

# CHAPTER 11: DATA MINING AND DATA VISUALIZATION

*DATA MINING*  
*SE4323*

## SEBUAH GAMBAR MENGUNGKAPKAN RIBUAN KATA

- DATA MINING adalah satu set dari aktivitas yang digunakan untuk mencari hal yang tersembunyi atau pola tak terduga dalam data.
- Teknik ini biasanya disebut sebagai *Knowledge Discovery in Database (KDD)*, dan termasuk analisis statistik, *neural/ fuzzy logic*, *intelligent agents* atau visualisasi data.
- Teknik KDD tak hanya menemukan pola yang berguna dari data, tapi juga digunakan untuk mengembangkan peramalan model.

# VERIFICATION VERSUS DISCOVERY

- Aktivitas pendukung keputusan terutama berdasarkan pada konsep *verification*/ pembuktian. Ini membutuhkan banyak pengetahuan sebelumnya pada bagian *decision-maker* dan dalam hal ini untuk **membuktikan satu hubungan yang diduga.**



## Hipotesis

adalah jawaban sementara terhadap masalah yang masih bersifat praduga karena masih harus dibuktikan kebenarannya

- Dengan bantuan teknologi, konsep dari pembuktian mulai bergeser ke *discovery*/ penemuan.

## MENGAPA DATA MINING TUMBUH DAN TERKENAL?

Kita sadar bahwa otak manusia mempunyai masalah dalam mengolah data yang multidimensional.

Teknik *machine learning* pada data mining menjadi lebih mampu, lebih beradab, lebih tajam hasil analisisnya dengan cara pembelajaran yang terus menerus

## DATA MINING PUNYA KETERBATASAN PULA

Walaupun pada literatur ada pernyataan sbb:

“data mining akan memungkinkan kita meramalkan siapa yang akan membeli satu produksi khusus”,  
namun ini harus kembali ke sifat manusia.

Dalam situasi tertentu, hasil dari data mining malah membuat masalah. Misalkan hasil mining menyatakan bahwa wanita lebih suka mendapat discount. Lalu dalam penerapannya, hanya wanita yang di discount sedang pria tidak, ini adalah pelanggaran diskriminasi jender, dan membuat permasalahan hukum/ sosial.

# TEKNIK YANG DIPAKAI DATA MINING

Data mining berjalan bersama secara paralel dengan perkembangan teknologi baru.

Teknik data-mining maka dapat dibagi jadi 4 kategori:

- |                   |                |
|-------------------|----------------|
| 1. classification | 2. association |
| 3. sequencing     | 4. clustering  |

## CLASSIFICATION METHODS

Hasil akhirnya adalah mencari aturan pendefinisian kondisi bahwa satu item akan dimiliki oleh satu subkelas atau kelas data.

Sebagai contoh, jika kita coba mendeteksi rumahtangga mana yang menanggapi *direct mail*, kita akan butuh aturan yang memisahkan yang termasuk dan yang tidak.

Ini adalah '*IF-THEN rules*' sering diungkapkan dengan struktur diagram pohon.

# ASSOCIATION METHODS

Teknik ini mencari semua transaksi dari satu sistem untuk mendapat pola dari kejadian.

Metode umum adalah analisis keranjang belanja (*Market Basket Analysis* – disingkat MBA), dimana kumpulan ribuan pembelian barang konsumen diuji.

Hasilnya diungkapkan dalam persentase.

Contoh: “30% dari pelanggan beli steak juga membeli arang”.



# SEQUENCING METHODS

Metode ini diterapkan pada *time series* data untuk mendapat pola yang tersembunyi.

Jika ketemu, dapat digunakan prediksi untuk kejadian berikutnya.

Contoh: grup pelanggan cenderung membeli produk yang tergabung dengan film yang populer → Asesoris berupa topi, tongkat, stiker, plakat, buku tulis bernuansa Harry Potter akan laris pada saat akan di release nya film Harry Potter.

# CLUSTERING TECHNIQUES

Teknik Clustering mencoba menciptakan partisi pada data menurut beberapa matrik jarak.

Cluster membentuk kelompok data bersama secara sederhana lewat kesamaan/ keserupaan dengan tetangganya.

Dengan menguji karakteristik setiap cluster, maka diharapkan dapat dibangun aturan untuk pengelompokan

# TEKNOLOGI DALAM DATA MINING

- *Statistik* – adalah teknologi data mining yang telah matang, tapi sering tidak dapat dipakai karena mereka butuh data yang bersih/clean. Sebagai tambahan, beberapa prosedur statistik berasumsi hubungan linear, yang membatasi penggunaannya.
- *Neural networks, genetic algorithms, fuzzy logic* – adalah teknologi yang dapat bekerja pada data yang sukar dan tidak tepat/teliti. Pemakaian yang luas membuat mereka populer dalam medan yang tidak pasti ini.

# TEKNOLOGI DALAM DATA MINING

(LANJUTAN...)

- *Decision trees* – teknologi ini secara konsep adalah sederhana dan mempunyai tambahan pada popularitas sebagai “tree growing software”. Karena cara yang dipakai, teknologi ini mungkin lebih baik disebut sebagai “*classification*” tree.

# NEW APPLICATIONS FOR DATA MINING

Sebagai teknologi yang telah matang, aplikasi baru muncul, terutama dalam 2 kategori: *text mining* dan *web mining*.

Beberapa contoh text mining :

- Menyaring/ *Distilling* arti dari teks
- Peringkasan akurat dari teks
- Menerangkan struktur tema teks
- Mengelompokan/ *Clustering* dari teks

# WEB MINING

*Web mining* kasus khusus dari *text mining* dimana penambangan terjadi pada *website*.

Menambah kepintaran website, seperti merekomendasikan link yang berelasi atau rekomendasi terhadap suatu produk yang baru.

# MARKET BASKET ANALYSIS: THE KING OF ALGORITHMS

Ini dipakai secara luas dan merupakan algoritma data mining yang cukup berhasil.

Tujuan: Mencari produk yang biasanya dibeli bersamaan.

Fungsi:

- Toko dapat menaruh barang tsb berdekatan.
- Direct marketers dapat memakai informasi untuk menentukan produk baru mana yang disarankan pelanggan saat ini.
- Kebijakan inventory dapat ditingkatkan jika reorder points merefleksikan kebutuhan untuk produk pelengkap.

# ASSOCIATION RULES FOR MARKET BASKET ANALYSIS

Associations menyarankan aturan formal

If *kondisi* then *hasilnya*.

If Thursday and diapers, then beer.

Aturan ditulis dalam bentuk “*left-hand side implies right-hand side*”  
dan contoh:

Yellow Peppers IMPLIES Red Peppers, Bananas, Bakery



# ASSOCIATION RULES FOR MARKET BASKET ANALYSIS

Untuk membuat aturan penggunaan yang efektif, tiga ukuran numerik tentang aturan itu berhubungan dengan :

- (1) dukungan/ *support*,
- (2) kepercayaan/ *confidence*
- (3) daya angkat/ *lift*

## MEASURES OF PREDICTIVE ABILITY

1. *Support* menunjuk pada persentase keranjang dimana aturan berlaku benar (kedua sisi produk kanan dan kiri keduanya hadir).
2. *Confidence* mengukur persentase dari keranjang belanja dimana “*the left-hand product also contained the right*”.
3. *Lift* mengukur berapa banyak frekuensi “*the left-hand item is found with the right than without the right*”.

# MEASURES OF PREDICTIVE ABILITY

## 3. Lift / Improvement Ratio

Lift Ratio adalah parameter penting selain support dan confidence dalam association rule. Lift Ratio mengukur seberapa penting rule yang telah terbentuk berdasarkan nilai support dan confidence. Lift Ratio merupakan nilai yang menunjukkan kevalidan proses transaksi dan memberikan informasi apakah benar produk A dibeli bersamaan dengan produk B.

Lift Ratio dapat dihitung dengan rumus

$$\text{SupportAB} / \text{Support(A)} \cdot \text{Support(B)}$$

# MEASURES OF PREDICTIVE ABILITY

If lift is greater than 1, it suggests that the presence of the items on the LHS has increased the probability that the items on the right hand side will occur on this transaction.

If the lift is below 1, it suggests that the presence of the items on the LHS make the probability that the items on the RHS will be part of the transaction *lower*.

Jika  $LIFT > 1$ , ini menunjukkan bahwa kehadiran item pada LHS telah meningkatkan probabilitas bahwa item pada sisi kanan akan terjadi pada transaksi ini.

Jika  $LIFT < 1$ , ini menunjukkan bahwa kehadiran item pada LHS membuat probabilitas bahwa item pada RHS akan menjadi bagian dari transaksi yang lebih rendah.

## MEASURES OF PREDICTIVE ABILITY

If the lift is 1, it suggests that the presence of items on the LHS and RHS really are independent: knowing that the items on the LHS are present makes **no** difference to the probability that items will occur on the RHS.

Jika  $\text{lift} = 1$ , ini menunjukkan bahwa kehadiran item pada LHS dan RHS benar-benar independen: mengetahui bahwa item pada LHS yang hadir **tidak membuat perbedaan** terhadap probabilitas barang-barang akan terjadi pada RHS.

## CONTOH

| <b>Rule:</b>      | <b>Green Peppers<br/>IMPLIES<br/>Bananas</b> | <b>Red Peppers<br/>IMPLIES<br/>Bananas</b> | <b>Yellow Peppers<br/>IMPLIES<br/>Bananas</b> |
|-------------------|--|--|---|
| <b>Lift</b>       | 1.37   | 1.43                                       | 1.17  |
| <b>Support</b>    | 3.77%  | 8.58%                                      | 22.12%  |
| <b>Confidence</b> | 85.96%                                       | 89.47%                                     | 73.09%  |

# MARKET BASKET ANALYSIS METHODOLOGY

Pertama kita membutuhkan transaksi pembelian. Ini sangatlah mudah didapat dari *cash register*.

Berikutnya, kita pilih daftar produk yang dianalisis, dan mentabulasikan berapa kali terjadi pembelian setiap barang bersama dengan barang lainnya.



Which items are frequently purchased together by my customers?

### Shopping Baskets



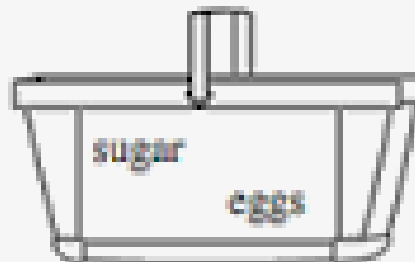
Customer 1



Customer 2



Customer 3



Customer n

Market Analyst

Market basket analysis.



## A CONVENIENCE STORE EXAMPLE (5 TRANSACTIONS)

Terdapat 5 contoh transaksi sbb:

Transaction 1: Pizza, cola, milk

Transaction 2: Milk, potato chips

Transaction 3: Cola, pizza

Transaction 4: Milk, pretzels

Transaction 5: Cola, pretzels

# Definisi Umum

**Itemset:** himpunan dari item-item yang muncul bersama-sama

**Kaidah asosiasi:** peluang bahwa item-item tertentu hadir bersama-sama.

$$X \rightarrow Y \text{ dimana } X \cap Y = \emptyset$$

**Support,  $\text{supp}(X)$**  dari suatu itemset  $X$  adalah rasio dari jumlah transaksi dimana suatu itemset muncul dari seluruh transaksi

# Definisi Umum

*Confidence* (keyakinan) dari kaidah  $X \rightarrow Y$ , ditulis  $\text{conf}(X \rightarrow Y)$  adalah

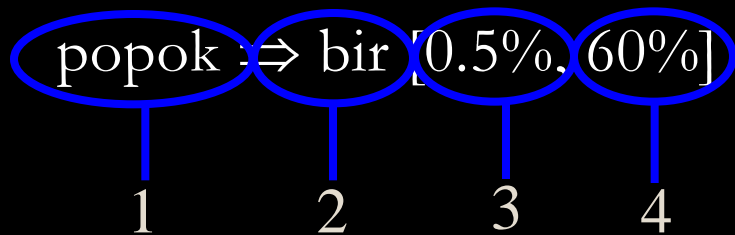
$$\text{conf}(X \rightarrow Y) = \text{supp}(X \cup Y) / \text{supp}(X)$$

*Confidence* bisa juga didefinisikan dalam terminologi peluang bersyarat:

$$\text{conf}(X \rightarrow Y) = P(Y | X) = P(X \cap Y) / P(X)$$

Database transaksi menyimpan data transaksi. Data transaksi bisa juga disimpan dalam suatu bentuk lain dari suatu database

# Kaidah Asosiasi: Dasar



"**IF** membeli popok,  
**THEN** membeli bir  
dalam 60% kasus  
dalam 0.5% dari baris-transaksi"

- 1 *Antecedent, left-hand side (LHS), body*
- 2 *Consequent, right-hand side (RHS), head*
- 3 *Support, frekuensi ("dalam berapa besar bagian dari data dalam LHS dan RHS terjadi bersama-sama")*
- 4 *Confidence, kekuatan ("jika LHS terjadi, seberapa yakin RHS akan terjadi")*

## CONFIDENCE LEVEL

| <u>Combination</u> | <u>Probability of Occurrence</u> |
|--------------------|----------------------------------|
| A                  | 45%                              |
| B                  | 42.5%                            |
| C                  | 40%                              |
| A and B            | 25%                              |
| A and C            | 20%                              |
| B and C            | 15%                              |
| A and B and C      | 5%                               |

If A and B, then C


$$\text{CL} = 0.05 / 0.25 = \mathbf{0.20}$$

If A and C, then B

$$\text{CL} = 0.05 / 0.20 = \mathbf{0.25}$$

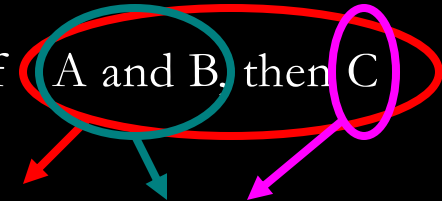
If B and C, then A

$$\text{CL} = 0.05 / 0.15 = \mathbf{0.33}$$

# LIFT

| <u>Combination</u> | <u>p</u> |
|--------------------|----------|
| A                  | 45%      |
| B                  | 42.5%    |
| C                  | 40%      |
| A and B            | 25%      |
| A and C            | 20%      |
| B and C            | 15%      |
| A and B and C      | 5%       |

If A and B, then C



$$I = 0.05 / (0.25 * 0.40) = \mathbf{0.50}$$

If A and C, then B

$$I = 0.05 / (0.20 * 0.425) = \mathbf{0.59}$$

If B and C, then A

$$I = 0.05 / (0.15 * 0.45) = \mathbf{0.74}$$

If A then B

$$I = 0.25 / (0.45 * 0.425) = \mathbf{1.31}$$

## THE PROBLEM OF BIG DATA

What if a store sell 100 items (n):

| # in combination (r) | # of combinations |
|----------------------|-------------------|
| 1                    | 100               |
| 2                    | 4,950             |
| 3                    | 161,700           |
| 4                    | 3,921,225         |
| 5                    | 75,287,520        |
| 6                    | 1,192,052,400     |
| 7                    | 16,007,560,800    |
| 8                    | 186,087,894,300   |

$$C_r^n = \frac{n!}{r! (n-r)!}$$

Transaction 1: Pizza, cola, milk  
Transaction 2: Milk, potato chips  
Transaction 3: Cola, pizza  
Transaction 4: Milk, pretzels  
Transaction 5: Cola, pretzels

## MBA EXAMPLE (5 TRANSACTIONS)

Support Cola = Support Milk =  $3/5 = 60\%$

Support Pizza = Support Pretzels =  $2/5 = 40\%$

Support Potato =  $1/5 = 20\%$

Support Cola & Pizza =  $2/5 = 40\%$

Support Potato & Milk =  $1/5 = 20\%$

Support Pizza, Cola, Milk =  $1/5 = 20\%$

Confidence Cola -> Pizza = (Support cola & pizza) / Support Cola =  $40/60 = 66\%$

Confidence Pizza -> Cola = (Support pizza & cola) / Support Pizza =  $40/40 = 100\%$

Confidence Potato -> Milk = (Support Potato & Milk) / Support Potato =  $20/20 = 100\%$

Confidence Cola, Piza -> Milk =  $20\%/40\% = 50\%$

Lift Cola And Pizza Then Milk =  $20\% / (40\% * 60\%) = 0.2 / (0.4 * 0.6) = 20/24 < 1$

(kurang bagus)