Discussion Section 1 (CS145)

2015-10-02

Week 01

Teaching Assistant

- TAs
  - DAS, ARIYAM：ariyam@ucla.edu
  - Li, Ruirui: goldenbearray@gmail.com
  - YU, WENCHAO: wenchao.ta@gmail.com
  - Office Hour: Tuesday 2-5pm @ 3551P Boelter Hall

Grading Scheme

- Homework: 30%
  - Three assignments.
- Project: 20%
- Midterm: 20%
  - Open book
  - Date: Wednesday, November 18
- Final Exam: 30%
  - Open book
  - Date: Tuesday, December 9

Homework Policy

- Hardcopy submission.
- No late submission will be accepted.
  - except that each student will a **one-day** extension for one of the three assignments without penalty if the student contacts the instructor and TA before the submission deadline to arrange a late submission

Discussion Forum

- https://ccle.ucla.edu/course/view/15F-COMSCI145-1

Outline

- Review:
  - Apriori rule.
  - Apriori-based frequent itemset mining.
  - FP-tree
  - Closed vs Max Patterns
- Problem discussion:
  - Frequent itemset mining using
    - Apriori rule.
    - FP-Growth

Apriori rule (property)

- The Apriori rule claims that if a itemset meets the minimum support threshold, then any subset of it also meets the minimum support threshold.
  - Anti-monotone property

Apriori-based Mining

- Framework
  - (1) Generate length (k+1) candidate itemsets from length k frequent items, and
  - (2) Test the candidates against the database.

Example

- Transaction database
  - Transactions consist of a set of items I={a, b, c, …}

TID	Items Bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

  - Problem: Find frequent item sets
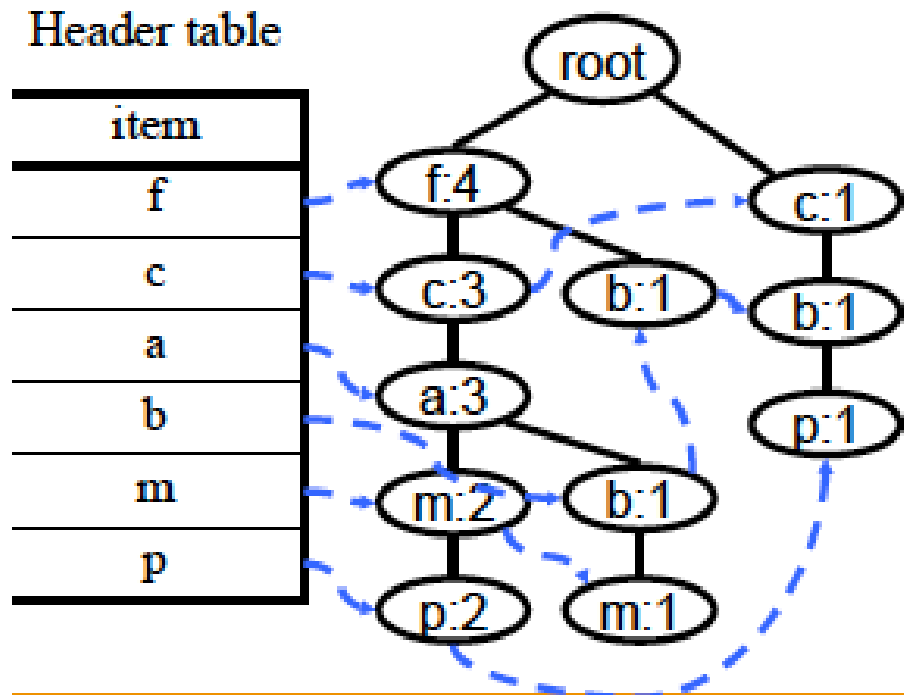   - Frequency >= minimum support threshold (3)

Apriori-based Mining

- Problem
  - Multiple scans of transaction database
  - Huge number of candidates
  - Tedious workload of support counting for candidates

FP-Tree

- Benefit
  - Preserve complete information for frequent pattern mining
  - Reduce irrelevant information
  - F-list sort the items in frequency descending order, so the more frequent items are more likely to be shared
  - Never be larger than the original database (not counting the node-links and the count fields)

Example

- Transaction database
  - Transactions consist of a set of items I={a, b, c, ...}

TID	Items Bought
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

  - Problem: Find frequent item sets
    - Frequency >= minimum support threshold (3)

Closed Patterns vs Max Patterns

- Closed Patterns:
  - A pattern P is a closed pattern if there is no super-pattern P' with the same support as P
- Max Patterns:
  - A pattern P is a max pattern if there exists no frequent superset P'
- Example:

  $TID_1 = \{a_1, a_2, a_{3\ldots} a_{50}\}$; $TID_2 = \{a_1, a_2, a_{3\ldots} a_{100}\}$
  min_sup = 1

  Closed Pattern = $\{a_1, a_2, a_{3\ldots} a_{50}\}$:2 and
  $\{a_1, a_2, a_{3\ldots} a_{100}\}$:1

  Max Pattern = $\{a_1, a_2, a_{3\ldots} a_{100}\}$:1