions 1, 3, and 5.
- e: Appears in transactions 2, 3, and 4.
- f: Appears in transactions 2, 3, 4, and 5.

This vertical format helps in quickly identifying the transactions associated with each item, facilitating operations like intersections for computing supports of larger itemsets.

---

*For exercise 5:

To determine all frequent patterns and maximal patterns at the specified support levels (3, 4, and 5) from the given transaction database, we'll first establish the support for all possible itemsets and then filter them by the desired thresholds. We will then identify the maximal patterns among those frequent patterns.

Step 1: Create Vertical Format Data

First, let's convert the transaction data to vertical format to facilitate support calculation:

```
| Items |           Appears           |
|-------|-----------------------------|
|   a   | {1, 2}                      |
|   b   | {3, 4, 5}                   |
|   c   | {1, 3, 6, 7}                |
|   d   | {1, 2, 3, 4, 6, 8}          |
|   e   | {1, 2, 3, 4, 5, 6, 7, 8}    |
|   f   | {2, 3, 4, 5, 7, 8}          |
```

Step 2: Compute Supports for Itemsets

We will use intersections of these sets to calculate supports for various itemsets:

- Single items:
  - a: 2
  - b: 3
  - c: 4
  - d: 6
  - e: 8
  - f: 6
- Pairs (calculated for selected frequent pairs based on individual item supports):
  - {d, e}: 6
  - {e, f}: 6
  - {d, f}: 5
- Triples and beyond (only those that can be potentially frequent):
  - {d, e, f}: 4

Step 3: Frequent Itemsets

Now we'll filter these by the specified support thresholds:

- Support ≥ 3:
  - {b}, {c}, {d}, {e}, {f}
  - {d, e}, {e, f}, {d, f}
  - {d, e, f}
- Support ≥ 4:
  - {c}, {d}, {e}, {f}
  - {d, e}, {e, f}, {d, f}
  - {d, e, f}
- Support ≥ 5:
  - {d}, {e}, {f}
  - {d, e}, {e, f}
- Note: {d, f} and {d, e, f} fall just below this threshold.

Step 4: Maximal Frequent Itemsets

To identify maximal patterns, we look for frequent itemsets that are not subsets of other frequent itemsets at the same support level:

- Maximal at Support ≥ 3:
  - {d, e, f} (it's the largest itemset frequent at this level, not contained fully within any other itemset)
- Maximal at Support ≥ 4:
  - {d, e, f} (still maximal as it's the largest frequent itemset not surpassed in frequency)
- Maximal at Support ≥ 5:
  - {d, e}, {e, f} (both are the largest combinations without being subsets at this threshold; {def} falls below this support)

Summary:

The calculated frequent and maximal itemsets reflect how sets of items appear together in transactions with at least the minimum frequency levels of 3, 4, and 5. These insights could be useful for identifying strong associations and potential rules in a market basket analysis or similar data mining tasks.

End.