# BÀI GIẢNG KHAI PHÁ DỮ LIỆU

# CHƯƠNG 4. PHÂN CỤM DỮ LiỆU

TS. Trần Mai Vũ
HÀ NỘI, 2020
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**
ĐẠI HỌC QUỐC GIA HÀ NỘI

# Nội dung

Giới thiệu bài toán phân cụm
Một số độ đo cơ bản cho phân cụm
Phân cụm phẳng
Phân cụm phân cấp
Phân cụm dựa trên mật độ
Phân cụm dựa trên mô hình
Gán nhãn cụm
Đánh giá phân cụm

**Charu C. Aggarwal, Chandan K. Reddy. *Data Clustering: Algorithms and Applications*. CRC Press 2014.**

**Israël César Lerman. *Foundations and Methods in Combinatorial and Statistical Data Analysis and Clustering*. Springer-Verlag London, 2016**

# 1. Giới thiệu bài toán phân cụm

- Bài toán
  - Tập dữ liệu D = $\{d_i\}$
  - Phân các dữ liệu thuộc D thành các cụm
    - Các dữ liệu trong một cụm: “tương tự” nhau (gần nhau)
    - Dữ liệu hai cụm: “không tương tự” nhau (xa nhau)
  - Đo “tương tự” (gần) nhau ?
    - *Tiên đề phân cụm:* Nếu người dùng lựa chọn một đối tượng *d* thì họ cũng lựa chọn các đối tượng cùng cụm với *d*
    - Khai thác “cách chọn lựa” của người dùng
    - Đưa ra một số độ đo “tương tự” theo biểu diễn dữ liệu
- Một số nội dung liên quan
  - Xây dựng độ đo tương tự
  - Khai thác thông tin bổ sung
  - Số lượng cụm cho trước, số lượng cụm không cho trước

# Sơ bộ tiếp cận phân cụm

- Phân cụm mô hình và phân cụm phân vùng
  - Mô hình: Kết quả là mô hình biểu diễn các cụm dữ liệu
  - Vùng: Danh sách cụm và vùng dữ liệu thuộc cụm
- Phân cụm đơn định và phân cụm xác suất
  - Đơn định: Mỗi dữ liệu thuộc duy nhất một cụm
  - Xác suất: Danh sách cụm và xác suất một dữ liệu thuộc vào các cụm
- Phân cụm phẳng và phân cụm phân cấp
  - Phẳng: Các cụm dữ liệu không giao nhau
  - Phân cấp: Các cụm dữ liệu có quan hệ phân cấp cha- con
- Phân cụm theo lô và phân cụm tăng
  - Lô: Tại thời điểm phân cụm, toàn bộ dữ liệu đã có
  - Tăng: Dữ liệu tiếp tục được bổ sung trong quá trình phân cụm

# Các phương pháp phân cụm

- Các phương pháp phổ biến
  - Phân vùng, phân cấp, dựa theo mật độ, dựa theo lưới, dựa theo mô hình, và phân cụm mờ
- Phân cụm phân vùng (phân cụm phẳng)
  - Xây dựng từng bước phân hoạch các cụm và đánh giá chúng theo các tiêu chí tương ứng
  - Tiếp cận: từ dưới lên (gộp dần), từ trên xuống (chia dần)
  - Độ đo tương tự / khoảng cách
  - K-mean, k-mediod, CLARANS, …
  - Hạn chế: Không điều chỉnh được lỗi
- Phân cụm phân cấp
  - Xây dựng hợp (tách) dần các cụm tạo cấu trúc phân cấp và đánh giá theo các tiêu chí tương ứng
  - Độ đo tương tự / khoảng cách
  - HAC: Hierarchical agglomerative clustering
  - CHAMELEON, BIRRCH và CURE, …

# Các phương pháp phân cụm

- Phân cụm dựa theo mật độ
  - Hàm mật độ: Tìm các phần tử chính tại nơi có mật độ cao
  - Hàm liên kết: Xác định cụm là lân cận phần tử chính
  - DBSCAN, OPTICS…
- Phân cụm dựa theo lưới
  - Sử dụng lưới các ô cùng cỡ: tuy nhiên cụm là các “ô” phân cấp
  - Tạo phân cấp ô lưới theo một số tiêu chí: số lượng đối tượng trong ô
  - STING, CLIQUE, WaweCluster…
- Phân cụm dựa theo mô hình
  - Giải thiết: Tồn tại một số mô hình dữ liệu cho phân cụm
  - Xác định mô hình tốt nhất phù hợp với dữ liệu
  - MCLUST…
- Phân cụm mờ
  - Giả thiết: không có phân cụm “cứng” cho dữ liệu và đối tượng có thể thuộc một số cụm
  - Sử dụng hàm mờ từ các đối tượng tới các cụm
  - FCM (Fuzzy CMEANS),…

# 2. Một số độ đo cơ bản

- Độ đo tương đồng
  - Biểu diễn: vector n chiều
  - Giá trị nhị phân: Ma trận kề, độ đo Jaccard
  - Giá trị rời rạc [0,m]: Chuyển m giá trị thành nhị phân, độ đo Jaccard
  - Giá trị thực : độ đo cosin hai vector
- Độ đo khác biệt
  - Đối ngẫu độ đo tương đồng
  - Thuộc tính nhị phân: đối cứng, không đối xứng
  - Giá trị rời rạc: hoặc tương tự trên hoặc dạng đơn giản (q thuộc tính giống nhau)
  - Giá trị thực: Khoảng cách Manhattan, Euclide, Mincowski
  - Tính xác định dương, tính đối xứng, tính bất đẳng thức tam giác

| Phần tử dữ liệu $p_2$ | Phần tử dữ liệu $p_1$ | | Tổng |
|---|---|---|---|
| | 1 | 0 | |
| 1 | $a$ | $b$ | $a+b$ |
| 0 | $c$ | $d$ | $c+d$ |
| Tổng | $a+c$ | $b+d$ | $a+b+c+d$ |

$$Jaccard(p_1, p_2) = \frac{a}{a+b+c}$$

$$\cos in(p_1, p_2) = \frac{p_1 \bullet p_2}{\|p_1\| \|p_2\|} = \frac{\sum_{i=1}^{n} p_{1i} p_{2i}}{\sqrt{\sum_{i=1}^{n} p_{1i}^2} \sqrt{\sum_{i=1}^{n} p_{2i}^2}}$$

$$d(p_1, p_2) = \frac{b+c}{a+b+c+d} \qquad d(p_1, p_2) = \frac{b+c}{a+b+c}$$

$$d(p_1, p_2) = \frac{n-q}{n}$$

$$d(p_1, p_2) = \sum_{i=1}^{n} | p_{1i} - p_{2i} | \qquad d(p_1, p_2) = \sqrt{\sum_{i=1}^{n} | p_{1i} - p_{2i} |^2}$$

$$d(p_1, p_2) = \sqrt[q]{\sum_{i=1}^{n} | p_{1i} - p_{2i} |^q}$$

# Một số độ đo cơ bản

- Ví dụ về độ khác biệt
  - CSDL xét nghiệm bệnh nhân
  - Quy về giá trị nhị phân: M/F, Y/N, N/P
  - Lập ma trận khác biệt cho từng cặp đối tượng.
  - Ví dụ, cặp (Nam, Vân): a=2, b=1, c=1, d=3

D(Nam, Vân) =(1+1)/(2+1+1)=0.5

**Bảng 5.2 Bảng kết quả xét nghiệm**

| *No* | *Tên* | *Giới tính* | *Chóng mặt* | *Ho* | *XN1* | *XN2* | *XN3* | *XN4* |
|---|---|---|---|---|---|---|---|---|
| 1 | Nam | M | Y | N | P | N | N | N |
| 2 | Vân | F | Y | N | P | N | P | N |
| 3 | Thắng | M | Y | P | N | N | N | N |

| *Phần tử dữ liệu $p_2$* | *Phần tử dữ liệu $p_1$* | | *Tổng* |
|---|---|---|---|
| | 1 | 0 | |
| 1 | *a* | *b* | *a+b* |
| 0 | *c* | *d* | *c+d* |
| Tổng | *a+c* | *b+d* | *a+b+c+d* |

$$d(p_1, p_2) = \frac{b+c}{a+b+c}$$

# 3. Thuật toán K-mean gán cứng

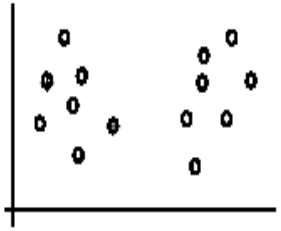
1. Khởi động: Chọn ngẫu nhiên *k* dữ liệu trong *S* làm trọng tâm (đại diện) cho các cụm $S_i = \{c_i: c_i \in S\}, \forall i=1,...,k$
2. Bước lặp:
   - 2.1. $S_i = \varnothing$ // Các cụm mới là rỗng
   - 2.2. $\forall d \in S$:
     - 2.2.1. Tính sim $(d, c_i), \forall i=1,...,k$
     - 2.2.2. $S_t = S_t \cup \{d\}$ nếu sim $(d, c_t) = \max \{\text{sim}(d, c_i) | i=1,...,k\}$
   - 2.3. $\forall i=1,...,k$ , tính lại trọng tâm các cụm $S_i$ : $c_i = \frac{1}{\|S_i\|} \sum_{d \in S_i} d$
3. Nếu chưa gặp Điều kiện dừng thì quay lại bước 2, ngược lại Dừng

- Một số lưu ý
  - Điều kiện dừng
    - Sau bước 2 không có sự thay đổi cụm
    - Điều kiện dừng cưỡng bức
      - Khống chế số lần lặp
      - Giá trị mục tiêu đủ nhỏ

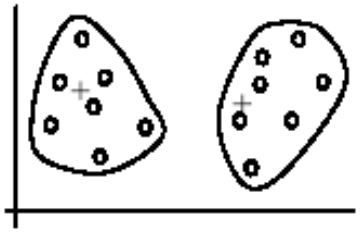
$$E = \sum_{C} \sum_{p \in C} | p - m_C |^2$$

  - Vấn đề chọn tập đại diện ban đầu ở bước Khởi động
  - Có thể dùng độ đo khoảng cách thay cho độ đo tương tự

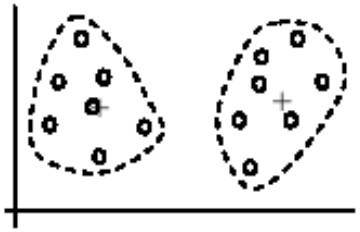
# a. Thuật toán K-mean gán cứng



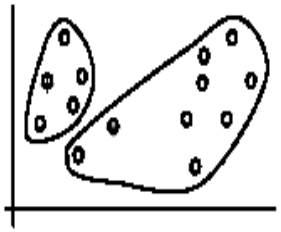
(A). Random selection of *k* centers



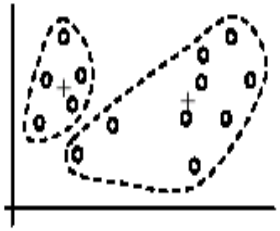
*Iteration* 2: (D). Cluster assignment



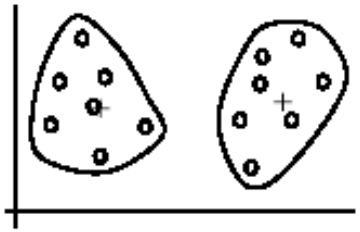
(E). Re-compute centroids



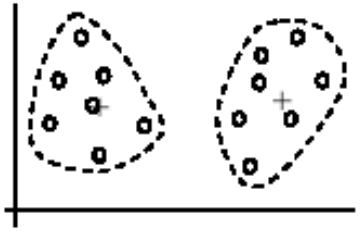
*Iteration* 1: (B). Cluster assignment



(C). Re-compute centroids



*Iteration* 3: (F). Cluster assignment



(G). Re-compute centroids

- Một số lưu ý (tiếp) và ví dụ
  - Trong bước 2: các trọng tâm có thể không thuộc S
  - Thực tế: số lần lặp  50
  - Thi hành k-mean với dữ liệu trên đĩa
    - Toàn bộ dữ liệu quá lớn: không thể ở bộ nhớ trong
    - Với mỗi vòng lặp: duyệt CSDL trên đĩa 1 lần
      - Tính được độ tương tự của d với các $c_i$.
      - Tính lại $c_i$ mới: bước 2.1 khởi động (tổng, bộ đếm); bước 2.2 cộng và tăng bộ đếm; bước 2.3 chỉ thực hiện k phép chia.



Bing Liu (2007), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, *Spinger*, 2007.

# Thuật toán K-mean mềm

- Input
  - Số nguyên k > 0: số cụm biết trước
  - Tập dữ liệu D (cho trước)
- Output
  - Tập k "đại diện cụm" $_c$ làm cực tiểu lỗi "lượng tử" $\sum_d \min_c |d - \mu_c|^2$
- Định hướng
  - Tinh chỉnh $_c$ dần với tỷ lệ học (learning rate) $\mu_c \leftarrow \mu_c + \Delta\mu_c$

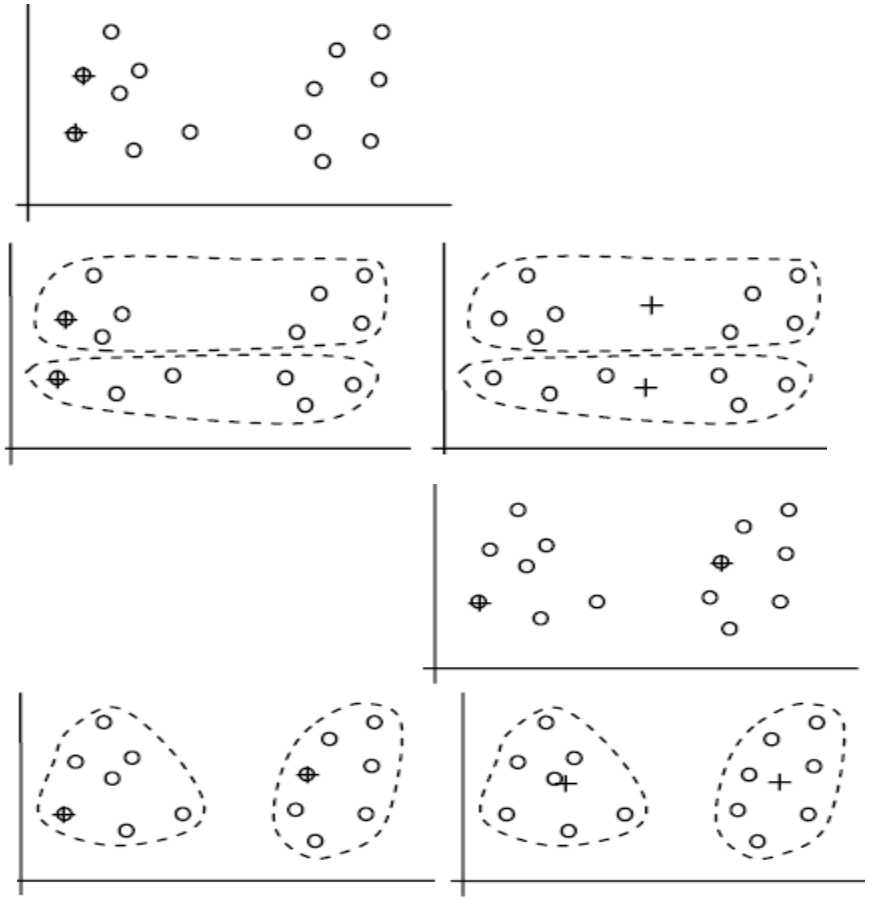
$$\Delta\mu_c = \sum_d \begin{cases} \eta(d - \mu_c) & \text{nếu } \mu_c \text{ gần } d \text{ nhất} \\ 0 & \text{các trường hợp khác} \end{cases}$$

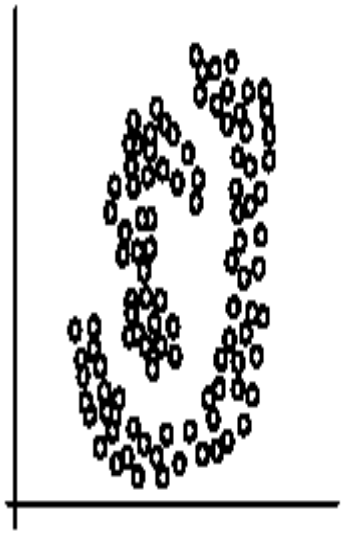
$$\Delta\mu_c = \eta \frac{1/|d-\mu_c|^2}{\sum_\gamma 1/|d-\mu_\gamma|^2}(d-\mu_c) \qquad \Delta\mu_c = \eta \frac{\exp(-|d-\mu_c|^2)}{\sum_\gamma \exp(-|d-\mu_\gamma|^2)}(d-\mu_c)$$

# Thuật toán K-mean

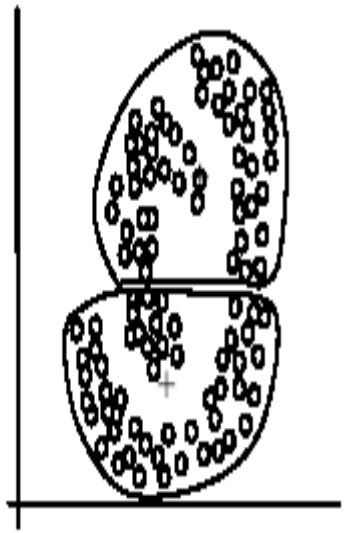
- Ưu điểm
  - Đơn giản, dễ sử dụng
  - Hiệu quả về thời gian: tuyến tính O(tkn), t số lần lặp, k số cụm, n là số phần tử
  - Một thuật toán phân cụm phổ biến nhất
  - Thường cho tối ưu cục bộ. Tối ưu toàn cục rất khó tìm
- Nhược điểm
  - Phải "tính trung bình được": dữ liệu phân lớp thì dựa theo tần số
  - Cần cho trước k : số cụm
  - Nhạy cảm với ngoại lệ (cách xa so với đại đa số dữ liệu còn lại): ngoại lệ thực tế, ngoại lệ do quan sát sai (làm sạch dữ liệu)
  - Nhạy cảm với mẫu ban đầu: cần phương pháp chọn mẫu thô tốt
  - Không thích hợp với các tập dữ liệu không siêu-ellip hoặc siêu cầu (các thành phần con không ellip/cầu hóa)

# Thuật toán K-mean





(A): Two natural clusters



(B): *k*-means clusters

*Trái*: Nhạy cảm với chọn mẫu ban đầu

*Phải*: Không thích hợp với bộ dữ liệu không siêu ellip/cầu hóa

Bing Liu (2007), Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, *Spinger*, 2007.

# b. Thuật toán PAM (K-mediod)

- K-mediod
  - Biến thể của K-mean: thay trọng tâm bằng một phần tử của D
  - Hàm mục tiêu
  - PAM: Partition Around Mediods

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - o_i |^2$$

- Input và Output
  - Input: D = {d} tập dữ liệu, độ đo tương tự sim, k>0
  - Output: Tập các cụm của D
- Thuật toán PAM
  1. Chọn ngẫu nhiên k phần từ trong D làm đại diện $c_i$.
  2. Gán các d D vào $C_d$ mà d tương tự $c_d$ nhất trong các $c_i$.
  3. Chọn ngẫu nhiên phần từ o không phải là đại diện cụm $c_j$.
  4. Tính hàm chi phí  (gia số hàm mục tiêu) nếu thay $c_i$ bằng o
  5. Nếu  <0 thay $c_i$ bằng o.
  6. Quy lại bước 2 cho đến khi quá trình hội tụ (không còn thay thế phần tử đại diện được).

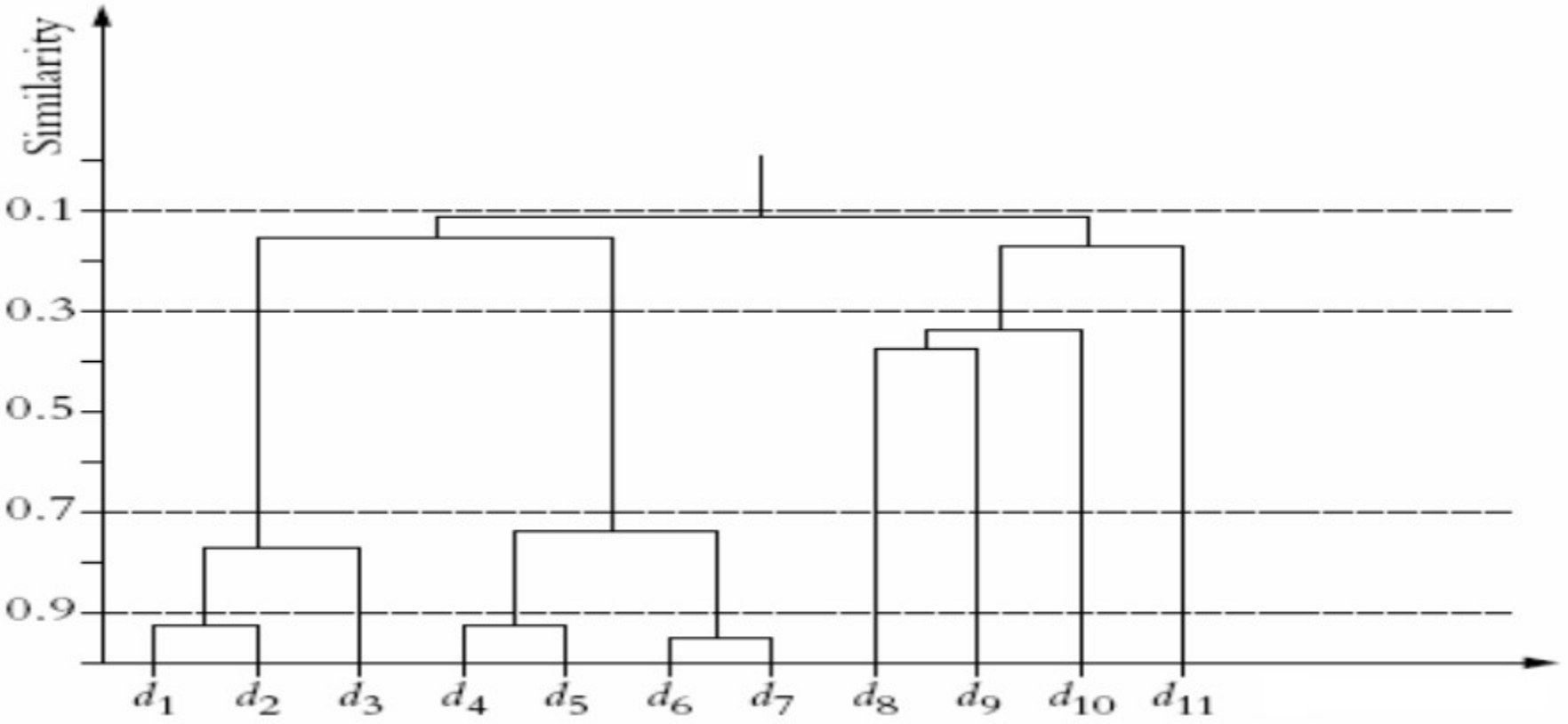
# 4. Phân cụm phân cấp

- HAC: Hierarchical agglomerative clustering
- Một số độ đo phân biệt cụm
  - Độ tương tự hai tài liệu
  - Độ tương tư giữa hai cụm
    - Độ tương tự giữa hai đại diện
    - Độ tương tự cực đại giữa hai dữ liệu thuộc hai cụm: single-link
    - Độ tương tự cực tiểu giữa hai dữ liệu thuộc hai cum: complete-link
    - Độ tương tự trung bình giữa hai dữ liệu thuộc hai cum
- Sơ bộ về thuật toán
  - Đặc điểm: Không cho trước số lượng cụm k, cho phép đưa ra các phương án phân cụm theo các giá trị k khác nhau
  - Lưu ý: k là một tham số  “tìm k tốt nhất”
  - Tinh chỉnh: Từ cụ thể tới khái quát

# a. Phân cụm phân cấp từ dưới lên

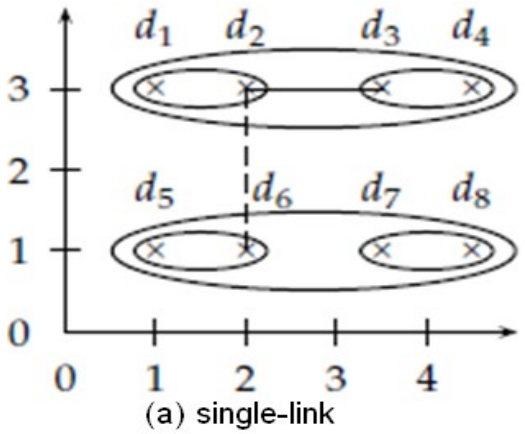
- Input và Output
  - Input: D = {d} tập dữ liệu, độ đo tương tự sim và có thể k>0 và q>0
  - Output: G: Tập các cụm phân cấp của D
- Thuật toán
  1. G  {{d}| d  D} // khởi tại G là tập các cụm chỉ một dữ liệu
  2. Nếu |G|<k thì dừng //đủ lượng cụm tối thiểu
  3. Tìm hai cụm $S_i$ và $S_j$ sao cho (I, j) = arg max (u,v) sim ($S_{u+,}$ $S_{v+)}$ // tìm hai cụm tương tự nhau nhất
  4. Nếu sim($S_i$, $S_j$)<q thì dừng //độ tương tự các cụm quá bé
  5. Loại bỏ $S_i$, $S_j$ khỏi G
  6. G  G  ($S_i$ $S_j$)
  7. Quay lại bước 2
- Giải thích
  - G là tập các cụm trong phân cụm
  - Điều kiện |G| < k có thể thay thế bằng |G|=1

# Phân cụm phân cấp từ dưới lên

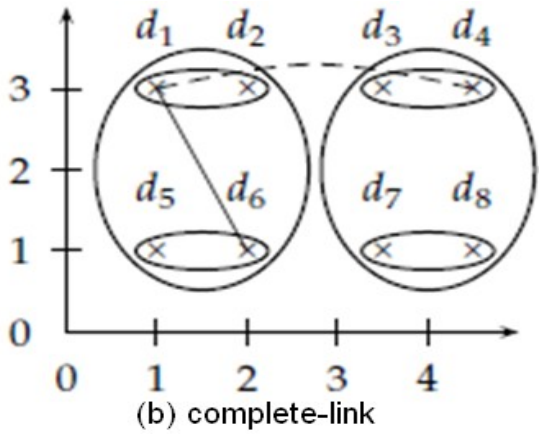



- Hoạt động HAC
  - Cho phép với mọi k
  - Chọn phân cụm theo “ngưỡng” về độ tương tự

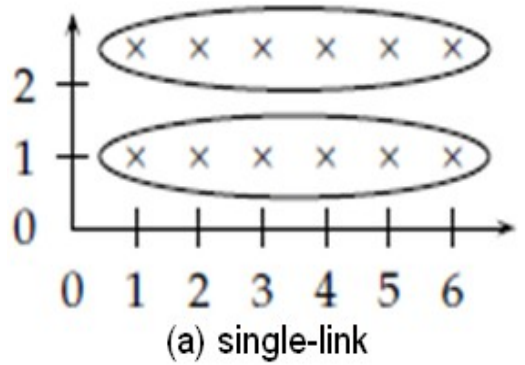
# HAC với các độ đo khác nhau



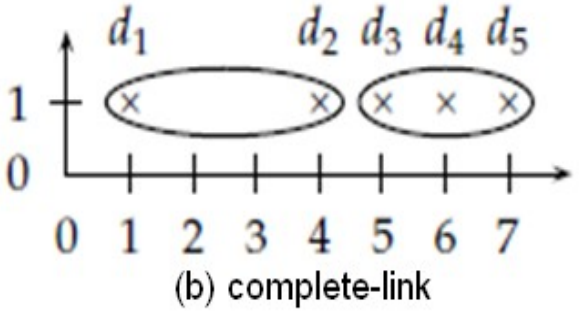
(a) single-link



(b) complete-link



(a) single-link



(b) complete-link

- Ảnh hưởng của các độ đo
  - Trên: Hoạt động thuật toán khác nhau theo các độ đo khác nhau: độ tương tự cực tiểu (complete-link) có tính cầu hơn so với cực đại
  - Dưới: Độ tương tự cực đại (Single-link) tạo cụm chuỗi dòng

# b. Phân cụm phân cấp BIRCH

- Balanced Iterative Reducing Clustering Using Hierarchies
  - Tính khả cỡ: Làm việc với tập dữ liệu lớn
  - Tính bất động: Gán không đổi đối tượng –> cụm
- Khái niệm liên quan
  - Đặc trưng phân cụm CF: tóm tắt của cụm
    - CF = <n, LS, SS>, n: số phần tử, LS: vector tổng các thành phần dữ liêu; SS : vector tổng bình phương các thành phần các đối tượng
    - <3, (9,10), (29,38)>. Khi ghép cụm không tính lại các tổng
  - Cây đặc trưng phân cụm CF Tree
    - Một cây cân bằng
    - Hai tham số: bề rộng b và ngưỡng t
    - Thuật toán xây dựng cây

# BIRCH: Năm độ đo khoảng cách

$$D0 = ((\vec{X0}_1 - \vec{X0}_2)^2)^{\frac{1}{2}}$$

$$D1 = |\vec{X0}_1 - \vec{X0}_2| = \sum_{i=1}^{d} |\vec{X0}_1^{(i)} - \vec{X0}_1^{(i)}|$$

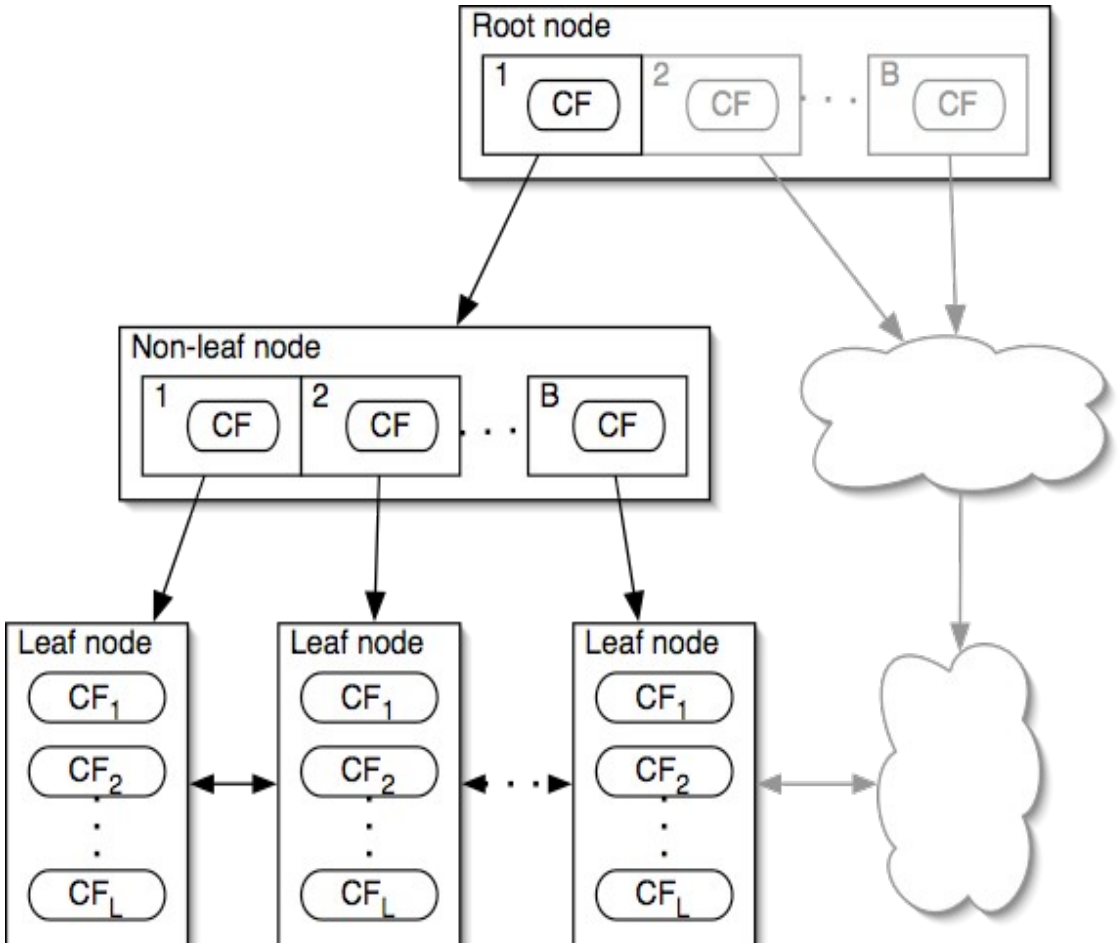
$$D2 = \left(\frac{\sum_{i=1}^{N_1}\sum_{j=N_1+1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{N_1 N_2}\right)^{\frac{1}{2}}$$

$$D3 = \left(\frac{\sum_{i=1}^{N_1+N_2}\sum_{j=1}^{N_1+N_2}(\vec{X}_i - \vec{X}_j)^2}{(N_1+N_2)(N_1+N_2-1)}\right)^{\frac{1}{2}}$$

$$D4 = \left(\sum_{k=1}^{N_1+N_2}\left(\vec{X}_k - \frac{\sum_{l=1}^{N_1+N_2}\vec{X}_l}{N_1+N_2}\right)^2 - \sum_{i=1}^{N_1}\left(\vec{X}_i - \frac{\sum_{l=1}^{N_1}\vec{X}_l}{N_1}\right)^2 - \sum_{j=N_1+1}^{N_1+N_2}\left(\vec{X}_j - \frac{\sum_{l=N_1+1}^{N_1+N_2}\vec{X}_l}{N_2}\right)^2\right)^{\frac{1}{2}}$$

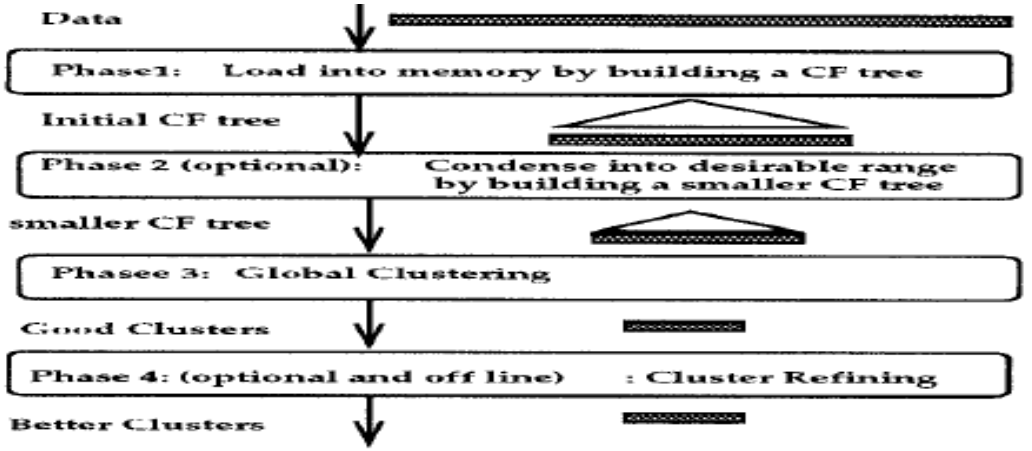
# Cây đặc trưng phân cụm CF Tree

- Mỗi nút không là lá có nhiều nhất là B cành
- Mỗi nút lá có nhiều nhất L đặc trưng phân cụm mà đảm bảo ngưỡng T
- Cỡ của nút được xác định bằng số chiều không gian dữ liệu và tham số P kích thước trang bộ nhớ



# Chèn vào CF Tree và BIRCH

- Cây ban đầu rỗng
- Chèn một "cụm" a vào cây
  - Xác định lá thích hợp: Duyệt từ gốc xuống một cách đệ quy để tới nút con gần a nhất theo 1 trong 5 khoảng cách nói trên
  - Biến đổi lá: Nếu gặp lá $L_1$ gần a nhất, kiểm tra xem $L_1$ có "hấp thụ" a không (chưa vượt ngưỡng); nếu có thì đặc trưng CF của $L_1$ bổ sung; Nếu không, tạo nút mới cho a; nếu không đủ bộ nhớ cho lá mới thì cần chia lá cũ
  - Biến đổi đường đi tới lá khi bổ sung phần tử mới
  - Tinh chỉnh việc trộn:




Tian Zhang, Raghu Ramakrishnan, Miron Livny (1996). BIRCH: An Efficient Data Clustering Method for Very Large Databases, *SIGMOD Conference 1996*: 103-114

# Các thuật toán phân cụm khác

- Nghiên cứu giáo trình

- Phân cụm phân cấp từ trên xuống DIANA
  - Đối ngẫu phân cụm phân cấp từ trên xuống: phần tử khác biệt -> cụm khác biệt S,
  - Thêm vào S các phần tử có d > 0

$$d_i = average(\sum_{x_j \notin S} | x_i - x_j |) - average(\sum_{x_j \in S} | x_i - x_j |)$$

- Phân cụm phân cấp ROCK
  - RObust Clustering using linKs: xử lý dữ liệu rời rạc, quyết định “gần” theo tập phần tử láng giềng sim (p, q) > >0.

- Phân cụm dựa trên mật độ DBSCAN
  - Density-Based Spatial Clustering of Application with Noise
  - #-neighborhood: vùng lân cận bán kính #
  - | #-neighborhood| > MinPts gọi đối tượng lõi
  - P đạt được trực tiếp theo mật độ từ q nếu q là đối tượng lõi và p thuộc #-neighborhood của q.
  - Đạt được nếu có dãy mà mỗi cái sau là đạt được trực tiếp từ cái trước

- Phân cụm phân cấp dựa trên mô hình
  - Làm phù hợp phân bố cụm với mô hình toán học
  - Phân cụm cực đại kỳ vọng, phân cụm khái niệm, học máy mạng nơron
  - Phân cụm cực đại kỳ vọng: khởi tạo, tính giá trị kỳ vọng, cực đại hóa kỳ vọng

# 7. Biểu diễn cụm và gán nhãn

- Các phương pháp biểu diễn điển dình
  - Theo đại diện cụm
    - Đại diện cụm làm tâm
    - Tính bán kính và độ lệch chuẩn để xác định phạm vi của cụm
    - Cụm không ellip/cầu hóa: không tốt
  - Theo mô hình phân lớp
    - Chỉ số cụm như nhãn lớp
    - Chạy thuật toán phân lớp để tìm ra biểu diễn cụm
  - Theo mô hình tần số
    - Dùng cho dữ liệu phân loại
    - Tần số xuất hiện các giá trị đặc trưng cho từng cụm
- Lưu ý
  - Dữ liệu phân cụm ellip/cầu hóa: đại diện cụm cho biểu diễn tốt
  - Cụm hình dạng bất thường rất khó biểu diễn

# Gán nhãn cụm

- Phân biệt các cụm (MU)

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

  - Chọn đặc trưng tương quan cụm
  - Nxy (x có đặc trưng t, y dữ liệu thuộc C)
    - N11 : số dữ liệu chứa t thuộc cụm C
    - N10 : số dữ liệu chứa t không thuộc cụm C
    - N01 : số dữ liệu không chứa t thuộc cụm C
    - N00 : số dữ liệu không chứa t không thuộc cụm C
    - N: Tổng số dữ liệu
- Hướng “trọng tâm” cụm
  - Dùng các đặc trưng tần số cao tại trọng tâm cụm
- Tiêu đề
  - Chọn đặc trưng của dữ liệu trong cụm gần trọng tâm nhất

# Ví dụ: Gán nhãn cụm văn bản

| | # docs | labeling method | | |
|---|---|---|---|---|
| | | centroid | mutual information | title |
| 4 | 622 | oil plant mexico production crude **power 000 refinery gas** bpd | plant oil production **barrels** crude bpd mexico **dolly capacity petroleum** | MEXICO: Hurricane Dolly heads for Mexico coast |
| 9 | 1017 | police security **russian** people military peace killed told **grozny court** | police killed military security peace told **troops forces rebels** people | RUSSIA: Russia's Lebed meets rebel chief in Chechnya |
| 10 | 1259 | 00 000 tonnes traders futures wheat prices **cents september** tonne | **delivery** traders futures tonne tonnes **desk** wheat prices 000 00 | USA: Export Business - Grain/oilseeds complex |

- Ví dụ
  - Ba phương pháp chọn nhãn cụm đối với 3 cụm là cụm 4 (622 tài liệu), cụm 9 (1017 tài liệu), cụm 10 (1259 tài liệu) khi phân cụm 10000 tài liệu đầu tiên của bộ Reuters-RCV1
  - centroid: các từ khóa có tần số cao nhất trong trọng tâm; mutual information (MU): thông tin liên quan phân biệt các cụm; title: tiêu đề tài liệu gần trọng tâm nhất.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.

# 8. Đánh giá phân cụm

- Đánh giá chất lượng phân cụm là khó khăn
  - Chưa biết các cụm thực sự
- Một số phương pháp điển hình
  - Người dùng kiểm tra
    - Nghiên cứu trọng tâm và miền phủ
    - Luật từ cây quyết định
    - Đọc các dữ liệu trong cụm
  - Đánh giá theo các độ đo tương tự/khoảng cách
    - Độ phân biệt giữa các cụm
    - Phân ly theo trọng tâm
  - Dùng thuật toán phân lớp
    - Coi mỗi cụm là một lớp
    - Học bộ phân lớp đa lớp (cụm)
    - Xây dựng ma trận nhầm lẫn khi phân lớp
    - Tính các độ đo: entropy, tinh khiết, chính xác, hồi tưởng, độ đo F và đánh giá theo các độ đo này

# Đánh giá theo độ đo tương tự

- Độ phân biệt các cụm
  - Cực đại hóa tổng độ tương tự nội tại của các cụm
  - Cực tiểu hóa tổng độ tương tự các cặp cụm khác nhau
  - Lấy độ tương tự cực tiểu (complete link), cực đại (single link)

$$J_e = \frac{1}{2}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{d_j,d_l\in S_i}\|d_j - d_l\|^2$$

$$J_s = \frac{1}{2}\sum_{i=1}^{k}\frac{1}{|S_i|}\sum_{d_j,d_l\in S_i} sim(d_j,d_l) = \frac{1}{2}\sum_{i=1}^{k}|S_i|sim(S_i)$$

- Một số phương pháp điển hình
  - Phân ly theo trọng tâm

$$J_e = \sum_{i=1}^{k}\sum_{d\in S_i}\|d - c_i\|^2$$

# Ví dụ: Chế độ, đặc điểm phân cụm web

- Hai chế độ
  - Trực tuyến: phân cụm kết quả tìm kiếm người dùng
  - Ngoại tuyến: phân cụm tập văn bản cho trước
- Đặc điểm
  - Chế độ trực tuyến: tốc độ phân cụm
    - Web số lượng lớn, tăng nhanh và biến động lớn
    - Quan tâm tới phương pháp gia tăng
  - Một lớp quan trọng: phân cụm liên quan tới câu hỏi tìm kiếm
    - Trực tuyến
    - Ngoại tuyến

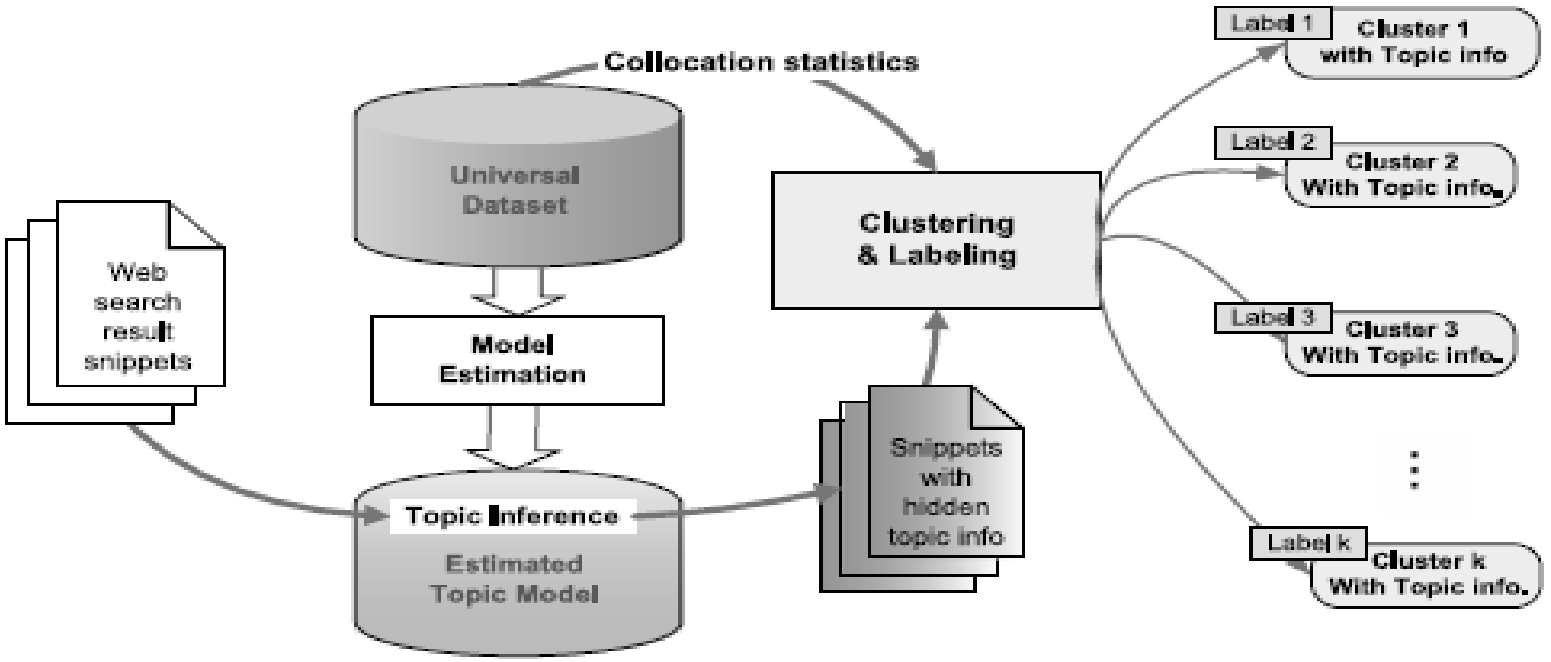
Carpineto C., Osinski S., Romano G., Weiss D. (2009). A survey of web clustering engines, *ACM Comput. Surv.* , **41**(3), Article 17, 38 pages.

# Ví dụ

| Bảng 7.2 Dữ liệu mẫu dành cho phân cụm phẳng | | | | | | |
|---|---|---|---|---|---|---|
| *Tên trang web* | *A1* | *A2* | *A3* | *A4* | *A5* | *A6* |
| Anthropology | 0 | 0.537 | 0.477 | 0 | 0.673 | 0.177 |
| Art | 0 | 0 | 0 | 0.961 | 0.195 | 0.196 |
| Biology | 0 | 0.347 | 0.924 | 0 | 0.111 | 0.112 |
| Chemistry | 0 | 0.975 | 0 | 0 | 0.155 | 0.158 |
| Communication | 0 | 0 | 0 | 0.78 | 0.626 | 0 |
| Computer Science | 0 | 0.989 | 0 | 0 | 0.13 | 0.067 |
| Criminal Justice | 0 | 0 | 0 | 0 | 1 | 0 |
| Economics | 0 | 0 | 1 | 0 | 0 | 0 |
| English | 0 | 0 | 0 | 0.98 | 0 | 0.199 |
| Geography | 0 | 0.849 | 0 | 0 | 0.528 | 0 |
| History | 0.991 | 0 | 0 | 0.135 | 0 | 0 |
| Mathematics | 0 | 0.616 | 0.549 | 0.49 | 0.198 | 0.201 |
| Modern Languages | 0 | 0 | 0 | 0.928 | 0 | 0.373 |
| Music | 0.97 | 0 | 0 | 0 | 0.17 | 0.172 |
| Philosophy | 0.741 | 0 | 0 | 0.658 | 0 | 0.136 |
| Physics | 0 | 0 | 0.894 | 0 | 0.315 | 0.318 |
| Political Science | 0 | 0.933 | 0.348 | 0 | 0.062 | 0.063 |
| Psychology | 0 | 0 | 0.852 | 0.387 | 0.313 | 0.162 |
| Sociology | 0 | 0 | 0.639 | 0.57 | 0.459 | 0.237 |
| Theatre | 0 | 0 | 0 | 0 | 0.967 | 0.254 |

| ***Bảng 7.6*** **Giá trị của hàm đánh giá dựa trên độ đo tương tự với giải thuật k-means** | | |
|---|---|---|
| *k*=2 | *k*=3 | *k*=4 |
| **1 [8.53381]**<br><br>Anthropology, Biology, Chemistry, Computer Science, Economics, Geography, Mathematics, Physics, Political Science, Psychology, Sociology<br><br>**2 [6.12743]**<br><br>Art, Communication, Criminal Justice, English, History, Modern Languages, Music, Philosophy, Theatre<br><br>**Σ= [12.0253]** | **1 [2.83806]**<br><br>History, Music, Philosophy<br><br>**2 [6.09107]**<br><br>Anthropology, Biology, Chemistry, Computer Science, Geography, Mathematics, Political Science<br><br>**3 [7.12119]**<br><br>Art, Communication, Criminal Justice, Economics, English, Modern Languages, Physics, Psychology, Sociology, Theatre<br><br>**Σ= [12.0253]** | **1 [3.81771]**<br><br>Art, Communication, English, Modern Languages<br><br>**2 [5.44416]**<br><br>Biology, Economics, Mathematics, Physics, Psychology, Sociology<br><br>**3 [2.83806]**<br><br>History, Music, Philosophy<br><br>**4 [5.64819]**<br><br>Anthropology, Chemistry, Computer Science, Criminal Justice, Geography, Political Science, Theatre<br><br>**Σ= [12.0253]** |

# Phân cụm kết quả tìm kiếm




(a) Choosing an appropriate "universal dataset."
(b) Performing topic analysis for the universal dataset.
(c) Finding collocations in the universal dataset.
(d) Performing topic inference for search snippets.
(e) Combining the original snippets with their hidden topics.
(f) Building a clustering/labeling system on the enriched snippets.