CSE 40647/60647 Data Science (Spring 2018)
Lecture 23: Frequent Pattern Mining: Evaluation

Quiz:

Given a transaction database:

Transaction ID	Items Bought
T1	{Mango, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T2	{Doll, Onion, Nintendo, Key-chain, Eggs, Yo-yo}
T3	{Mango, Apple, Key-chain, Eggs}
T4	{Mango, Umbrella, Corn, Key-chain, Yo-yo}
T5	{Corn, Onion, Onion, Key-chain, Ice-cream, Eggs}

or simplified as follows:

Transaction ID	**Items Bought**
T1	MONKEY
T2	DONKEY
T3	MAKE
T4	MUCKY
T5	COKIE

Use **FP-Growth** to find all frequent itemsets and their support if min_sup = 60%:

Solution:

Solution (continue):

Use **Apriori** to double check if your answer is correct.

Solution:

This lecture:

- Interestingness measures
  - Basic measures for association rules
    - Support
    - Confidence
  - Null-variant measures
    - Chi-square
    - Lift
  - Null-invariant measures
    - AllConf
    - Jaccard
    - Cosine
    - Kulczynski
    - MaxConf
    - Imbalance Ratio

Measure	Definition	Range	Null-Invariant
$\chi^2(A, B)$	$\sum_{i,j=0,1} \frac{(e(a_i b_j) - o(a_i b_j))^2}{e(a_i b_j)}$	$[0, \infty]$	No
$Lift(A, B)$	$\frac{s(A \cup B)}{s(A) \times s(B)}$	$[0, \infty]$	No
$AllConf(A, B)$	$\frac{s(A \cup B)}{max\{s(A), s(B)\}}$	$[0, 1]$	Yes
$Jaccard(A, B)$	$\frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)}$	$[0, 1]$	Yes
$Cosine(A, B)$	$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}}$	$[0, 1]$	Yes
$Kulczynski(A, B)$	$\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)})$	$[0, 1]$	Yes
$MaxConf$(A, B)	$max\{\frac{s(A)}{s(A \cup B)}, \frac{s(B)}{s(A \cup B)}\}$	$[0, 1]$	Yes

X^2 *and lift are not null-invariant*

Jaccard, consine, AllConf, MaxConf, and *Kulczynski* are *null-invariant measures*

max{ s(AUB) / s(A) , s(AUB) / s(B) }

Notes:

Question: Given a student data base of 100 students:

Student ID	Items
001	Coffee, Tea, Algo.
002	Coffee, Tea, Algo., OS
003	Coffee, Algo., OS
004	Coffee, OS, Data Science
005	Tea, Data Science
006-**100**	None from {Coffee, Tea, Algo., OS, Data Science}

Fill in the tables below:

	{Algo.}	No {Algo.}
{Coffee}		
No {Coffee}		

	{Algo., OS}	No {Algo., OS}
{Coffee, Tea}		
No {Coffee ,Tea}		

For each of the following association rules, calculate the measures:

Association rule	**Support & Confidence**	**Lift**	**AllConf**	**Kulc**
{Coffee} →{Algo.}				
{Coffee, Tea} →{Algo., OS}				

- Data cube
  - Dimension, dimension level, dimension value, and cells
  - Basic cells, aggregate cells
  - Basic cuboids, aggregate cuboids
  - Schemas for data cube: Star, Snowflake, Constellation
  - Measures for data cube: Distributive, Algebraic, Holistic
  - Operations in data cube: roll up, drill down, slice and dice, pivot
  - Iceberg cube, iceberg cells
  - Closed cell, closed cube

Notes:

[30] Data warehousing, OLAP, and data cube computation.
Suppose the base cuboid of a data cube contains two cells

- $(a_1, a_2, a_3, a_4, a_5, a_6) : 1$,
- $(a_1, \mathbf{b_2}, a_3, \mathbf{b_4}, a_5, \mathbf{b_6}) : 1$.

where $a_i \neq b_i$ for any dimension $i \in \{2, 4, 6\}$. Assume each dimension contains no concept hierarchy (i.e., has a single level). (Hint: $2^3 = 8$, $2^4 = 16$, $2^5 = 32$, $2^6 = 64$)

(a) [6] How many **nonempty cuboids** are there in this data cube?

(b) [6] How many **nonempty closed cells** are there in this data cube?

(c) [6] How many **nonempty aggregated closed cells** are there in this data cube? What are they?

(d) [6] How many **nonempty aggregated cells** are there in this data cube?

(e) [6] If we set **minimum support = 2**, how many **nonempty aggregated cells** are there in the corresponding **iceberg cube**?