A model to discover rare patterns using ML techniques

DBSE - TeamProject-2020
Supervisor - Sadeq Darrab
Team - 11
[Madhu,Deeksha,Seles, Surabhi, Stanley, Priyam]
OVGU

Agenda

1. Motivation
2. Research Aim
3. Early literature review results
4. Tentative Timeline

Motivation

Fig 1[1]

Fig 2[1]

Rare Patterns & Challenges

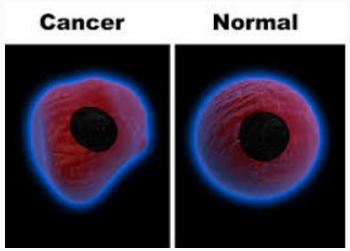


Fig 4[1]

Rare patterns appears in less percentage of data

Hard to generalise

State of the art algorithms miss rare patterns because these rules have less than min support

If we lower the minsupport, number of rules generated explodes

Rare Item sets, more formally...

Definition 2.1 (Rare Itemset). Given a user-specified minsup threshold minsup ∈ [0, 1], X is called a rare itemset or rare pattern in D if sup(X, D) ≤ minsup.[2]

Let I = i1, i2, . . . , im be the universe of items. A set X ⊆ I of items is called an itemset or pattern.

D- Database with set of transactions,|D| is the total number of transactions in D, sup(X,D) of the itemset X in D, and confidence conf(X->Y,D) X and Y are bought together,

$$\mathrm{sup}(X, D) = \frac{\mathrm{count}(X, D)}{|D|},$$

$$\mathrm{conf}(X \to Y, D) = \frac{\mathrm{sup}(XY, D)}{\mathrm{sup}(X, D)}.$$

[2]

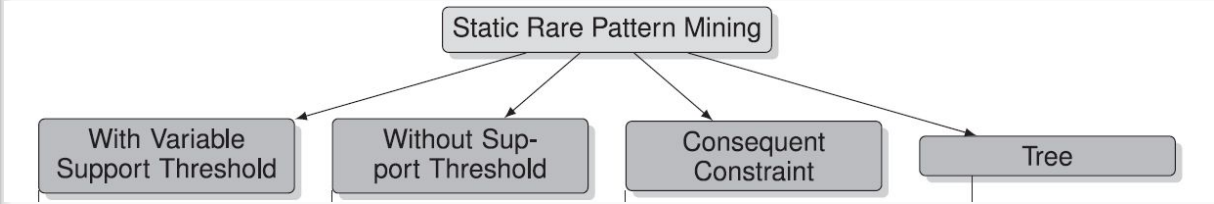
So far >>>>>>>>>

Static- Data is not streamed

Dynamic- Data is coming in continuous streams

Drifting-

What is data distribution is changed from one batch to another in continuous streams?



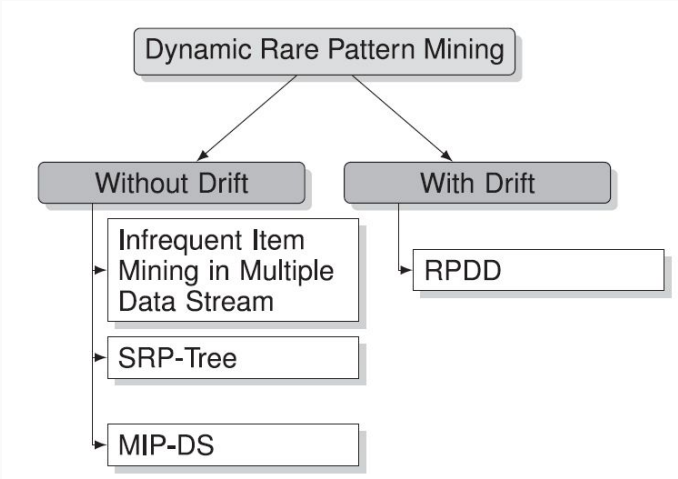
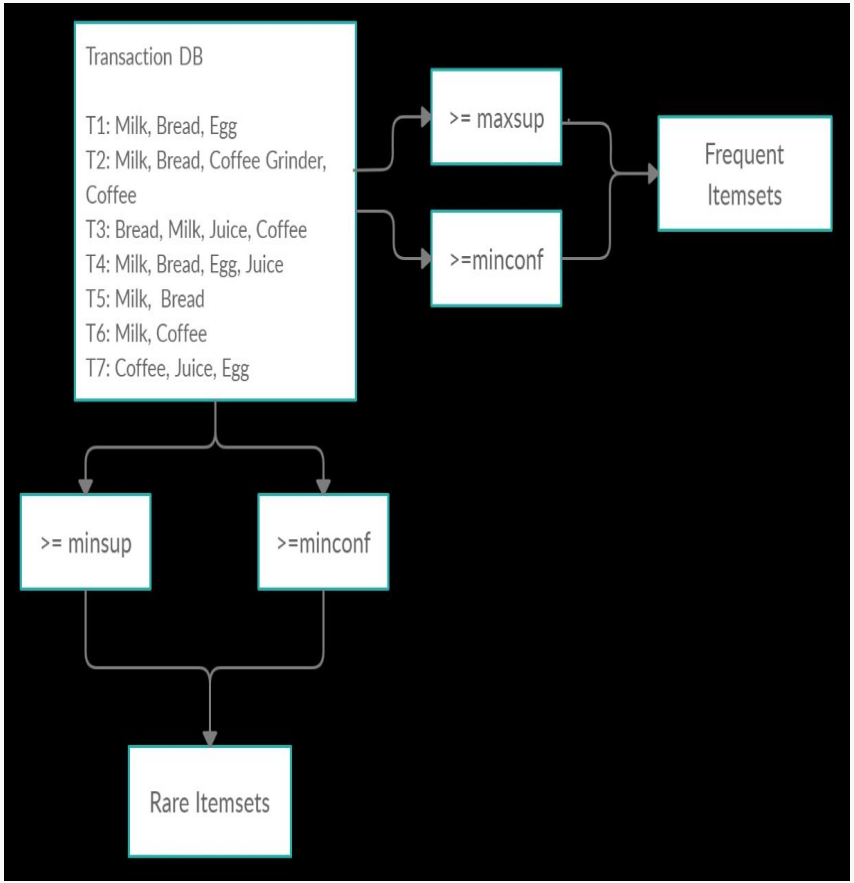


Fig 5,6 [2]

Inverse Apriori[3]



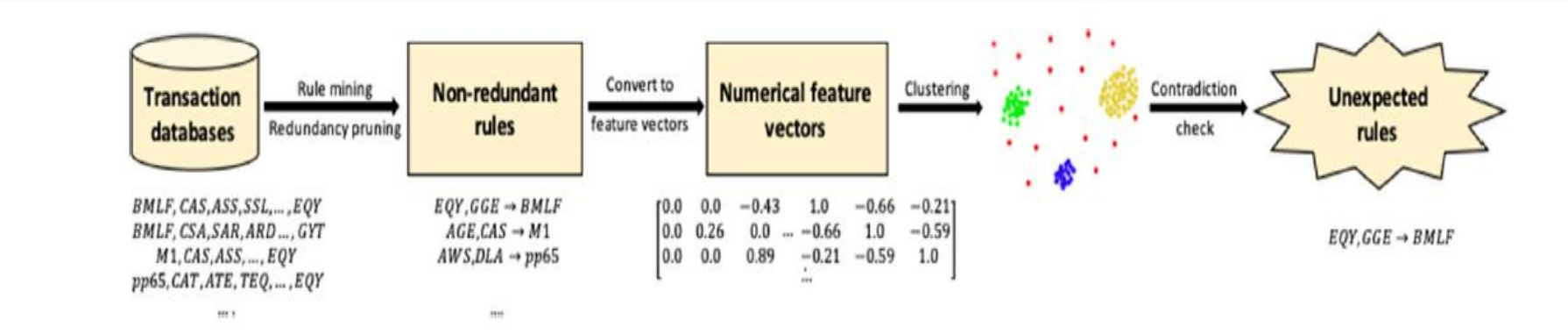
Finds “perfectly rare itemsets”

->Support is lower than maxsup and higher than minsup

Suffers from same drawbacks as Apriori, e.g.

-> Generate too many rules
-> Generate candidates which are not in database

Clustering association rules[4]



Generating unexpected rules

Rules are generated using Apriori then Redundant Rules are pruned

Patterns/Rules are converted into feature vectors to calculate distance between rules

DBSCAN used as clustering technique

RPP[5]

Table 1. A simple dataset.

TID	Items	Ordered items
1	a, b, c, d	b, c, a, d
2	b, d	b, d
3	a, b, c, e	b, c, a, e
4	c, d, e, h	c, d, e
5	a, b, c, g	b, c, a

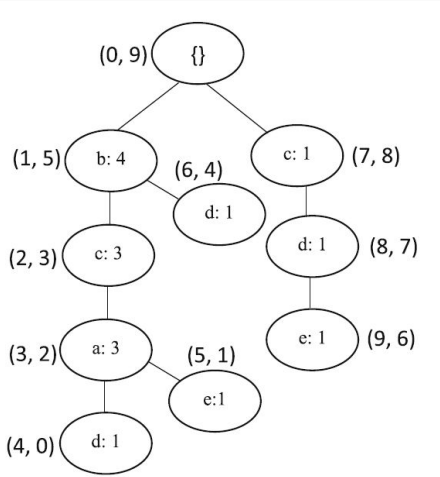


Table 2. RN-lists of interesting rare items.

Item	RPP-codes	Support
b	{(1, 5): 4}	4
c	{(2, 3):3, (7, 8):1}	4
a	{(3, 2):3}	3
d	{(4, 0):1, (6, 4):1, (8, 7):1}	3
e	{(5, 1):1, (9, 6):1}	2

Omits useless candidate itemsets

Tree is Constructed from transactions that contain at least one rare item,{(X-pre-order, X-post-order): count}

Novel DS, RN-List of all interesting rare items

After intersection - {a: 3, e: 2, d: 2, ba:3, bd: 2, ca:3, cd: 2, ce: 2, bca: 3}

Rare items using only a,d,e since sup< max(sup)

[5]

Why it's still an open problem

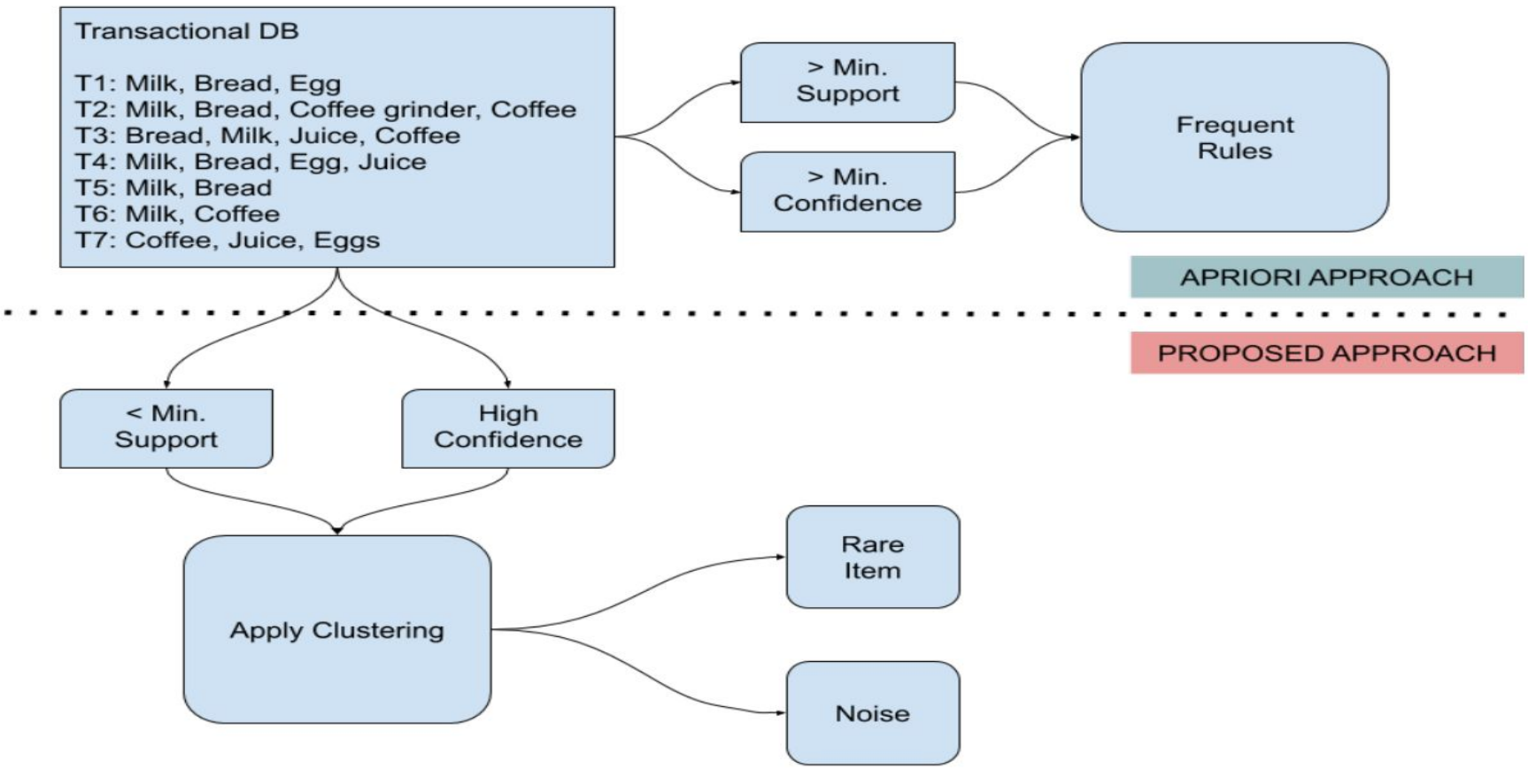
->No single method is guaranteed to work in all situations!

->In case of online continuous streams of data its inefficient to apply existing algorithms.

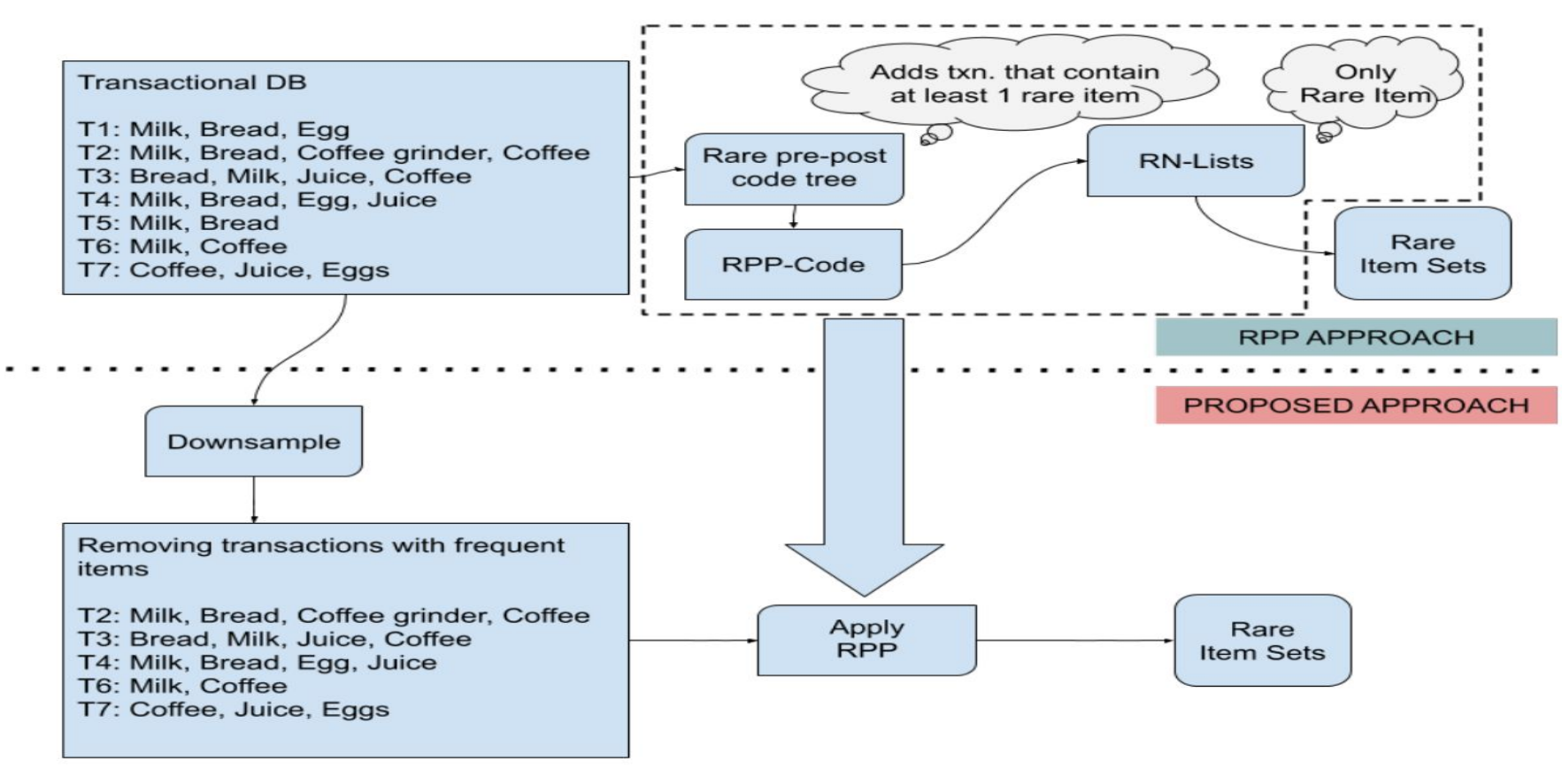
-> Noise or rare? - Rare Rules generated are really meaningful or just random patterns?

-> Scalability and performance - In both static and streaming data

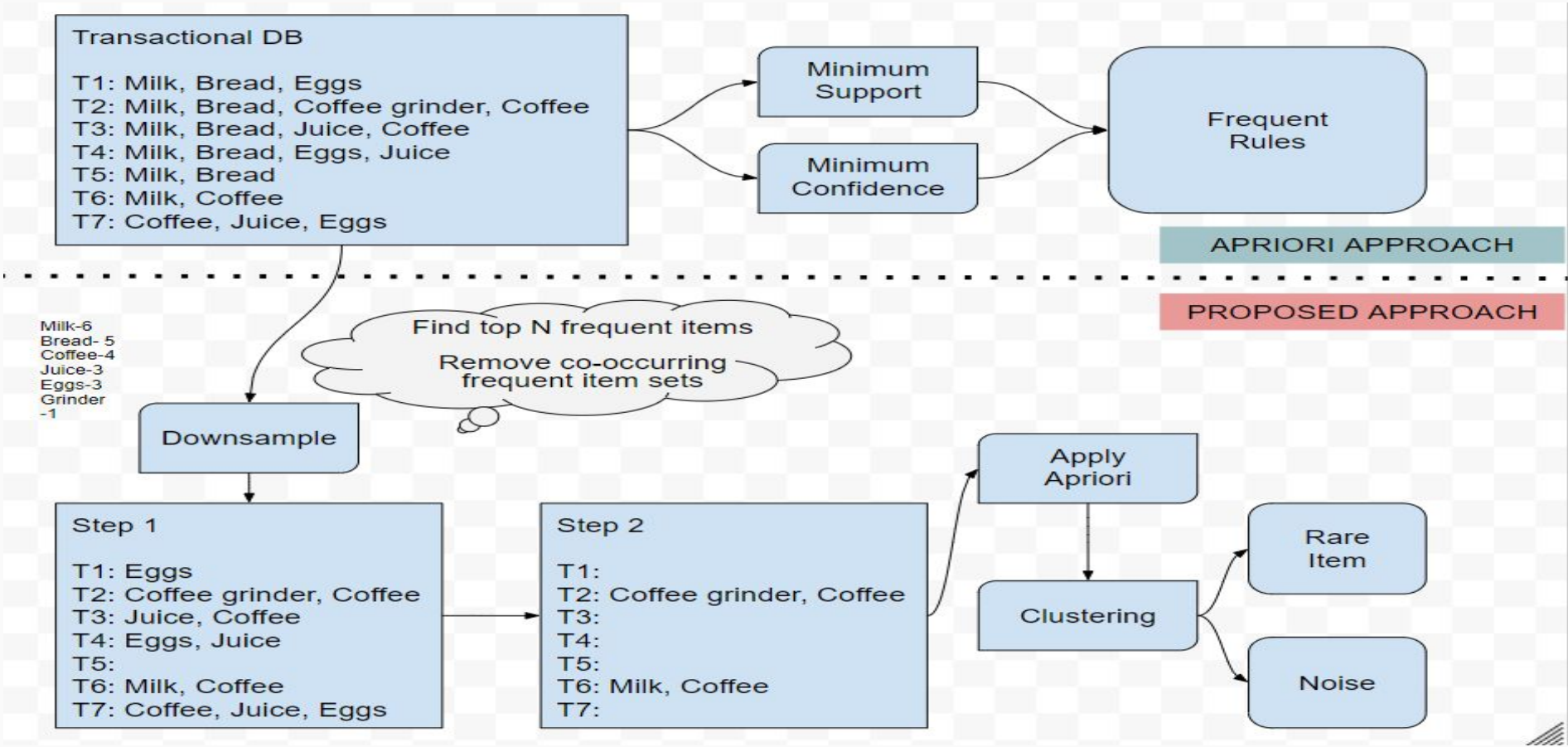
Approach 1- Consider data below min support with high confidence



Approach 2- First downsample then apply RPP[5]



Approach 3 - Only downsampling + Apriori



Timeline


Project Timeline
Nov-20
Dec-20
Jan-21
Feb-21
TASK
W1
W2
W3
W4
W5
W6
W7
W8
W9
W10
W11
W13
W14
W15
Literature Review/ Current Workflow
Conceptualised design of proposed approach
Implementation of approach
Evaluation
Testing and Optimization
Documentation and Report Writing

References-

[1] Fig 1 to 4

www.google.com

[2] Fig , 5,6

Yun Sing Koh and Sri Devi Ravana. 2016. Unsupervised rare pattern mining: A survey. ACM Trans. Knowl. Discov. Data 10, 4, Article 45 (May 2016), 29 pages.
DOI: http://dx.doi.org/10.1145/2898359

[3] Inverse Apriori

Finding Sporadic Rules Using Apriori-Inverse,Yun Sing Koh and Nathan Rountree ,Department of Computer Science, University of Otago, New Zealand {ykoh, rountree}@cs.otago.ac.nz

[4] Clustering association rules
Clustering association rules to build beliefs and discover unexpected patterns,Danh Bui-Thi1 · Pieter Meysman1 · Kris Laukens1,https://doi.org/10.1007/s10489-020-01651-1

[5] RPP RPP Algorithm: A Method for Discovering Interesting Rare Itemsets.Sadeq Darrab(B), David Broneske, and Gunter Saake.University of Magdeburg, Magdeburg, Germany
{sadeq.darrab,david.broneske,gunter.saake}@ovgu.de