IF5181 Pengenalan Pola

Mining Frequent Pattern

Masayu Leylia Khodra

Slide diambil dari kuliah Yudi Wibisono & Han dkk. (2011)

Referensi

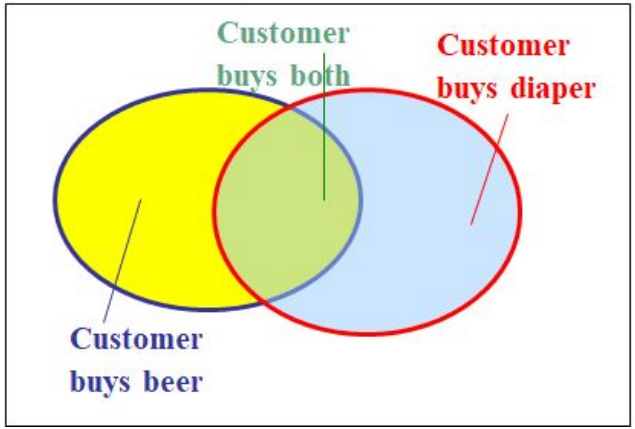
- Bab 6 & 7 dari Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- Yudi Wibisono. Mining Association Rule
  https://docs.google.com/presentation/d/1UimlWVRkI6_OFgXYjBua77Q2IqK1AjWcUA1lQhx65qs/edit?usp=sharing

Frequent Pattern

- Pattern yang sering muncul.
  - Pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set (Han dkk, 2011)
- Frequent itemset: selai+roti, gula+telor
- Frequent sequential pattern: beli PC lalu kamera lalu mem card
- Frequent structured pattern: subgraphs / subtrees

Frequent Itemset

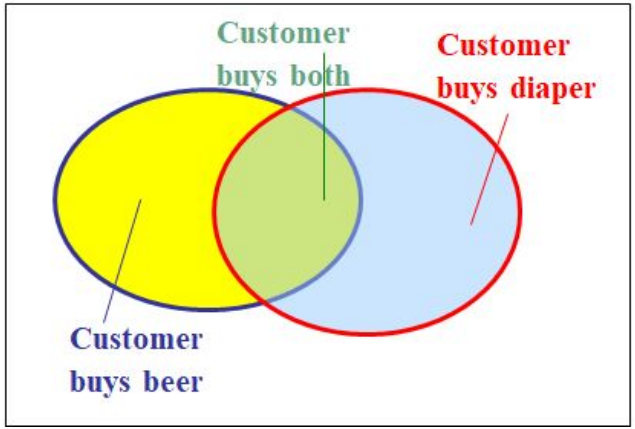
Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- An itemset X is *frequent* if X's support is no less than a *minsup* threshold

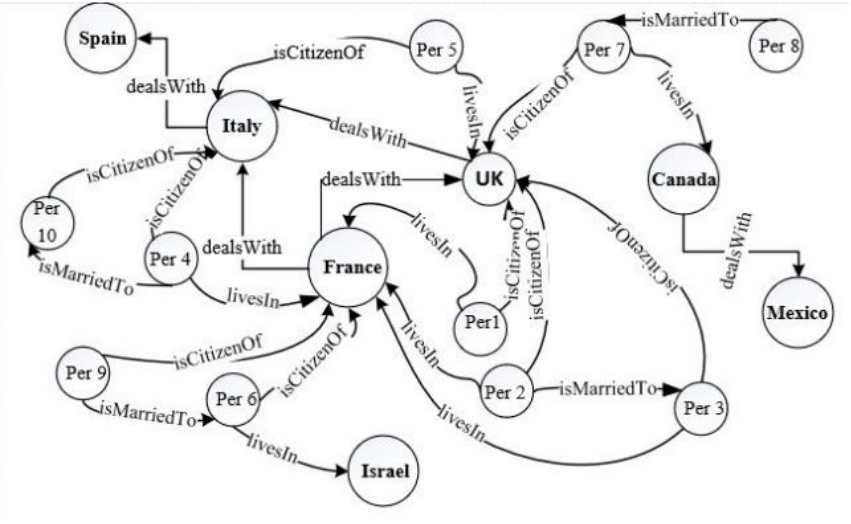
Frequent Itemset (lanj)

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



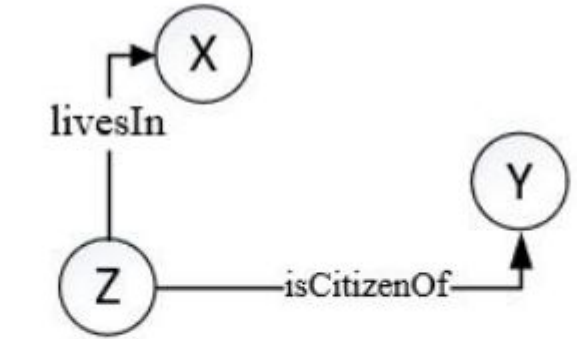
- *(absolute) support*, or, *support count* of X: Frequency or occurrence of an itemset X
- *(relative) support*, *s*, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- Frequent Pattern (minsup=3) :
  - Beer:3,
  - Nuts:3,
  - Diaper:4,
  - Eggs:3,
  - {Beer, Diaper}:3

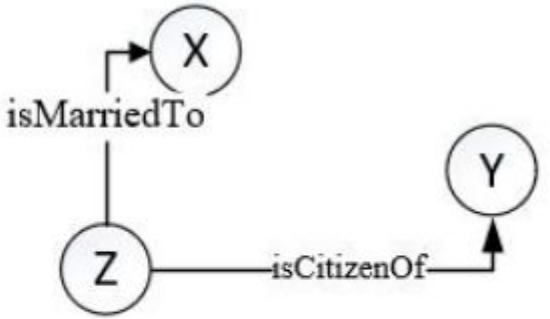
Frequent (Structured) Pattern



G

Figure 1: Graf Properti Yago KB





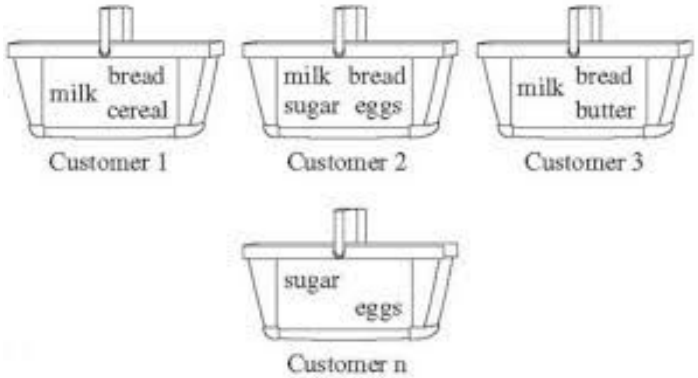
Frequent Pattern Analysis

- Motivasi: Menemukan pola yang menarik pada data
  - What products were often purchased together?
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?

https://medium.com/@niharika.goel/market-basket-analysis-association-rules-e7c27b377bd8

Frequent Pattern Analysis: Aplikasi

- Market Basket Analysis
  Jika customer membeli susu,
  seberapa mungkin ia juga membeli roti



- Aplikasi lain:
  - catalog design, sale campaign analysis, Web log (click stream) analysis, dan DNA sequence analysis.
- Praproses task mining yang lain:
  - Cluster, klasifikasi, semantic data compression

Support dan Confidence

- Besaran “kemenarikan” (interestingness) dari sebuah pola
- Hanya pola yang melewati nilai dan support tertentu saja yang diperhitungkan.
- Contoh:
  - Beli Sabun → Beli Pasta gigi
    (Support: 60%, Confidence 70%)

Support dan Confidence: Contoh

- A → B (support 50%, Confidence 75%)
  - Support 50% : 50% transaksi, A dan B dibeli bersama
  - Confidence 75%: 75% transaksi saat seseorang membeli A, dia juga membeli B

Support dan Confidence: Contoh

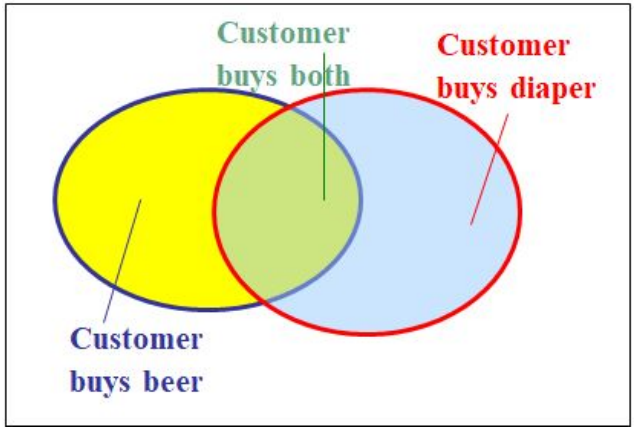
Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, P(A) > 0$$

- $A \rightarrow D$ (support:60%, confidence:100%)
- $D \rightarrow A$ (support:60%, confidence:75%)
- Support: A dan D muncul 3 dari 5 transaksi.
- Conf A→D: 3 kali beli A, 3 kali mengandung D.
- Conf D→A: 4 kali beli D, hanya 3 yg mengandung A

Association Rule

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Association Rule: bentuk aturan item muncul bersama
- Itemset X = {x1, …, xk}
- Temukan rules $X \rightarrow Y$ dengan min support and confidence
  - support, *s*, probabilitas transaksi mengandung $X \cup Y$
  - confidence, *c,* conditional probability transaksi memiliki X juga mengandung *Y*

Apriori for Boolean Association Rule

- Tujuan: finding frequent itemset
- Pendekatan Generate and Test
- Apriori property: All nonempty subsets of a frequent itemset must also be frequent
- Metode: Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k + 1)-itemsets.

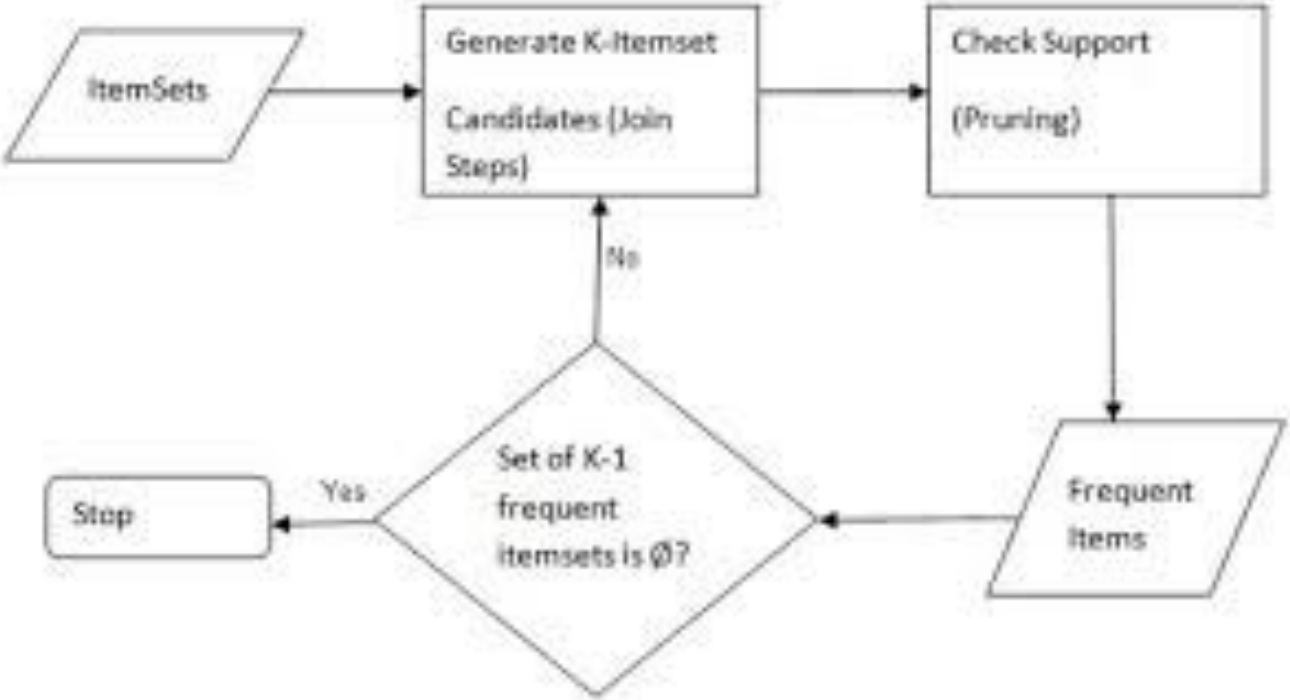
Apriori: Two-step Process

- Joint-step: To find L_k, a set of candidate k-itemsets is generated by joining L_{k-1} with itself. This set of candidates is denoted C_k.
  - k=1: scan DB untuk mendapat 1-item set (C1)
  - k>1: bangkitkan k+1 itemset dari itemset yang ada sebelumnya (C_{k+1}).

Apriori: Two-step Process

- Prune-step: C_k is a superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k.
  - A database scan to determine the count of each candidate in C_k would result in the determination of L_k (i.e., all candidates having a count no less than the minimum support count are frequent by definition, and therefore belong to L_k).
  - Hitung support elemen kandidat C_k, filter kandidat dengan frekuensi (support) memenuhi minsupport.

Apriori


ItemSets
Generate K-Itemset
Candidates (Join Steps)
Check Support
(Pruning)
No
Set of K-1 frequent itemsets is Ø?
Yes
Stop
Frequent Items

Apriori: Contoh

$Sup_{min} = 2$

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

C_1

1st scan

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

Contoh (lanj) Hitung Confidence

- Confidence (A ➔ B) = P(B|A) = support_count(A U B) / support_count(A)
- Itemset yang dihasilkan
  - 1 elemen: {A}, {B}, {C}, {E}
  - 2 elemen: {AC}, {BC}, {BE}, {CE}
  - 3 elemen : {BCE}

    E → B : conf = (3 / 3 = 100%)

    B C → E : conf = (2 / 2 =100 %)

    B E → C : conf = (2 / 3 = 66.6%)

    C E → B

    B → C E dst...

Generate Association Rule

- Strong association rules satisfy both minimum support and minimum confidence.
- Generate Association Rule:
  - For each frequent itemset l, generate all nonempty subsets of l.
  - For every nonempty subset s of l, output the rule “s→(l-s) ” if conf(s→(l-s))>= min_conf, where min_conf is the minimum confidence threshold.

Generate Association Rule: Contoh

- Itemset yang dihasilkan
  - 1 elemen: {A}, {B}, {C}, {E}
  - 2 elemen: {AC}, {BC}, {BE}, {CE}
  - 3 elemen : {BCE}
- Generate rule:

$E \rightarrow B$: conf = (3 / 3 = 100%)

$B, C \rightarrow E$: conf = (2 / 2 =100 %)

$B, E \rightarrow C$: conf = (2 / 3 = 66.6%)

$C, E \rightarrow B$

$B \rightarrow C, E$ dst...

Meningkatkan Efisiensi Metode A Priori

- Teknik Hash-based, mengurangi jumlah itemset.
- Transaction Reduction
- Partisi data
- Sampling
- Dynamic itemset counting

Mining Freq. Itemset tanpa mengenerate Kandidat

- Masalah apriori:
  - Dapat menghasilkan jumlah kandidat yang sangat besar.
  - Harus menscan database berulang kali dengan pattern matching.
  - Metode tanpa generate kandidat: Frequent Pattern growth → FP-growth
    - Teknik: compress database ke dalam FP-tree, bagi dalam conditional database, dan mine secara terpisah.

Buat FP-tree

TID	*Item yang dibeli*	*(ordered) frequent items*
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

min_support = 3

F-list = f-c-a-b-m-p

- Scan DB untuk mencari frekuensi
- Sort dan dijadikan F-List

Item	*frequency*
f	4
c	4
a	3
b	3
m	3
p	3

Buat FP-Tree (lanj):

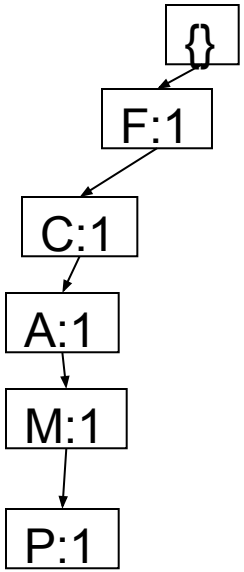
Atur ulang item (berdasarkan Flist)

TID	*Items bought*	*(ord) frequent items*
100	{*f, a, c, d, g, i, m, p*}	{*f, c, a, m, p*}
200	{*a, b, c, f, l, m, o*}	{*f, c, a, b, m*}
300	{*b, f, h, j, o, w*}	{*f, b*}
400	{*b, c, k, s, p*}	{*c, b, p*}
500	{*a, f, c, e, l, p, m, n*}	{*f, c, a, m, p*}

F-list = f-c-a-b-m-p

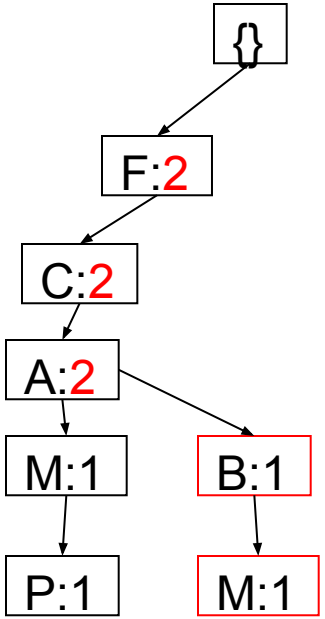
Pembentukan Tree (1)

100 **{*f, a, c, d, g, i, m, p*}** **{*f, c, a, m, p*}**



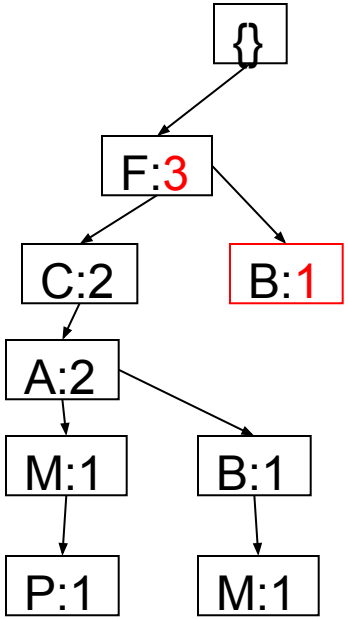
Pembentukan Tree (2)

200 ***{a, b, c, f, l, m, o}*** ***{f, c, a, b, m}***



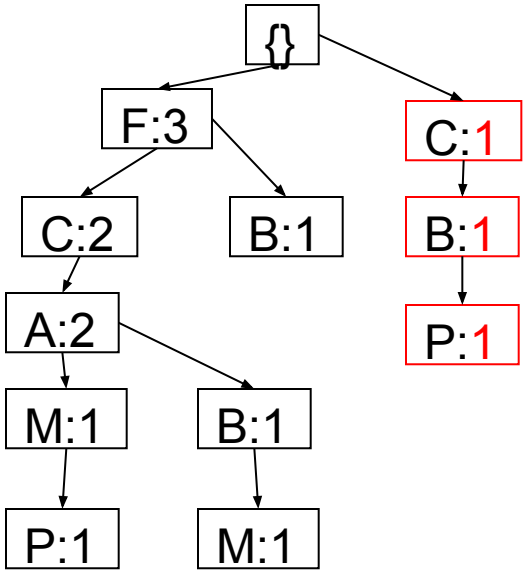
Pembentukan Tree(3)

300 ***{b, f, h, j, o, w}*** ***{f, b}***



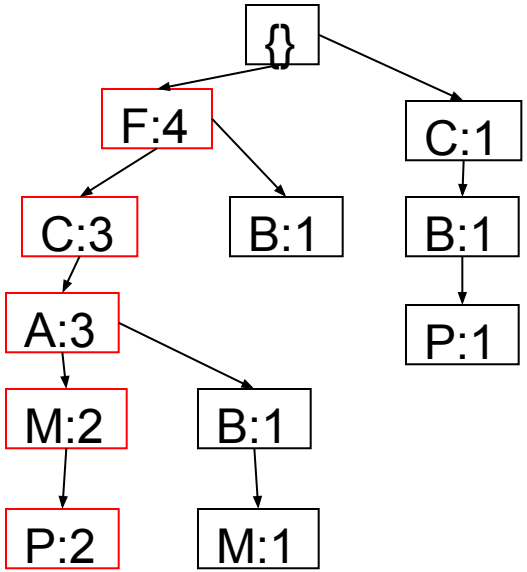
Pembentukan Tree (4)

400 **{*b, c, k, s, p*}** **{*c, b, p*}**

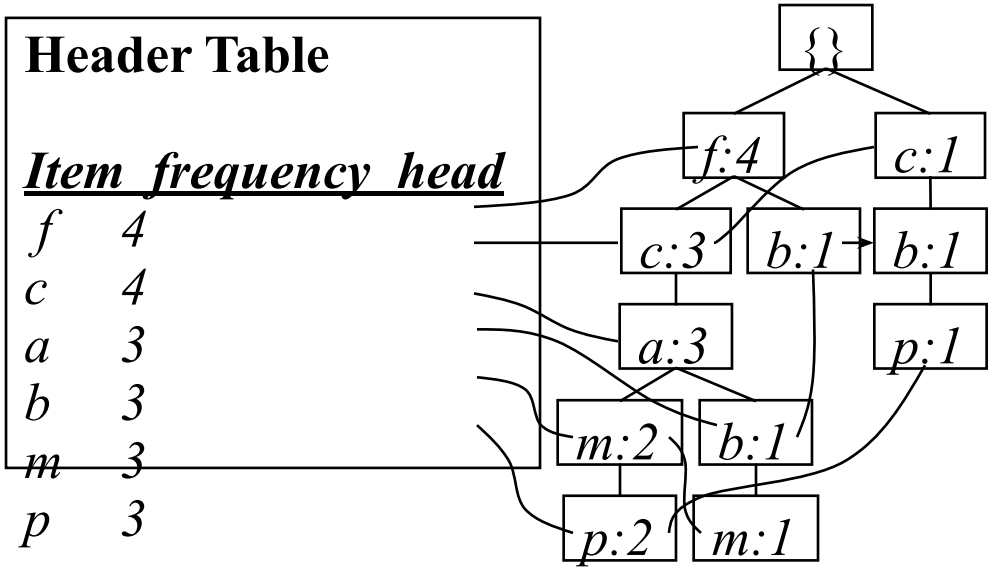


Pembentukan Tree (5)

500 **{*a, f, c, e, l, p, m, n*}** **{*f, c, a, m, p*}**

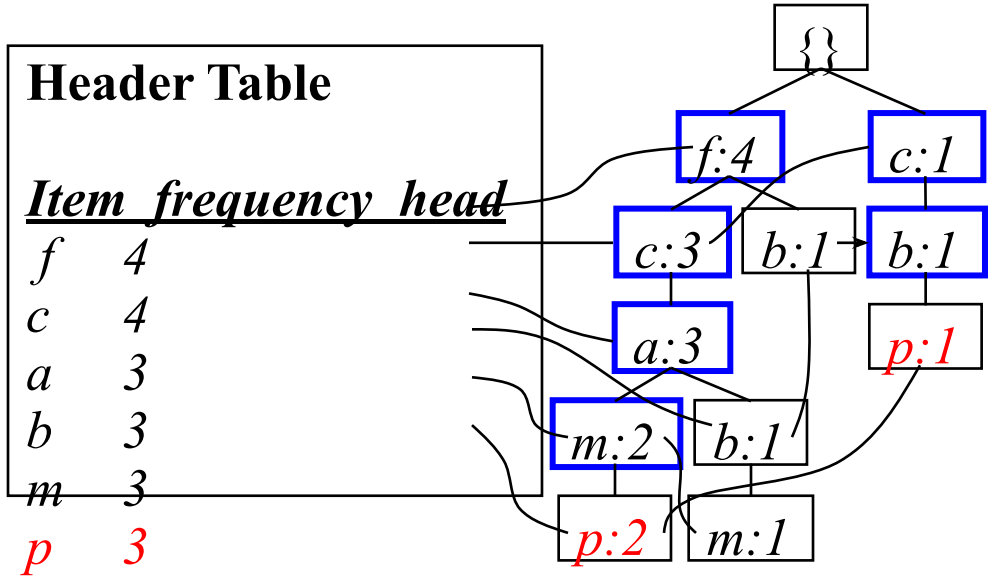


Item Header Table

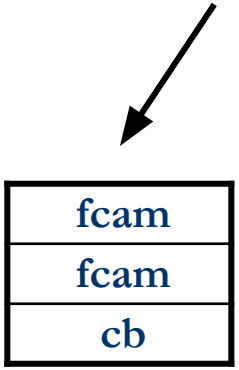


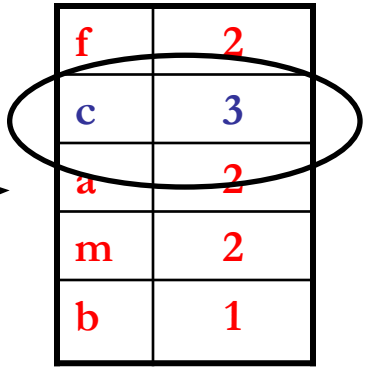
Conditional Pattern (1)

- Mulai dari freq yg paling rendah: p



p: fcam:2, cb:1





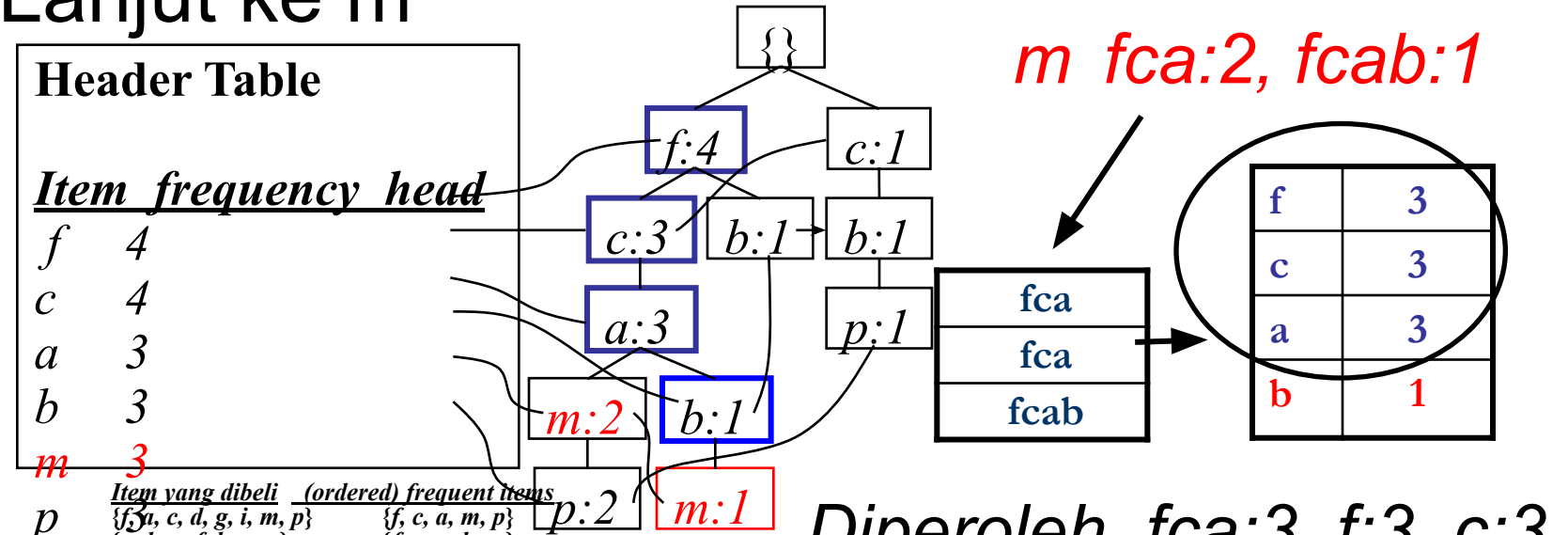
TID	Item yang dibeli	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Diperoleh, c:3

(yg lain tdk lolos support)

Conditional Pattern (2)

- Lanjut ke m



TID	Item yang dibeli	(ordered) frequent items
100	{f, a, c, d, g, i, m, p}	{f, c, a, m, p}
200	{a, b, c, f, l, m, o}	{f, c, a, b, m}
300	{b, f, h, j, o, w}	{f, b}
400	{b, c, k, s, p}	{c, b, p}
500	{a, f, c, e, l, p, m, n}	{f, c, a, m, p}

Diperoleh, fca:3. f:3, c:3, a:3 (yg lain tdk lolos support)

Diperoleh set yang paling banyak muncul:
fca, f, c, a

fca → f : conf = (..)

fca → c : conf = (..)

fca → a : conf = (..)

f → a dst...

Latihan FP-Tree

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

Support: 50%
Conf: 50%

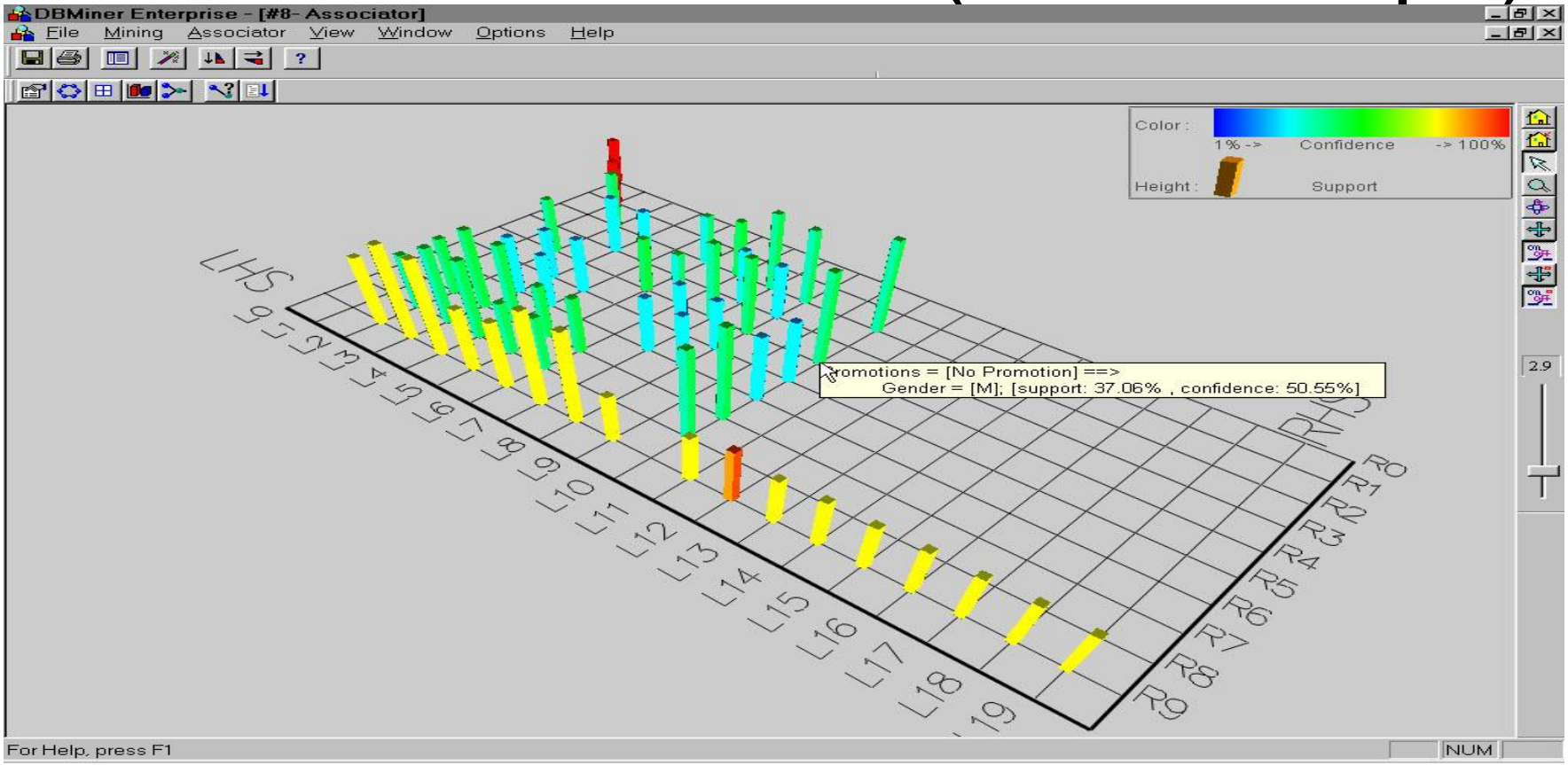
Latihan

***Exercise 1.* Apriori**

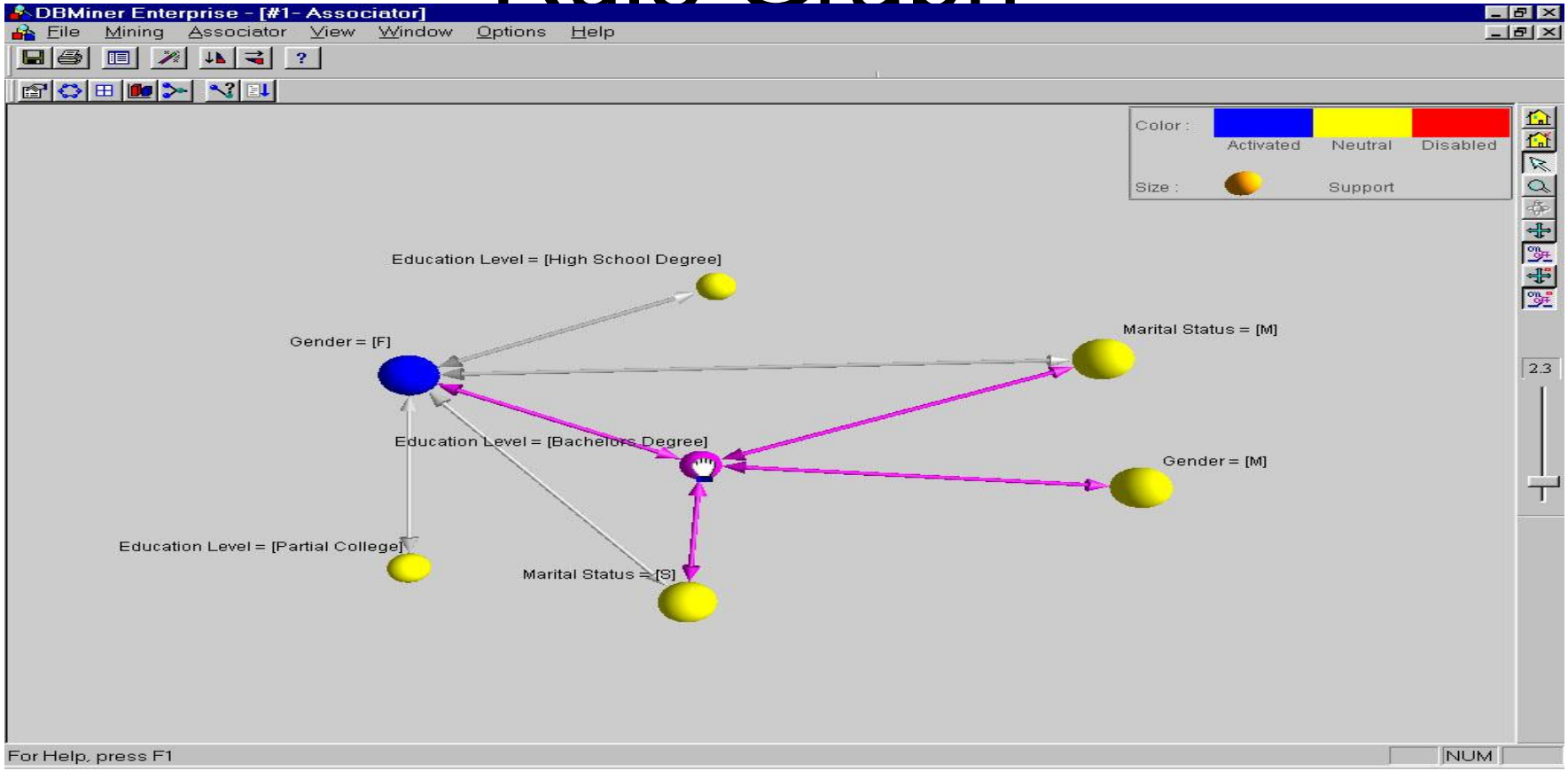
Trace the results of using the Apriori algorithm on the grocery store example with support threshold s=33.34% and confidence threshold c=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence.

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

Visualisasi Ass. Rule (Plane Graph)



Rule Graph


DBMiner Enterprise - [#1- Associator]
File Mining Associator View Window Options Help
Color :
Activated
Neutral
Disabled
Size :
Support
Education Level = [High School Degree]
Gender = [F]
Marital Status = [M]
Education Level = [Bachelors Degree]
Gender = [M]
Education Level = [Partial College]
Marital Status = [S]
2.3
For Help, press F1
NUM

Mining Berbagai Ass. Rule

- Multilevel
- Multidimensi
- Quantitative
- interesting correlation patterns