Istraživanje podataka 1 - vežbe 11, 2020.

1 Zadaci

1. Nacrtati rešetku skupova stavki koja odgovara skupu podataka datom u tabeli 1.

Id transakcije	Stavke u transakciji
1	$\{Jabuke, Lubenica, Maline\}$
2	$\{Jabuke, Maline\}$
3	$\{Jabuke, Banane\}$
4	$\{Jabuke, Banane\}$
5	$\{Banane\}$
6	$\{Banane, Lubenica, Maline\}$

Tabela 1: Skup podataka sadrži podatke o kupovini namirnica na pijaci voća, jedna transakcija je jedna kupovina jednog kupca.

Označiti čvorove u rešetki sledećim slovima:

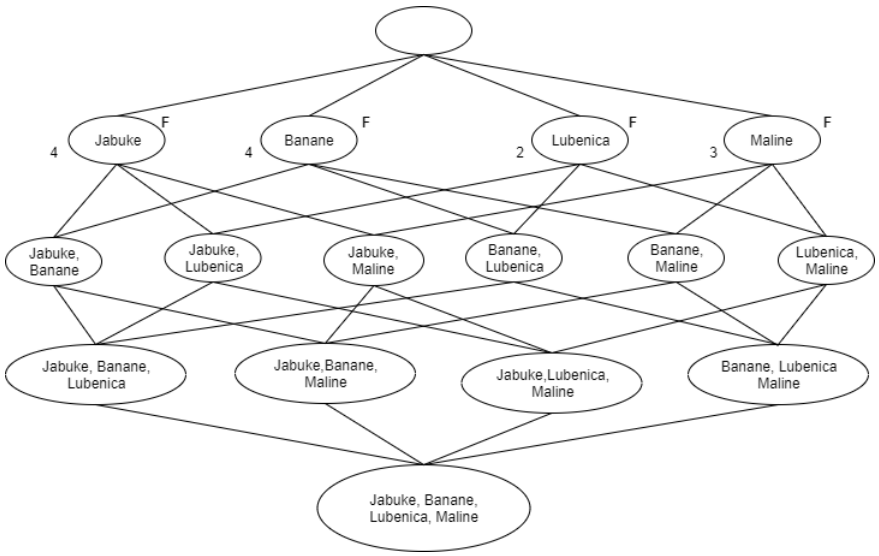
- N: ako se skup stavki ne smatra kandidatom po Apriori algoritmu, tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
- F: kandidat skupa stavki je čest po Apriori algoritmu
- I: Ako se skup stakvi smatra retkim posle određivanja podrške

Minimalna podrška za česte skupove je 0,3.

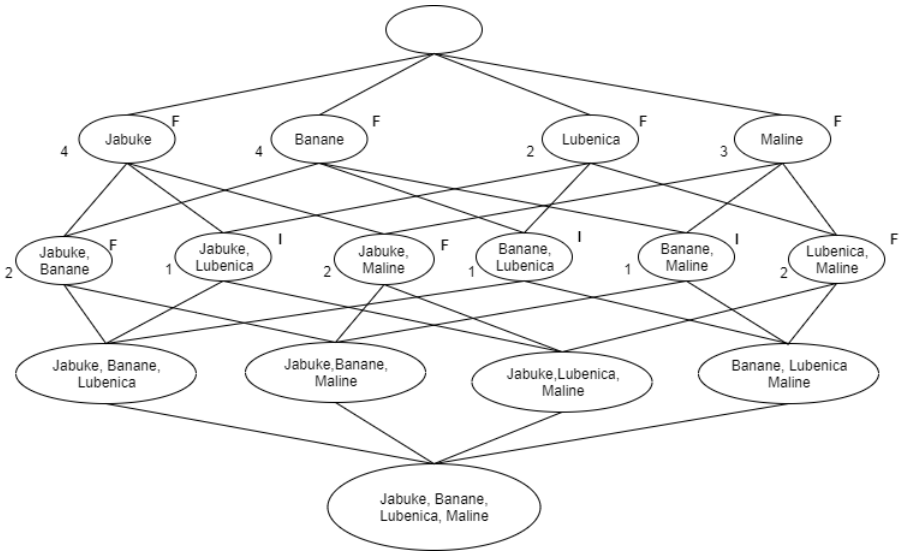
Rešenje

Pošto je $min_{sup} = 0,3$, skupovi stavki koji se pojavljuju u bar 2 transakcije su česti. Pored svakog čvora u rešetkama koje su deo rešenja je u gornjem desnom uglu označeno da li je pridruženi skup podataka čest (F), nije čest ali je bilo neophodno prebrojavanje u koliko transakcija se javlja jer su mu svi podskupovi česti (I), ili nije čest i nije bilo potrebno računati u koliko transakcija se javlja jer mu bar jedan podskup nije čest (N). U donjem levom uglu čvora je ispisan broj transakcija u kojima se javlja pridruženi skup stavki ukoliko je bilo potrebno izračunati tu vrednost.

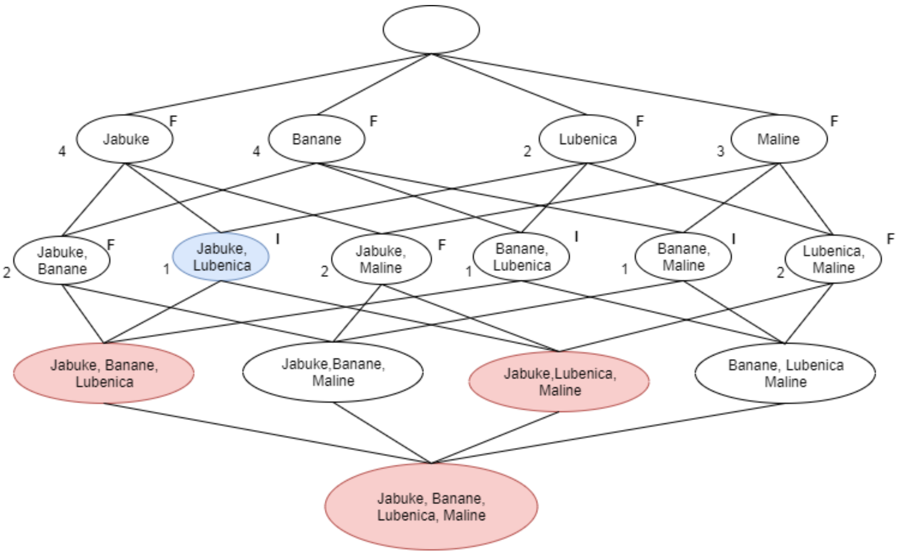
Prvo se za 1-skupove stavki prebrojava u koliko transakcija se pojavljuju i za svaki 1-skup stavki određuje da li je F ili I.

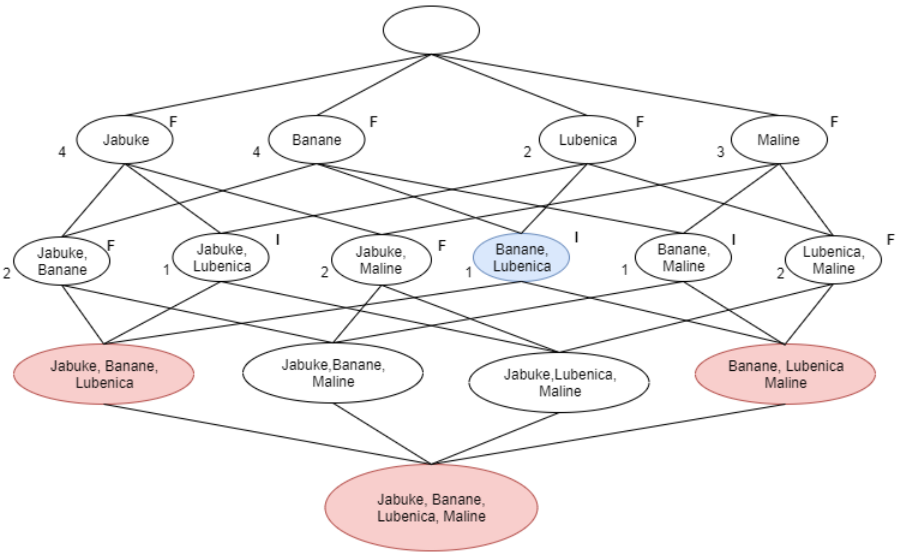


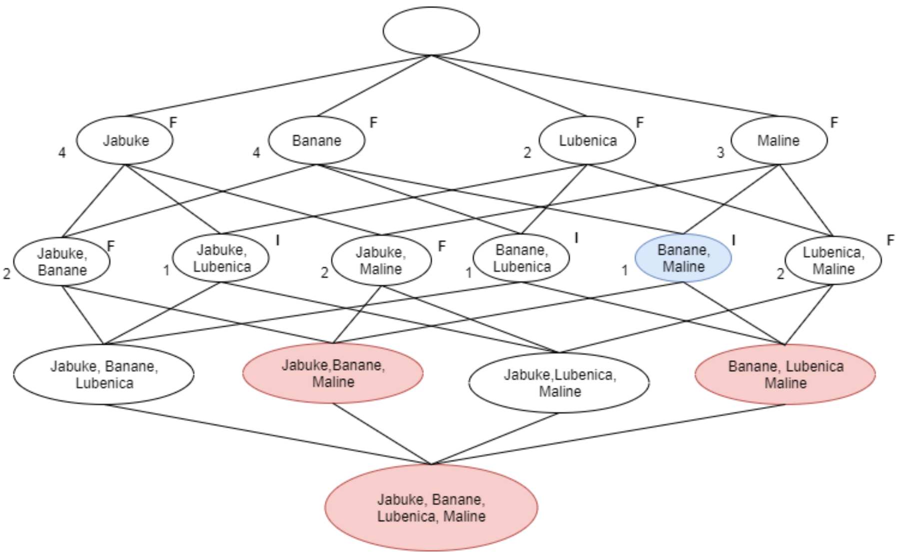
Pošto su svi 1-skupovi stavki česti, za svaki 2-skup stavki se mora izračunati broj transakcija u kojima se javlja da bi se odredilo da li je skup čest ili ne.



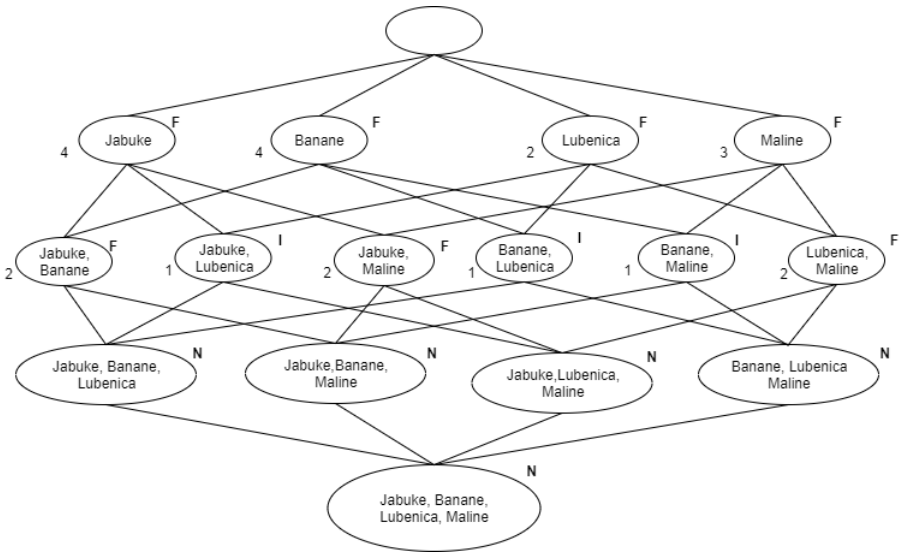
Za 2-skupove stavki koji nisu česti (označeni sa I) mogu se odmah pronaći svi nadskupovi i označiti kao retki, bez računanja broja transakcija u kojima se javljaju. Na narednim rešetkama je po jedan nečest 2-skup stavki označen plavom bojom, a crvenom bojom njegovi nadskupovi za koje se zna da nisu česti bez računanja broja transakcija u kojima se javljaju i koji mogu da se označe sa N.







Pošto svaki 3-skup stavki ima podskup koji nije čest, ni on ne može biti čest. Isto važi i za 4-skup stavki. Rešetka sa pridruženim oznakama za svaki čvor izgleda:



2. Nacrtati rešetku skupova stavki koja odgovara datom skupu podataka na slici. Označiti čvorove u rešetki sledećim slovima:

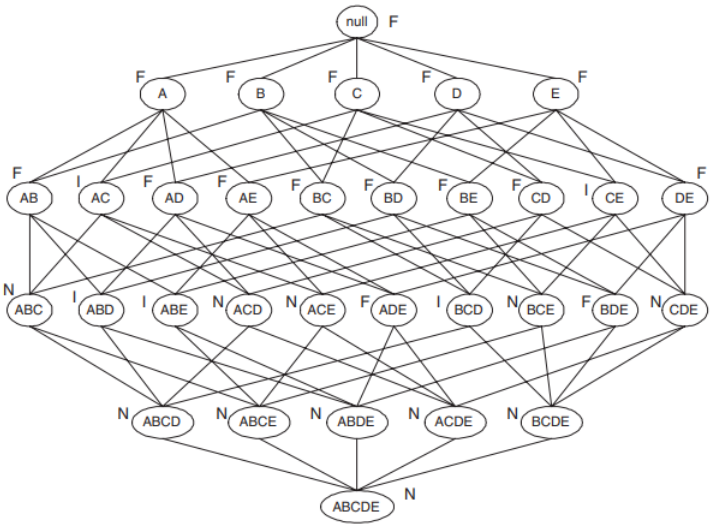
   - N: ako se skup stavki ne smatra kandidatom po Apriori algoritmu. Tj. ako 1) skup stavki nije generisan u koraku generisanja kandidata, ili 2) je generisan u koraku generisanja kandidata ali je kasnije uklonjen u koraku čišćenja kandidata jer neki njegov podskup nije čest.
   - F: kandidat skupa stavki je čest po Apriori algoritmu.
   - I: Ako se skup stakvi smatra retkim posle određivanja podrške.

Transaction ID	Items Bought
1	$\{a,b,d,e\}$
2	$\{b,c,d\}$
3	$\{a,b,d,e\}$
4	$\{a,c,d,e\}$
5	$\{b,c,d,e\}$
6	$\{b,d,e\}$
7	$\{c,d\}$
8	$\{a,b,c\}$
9	$\{a,d,e\}$
10	$\{b,d\}$

   Minimalna podrška za česte skupove je 0,3.

   - Koliki je procenat čestih skupova stavki?
   - Koliki je odnos čišćenja Apriori algoritma za ovaj skup podataka? Odnos čišćenja je definisan kao procenat skupova stavki koji nisu generisani za vreme generisanja kandidata ili su eliminisani u koraku čišćenja kandidata.
   - Koliki je odnos *lažnog alarma* (procenat kandidatskih skupova stavki koji su obeleženi kao retki posle prebrojavanja podrške)?

Rešenje



Odgovori:

- $\frac{16}{32} = 50\%$
- $\frac{11}{32} = 34,4\%$
- $\frac{5}{32} = 15,6\%$