


Pattern Mining: Konsep dan Metode Dasar

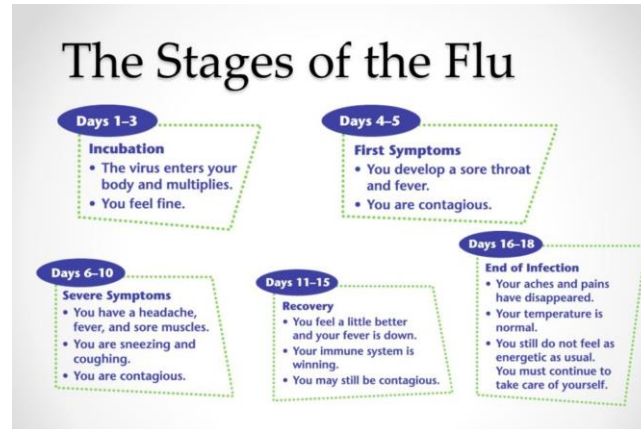
- ☐ Konsep dasar 
- ☐ Metode Frequent Itemset Mining
- ☐ Metode Evaluasi Pola
- ☐ Simpulan

Apa itu Pattern?

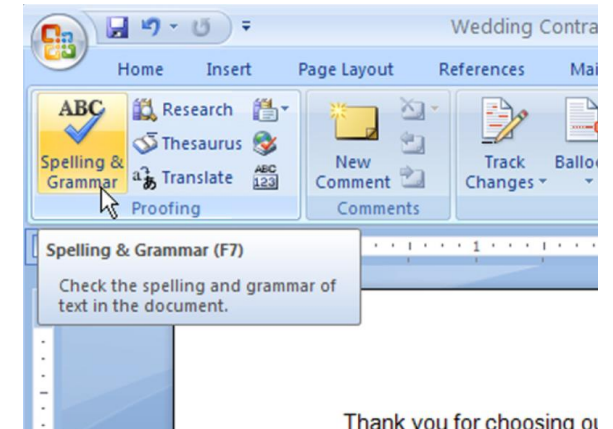
- ❑ **Patterns:** Sebuah kumpulan item, subsekuens, atau substruktur yang sering muncul bersamaan (atau memiliki korelasi yang kuat) dalam suatu dataset
- ❑ Pattern mewakili sifat-sifat intrinsik dan penting dari dataset.



Frequent item set



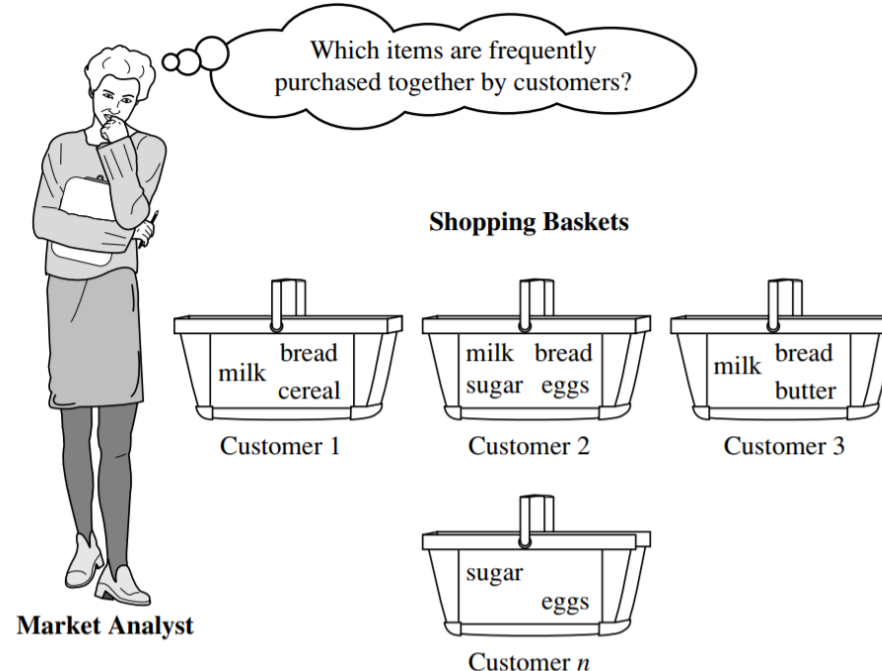
Frequent sequences



Frequent structures

Apa itu **Pattern discovery** ?

- ❑ **Pattern discovery**: Mengungkap pola dari kumpulan data besar
- ❑ Ini dapat menjawab pertanyaan seperti:
 - ❑ Produk apa yang sering dibeli bersama?
 - ❑ Apa pembelian selanjutnya setelah membeli iPad?



Pattern Discovery: Mengapa Ini Penting?

- ❑ **Fondasi** untuk tugas data mining penting
 - ❑ Association, correlation, dan causality analysis
 - ❑ Mining **sequential**, Pola Struktural (misalnya, sub-grafik)
 - ❑ **Classification**: Analisis berbasis pola diskriminatif
 - ❑ **Cluster** analysis: Pengklusteran subruang berbasis pola
- ❑ Aplikasi yang lebih Luas
 - ❑ Market basket analysis, cross-marketing, catalog design, sale campaign analysis, Web log analysis, biological sequence analysis
 - ❑ Berbagai Jenis data: spatiotemporal, multimedia, time-series, and stream data

Konsep dasar: Database Transaksional

- ❑ Transactional Database (TDB)
 - ❑ Setiap transaksi dikaitkan dengan pengidentifikasi, yang disebut Transaction ID (TID).
 - ❑ Mungkin juga memiliki jumlah yang terkait dengan setiap barang yang terjual

Tid	Barang yang dibeli
1	Bir, kacang, popok
2	Bir, Kopi, Popok
3	Bir, Popok, Telur
4	Kacang, Telur, Susu
5	Kacang, Kopi, Popok, Telur, Susu

Konsep dasar: k-Itemsets dan Supports

- Itemset: satu atau beberapa item

$$I = \{I_1, I_2, \dots, I_m\}$$

- k-itemset: Kumpulan item yang berisi k item :

$$X = \{x_1, \dots, x_k\}$$

- Contoh. {Bir, Kacang, Popok} adalah 3 item set

- Absolute support (count)

- $\text{sup}\{X\}$ = kemunculan Kumpulan item X

- Ex. $\text{sup}\{\text{Bir}\} = 3$
- Ex. $\text{sup}\{\text{Popok}\} = 4$
- Ex. $\text{sup}\{\text{Bir, Popok}\} = 3$
- Ex. $\text{sup}\{\text{Bir, Telur}\} = 1$

Tid	Barang yang dibeli
1	Bir, kacang, popok
2	Bir, Kopi, Popok
3	Bir, Popok, Telur
4	Kacang, Telur, Susu
5	Kacang, Kopi, Popok, Telur, Susu

- Relative support

- $s\{X\}$ = Bagian Transaksi yang berisi X (yaitu, probabilitas bahwa transaksi berisi X)
- Ex. $s\{\text{Bir}\} = 3/5 = 60\%$
- Ex. $s\{\text{Popok}\} = 4/5 = 80\%$
- Ex. $s\{\text{Bir, Telur}\} = 1/5 = 20\%$

Konsep dasar : Frequent Itemsets (Patterns)

- Sebuah itemset(atau Pattern) X dikatakan sering (Frequent) jika support X tidak kurang dari *minsup* threshold σ

- $\sigma = 50\%$ (σ : *minsup* threshold)
Untuk 5 transaksi yang diberikan

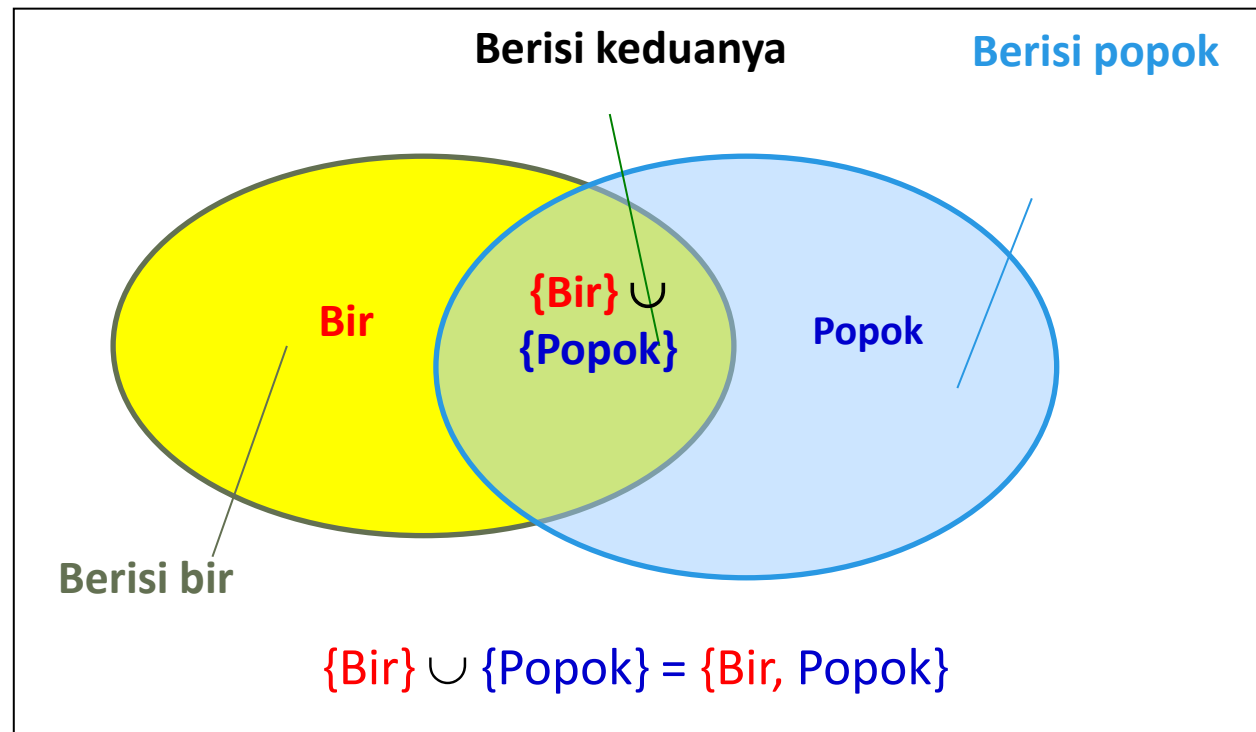


Tid	Barang yang dibeli
1	Bir, kacang, popok
2	Bir, Kopi, Popok
3	Bir, Popok, Telur
4	Kacang, Telur, Susu
5	Kacang, Kopi, Popok, Telur, Susu

- frequent 1-itemsets:
 - Bir: 3/5 (60%); Kacang: 3/5 (60%);
Popok: 4/5 (80%); Telur: 3/5 (60%)
- frequent 2-itemsets:
 - {Bir, Popok}: 3/5 (60%)
- frequent 3-itemsets?
 - None

Frequent Itemsets ke Association Rules

- Dibandingkan dengan itemset, association rules bisa lebih jelas
 - misal. Popok \rightarrow Bir
 - *Membeli popok kemungkinan besar dapat menyebabkan pembelian bir*



Association Rules

- ❑ Bagaimana kita menghitung strength dari association rule $X \rightarrow Y$ (X dan Y adalah itemsets)?
- ❑ Pertama-tama kita menghitung dua metrik berikut, s dan c.
 - ❑ **Support $X \cup Y$**
 - ❑ Ex. $s\{\text{Popok, Bir}\} = 3/5 = 0.6$ (60%)
 - ❑ **Confidence $X \rightarrow Y$**
 - ❑ **conditional probability** bahwa transaksi yang berisi X juga berisi Y:
 - ❑ $c = \text{sup}(X, Y) / \text{sup}(X)$
 - ❑ Ex. $c = \text{sup}\{\text{Popok, Bir}\} / \text{sup}\{\text{Popok}\} = 3/4 = 0.75$

Tid	Barang yang dibeli
1	Bir, kacang, popok
2	Bir, Kopi, Popok
3	Bir, Popok, Telur
4	Kacang, Telur, Susu
5	Kacang, Kopi, Popok, Telur, Susu

- ❑ Dalam analisis pola, kita sering tertarik pada aturan yang mendominasi database, dan kedua metrik ini memastikan popularitas dan korelasi X dan Y.

Mining Frequent Itemsets dan Association Rules

□ Association rule mining

- Diberikan dua thresholds: (*minimum support*) *minsup*, (*minimum confidence*) *minconf*
- Temukan semua rules, $X \rightarrow Y (s, c)$ sedemikian rupa sehingga $s \geq \text{minsup}$ dan $c \geq \text{minconf}$

□ *minsup* = 50%

- Freq. 1-itemsets: Bir: 3, Kacang: 3, Popok: 4, Telur: 3
- Freq. 2-itemsets: {Bir, Popok}: 3

□ *minconf* = 50%

- $Bir \rightarrow Popok$ (60%, 100%)
- $Popok \rightarrow Bir$ (60%, 75%)

Tid	Barang yang dibeli
1	Bir, kacang, popok
2	Bir, Kopi, Popok
3	Bir, Popok, Telur
4	Kacang, Telur, Susu
5	Kacang, Kopi, Popok, Telur, Susu



- Mining association rules dan mining frequent patterns berkaitan erat

Challenge: Ada banyak Frequent Patterns!

- Berapa banyak frequent itemset terdapat dalam TDB_1 berikut (minsup = 1)?

- TDB_1 : $T_1: \{a_1, \dots, a_{50}\}; T_2: \{a_1, \dots, a_{100}\}$



1-itemsets: $\{a_1\}: 2, \{a_2\}: 2, \dots, \{a_{50}\}: 2, \{a_{51}\}: 1, \dots, \{a_{100}\}: 1,$

2-itemsets: $\{a_1, a_2\}: 2, \dots, \{a_1, a_{50}\}: 2, \{a_1, a_{51}\}: 1 \dots, \dots, \{a_{99}, a_{100}\}: 1,$

$\dots, \dots, \dots, \dots$

99-itemsets: $\{a_1, a_2, \dots, a_{99}\}: 1, \dots, \{a_2, a_3, \dots, a_{100}\}: 1$

100-itemset: $\{a_1, a_2, \dots, a_{100}\}: 1$

- Jumlah total frequent itemset :

$$\binom{100}{1} + \binom{100}{2} + \binom{100}{3} + \dots + \binom{100}{100} = 2^{100} - 1$$

Closed Patterns

- ❑ Solusi 1: **Closed patterns**: Suatu pola (itemset) X disebut tertutup (**closed**) jika X sering muncul, dan tidak ada pola superset Y yang lebih besar dari X, dengan frekuensi yang sama seperti X.
- ❑ TDB₁: T₁: {a₁, ..., a₅₀}; T₂: {a₁, ..., a₁₀₀}
- ❑ anggap *minsup* = 1. Berapa banyak closed patterns pada TDB₁ berikut?
 - ❑ Dua= P₁: "{a₁, ..., a₅₀}: 2"; P₂: "{a₁, ..., a₁₀₀}: 1"
- ❑ **Closed pattern** merupakan **Kompresi lossless** dari frequent patterns
 - ❑ Mengurangi jumlah pola yang perlu dianalisa tanpa kehilangan informasi tentang frekuensi kemunculan mereka
 - ❑ {a₂, ..., a₄₀}: 2", "{a₅, a₅₁}: 1"

Max-Patterns

- ❑ Solution 2: **Max-patterns**: Pola X adalah **max-pattern** jika X sering muncul dan tidak ada pola superset Y yang lebih besar dari X yang juga sering muncul.
- ❑ Perbedaan dengan close-patterns?
 - ❑ Tidak memperhatikan support sebenarnya dari sub-pola dalam max-pattern
 - ❑ $TDB_1: T_1: \{a_1, \dots, a_{50}\}; T_2: \{a_1, \dots, a_{100}\}$
 - ❑ anggap $minsup = 1$. Berapa banyak max-patterns pada TDB_1 berikut?
 - ❑ satu: $P: \{\{a_1, \dots, a_{100}\}: 1\}$
- ❑ **Max-pattern** merupakan kompresi **lossy**!
 - ❑ Kita hanya tahu bahwa $\{a_1, \dots, a_{40}\}$ adalah pola yang sering muncul, tetapi kita tidak lagi mengetahui support sebenarnya dari $\{a_1, \dots, a_{40}\}$, dan seterusnya!
- ❑ Jadi dalam penerapannya, close patterns lebih diinginkan daripada max patterns

-
- ❑ TDB2 = T1:{a, b, c, d};T2:{a, b, c};T3:{a, b, d};T4:{a, b};T5:{a, c} ;minsup=2
 - ❑ closed patterns = {a, b, c}, {a, b, d}, {a, b}, {a, c}, {b, c}
 - ❑ {a, b, c}: Frekuensi 2, dan tidak ada superset dengan frekuensi yang sama
 - ❑ {a,d} = bukan closed patterns ?
 - ❑ Max patterns = {a, b, c}, {a, b, d}
 - ❑ {a, b, c}: Frekuensi 2, dan tidak ada superset yang frekuensi
 - ❑ {a, b} = bukan max patterns

Pattern Mining: Konsep dan Metode Dasar

☐ Konsep dasar

☐ Metode Frequent Itemset Mining



☐ Metode Evaluasi Pola

☐ Simpulan

Downward Closure

- ❑ Observasi: $TDB_1: T_1: \{a_1, \dots, a_{50}\}; T_2: \{a_1, \dots, a_{100}\}$
 - ❑ frequent itemset: $\{a_1, \dots, a_{50}\}$
 - ❑ subsets semua frequent: $\{a_1\}, \{a_2\}, \dots, \{a_{50}\}, \{a_1, a_2\}, \dots, \{a_1, \dots, a_{49}\}, \dots$
 - ❑ Ada beberapa hubungan tersembunyi di antara frequent patterns!
- ❑ downward closure (disebut “Apriori”)
 - ❑ Jika **{bir,popok,kacang}** adalah frequent, begitu pula **{bir,popok}**
 - ❑ Setiap transaksi yang berisi **{bir,popok,kacang}** juga berisi **{bir,popok}**
 - ❑ Apriori: setiap subset dari frequent itemset harus frequent
- ❑ Metodologi mining yang efisien
 - ❑ Jika ada subset dari kumpulan item S adalah infrequent, maka tidak ada kemungkinan bagi S untuk menjadi frequent—mengapa kita bahkan harus mempertimbangkan S!?

Algoritma Apriori

- ❑ Algoritma Apriori adalah metode dalam data mining yang digunakan untuk menemukan frequent itemsets dan association rules dalam dataset transaksi. Algoritma ini adalah salah satu teknik dasar dalam market basket analysis dan sering digunakan untuk mengidentifikasi pola item yang sering muncul bersama dalam transaksi.
- ❑ Konsep Dasar
 - ❑ Konsep Dasar Frequent Itemsets: Kumpulan item yang sering muncul bersama dalam transaksi, melebihi batas minimum support yang ditetapkan.
 - ❑ Association Rules: Aturan yang menggambarkan hubungan antara itemsets, misalnya, "Jika A dibeli, maka B juga cenderung dibeli," yang diukur dengan confidence dan support.

Langkah Apriori

- ❑ Inisialisasi:
 - ❑ Tentukan minimum support dan minimum confidence.
 - ❑ Hitung frekuensi setiap itemset 1-item (item tunggal) dalam dataset transaksi.
- ❑ Generasi Kandidat dan Pengujian:
 - ❑ Level 1: Temukan semua itemsets frekuen dari itemsets 1-item.
 - ❑ Level 2 dan seterusnya: Buat kandidat itemsets k-item dengan menggabungkan itemsets (k-1)-item yang sudah frekuen.
 - ❑ Periksa frekuensi kandidat itemsets tersebut dan simpan itemsets yang memenuhi minimum support.
 - ❑ Proses ini diulang untuk itemsets dengan jumlah item yang lebih besar hingga tidak ada itemsets baru yang memenuhi minimum support.

The Apriori Algorithm—An Example

Database TDB

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

minsup = 2

C_1

1st scan

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

F_1

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

C_2

F_2

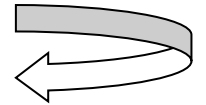
Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

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

2nd scan

C_2

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



C_3

Itemset	sup
{B, C, E}	2

3rd scan

F_3

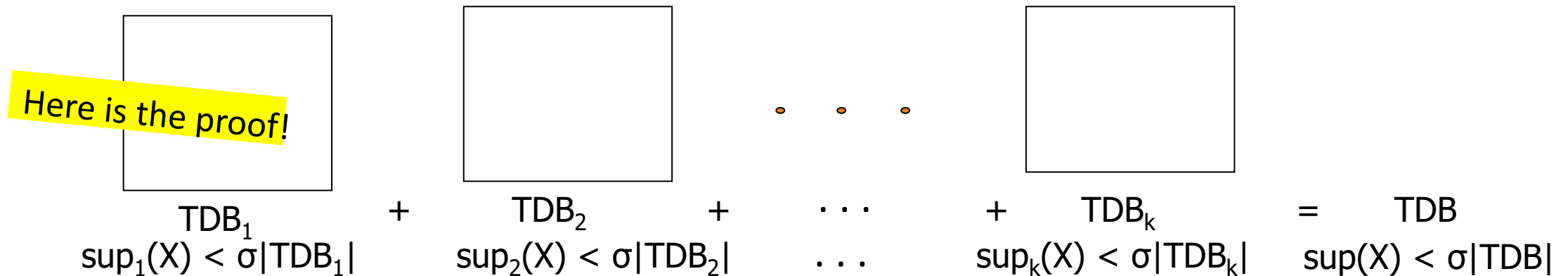
Itemset	sup
{B, C, E}	2

Apriori: Improve dan Alternatif

- ❑ Mengurangi proses scan database
 - ❑ **Partitioning** (e.g., Savasere, et al., 1995)
 - ❑ Dynamic itemset counting (Brin, et al., 1997)
- ❑ Mengurangi jumlah kandidat
 - ❑ **Hashing** (e.g., DHP: Park, et al., 1995)
 - ❑ Pruning by support lower bounding (e.g., Bayardo 1998)
 - ❑ Sampling (e.g., Toivonen, 1996)
- ❑ Eksplorasi special data structures
 - ❑ Tree projection (Agarwal, et al., 2001)
 - ❑ H-miner (Pei, et al., 2001)
 - ❑ Hypercube decomposition (e.g., LCM: Uno, et al., 2004)

Partitioning: Scan Database hanya dua kali

- teorema: *Setiap itemset yang berpotensi frequent di TDB harus frequent di setidaknya salah satu partisi TDB*



- Metode: Scan DB Dua kali (A. Savasere, E. Omiecinski and S. Navathe, *VLDB'95*)
 - Scan 1: Partisi database sehingga setiap partisi dapat muat di memori utama
 - menambang local frequent patterns di partisi ini
 - menghitung frekuensi setiap item dan itemsets kecil di setiap partisi
 - Scan 2: Konsolidasi global frequent patterns
 - Mencari kandidat global frequent itemset (frequent setidaknya satu partisi)
 - Mencari frekuensi sebenarnya dari kandidat tersebut, dengan scan TDB_i sekali lagi

Direct Hashing and Pruning (DHP)

- ❑ **Hashing:** $v = \text{hash}(\text{itemset})$
- ❑ **Scan 1:** menghitung 1-itemset, hash 2-itemset untuk menghitung jumlah bucket
- ❑ Contoh: scan 1 TDB, hitung 1-itemset, dan hash 2-itemsets dalam transaksi ke bucket-nya
 - ❑ {ab, ad, ce}
 - ❑ {bd, be, de}
 - ❑ ...
- ❑ Setelah di scan *Jika minsup = 80, hapus ab, ad, ce, karena $\text{count}\{ab, ad, ce\} < 80$*

Itemsets	Count
{ab, ad, ce}	35
{bd, be, de}	298
.....	...
{yz, qs, wt}	58

Hash Table



Check the minsup

K-itemset tidak dapat sering terjadi jika jumlah bucket hashing yang sesuai berada di bawah ambang batas minsup

Exploring Vertical Data Format: ECLAT

- ❑ ECLAT (Equivalence Class Transformation): algoritma **depth-first search** menggunakan set intersection [Zaki et al. @KDD'97]
- ❑ Vertical format
- ❑ Tid-Lists
 - ❑ $t(X) = t(Y)$: X dan Y selalu terjadi secara bersamaan (misalnya, $t(ac) = t(d)$)
 - ❑ $t(X) \subset t(Y)$: transaksi yang memiliki X selalu memiliki Y (Misalnya, $t(ac) \subset t(ce)$)
- ❑ Frequent patterns: vertical intersections
- ❑ Menggunakan diffset untuk mempercepat mining
 - ❑ Hanya melacak perbedaan tids
 - ❑ $t(e) = \{T_{10}, T_{20}, T_{30}\}$, $t(ce) = \{T_{10}, T_{30}\} \rightarrow \text{Diffset}(ce, e) = \{T_{20}\}$

A transaction DB in Horizontal Data Format

Tid	Itemset
10	a, c, d, e
20	a, b, e
30	b, c, e

The transaction DB in Vertical Data Format

Item	TidList
a	10, 20
b	20, 30
c	10, 30
d	10
e	10, 20, 30

$t(e) = \{T_{10}, T_{20}, T_{30}\};$
 $t(a) = \{T_{10}, T_{20}\};$
 $t(ae) = \{T_{10}, T_{20}\}$

Contoh: Dari Transactional DB ke Ordered Frequent Itemlist

Example: A Sample Transactional Database

TID	Items in the Transaction
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o, w}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

min_support = 3

- Scan DB sekali, temukan satu item frequent pattern:

f:4, a:3, c:4, b:3, m:3, p:3

- Urutkan frequent items berdasarkan frekuensi dari yang terbesar, f-list

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

- Scan DB lagi, gunakan frequent itemlist untuk setiap transaksi untuk membangun FP-tree

TID	Items in the Transaction	Ordered, frequent itemlist
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

Contoh: Membangun FP-tree dari Transaction DB

TID	Ordered, frequent itemlist
100	<u>f, c, a, m, p</u>
200	<u>f, c, a, b, m</u>
300	f, b
400	c, b, p
500	f, c, a, m, p

frequent Itemlist 1: "f, c, a, m, p"

Item	Frqncy	hdr
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	

Header Table

frequent itemlist 2 "f, c, a, b, m"

itm	hdr
f	
c	
a	
b	
m	
p	

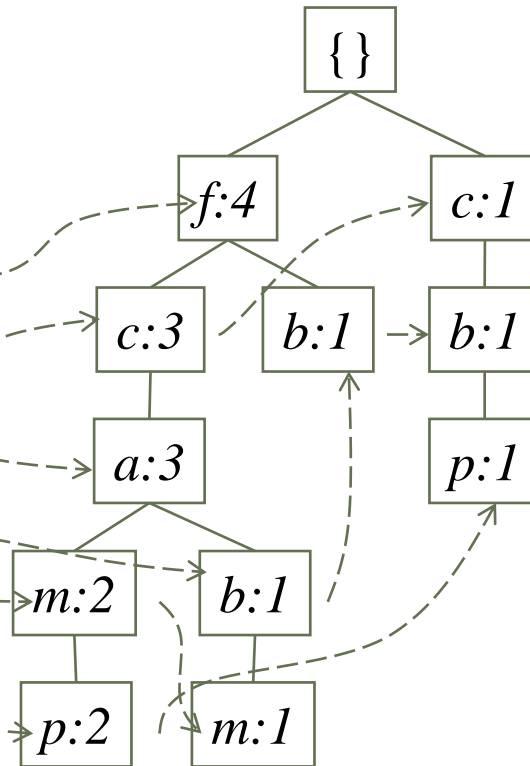
Semua frequent itemlists

itm	hdr
f	
c	
a	
b	
m	
p	

Mining FP-Tree

min_support = 3

Item	Frequency	Header
f	4	
c	4	
a	3	
b	3	
m	3	
p	3	



Conditional database of each pattern

Item	Conditional database
c	f:3
a	fc:3
b	fca:1, f:1, c:1
m	fca:2, fcab:1
p	fcam:2, cb:1

Mine Each Conditional Database Recursively

min_support = 3

Conditional Data Bases

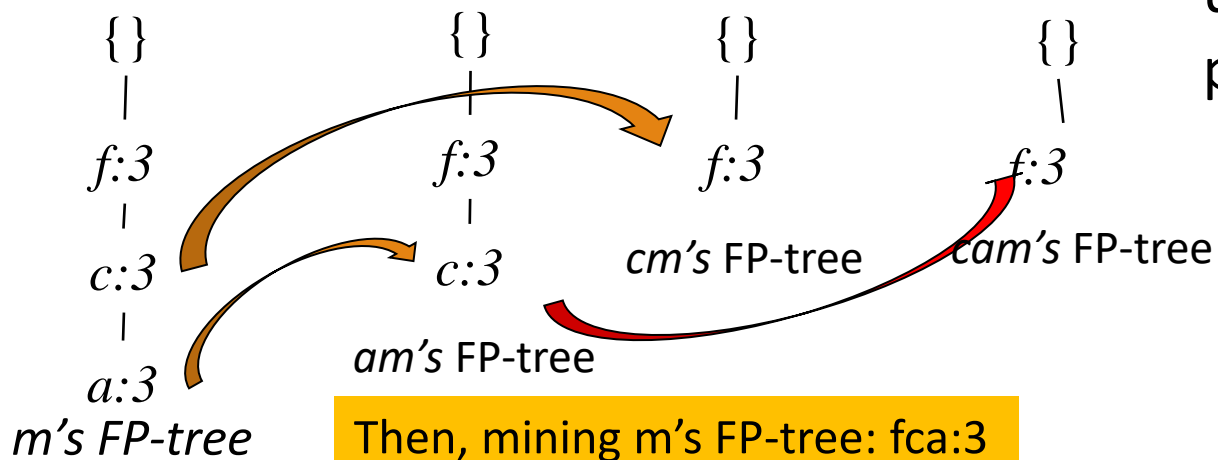
<i>item</i>	<i>cond. data base</i>
<i>c</i>	<i>f:3</i>
<i>a</i>	<i>fc:3</i>
<i>b</i>	<i>fca:1, f:1, c:1</i>
<i>m</i>	<i>fca:2, fcab:1</i>
<i>p</i>	<i>fcam:2, cb:1</i>

***p*'s conditional DB: $fcam:2, cb:1 \rightarrow c:3$**

m's conditional DB: $fca:2, fcab:1 \rightarrow fca:3$

b 's conditional DB: $fca:1, f:1, c:1 \rightarrow \phi$

untuk cabang Tunggal FP-tree, semua frequent patterns dapat dihasilkan dalam satu tembakan



m: 3

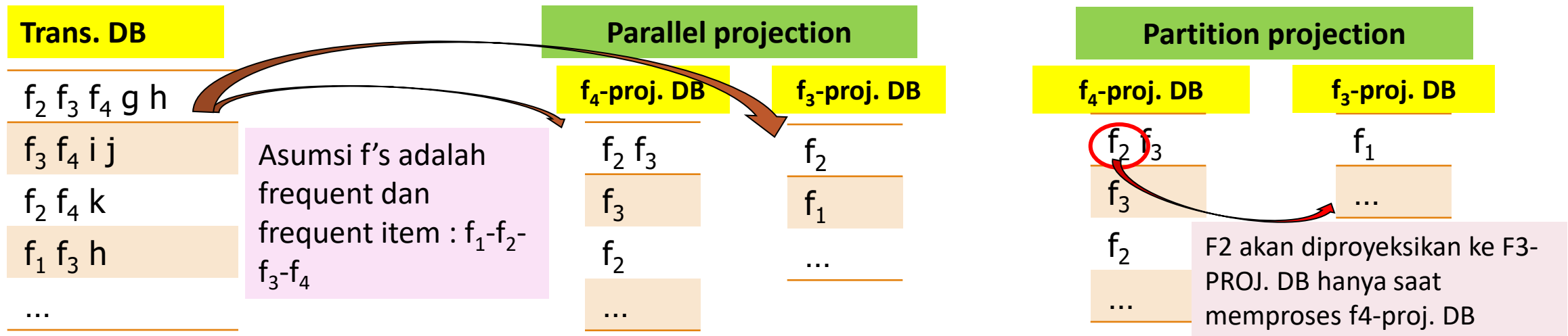
fm: 3, cm: 3, am: 3

fcm: 3, fam:3, cam: 3

fcam: 3

Scaling FP-growth by Item-Based Data Projection

- Bagaimana jika FP-tree tidak dapat muat dalam memori?
 - “proyek” database berdasarkan frequent single items
 - Membangun & menambang FP-tree untuk setiap DB yang diproyeksikan
- **Parallel projection** vs. **partition projection**
 - Parallel projection: Memproyeksikan DB pada masing-masing frequent item
 - Semua partisi dapat diproses secara paralel
 - Partition projection: Partisi DB secara berurutan
 - Meneruskan bagian yang belum diproses ke partisi berikutnya



Pattern Mining: Konsep dan Metode Dasar

- ❑ Konsep dasar
- ❑ Metode Frequent Itemset Mining
- ❑ Metode Evaluasi Pola
- ❑ Simpulan



Bagaimana menilai Rule/Pattern?

- ❑ Penambangan pola akan menghasilkan sekumpulan besar patterns/rules
 - ❑ Tidak semua patterns/rules yang dihasilkan menarik
- ❑ Ukuran kemenarikan: **Objective** vs. **subjective**
 - ❑ **Objective**
 - ❑ Support, confidence, correlation, ...
 - ❑ **Subjective**
 - ❑ Pengguna yang berbeda mungkin menilai kemenarikan secara berbeda
 - ❑ Permintaan pengguna
 - ❑ Query-based: Relevan dengan permintaan khusus pengguna
 - ❑ Menilai berdasarkan basis pengetahuan seseorang
 - ❑ unexpected, freshness, timeliness

Support-Confidence Framework

- Apakah s dan c menarik di association rules: “ $A \Rightarrow B$ ” [s, c]?
- Contoh: Misalkan satu sekolah mungkin memiliki statistik berikut tentang # siswa yang mungkin bermain bola basket dan/atau makan sereal:

	play-basketball	not play-basketball	sum (row)
eat-cereal	400	350	750
not eat-cereal	200	50	250
sum(col.)	600	400	1000

- Association rule mining :
 - $play-basketball \Rightarrow eat-cereal$ [40%, 66.7%] (higher s & c)
 - Support (40%): Artinya, 40% dari keseluruhan siswa dalam dataset bermain basket dan makan sereal
 - Confidence (66.7%): Artinya, dari semua siswa yang bermain basket, 66.7% juga makan sereal.
 - Rule ini memberikan kesan bahwa siswa yang bermain basket cenderung makan sereal, tetapi faktanya, secara umum lebih banyak siswa yang makan sereal dibandingkan mereka yang bermain basket. Jadi, aturan ini mungkin tidak signifikan kelihatannya.

Support-Confidence Framework

- ❑ association rule yang sebelumnya menyesatkan: Secara keseluruhan % siswa yang makan sereal adalah $75\% > 66.7\%$,
- ❑ Aturan yang signifikan:
 - ❑ $\neg \text{play-basketball} \Rightarrow \text{eat-cereal}$ [35%, 87.5%] (high s & c)
 - ❑ Support (35%): Artinya, 35% dari keseluruhan siswa tidak bermain basket dan makan sereal.
 - ❑ Confidence (87.5%): Artinya, dari semua siswa yang tidak bermain basket, 87.5% dari mereka makan sereal.
 - ❑ siswa yang tidak bermain basket lebih mungkin untuk makan sereal dibandingkan siswa yang bermain basket

Interestingness Measure: Lift

- **lift** : Ukuran yang digunakan untuk mengukur dependensi atau korelasi antara dua even

$$lift(B, C) = \frac{c(B \rightarrow C)}{s(C)} = \frac{s(B \cup C)}{s(B) \times s(C)}$$

Lift is more telling than s & c

- Lift(B, C) dapat memberi tahu bagaimana B dan C berkorelasi

- Lift(B, C) = 1: B dan C independen
- > 1: berkorelasi positif
- < 1: berkorelasi negatif

- contoh

$$lift(B, C) = \frac{400/1000}{600/1000 \times 750/1000} = 0.89$$

$$lift(B, \neg C) = \frac{200/1000}{600/1000 \times 250/1000} = 1.33$$

- Dengan demikian, B dan C berkorelasi negatif karena lift(B, C) < 1;

- B dan ¬C berkorelasi positif karena lift(B, ¬C) > 1

	B	¬B	Σ _{row}
C	400	350	750
¬C	200	50	250
Σ _{col.}	600	400	1000

Interestingness Measure: Chi-square χ^2

- Ukuran lain untuk menguji korelasi events: χ^2

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- Perhatikan tabel sebelah kanan:

$$\chi^2 = \frac{(400 - 450)^2}{450} + \frac{(350 - 300)^2}{300} + \frac{(200 - 150)^2}{150} + \frac{(50 - 100)^2}{100} = 55.56$$

- Mencari tabel distribusi $\chi^2 \rightarrow$ B dan C berkorelasi
- Uji χ^2 menunjukkan bahwa B dan C berkorelasi negatif karena nilai ekspektasinya adalah 450 tetapi yang diamati hanya 400
- Dengan demikian, χ^2 lebih informatif dibandingkan dengan support-confidence framework

	B	$\neg B$	Σ_{row}
C	400 (450)	350 (300)	750
$\neg C$	200 (150)	50 (100)	250
Σ_{col}	600	400	1000

Nilai diharapkan

Nilai diamati

Nilai diharapkan =

$$\frac{750 \times 600}{1000} = 450$$

Interestingness Measure: Chi-square χ^2

□ Interpretasi:

- Jika χ^2 yang dihitung lebih besar dari **nilai kritis**: Hipotesis nol ditolak. Ini berarti ada bukti yang cukup untuk menyatakan bahwa ada perbedaan signifikan antara frekuensi yang diobservasi dan frekuensi yang diharapkan.
- Jika χ^2 yang dihitung lebih kecil dari **nilai kritis**: Hipotesis nol tidak ditolak. Ini berarti tidak ada bukti yang cukup untuk menyatakan adanya perbedaan signifikan antara frekuensi yang diobservasi dan frekuensi yang diharapkan.

□ Derajat Kebebasan (df):

- Derajat kebebasan biasanya dihitung sebagai: $df = (r-1) \times (c-1)$
 - Dimana r adalah jumlah baris dalam tabel kontingensi dan c adalah jumlah kolom.

□ Tingkat Signifikansi (α):

- Tingkat signifikansi adalah probabilitas kesalahan tipe I yang diterima, biasanya 0.05 (5%) atau 0.01 (1%).

df	$\alpha = 0,10$	$\alpha = 0,05$
1	2,706	3,841
2	4,605	5,991
3	6,251	7,815
4	7,779	9,488
5	9,236	11,070

Lift and χ^2 : Apakah Solusi yang baik?

- ❑ Null transactions: Transaksi yang tidak mengandung B atau C
- ❑ new dataset D
 - ❑ BC (100) jauh lebih langka daripada B¬C (1000) dan ¬BC (1000), tetapi ada banyak¬B¬C (100000)
 - ❑ Kemungkinan B & C akan terjadi bersama!
- ❑ $\text{Lift}(B, C) = 8.44 \gg 1$ (Lift menunjukkan B dan C sangat berkorelasi positif!)
- ❑ $\chi^2 = 670$: Observed(BC) \gg expected value (11.85)
- ❑ Terlalu banyak *null transactions* dapat “spoil the soup”!

	B	¬B	Σ_{row}
C	100	1000	1100
¬C	1000	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

null transactions

Tabel kontingensi dengan penambahan nilai diharapkan

	B	¬B	Σ_{row}
C	100 (11.85)	1000	1100
¬C	1000 (988.15)	100000	101000
$\Sigma_{\text{col.}}$	1100	101000	102100

Interestingness Measures & Null-Invariance

- ❑ **Null invariance** berarti: Jumlah **null transactions** tidak dipermasalahkan.
Tidak mengubah nilai pengukuran.
- ❑ beberapa interestingness measures: Beberapa invarian nol

Measure	Definition	Range	Null-Invariant?
$\chi^2(A, B)$	$\sum_{i,j} \frac{(e(a_i, b_j) - o(a_i, b_j))^2}{e(a_i, b_j)}$	$[0, \infty]$	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
$Kulczynski(A, B)$	$\frac{1}{2} \left(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)} \right)$	$[0, 1]$	Yes
$MaxConf(A, B)$	$\max\left\{ \frac{s(A \cup B)}{s(A)}, \frac{s(A \cup B)}{s(B)} \right\}$	$[0, 1]$	Yes

Let

$$p = \frac{s(A \cup B)}{s(A)} = P(B|A)$$

$$q = \frac{s(A \cup B)}{s(B)} = P(A|B)$$

p, q are null invariant

Essentially min,
max, mean variants
of p, q

Null Invariance

- Mengapa invarian nol sangat penting untuk analisis data transaksi besar-besaran?
 - Many transactions may contain neither milk nor coffee!

milk vs. coffee contingency table

	<i>milk</i>	$\neg milk$	Σ_{row}
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	$\neg c$
Σ_{col}	<i>m</i>	$\neg m$	Σ

- Lift and χ^2 bukan null-invariant: Tidak baik untuk mengevaluasi data yang berisi terlalu banyak atau terlalu sedikit transaksi null!
- Many measures are not null-invariant!

Null-transactions
w.r.t. *m* and *c*

Data set	<i>mc</i>	$\neg mc$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	χ^2	<i>Lift</i>
<i>D</i> ₁	10,000	1,000	1,000	100,000	90557	9.26
<i>D</i> ₂	10,000	1,000	1,000	100	0	1
<i>D</i> ₃	100	1,000	1,000	100,000	670	8.44
<i>D</i> ₄	1,000	1,000	1,000	100,000	24740	25.75
<i>D</i> ₅	1,000	100	10,000	100,000	8173	9.18
<i>D</i> ₆	1,000	10	100,000	100,000	965	1.97

Comparison of Null-Invariant Measures

- ❑ Tidak semua null-invariant measures hasilnya sama
- ❑ Mana yang lebih baik?
 - ❑ $D_4 - D_6$ bedakan null-invariant measures
 - ❑ Kulc (Kulczynski 1927) stabil dan seimbang

2-variable contingency table

	<i>milk</i>	$\neg milk$	Σ_{row}
<i>coffee</i>	<i>mc</i>	$\neg mc$	<i>c</i>
$\neg coffee$	$m\neg c$	$\neg m\neg c$	$\neg c$
Σ_{col}	<i>m</i>	$\neg m$	Σ

All 5 are null-invariant

Data set	<i>mc</i>	$\neg mc$	$m\neg c$	$\neg m\neg c$	<i>AllConf</i>	Jaccard	<i>Cosine</i>	<i>Kulc</i>	<i>MaxConf</i>
D_1	10,000	1,000	1,000	100,000	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	0.09	0.05	0.09	0.09	0.09
D_4	1,000	1,000	1,000	100,000	0.5	0.33	0.5	0.5	0.5
D_5	1,000	100	10,000	100,000	0.09	0.09	0.29	0.5	0.91
D_6	1,000	10	100,000	100,000	0.01	0.01	0.10	0.5	0.99

Subtle: They disagree on those cases

Imbalance Ratio with Kulczynski Measure

- IR (Imbalance Ratio): mengukur imbalance dari dua item set A and B dalam rule implications:

$$IR(A, B) = \frac{|s(A) - s(B)|}{s(A) + s(B) - s(A \cup B)}$$


- Kulczynski and Imbalance Ratio (IR) bersama-sama menyajikan gambaran yang jelas untuk ketiga kumpulan data D4 hingga D6
 - D₄ neutral & balanced; D₅ neutral tetapi imbalanced
 - D₆ neutral tetapi very imbalanced

Data set	<i>mc</i>	$\neg mc$	<i>m</i> $\neg c$	$\neg m$ $\neg c$	Jaccard	<i>Cosine</i>	<i>Kulc</i>	IR
<i>D</i> ₁	10,000	1,000	1,000	100,000	0.83	0.91	0.91	0
<i>D</i> ₂	10,000	1,000	1,000	100	0.83	0.91	0.91	0
<i>D</i> ₃	100	1,000	1,000	100,000	0.05	0.09	0.09	0
<i>D</i> ₄	1,000	1,000	1,000	100,000	0.33	0.5	0.5	0
<i>D</i> ₅	1,000	100	10,000	100,000	0.09	0.29	0.5	0.89
<i>D</i> ₆	1,000	10	100,000	100,000	0.01	0.10	0.5	0.99

Measures untuk memilih Evaluasi Pattern yang Efektif?

- ❑ Null value kasus dominan di banyak kasus datasets besar
- ❑ *Null-invariance* is an important property
- ❑ Lift, χ^2 and cosine adalah ukuran yang baik jika transaksi null tidak dominan
 - ❑ Sebaliknya, *Kulczynski + Imbalance Ratio* harus digunakan untuk menilai kemenarikan suatu pola

Pattern Mining: Basic Concepts and Methods

- ❑ Basic Concepts
- ❑ Frequent Itemset Mining Methods
- ❑ Which Patterns Are Interesting?—Pattern Evaluation Methods
- ❑ Summary 

Simpulan

- ❑ Basic Concepts
 - ❑ What Is Pattern Discovery? Why Is It Important?
 - ❑ Basic Concepts: Frequent Patterns and Association Rules
 - ❑ Compressed Representation: Closed Patterns and Max-Patterns
- ❑ Efficient Pattern Mining Methods
 - ❑ The Downward Closure Property of Frequent Patterns
 - ❑ The Apriori Algorithm
 - ❑ Extensions or Improvements of Apriori
 - ❑ Mining Frequent Patterns by Exploring Vertical Data Format
 - ❑ FPGrowth: A Frequent Pattern-Growth Approach
 - ❑ Mining Closed Patterns
- ❑ Pattern Evaluation
 - ❑ Interestingness Measures in Pattern Mining
 - ❑ Interestingness Measures: Lift and χ^2
 - ❑ Null-Invariant Measures
 - ❑ Comparison of Interestingness Measures