

IF5181 Pengenalan Pola

Clustering

Masayu Leylia Khodra

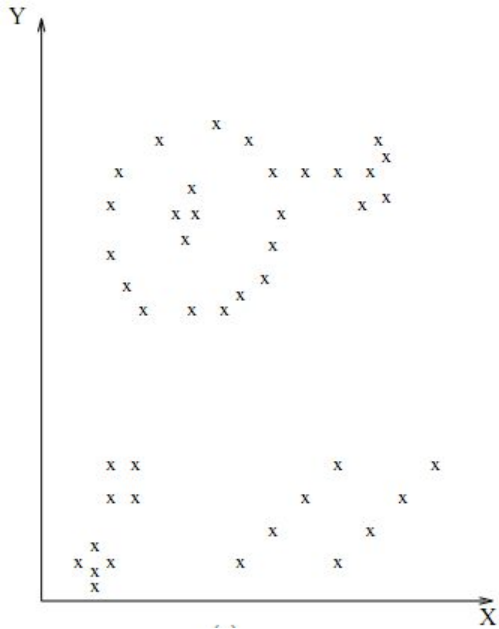
Referensi

- Bab 10 & 11 dari Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
- Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2), 165-193.
- Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., ... & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE transactions on emerging topics in computing*, 2(3), 267-279.
- Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, 664-681.
- A.K. Jain, M.N. Murty, P.J. FLYNN (1999), Data Clustering: A Review. ACM computing surveys
- Pengyu Hong (2005), Introduction to Hierarchical Clustering Analysis
- Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data* (pp. 25-71). Springer, Berlin, Heidelberg.
- Rui Xu, Donald Wunsch (2005), Survey of Clustering Algorithm
- DBSCAN
<http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=72F8D8F33A502FAB448D4C13809D83C3?doi=10.1.1.71.1980&rep=rep1&type=pdf>
- http://www.cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_density.pdf

Outline

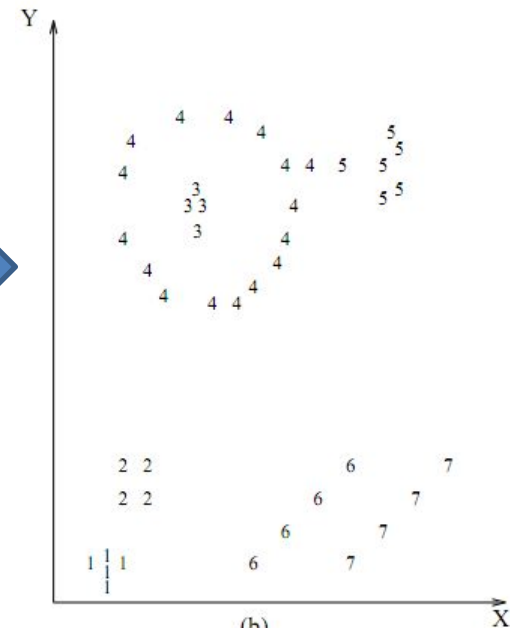
- Clustering: what, why

Clustering: What ?



Input Data

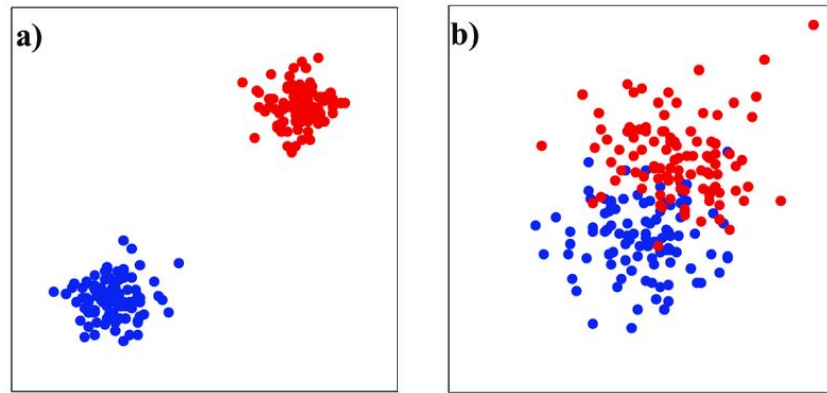
Proses pengelompokan data menjadi clusters berbasis kesamaan data



Desired cluster

Clustering: Finding natural groups

- High intra-cluster similarity/
low intra-cluster variance
 - Data pd cluster yang sama harus semirip mungkin
- Low inter-cluster similarity /
high inter-cluster variance
 - Data pd cluster yang berbeda harus sejauh mungkin
- Pengukuran kemiripan dan jarak harus jelas dan punya semantik praktis (sesuai domain)



a) **Low** intra-class variance and **high** inter-class variance: compact well separated clusters. b) **High** intra-class variance and **low** inter-class variance: wide clusters without a clear frontier.

https://www.researchgate.net/figure/Inter-class-and-Intra-class-variances-concept-a-Low-intra-class-variance-and-high_fig2_278382762

Clustering: Why ?

- Data discovery (cluster = struktur internal data)
 - Contoh: search engine, news aggregator, gen
- Tujuan awalnya partisi / pengelompokan
 - Contoh: segmentasi pasar
- Bagian dari teknik lainnya
 - Contoh: peringkasan berbasis clustering

Why: Clustering pada Search Engine



clustering

Results 1-5 of 5 in Natural language

[Sources](#) [Sites](#) [Time](#) [Topics](#)

Top 284 Results

remix

- Search, Engine (27)
 - + Yippy, Concept Clustering (5)
 - Meta Search (7)
 - **Natural language (5)**
 - Classification, Clustering (3)
 - Theory (2)
 - Relational (3)
 - Demonstration (2)
 - Other Topics (7)
- + Technology (25)
- + Algorithms (26)
- + Cluster Analysis (18)
- + Methods (20)
- + Blog (12)
- + Definition (9)
- + Machine Learning (16)
- + Windows (15)

[Inbenta - Artificial Intelligence | Enterprise Search | Chatbots | Ticketing](#) [new window](#) [preview](#)

... Inbenta Meaning-Text Theory **Natural Language** Processing Semantic **Clustering** & Gap Analysis Schedule a Demo About us Leadership ... find answers? Integrating Inbenta **Natural Language** Technology Semantic **Clustering** The Meaning-Text Theory Resources eBooks Videos Webinars ...

<https://www.inbenta.com/en> - [cache](#) - Yippy Index

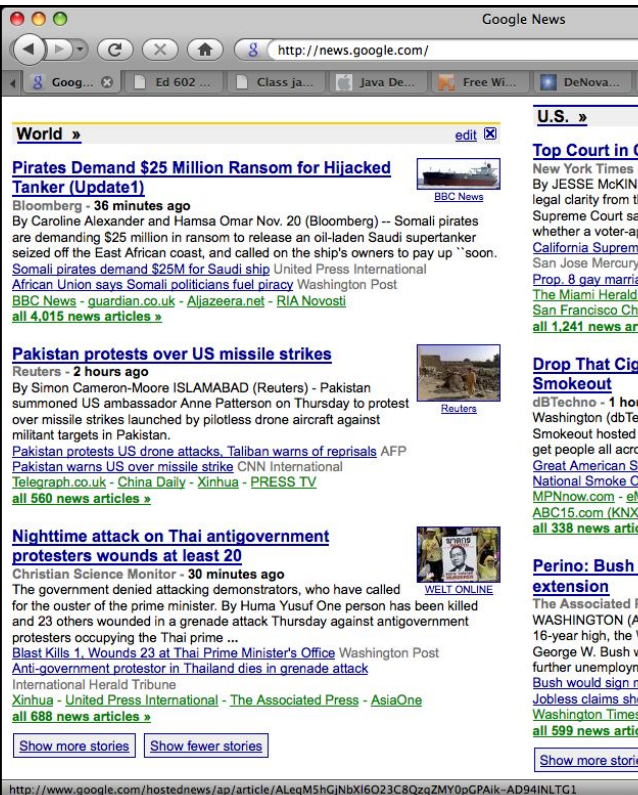
[\(GSA\) Google Search Appliance Replacement | Yippy](#) [new window](#) [preview](#)

... **Search Appliance**, including analytic NLP, email discovery, concept **clustering**, classification, user **search** ranking, tagging and saving. Security ... link analysis, and freshness. Coupled with analytics, concept **clustering**, sentiment analysis, and **natural language** processing makes the ... yippyinc.com/google-search-appliance-replacement - [cache](#) - yippyincweb

[LingPipe Blog | Natural Language Processing and Text Analytics](#) [new window](#) [preview](#)

Struktur internal
hasil pencarian

Why: Clustering pada News Aggregator



The screenshot shows a Google News browser window with the URL <http://news.google.com/>. The page is organized into sections: 'World', 'U.S.', and 'Top Court in C'. Under 'World', there are articles about pirates demanding a \$25 million ransom for a hijacked tanker and Pakistan protesting over US missile strikes. Under 'U.S.', there are articles about a nighttime attack on Thai antigovernment protesters and a drop in cigarette smokeouts. The 'Top Court in C' section features a headline about John McCain's death. The browser's address bar and tabs are visible at the top.

World [edit](#)

Pirates Demand \$25 Million Ransom for Hijacked Tanker (Update1)
Bloomberg - 36 minutes ago
By Caroline Alexander and Hamsa Omar Nov. 20 (Bloomberg) -- Somali pirates are demanding \$25 million in ransom to release an oil-laden Saudi supertanker seized off the East African coast, and called on the ship's owners to pay up "soon."
[Somali pirates demand \\$25M for Saudi ship](#) United Press International
[African Union says Somali politicians fuel piracy](#) Washington Post
[BBC News](#) - [guardian.co.uk](#) - [Aljazeera.net](#) - [RIA Novosti](#)
[all 4,015 news articles](#)

Pakistan protests over US missile strikes
Reuters - 2 hours ago
By Simon Cameron-Moore ISLAMABAD (Reuters) - Pakistan summoned US ambassador Anne Patterson on Thursday to protest over missile strikes launched by pilotless drone aircraft against militant targets in Pakistan.
[Pakistan protests US drone attacks, Taliban warns of reprisals](#) AFP
[Pakistan warns US over missile strike](#) CNN International
[Telegraph.co.uk](#) - [China Daily](#) - [Xinhua](#) - [PRESS TV](#)
[all 560 news articles](#)

Nighttime attack on Thai antigovernment protesters wounds at least 20
Christian Science Monitor - 30 minutes ago
The government denied attacking demonstrators, who have called for the ouster of the prime minister. By Huma Yusuf One person has been killed and 23 others wounded in a grenade attack Thursday against antigovernment protesters occupying the Thai prime ...
[Blast Kills 1, Wounds 23 at Thai Prime Minister's Office](#) Washington Post
[Anti-government protest in Thailand dies in grenade attack](#) International Herald Tribune
[Xinhua](#) - [United Press International](#) - [The Associated Press](#) - [AsiaOne](#)
[all 688 news articles](#)

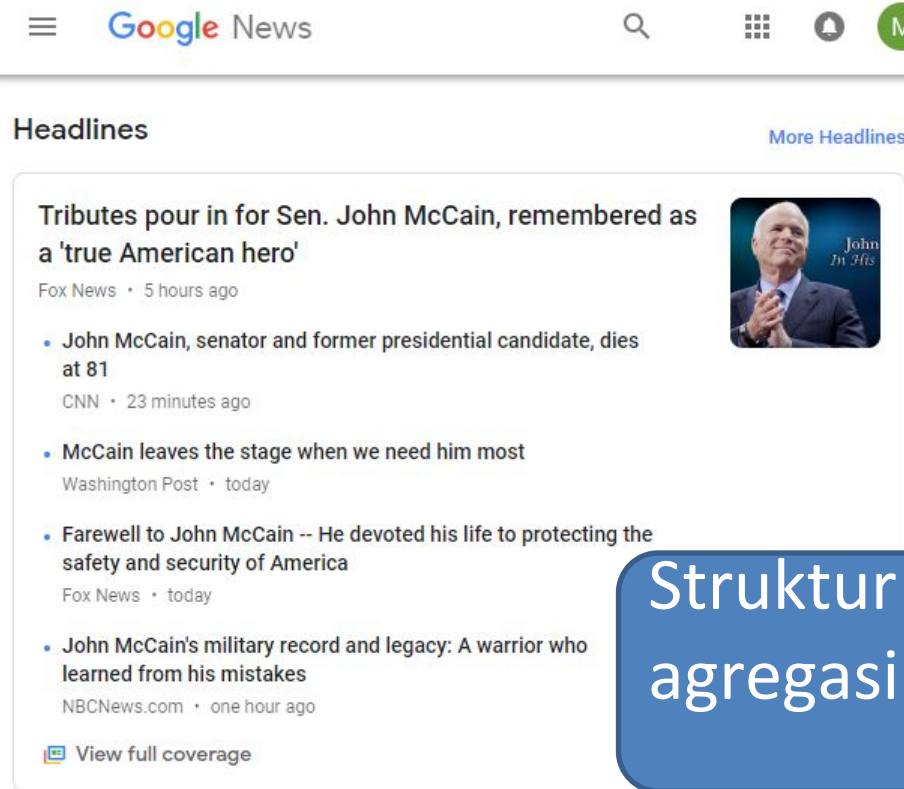
[Show more stories](#) [Show fewer stories](#)

U.S. [»](#)

Drop That Cig Smokeout
dBTechno - 1 hou
Washington (dbTex Smokeout hosted I get people all acro
[Great American Sr National Smoke O](#)
[MPNnow.com](#) - [eN ABC15.com \(KNX\)](#)
[all 338 news artic](#)

Perino: Bush extension
The Associated P
WASHINGTON (A 16-year high, the V George W. Bush w further unemploy
[Bush would sign n Jobless claims shc](#)
[Washington Times](#)
[all 599 news artic](#)

[Show more storie](#)



The screenshot shows a Google News page with the Google News logo and a search bar. The main headline is 'Tributes pour in for Sen. John McCain, remembered as a 'true American hero''. Below the headline, there are several bullet points listing news stories about John McCain's death and legacy. A small image of John McCain is visible on the right side of the page. The page is organized into sections: 'Headlines' and 'More Headlines'.

Headlines [More Headlines](#)

Tributes pour in for Sen. John McCain, remembered as a 'true American hero'
Fox News • 5 hours ago

- John McCain, senator and former presidential candidate, dies at 81
CNN • 23 minutes ago
- McCain leaves the stage when we need him most
Washington Post • today
- Farewell to John McCain -- He devoted his life to protecting the safety and security of America
Fox News • today
- John McCain's military record and legacy: A warrior who learned from his mistakes
NBCNews.com • one hour ago

[View full coverage](#)

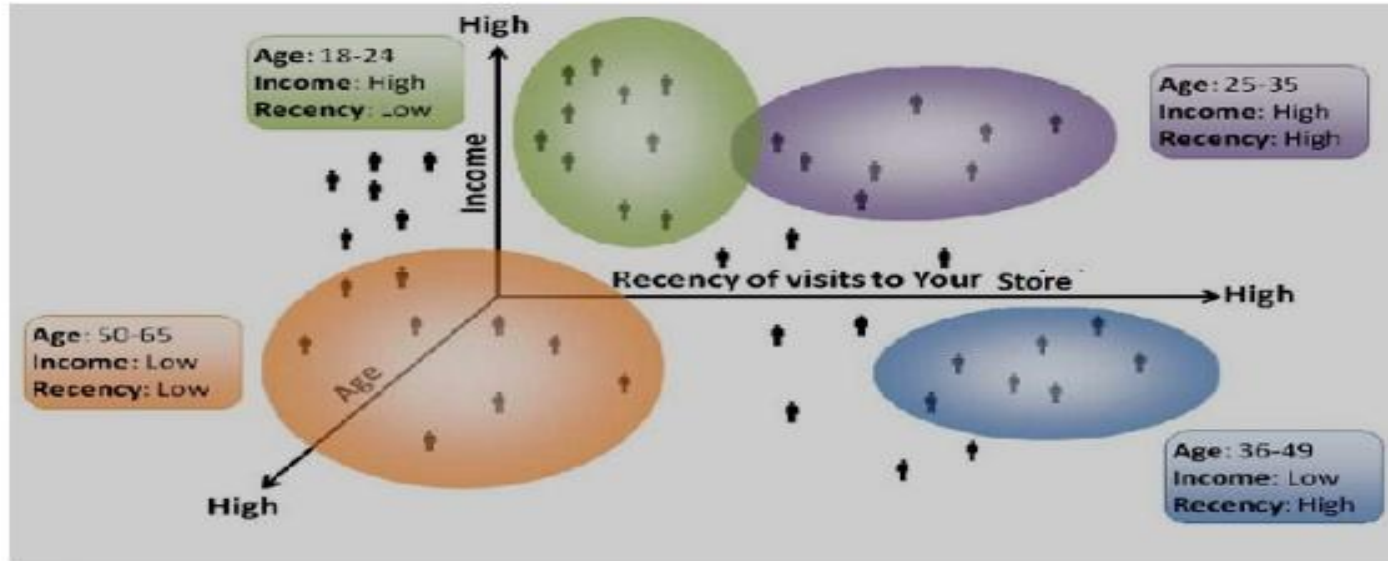
Struktur internal agregasi berita

http://genome-www.stanford.edu/sarcoma/supplemental_data.html



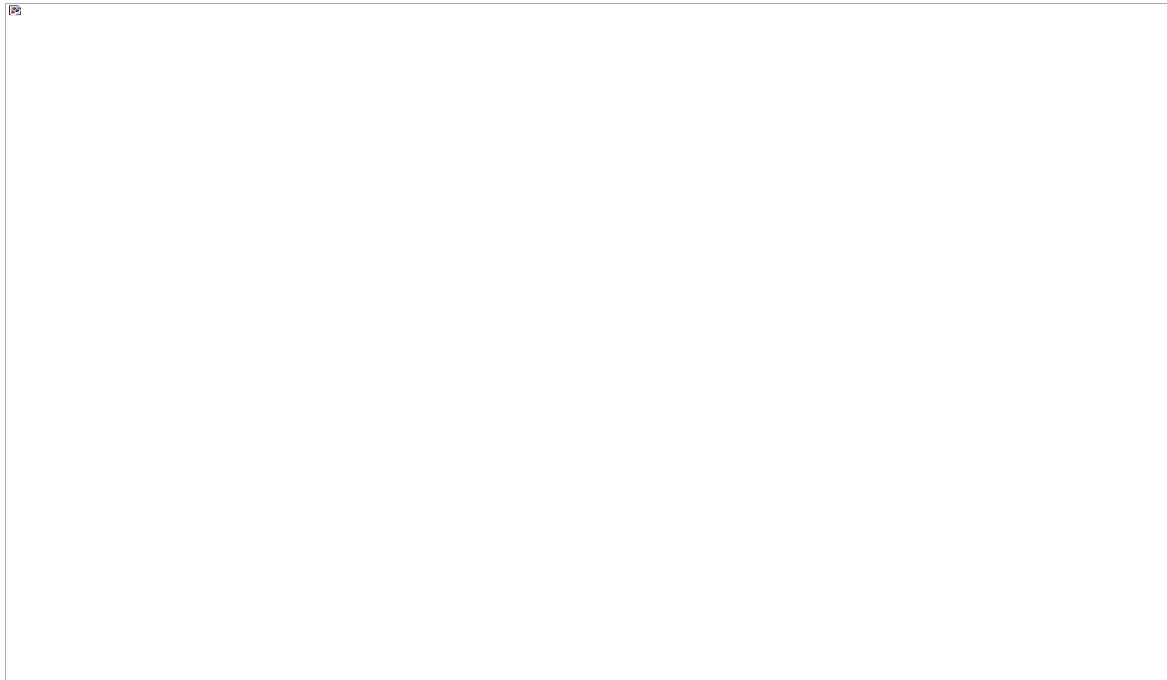
Why: Clustering untuk Segmentasi

Example - Clusters using Age, Income & Recency



Copyright: Canvass 2013-2016

Why: Clustering-based Approach

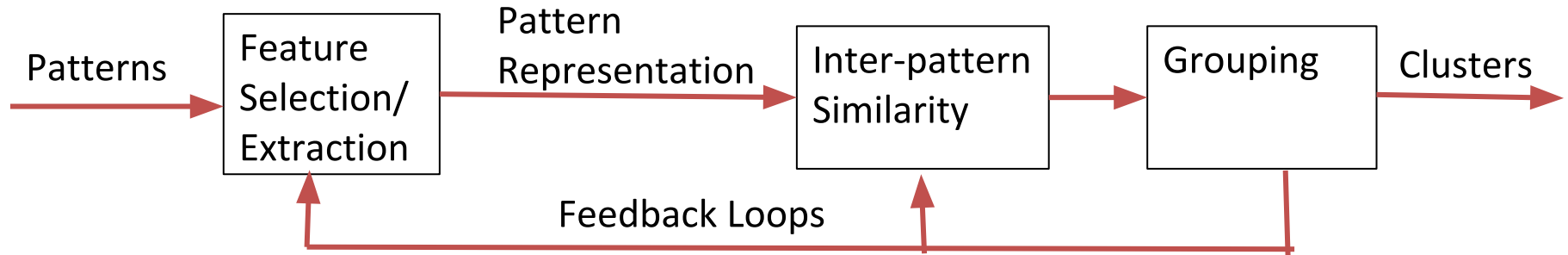


Clustering-based
summarization

Clustering-based
outlier detection

Clustering-based
analysis

Tahapan Clustering



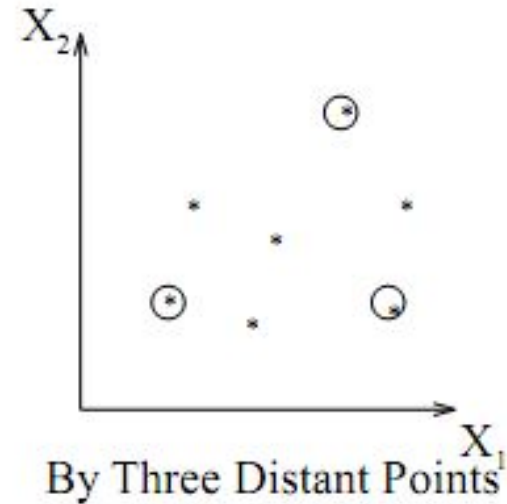
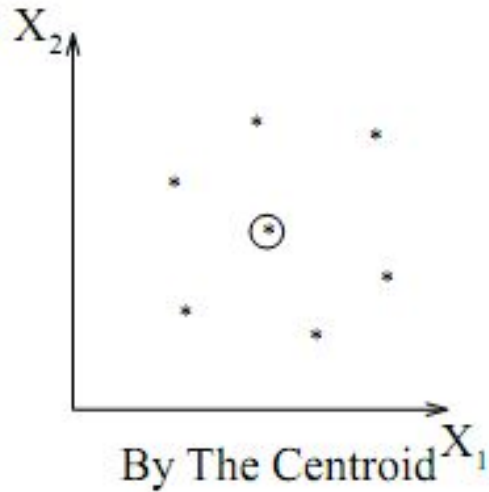
- Tahapan utama:
 - 1) Feature selection: original features → subset of features
Feature extraction: transformation into new features
 - 2) pattern proximity/similarity measure
 - 3) Grouping
- Clustering output: hard atau soft (membership degree)

Tahapan Clustering (lanjutan)

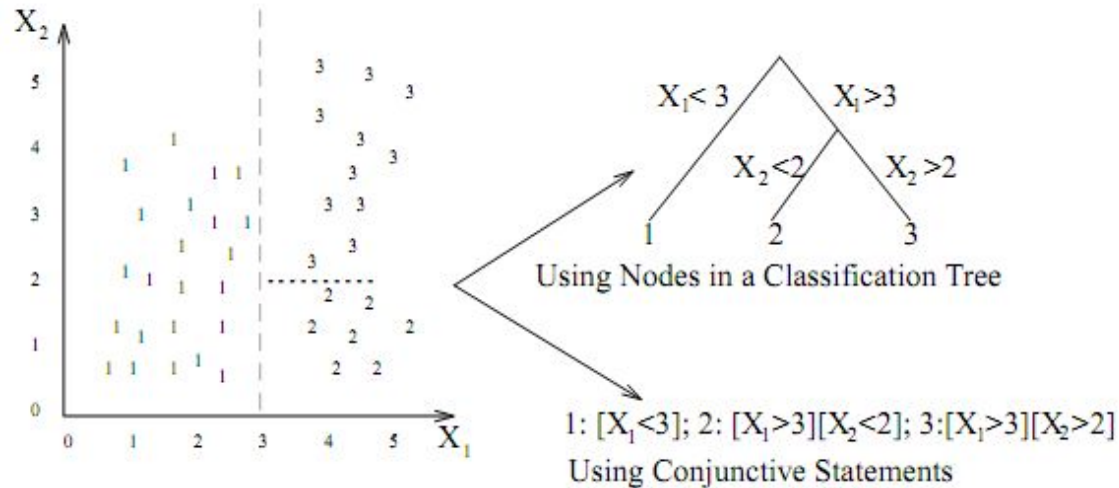
- Tahapan opsional:
 - 4) data abstraction
 - 5) assessment of output (good or poor)

Representasi Cluster (1)

- Centroid atau set of distant point



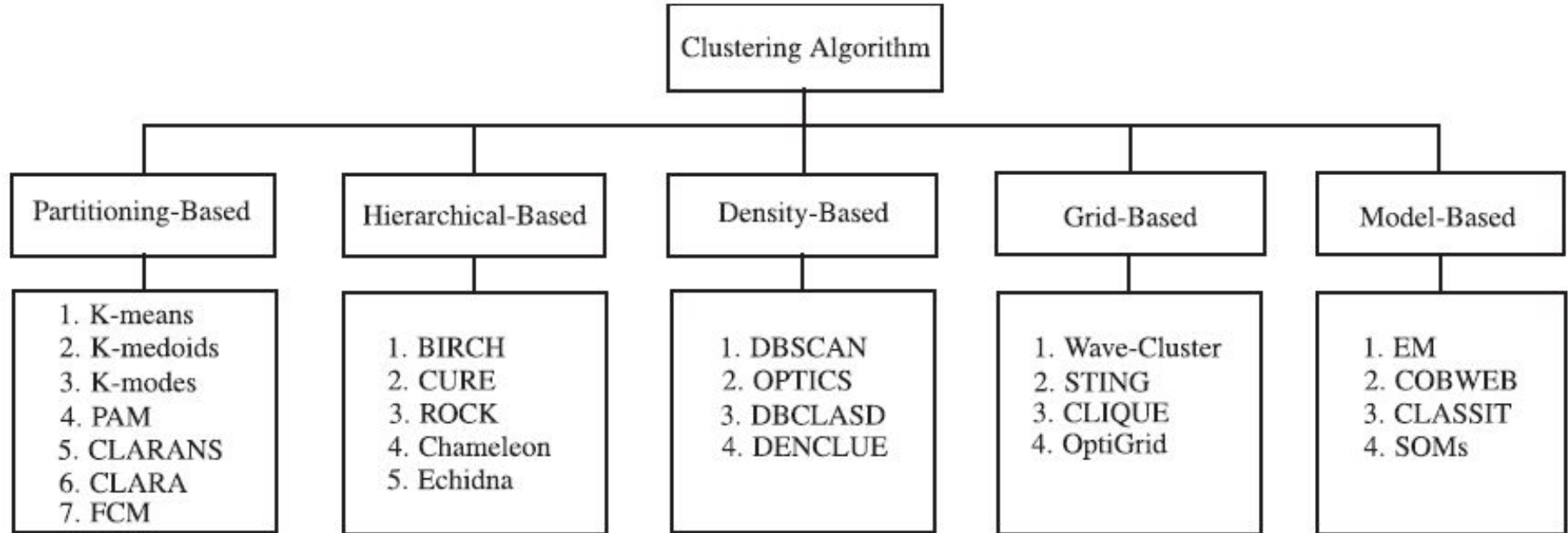
Representasi Cluster (2)



- Pohon klasifikasi
- Conjunctive statements

Kategori Metode Clustering

(Fahad, 2014; Han & Kamber, 2006)



Kategori Metode Clustering

(Han & Kamber, 2006)

1. Metode *partitioning*

- mengidentifikasi partisi yang mengoptimalkan kriteria pengelompokan (squared error, absolute error)
- Konstruksi k-partisi data (partisi \sim cluster); $k \leq$ jumlah data
- Contoh: K-means, k-medoids

2. Metode *hierarchical*

- menghasilkan rangkaian partisi bersarang
- Agglomerative (bottom-up, merge):
1 object \sim 1 cluster \rightarrow 1 cluster n-object
- Divisive (top-down, split):
1 cluster n-object \rightarrow 1 object \sim 1 cluster

Kategori Metode Clustering (lanj)

(Han & Kamber, 2006)

3. Metode berbasis density

- Densitas: jumlah objek
- Contoh: DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

4. Metode berbasis grid

- Struktur grid, cepat, bergantung jumlah sel, tidak dipengaruhi jumlah objek, perhitungan bisa dilakukan secara paralel
- Contoh: STING (STatistical INformation Grid)

5. Metode berbasis model

- Contoh: EM (Expectation-Maximization), SOM (self-organizing map)

Evaluasi Cluster: Purity

Semakin besar nilai purity, semakin baik solusi clustering yang dihasilkan.

- Purity untuk setiap cluster C_r dengan ukuran n_r

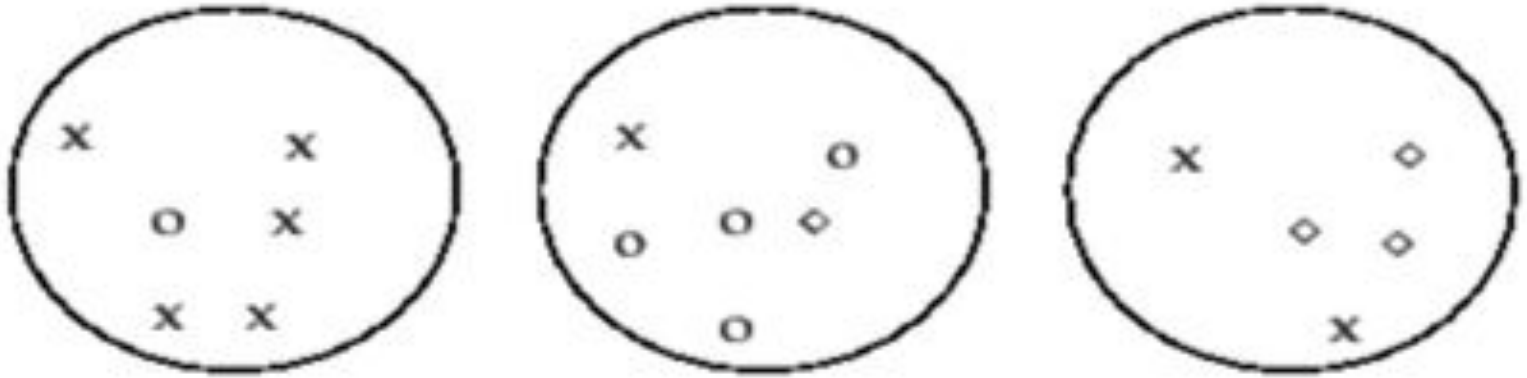
$$P(C_r) = \frac{1}{n_r} \max_i n_r^i$$

- Purity dari keseluruhan clustering

$$Purity(C) = \sum_{r=1}^k \frac{n_r}{n} P(C_r)$$

$$\text{or} = \frac{1}{n} \sum_{r=1}^k \max_i (n_r^i)$$

Purity: Contoh



► $\text{Purity} = 1/17 * (5 + 4 + 3) \approx 0.71$

Tugas 2

- Gunakanlah dataset yang sama dengan Tugas1
- Tools: Jupyter Notebook dengan python. Gunakan library untuk konstruksi model clustering.
- Skenario:
 - Gunakanlah hasil split dataset (train:test) dari tugas1
 - Lakukanlah clustering training data , hitunglah purity dengan menggunakan. Pilihlah satu algoritma dari setiap kategori algoritma clustering.
 - Gunakanlah test data untuk mendapatkan akurasi dari setiap algoritma clustering.
- Lakukanlah analisis hasil testing.
- Tugas dikumpulkan Rabu 25 September 2019 jam 23.55