


Tìm kiếm thông tin

Chương 8. Phân lớp và phân cụm văn bản

Soạn bởi: TS. Nguyễn Bá Ngọc

2021

Nội dung

- 
1. Phân lớp văn bản và ứng dụng trong TKTT
 2. Các phương pháp phân lớp văn bản
 3. Các phương pháp trích chọn đặc trưng
 4. Đánh giá kết quả phân lớp
 5. Chia cụm văn bản và ứng dụng trong TKTT
 6. Các phương pháp chia cụm văn bản
 7. Đánh giá kết quả chia cụm

Bài toán Phân lớp văn bản

Ở mức sơ lược nhất bài toán phân lớp văn bản có thể được phát biểu như sau:

- Cho tập lớp $C = \{c_1, c_2, \dots, c_J\}$
- Yêu cầu: Xác định lớp của văn bản d bất kỳ.

TKTT và phân lớp về bản chất có nhiều điểm tương đồng. Nếu coi câu truy vấn là biểu diễn của 1 lớp thì có thể coi hệ thống tìm kiếm thông tin như 1 hệ thống phân lớp: Các văn bản được trả về được coi như thuộc lớp, các văn không được trả về được coi như không thuộc lớp.

Các truy vấn cố định

- Kịch bản sử dụng:


- Người dùng có nhu cầu theo dõi thông tin mới về một chủ đề, ví dụ:
 - Theo dõi diễn biến *tìm kiếm dấu vết sự sống trên sao Hỏa*
 - Theo dõi các *bài viết mới về 1 thương hiệu được quan tâm*
 - v.v..
- Người dùng thiết lập mô tả nhu cầu thông tin, hệ thống sẽ gửi các kết quả mới được phát hiện
 - Mô tả của người dùng có vai trò giống như biểu diễn của lớp
 - Hệ thống phân chia các nội dung mới theo 2 lớp: Phù hợp/không phù hợp

- Ví dụ triển khai thực tế:

- Google Alerts - Hoạt động trong môi trường Web, quy mô lớn
- Awario - Theo dõi thông tin về các thương hiệu
- v.v.

Alerts

Monitor the web for interesting new content

 information retrieval



How often

As-it-happens



Sources

Automatic



Language

Vietnamese



Region

Any Region



How many

Only the best results



Deliver to

RSS feed



Update alert

Hide options ▲

Alert preview

There are no recent results for your search query. Below are existing results that match your search query.

Lọc nội dung rác

From: "" <takworldd@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

=====

Click Below to order:

<http://www.wholesaledaily.com/sales/nmd.htm>

=====

Phân lớp tin tức

Google News

Search for topics, locations & sources

Top stories

For you

Following

Saved searches

COVID-19

U.S.

World

Your local news

Business

Technology

Entertainment

Sports

Science

Health

Language & region
English (United States)

Settings

Get the Android app



Technology

Follow

Share

Latest

Mobile

Gadgets

Internet

Virtual reality

Artificial in

Round Up: The Reviews Of Valve's Steam Deck Are In - What's It Like Compared To Switch?

Nintendo Life · 5 hours ago

- Steam Deck review: it's not ready
The Verge · 16 hours ago

View Full Coverage



Apple patents Magic Keyboard with integrated Mac inside to bring macOS to any display

9to5Mac · 13 hours ago

- Apple Imagines Mac-Inside-a-Keyboard Device Evocative of 80s Home Computers
MacRumors · 21 hours ago

View Full Coverage



8 Minutes Of Elden Ring Running On Steam Deck

GameSpot · 7 hours ago

- Elden Ring: Bandai Namco Apologizes for Performance Issues - IGN
IGN · 21 hours ago

View Full Coverage



Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
7. Đánh giá kết quả chia cụm

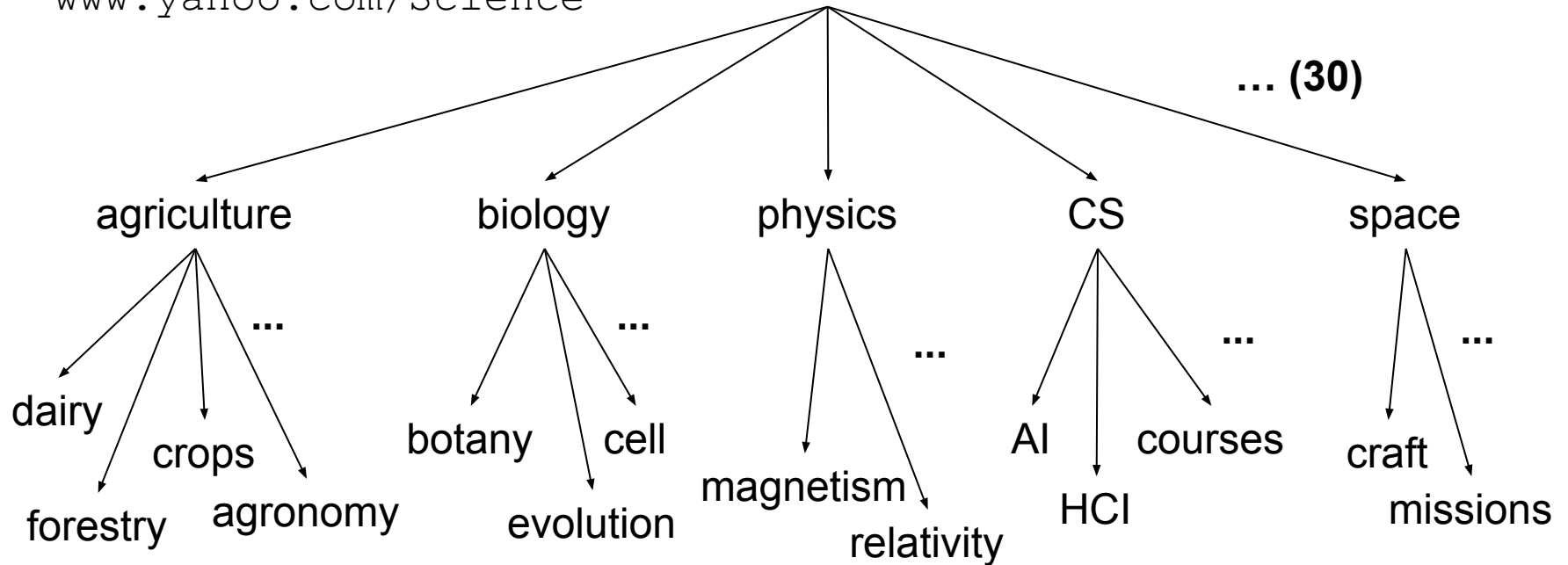
Phương pháp thủ công

Người phân lớp rà soát nội dung và phân chia nội dung theo 1 hệ thống danh mục

- Ví dụ triển khai thực tế:
 - Yahoo! Directory
 - Looksmart, about.com, ODP, PubMed
 - Tổ chức tin bài theo danh mục trong các trang tin tức.
- Một số đặc điểm:
 - Chất lượng phụ thuộc vào kỹ năng của người thực hiện.
 - Kết quả do nhiều người khác nhau thực hiện có thể khác nhau
 - Khó mở rộng và thực hiện trên quy mô lớn.

Cấu trúc phân cấp của Yahoo!

www.yahoo.com/Science



Phân lớp bán tự động dựa trên luật

Người phân lớp thiết lập các luật phân lớp trong 1 môi trường chuyên dụng. Hệ thống thực hiện phân lớp theo các tập luật được cung cấp.

- Hệ thống có thể cung cấp các công cụ hỗ trợ xây dựng luật:
 - Các công cụ phân tích dữ liệu;
 - Ngôn ngữ truy vấn tiên tiến với các cấu trúc bậc cao.
 - v.v..
- Một số đặc điểm tiêu biểu:
 - Chất lượng phân lớp phụ thuộc vào luật;
 - Luật cần được duy trì và cập nhật theo thời gian;
 - Luật có thể được áp dụng cho dữ liệu ở quy mô lớn.

Ví dụ 8.1. Luật phân lớp trong Verity

Định nghĩa lớp art (nghệ thuật)

```
comment line      # Beginning of art topic definition
top-level topic   art ACCRUE
topic definition modifiers {
    /author = "fsmith"
    /date  = "30-Dec-01"
    /annotation = "Topic created
                        by fsmith"
subtopic topic    * 0.70 performing-arts ACCRUE
evidence topic    ** 0.50 WORD
topic definition modifier /wordtext = ballet
evidence topic    ** 0.50 STEM
topic definition modifier /wordtext = dance
evidence topic    ** 0.50 WORD
topic definition modifier /wordtext = opera
evidence topic    ** 0.30 WORD
topic definition modifier /wordtext = symphony
subtopic          * 0.70 visual-arts ACCRUE
                  ** 0.50 WORD
                  /wordtext = painting
                  ** 0.50 WORD
                  /wordtext = sculpture
subtopic          * 0.70 film ACCRUE
                  ** 0.50 STEM
                  /wordtext = film
subtopic          ** 0.50 motion-picture PHRASE
                  *** 1.00 WORD
                  /wordtext = motion
                  *** 1.00 WORD
                  /wordtext = picture
                  ** 0.50 STEM
                  /wordtext = movie
subtopic          * 0.50 video ACCRUE
                  ** 0.50 STEM
                  /wordtext = video
                  ** 0.50 STEM
                  /wordtext = vcr
# End of art topic
```

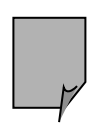
Có thể sử dụng câu truy vấn trong nền tảng tìm kiếm thông tin như luật phân lớp.

Bài toán phân lớp văn bản tự động

- Cho:
 - Tập lớp $C = \{c_1, c_2, \dots, c_J\}$
 - Dữ liệu huấn luyện D gồm các văn bản được gán nhãn theo các lớp trong C
- Yêu cầu:
 - Phương pháp huấn luyện/giải thuật để học hàm phân lớp γ mô phỏng theo D .
 - Với mỗi văn bản d cần được phân lớp, $\gamma(d)$ cho biết lớp của văn bản d , $\gamma(d) \in C$

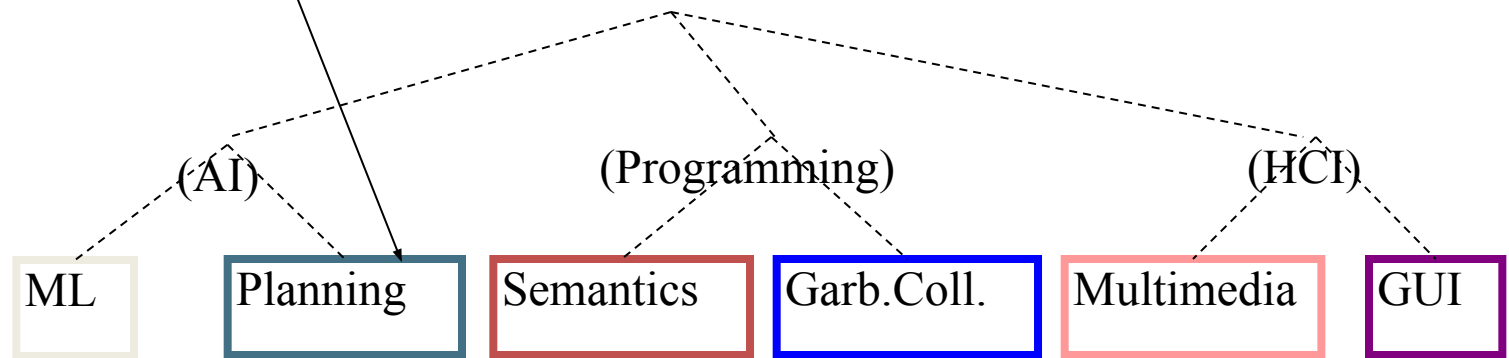
Ví dụ 8.2. Bài toán phân lớp văn bản tự động

*Dữ liệu
kiểm thử:*



"planning language
proof intelligence"

Các lớp:



*Dữ liệu
huấn luyện:*

learning
intelligence
algorithm
reinforcement
network...

planning
schedule
reasoning
plan
language...

programming
semantics
language
proof...

garbage
collection
memory
optimization
region...

...

...

Các phương pháp phân lớp tự động

Một số phương pháp thông dụng:

- Rocchio
- kNN/k-Nearest Neighbors/k-Láng giềng gần nhất
- SVM/Support-vector Machines/Máy vec-tơ hỗ trợ
- Naïve Bayes
- Decision trees/Cây quyết định
- ... về cơ bản phân lớp tự động là vấn đề học có giám sát.
- Trong thực tế có thể kết hợp đồng thời nhiều phương pháp

Giám sát được thực hiện qua dữ liệu huấn luyện

Các đặc trưng

- Phương pháp học có giám sát sử dụng nhiều loại đặc trưng khác nhau: URL, địa chỉ Email, từ khóa, v.v...
- Đối với phân lớp văn bản có thể coi mỗi từ như 1 đặc trưng.
- Trong bài giảng này chúng ta sẽ biểu diễn các văn bản theo định dạng túi từ.

Ví dụ 8.3. Biểu diễn túi từ

Y (

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.

) = C

Ví dụ 8.3. Biểu diễn túi từ²₍₂₎

Y (

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

) = C

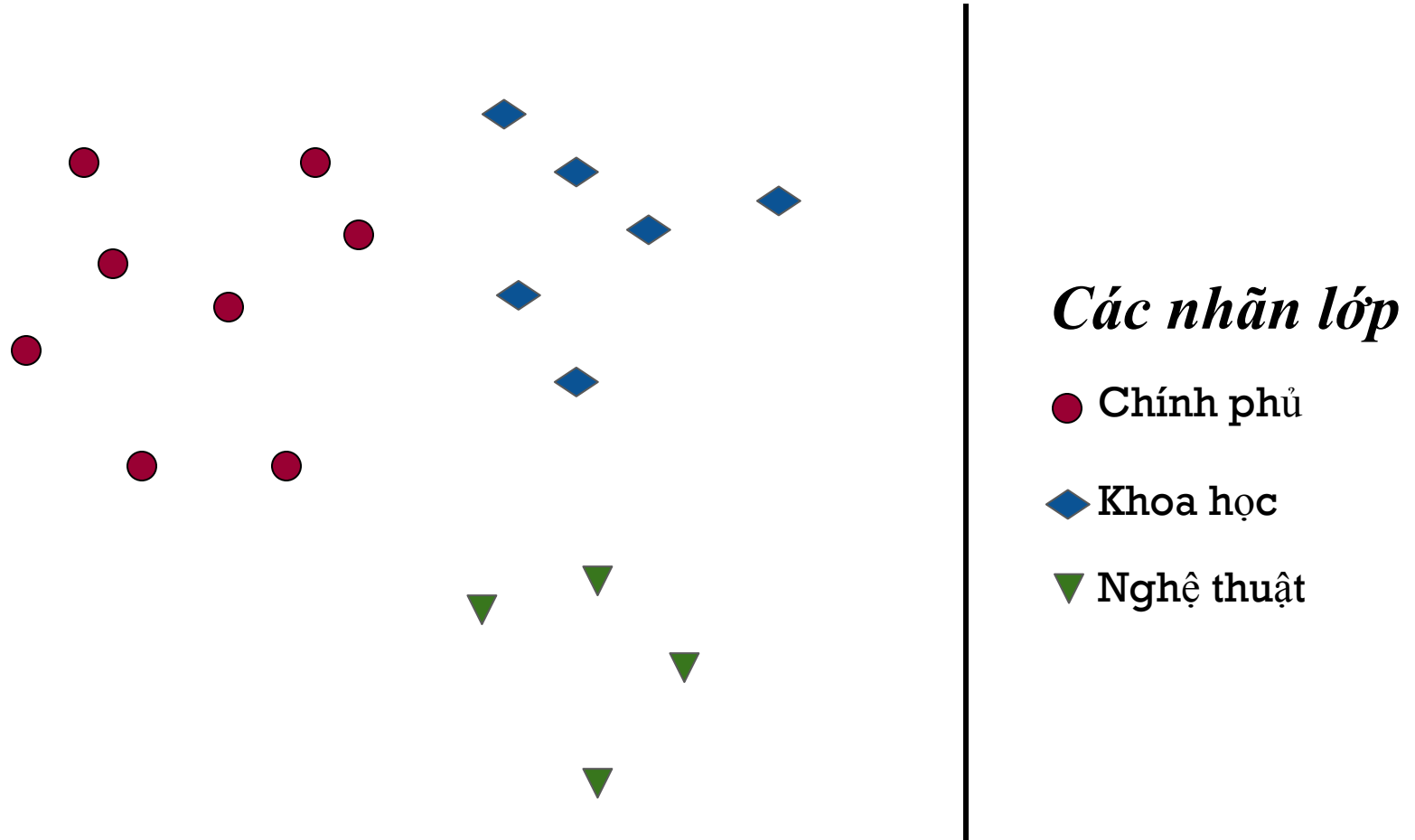
Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
- 2.1. Phân lớp trong VSM
- 2.2. Phân lớp dựa trên xác suất
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
7. Đánh giá kết quả chia cụm

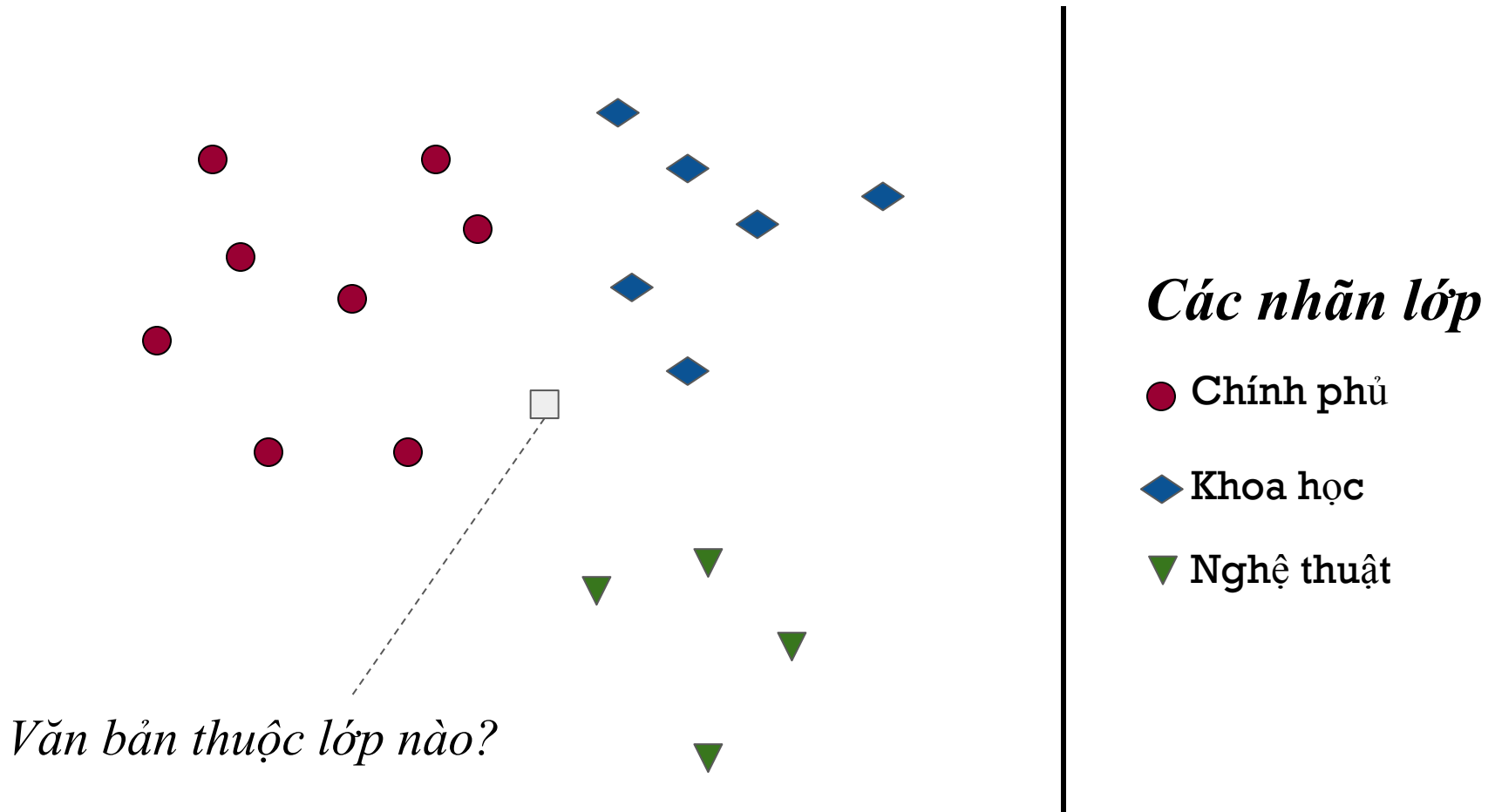
Phân lớp trong mô hình không gian vec-tơ

- Trong VSM: Mỗi văn bản được biểu diễn như 1 vec-tơ, tương đương với 1 điểm trong không gian.
- Các mục tiêu phân lớp:
 - Mục tiêu 1: Các văn bản trong cùng một lớp tạo thành 1 vùng liên tục trong không gian
 - Mục tiêu 2: Tập văn bản của các lớp khác nhau không chồng lấn hoặc chồng lấn ở mức độ hạn chế.
- Học hàm phân lớp: Có bản chất toán học giống như xây dựng các mặt phân tách các lớp trong không gian

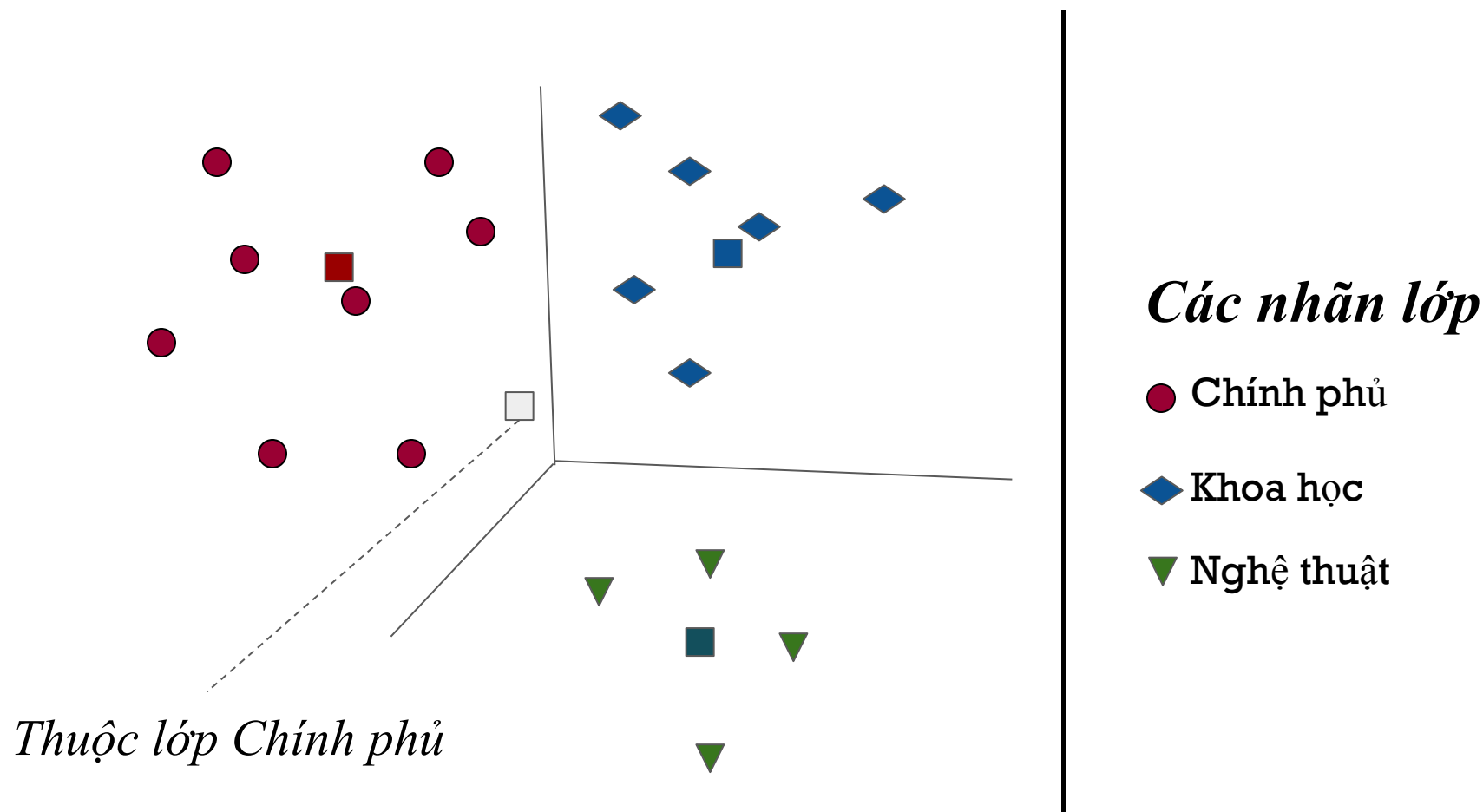
Ví dụ 8.4. Văn bản trong không gian vec-tơ



Ví dụ 8.4. Văn bản trong không gian vec-tơ⁽²⁾



Ví dụ 8.4. Văn bản trong không gian vec-tơ⁽³⁾



Phân lớp Rocchio: Khái niệm trọng tâm

$$\vec{\mu}(c) = \frac{1}{|D_c|} \sum_{d \in D_c} \vec{v}(d)$$

- Trong đó D_c là tập hợp tất cả các văn bản trong tập văn bản huấn luyện thuộc lớp c và $\vec{v}(d)$ là biểu diễn của văn bản d trong không gian vec-tơ
- (*Tương tự giải thuật Rocchio, Chương 6*)

Phân lớp Rocchio

- Huấn luyện: Tính vec-tơ trọng tâm cho mỗi lớp dựa trên dữ liệu huấn luyện;
- Phân lớp: Lựa chọn trọng tâm gần nhất
 - *Vec-tơ trọng tâm có vai trò như nguyên mẫu của lớp.*
- Kết quả phân lớp có thể khác với dữ liệu huấn luyện.

Thử lấy ví dụ minh họa trường hợp phân lớp Rocchio cho kết quả khác với dữ liệu huấn luyện

Phân lớp Rocchio 2-lớp

Có thể được biểu diễn như 1 phân lớp tuyến tính:

- Có thể định nghĩa đường/mặt phân lớp theo công thức:

$$\sum_{i=1}^M w_i d_i = \theta$$

- Đối với Rocchio, thiết lập:

$$\vec{w} = \vec{\mu}(c_1) - \vec{\mu}(c_2)$$

$$\theta = 0.5 \times (|\vec{\mu}(c_1)|^2 - |\vec{\mu}(c_2)|^2)$$

Ví dụ 8.5. Phân lớp tuyến tính

- Lớp: "interest"
- Ví dụ các đặc trưng của bộ phân lớp tuyến tính

w_i	t_i	w_i	t_i
0.70	prime	-0.71	dlrs
0.67	rate	-0.35	world
0.63	interest	-0.33	sees
0.60	rates	-0.25	year
0.46	discount	-0.24	group
0.43	bundesbank	-0.24	dlr

- Tính tích vô hướng của vec-tơ đặc trưng và vec-tơ trọng số, sau đó so sánh giá trị thu được với θ

Phân lớp Rocchio: Một số đặc điểm

- Chủ yếu được sử dụng để phân lớp văn bản
 - Tỏ ra khá hiệu quả trong phân lớp văn bản
 - Nhưng trong trường hợp tổng quát kém hiệu quả hơn Naïve Bayes
 - Ít được sử dụng với các loại dữ liệu khác.
- Giải thuật huấn luyện và phân lớp đơn giản.

Phân lớp k-Láng giềng gần nhất

- kNN = k Nearest Neighbors
- Các bước phân lớp văn bản d:
 - Xác định k láng giềng gần nhất của d
 - Chọn lớp chiếm đa số trong k láng giềng
 - Với giá trị k lớn có thể ước lượng
$$P(c|d) \propto (\text{số lượng văn bản thuộc lớp } c)/k$$

kNN

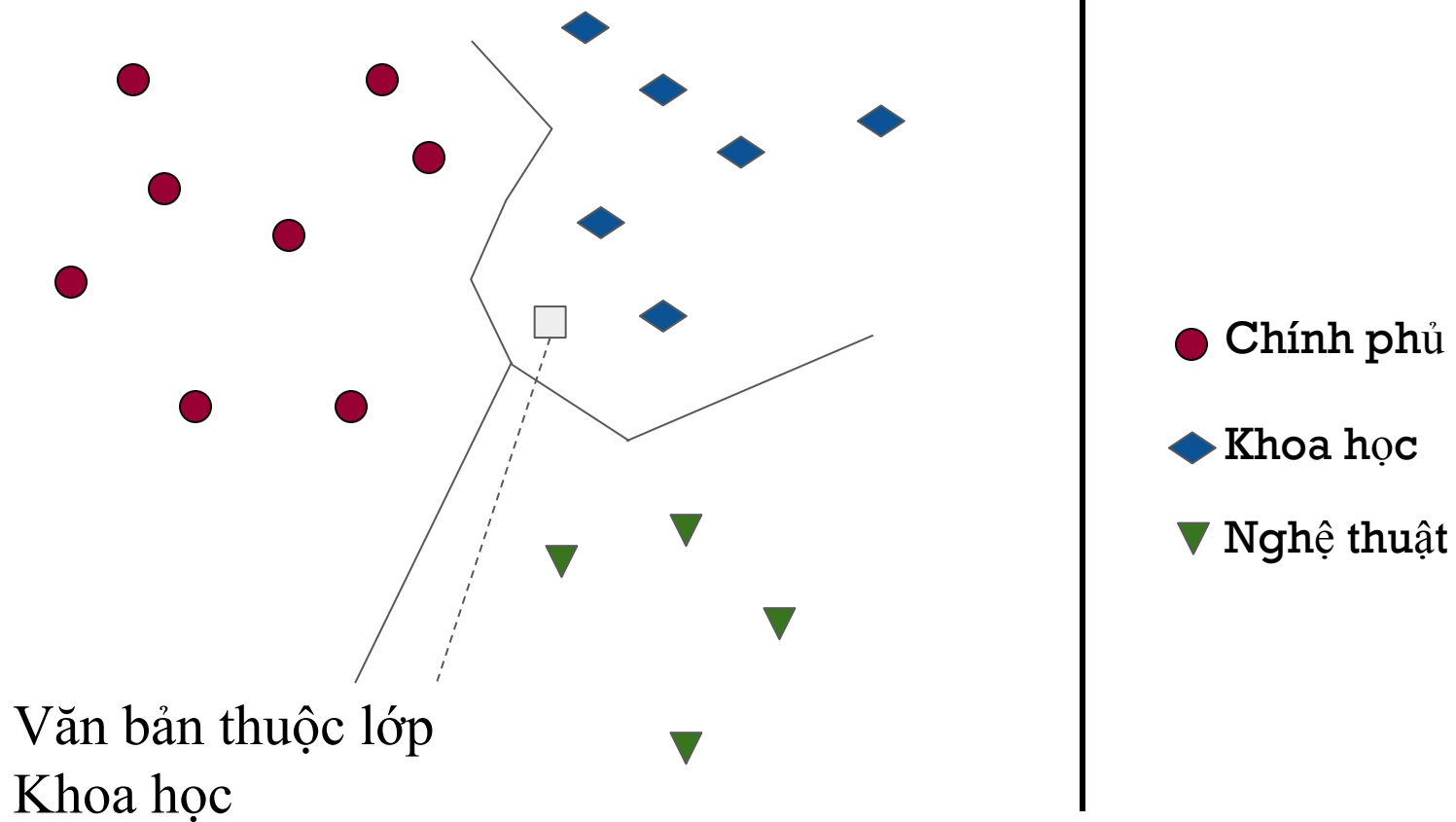
- Huấn luyện: Chỉ lưu các mẫu đã được gán nhãn/Tạo biểu diễn văn bản cho tập huấn luyện D
- Phân lớp văn bản d (với 1NN):
 - Tính độ tương đồng giữa d và tất cả các văn bản trong D ;
 - Gán d vào lớp của văn bản trong D có độ tương đồng cao nhất.
- kNN còn được gọi là:
 - Học dựa trên trường hợp;
 - Học dựa trên bộ nhớ.
- Cơ sở của kNN: Giả thuyết liên tục

kNN₍₂₎

- 1NN có thể mắc nhiều lỗi do:
 - Quyết định dựa trên một trường hợp duy nhất
 - Nhiễu (ví dụ, một lỗi) trong dữ liệu huấn luyện
- Các giá trị $k > 1$ có thể cho kết quả tốt hơn.
- k thường được chọn là số lẻ; 3 và 5 là các giá trị thường được sử dụng.

Ví dụ 8.6. Văn bản trong không gian vec-tơ

Biểu đồ Voronoi



Triển khai kNN với chỉ mục ngược

- Tìm kiếm các láng giềng gần nhất bằng phương pháp vét cạn trên toàn tập văn bản D tốn nhiều thời gian.
- Tìm kiếm k láng giềng gần nhất tương tự tìm Top k kết quả tìm kiếm với câu truy vấn là văn bản đang được phân lớp.
 - Sử dụng chỉ mục ngược tương tự mô hình không gian vec-tơ.
- Độ phức tạp phân lớp: $O(B|V_t|)$ trong đó B là số lượng văn bản trung bình có chứa các từ của văn bản được phân lớp.
 - Thông thường $B \ll |D|$

kNN: Các đặc điểm

- Không cần trích chọn đặc trưng
- Mở rộng tốt với số lượng lớp lớn
 - Không cần huấn luyện n bộ phân lớp cho n lớp
- Các lớp có thể ảnh hưởng lẫn nhau
 - Các thay đổi nhỏ với một lớp có thể tạo hiệu ứng lan truyền
- Có thể tốn nhiều chi phí ở bước phân lớp
- Trong hầu hết các trường hợp chính xác hơn Naïve Bayes và Rocchio
- Với lượng lớn dữ liệu có thể là bộ phân lớp lý tưởng (tối ưu Bayes)

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
 - 2.1. Phân lớp trong VSM
 - 2.2. Phân lớp dựa trên xác suất
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
7. Đánh giá kết quả chia cụm

Naïve Bayes: Suy diễn công thức

Ước lượng xác suất văn bản thuộc lớp và chọn lớp có xác suất thuộc lớn nhất:

- Lớp có xác suất thuộc lớn nhất:

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} P(c|d)$$

- Áp dụng luật Bayes $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$:

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} \frac{P(d|c)P(c)}{P(d)}$$

- Có thể giản lược $P(d)$ (giống nhau cho tất cả các lớp)

$$c_{\text{map}} = \arg \max_{c \in \mathbb{C}} P(d|c)P(c)$$

Vấn đề quá nhiều tham số

$$\begin{aligned}c_{\text{map}} &= \arg \max_{c \in \mathbb{C}} P(d|c)P(c) \\ &= \arg \max_{c \in \mathbb{C}} P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)P(c)\end{aligned}$$

- Nếu coi mỗi tổ hợp n_d từ của văn bản d là 1 tham số, thì
 - Sẽ có rất nhiều tham số cho $P(\langle t_1, \dots, t_k, \dots, t_{n_d} \rangle | c)P(c)$.
 - Cần rất nhiều mẫu huấn luyện để ước lượng các tham số đó.
- *(Vấn đề với số lượng tham số lớn còn được gọi là vấn đề dữ liệu thừa)*

Giả thuyết độc lập từ vựng

- Có thể sử dụng giả thuyết độc lập từ vựng để giảm số lượng tham số tới mức có thể quản lý:

$$P(d|c) = P(\langle t_1, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

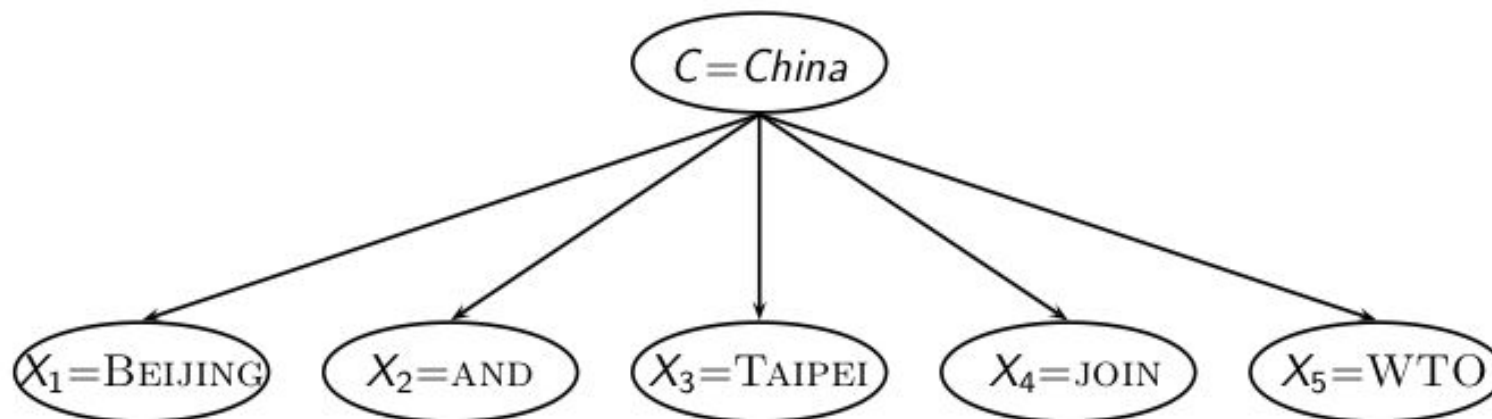
- Xác suất quan sát một dãy từ tỉ lệ với tích xác suất quan sát các từ trong dãy (thành phần $P(X_k = t_k | c)$).

Giả thuyết độc lập vị trí

$$\hat{P}(t_{k_1}|c) = \hat{P}(t_{k_2}|c)$$

- Ví dụ, đối với một văn bản trong lớp UK, xác suất sinh từ QUEEN ở vị trí đầu tiên trong văn bản giống với xác suất sinh từ đó ở vị trí cuối.
- Biểu diễn văn bản với các giả thuyết độc lập từ vựng và độc lập vị trí được gọi là biểu diễn túi từ

Mô hình sinh: Mô hình đa thức



$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

Tương tự mô hình ngôn ngữ

- Lớp c được chọn với xác suất $P(c)$
- Lớp c sinh từ t_k với xác suất $P(t_k|c)$, giá trị xác suất độc lập với các từ còn lại;
- Kết quả phân lớp được xác định theo lớp với xác suất lớn nhất.

Lấy log

- Tích gộp nhiều đại lượng xác suất nhỏ có thể gây tràn độ chính xác số học;
- Kết quả phân lớp tương đương về mặt lý thuyết nếu sử dụng tổng gộp log của các đại lượng xác suất
- Trong thực tế sử dụng công thức sau:

$$\begin{aligned}\gamma(d) &= \arg \max_{c \in C} [\log p(c|d)] \\ &= \arg \max_{c \in C} \left[\log p(c) + \sum_{1 \leq k \leq n_d} \log p(t_k|c) \right]\end{aligned}$$

Ước lượng các tham số

- Các thành phần xác suất $P(c)$ và $P(t_k|c)$ được ước lượng theo khả năng cực đại trên dữ liệu huấn luyện:
 - $P(c) = N_c/N$, trong đó N_c là số lượng văn bản thuộc lớp c , N là số lượng văn bản trong bộ dữ liệu huấn luyện
- Xác suất có điều kiện:

$$p(t_k|c) = \frac{cf_{c,t_k}}{\sum_{t \in V} cf_{c,t}}$$

Trong đó $cf_{c,t}$ là số lần từ t xuất hiện trong lớp c .

Các giá trị 0 và làm mịn

- Nếu có một từ $t \in d$ nhưng không xuất hiện trong bất kỳ văn bản nào của lớp c thì $P(t|c) = 0$, kéo theo xác suất $P(c|d) = 0$
 - Chúng ta không mong muốn sự thay đổi đột ngột này
- Làm mịn Laplace: cộng thêm 1 vào mỗi sự kiện
 - Tương tự như trong mô hình ngôn ngữ

$$p(t_k|c) = \frac{cf_{c,t_k} + 1}{\sum_{t \in V} (cf_{c,t} + 1)} = \frac{cf_{c,t_k} + 1}{\sum_{t \in V} cf_{c,t} + |V|}$$

Naïve Bayes đa thức: Huấn luyện

TRAINMULTINOMIALNB(\mathbb{C}, \mathbb{D})

```
1   $V \leftarrow \text{EXTRACTVOCABULARY}(\mathbb{D})$ 
2   $N \leftarrow \text{COUNTDOCS}(\mathbb{D})$ 
3  for each  $c \in \mathbb{C}$ 
4  do  $N_c \leftarrow \text{COUNTDOCSINCLASS}(\mathbb{D}, c)$ 
5       $prior[c] \leftarrow N_c / N$ 
6       $text_c \leftarrow \text{CONCATENATETEXTOFALLDOCSINCLASS}(\mathbb{D}, c)$ 
7      for each  $t \in V$ 
8      do  $cf_{c,t} \leftarrow \text{COUNTTOKENSOFTERM}(text_c, t)$ 
9      for each  $t \in V$ 
10     do  $condprob[t][c] \leftarrow \frac{cf_{c,t} + 1}{\sum_{t'} (cf_{c,t'} + 1)}$ 
11 return  $V, prior, condprob$ 
```

Naïve Bayes đa thức: Phân lớp

APPLYMULTINOMIALNB(\mathbb{C} , V , $prior$, $condprob$, d)

1 $W \leftarrow \text{EXTRACTTOKENSFROMDOC}(V, d)$

2 **for each** $c \in \mathbb{C}$

3 **do** $score[c] \leftarrow \log prior[c]$

4 **for each** $t \in W$

5 **do** $score[c] + = \log condprob[t][c]$

6 **return** $\arg \max_{c \in \mathbb{C}} score[c]$

Naïve Bayes đa thức: Độ phức tạp giải thuật

mode	time complexity
training	$\Theta(\mathbb{D} L_{ave} + \mathbb{C} V)$
testing	$\Theta(L_a + \mathbb{C} M_a) = \Theta(\mathbb{C} M_a)$

- L_{ave} : Độ dài trung bình của văn bản huấn luyện, L_a : Độ dài văn bản cần phân lớp, M_a : Số lượng từ duy nhất trong văn bản phân lớp;
- Naïve Bayes đa thức có độ phức tạp tuyến tính
 - Hiệu quả & dễ mở rộng

Ví dụ 8.7. Phân lớp Naïve Bayes đa thức

- Cho bộ dữ liệu:

	docID	words in document	in $c = \textit{China}$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

- Yêu cầu: Ước lượng các tham số và phân lớp văn bản kiểm thử (docID = 5)

Ví dụ 8.7. Phân lớp Naïve Bayes đa thức₍₂₎

Priors: $\hat{P}(c) = 3/4$ and $\hat{P}(\bar{c}) = 1/4$ Conditional probabilities:

$$\hat{P}(\text{CHINESE}|c) = (5 + 1)/(8 + 6) = 6/14 = 3/7$$

$$\hat{P}(\text{TOKYO}|c) = \hat{P}(\text{JAPAN}|c) = (0 + 1)/(8 + 6) = 1/14$$

$$\hat{P}(\text{CHINESE}|\bar{c}) = (1 + 1)/(3 + 6) = 2/9$$

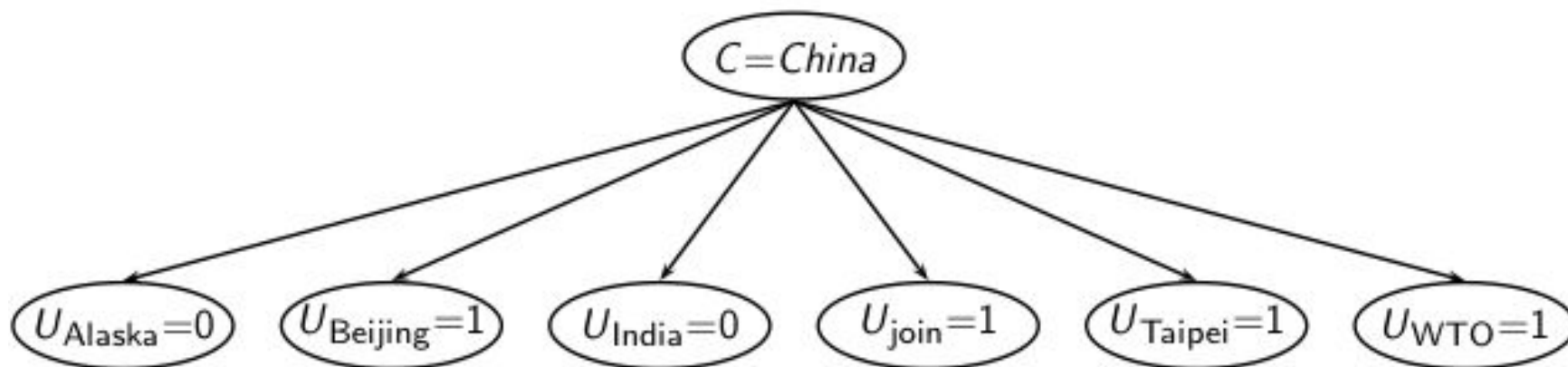
$$\hat{P}(\text{TOKYO}|\bar{c}) = \hat{P}(\text{JAPAN}|\bar{c}) = (1 + 1)/(3 + 6) = 2/9$$

$$\hat{P}(c|d_5) \propto 3/4 \cdot (3/7)^3 \cdot 1/14 \cdot 1/14 \approx 0.0003$$

$$\hat{P}(\bar{c}|d_5) \propto 1/4 \cdot (2/9)^3 \cdot 2/9 \cdot 2/9 \approx 0.0001$$

d_5 Thuộc lớp China

Mô hình sinh: Mô hình Bernoulli/Nhị phân



$$P(d|c) = P(\langle e_1, \dots, e_i, \dots, e_M \rangle | c)$$

$$P(d|c) = P(\langle e_1, \dots, e_M \rangle | c) = \prod_{1 \leq i \leq M} P(U_i = e_i | c)$$

Các tham số:

$$P(U_t = 0|c) = 1 - P(U_t = 1|c)$$

$P(U_t = 1|c) = N_{ct}/N_c$ - Trong đó N_{ct} là số lượng văn bản thuộc lớp c có chứa t .

Làm mịn Laplace: $P(U_t = 1|c) = (N_{ct} + 1)/(N_c + 2)$

Naïve Bayes nhị phân: Huấn luyện

TRAINBERNOULLINB(\mathbb{C}, \mathbb{D})

1 $V \leftarrow \text{EXTRACTVOCABULARY}(\mathbb{D})$

2 $N \leftarrow \text{COUNTDOCS}(\mathbb{D})$

3 **for each** $c \in \mathbb{C}$

4 **do** $N_c \leftarrow \text{COUNTDOCSINCLASS}(\mathbb{D}, c)$

5 $\text{prior}[c] \leftarrow N_c / N$

6 **for each** $t \in V$

7 **do** $N_{ct} \leftarrow \text{COUNTDOCSINCLASSCONTAININGTERM}(\mathbb{D}, c, t)$

8 $\text{condprob}[t][c] \leftarrow (N_{ct} + 1) / (N_c + 2)$

9 **return** $V, \text{prior}, \text{condprob}$

Naïve Bayes nhị phân: Phân lớp

```
APPLYBERNOULLINB( $\mathbb{C}$ ,  $V$ ,  $prior$ ,  $condprob$ ,  $d$ )  
1   $V_d \leftarrow \text{EXTRACTTERMSFROMDOC}(V, d)$   
2  for each  $c \in \mathbb{C}$   
3  do  $score[c] \leftarrow \log prior[c]$   
4      for each  $t \in V$   
5      do if  $t \in V_d$   
6          then  $score[c] += \log condprob[t][c]$   
7          else  $score[c] += \log(1 - condprob[t][c])$   
8  return  $\arg \max_{c \in \mathbb{C}} score[c]$ 
```

Ví dụ 8.8. Phân lớp Naïve Bayes nhị phân

- Cho bộ dữ liệu:

	docID	words in document	in $c = \textit{China}$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

- Yêu cầu: Ước lượng các tham số và phân lớp văn bản kiểm thử (docID = 5) theo phương pháp Bernoulli Naïve Bayes

Ví dụ 8.8. Phân lớp Naïve Bayes nhị phân₍₂₎

$$\hat{P}(\text{Chinese}|c) = (3 + 1)/(3 + 2) = 4/5$$

$$\hat{P}(\text{Japan}|c) = \hat{P}(\text{Tokyo}|c) = (0 + 1)/(3 + 2) = 1/5$$

$$\hat{P}(\text{Beijing}|c) = \hat{P}(\text{Macao}|c) = \hat{P}(\text{Shanghai}|c) = (1 + 1)/(3 + 2) = 2/5$$

$$\hat{P}(\text{Chinese}|\bar{c}) = (1 + 1)/(1 + 2) = 2/3$$

$$\hat{P}(\text{Japan}|\bar{c}) = \hat{P}(\text{Tokyo}|\bar{c}) = (1 + 1)/(1 + 2) = 2/3$$

$$\hat{P}(\text{Beijing}|\bar{c}) = \hat{P}(\text{Macao}|\bar{c}) = \hat{P}(\text{Shanghai}|\bar{c}) = (0 + 1)/(1 + 2) = 1/3$$

$$\begin{aligned}\hat{P}(c|d_5) &\propto \hat{P}(c) \cdot \hat{P}(\text{Chinese}|c) \cdot \hat{P}(\text{Japan}|c) \cdot \hat{P}(\text{Tokyo}|c) \\ &\quad \cdot (1 - \hat{P}(\text{Beijing}|c)) \cdot (1 - \hat{P}(\text{Shanghai}|c)) \cdot (1 - \hat{P}(\text{Macao}|c)) \\ &= 3/4 \cdot 4/5 \cdot 1/5 \cdot 1/5 \cdot (1 - 2/5) \cdot (1 - 2/5) \cdot (1 - 2/5) \\ &\approx 0.005\end{aligned}$$

$$P(\text{NOT } c|d_5) = ?$$

SpamAssassin

- Sử dụng Naïve Bayes để lọc nội dung rác
- Bộ lọc được kết hợp từ nhiều thành phần
 - Xác suất phân lớp theo Naïve Bayes
 - Có xuất hiện: Bitcoin
 - Biểu thức chính quy: Triệu đô la ((đô la) NN, NNN, NNN.NN)
 - Câu: impress ... girl
 - Từ: Bắt đầu với nhiều số
 - Chủ đề được viết hoa hoàn toàn
 - HTML có tỉ lệ văn bản nhỏ so với vùng hình ảnh
 - v.v..

https://spamassassin.apache.org/old/tests_3_3_x.html

Các vi phạm giả thuyết: Naïve Bayes

Các giả thuyết độc lập không đúng với các văn bản thực tế

- Độc lập từ:

$$P(\langle t_1, \dots, t_{n_d} \rangle | c) = \prod_{1 \leq k \leq n_d} P(X_k = t_k | c)$$

- Độc lập vị trí:

$$\hat{P}(t_{k_1} | c) = \hat{P}(t_{k_2} | c)$$

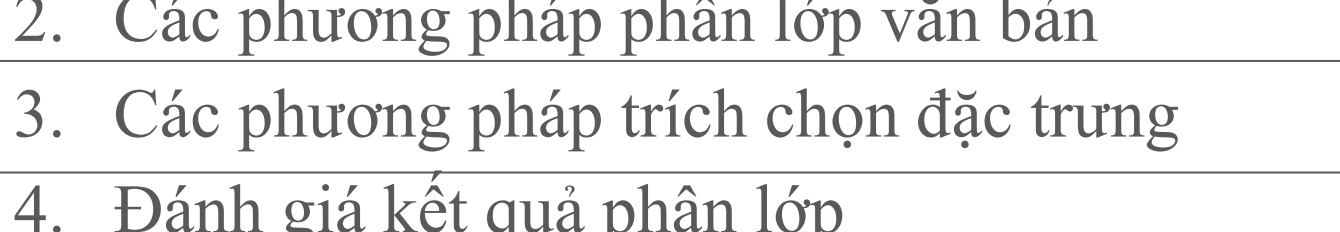
- Các câu hỏi:

- Thử lấy ví dụ trường hợp giả thuyết độc lập từ không đúng?
- Thử lấy ví dụ trường hợp giả thuyết độc lập vị trí không đúng?
- Vì sao Naïve Bayes vẫn có thể cho kết quả tốt dù các giả thuyết có thể không đúng?

Vì sao khó đánh lừa Naïve Bayes

- Huấn luyện và kiểm thử rất nhanh (chủ yếu chỉ đếm từ)
- Yêu cầu lưu trữ không cao.
- Hoạt động rất tốt trong lĩnh vực với nhiều đặc trưng quan trọng như nhau
- Chắc chắn trước các đặc trưng không phù hợp hơn nhiều phương pháp học
 - Các đặc trưng không phù hợp bị loại bỏ không ảnh hưởng tới kết quả
- Ít bị ảnh hưởng bởi vấn đề chuyển dịch khái niệm (concept drift) (định nghĩa lớp thay đổi theo thời gian)
- Phương pháp nền tốt cho phân loại văn bản (tuy không phải là tốt nhất)

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
 2. Các phương pháp phân lớp văn bản
 3. Các phương pháp trích chọn đặc trưng
 4. Đánh giá kết quả phân lớp
 5. Chia cụm văn bản và ứng dụng trong TKTT
 6. Các phương pháp chia cụm văn bản
 7. Đánh giá kết quả chia cụm
- 

Vì sao cần trích chọn đặc trưng?

- Số lượng từ trong tập văn bản rất lớn
 - 10000 - 1000000 từ khác nhau ... và nhiều hơn
- Trích chọn đặc trưng giúp giảm kích thước dữ liệu
 - Một số bộ phân lớp không thể xử lý ≥ 1000000 đặc trưng.
- Rút ngắn thời gian huấn luyện
 - Thời gian huấn luyện có thể phụ thuộc bậc 2 hoặc cao hơn vào số lượng đặc trưng
- Giúp mô hình hoạt động nhanh hơn, giảm thời gian phân lớp.
- Có thể cho kết quả phân lớp tốt hơn, do:
 - Loại bỏ đặc trưng nhiễu
 - Tránh hiện tượng quá vừa

Đặc trưng nhiễu

- Đặc trưng nhiễu làm phát sinh lỗi phân lớp
 - Giả sử một từ hiếm t không chứa thông tin liên quan đến lớp c nhưng lại chỉ xuất hiện trong các văn bản của lớp c trong dữ liệu huấn luyện.
 - Vì t là từ hiếm nên bộ phân lớp sau huấn luyện có thể coi t như một tín hiệu mạnh để xếp các văn bản chứa t vào lớp c
 - Hiện tượng này được gọi là hiện tượng quá vừa *overfitting*

Trích chọn đặc trưng

*Chỉ giữ lại các đặc trưng hữu ích, loại bỏ các đặc trưng nhiễu.
Lựa chọn được thực hiện dựa trên các đại lượng thể hiện tính hữu ích của từ cho phân lớp văn bản/giá trị phân lớp của từ.*

SELECTFEATURES(\mathbb{D} , c , k)

1 $V \leftarrow \text{EXTRACTVOCABULARY}(\mathbb{D})$

2 $L \leftarrow []$

3 **for each** $t \in V$

4 **do** $A(t, c) \leftarrow \text{COMPUTEFEATUREUTILITY}(\mathbb{D}, t, c)$

5 APPEND(L , $\langle A(t, c), t \rangle$)

6 **return** FEATURESWITHLARGESTVALUES(L , k)

How do we compute A , the feature utility?

Giá trị phân lớp của từ

Giá trị phân lớp có thể được đánh giá qua nhiều đại lượng thống kê khác nhau:

- Tần suất - Lựa chọn những từ xuất hiện thường xuyên nhất.
- Hàm lượng thông tin - Lựa chọn từ chứa nhiều thông tin nhất.
- X^2 (Chi bình phương) - Lựa chọn từ có xác suất độc lập với lớp thấp nhất, hay nói cách khác gắn kết với lớp chặt nhất

Tần suất từ

- Đếm số lần từ xuất hiện và chỉ sử dụng những từ phổ biến nhất
- Một số đặc điểm:
 - Đơn giản và dễ thực hiện;
 - Tương đối hiệu quả;
 - Trong 1 số nghiên cứu, hiệu quả có thể đạt 90% so với các phương pháp phức tạp hơn.

Hàm lượng thông tin

- Hàm lượng thông tin I được tính theo công thức:

$$I(U; C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

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

- N_{11} là số lượng văn bản có chứa t và thuộc lớp c; N_{10} số lượng văn bản có chứa t không thuộc lớp c; N_{01} - # không chứa t, thuộc lớp c; N_{00} - # không chứa t, không thuộc lớp c.

$$N = N_{11} + N_{10} + N_{01} + N_{00} = \text{Tổng số lượng văn bản.}$$

Ví dụ 8.9. Tính hàm lượng thông tin

Từ export và lớp poultry

$$\begin{array}{l}
 e_t = e_{\text{EXPORT}} = 1 \\
 e_t = e_{\text{EXPORT}} = 0
 \end{array}
 \begin{array}{cc}
 e_c = e_{\text{poultry}} = 1 & e_c = e_{\text{poultry}} = 0 \\
 \hline
 N_{11} = 49 & N_{10} = 27,652 \\
 N_{01} = 141 & N_{00} = 774,106
 \end{array}
 \text{ Plug}$$

these values into formula:

$$\begin{aligned}
 I(U; C) &= \frac{49}{801,948} \log_2 \frac{801,948 \cdot 49}{(49 + 27,652)(49 + 141)} \\
 &+ \frac{141}{801,948} \log_2 \frac{801,948 \cdot 141}{(141 + 774,106)(49 + 141)} \\
 &+ \frac{27,652}{801,948} \log_2 \frac{801,948 \cdot 27,652}{(49 + 27,652)(27,652 + 774,106)} \\
 &+ \frac{774,106}{801,948} \log_2 \frac{801,948 \cdot 774,106}{(141 + 774,106)(27,652 + 774,106)} \\
 &\approx 0.000105
 \end{aligned}$$

Ví dụ 8.9. Trích chọn đặc trưng theo I (MI)

Thử nghiệm với bộ dữ liệu Reuters

Class: *coffee*

term	MI
COFFEE	0.0111
BAGS	0.0042
GROWERS	0.0025
KG	0.0019
COLOMBIA	0.0018
BRAZIL	0.0016
EXPORT	0.0014
EXPORTERS	0.0013
EXPORTS	0.0013
CROP	0.0012

Class: *sports*

term	MI
SOCCER	0.0681
CUP	0.0515
MATCH	0.0441
MATCHES	0.0408
PLAYED	0.0388
LEAGUE	0.0386
BEAT	0.0301
GAME	0.0299
GAMES	0.0284
TEAM	0.0264

X^2

- Được sử dụng để đánh giá tính độc lập của hai sự kiện
 - Sự kiện xuất hiện lớp và sự kiện xuất hiện từ
- Lựa chọn từ có X^2 lớn:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

- X^2 lớn thể hiện sự gắn kết chặt giữa từ và lớp, vì vậy từ có khả năng cao là một đặc trưng tốt để phân lớp.
- Công thức tương đương về mặt toán học

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

Ví dụ 8.10. Trích chọn đặc trưng với X^2

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$	
$e_t = e_{EXPORT} = 1$	$N_{11} = 49$	$N_{10} = 27,652$	Plug
$e_t = e_{EXPORT} = 0$	$N_{01} = 141$	$N_{00} = 774,106$	

$$\begin{aligned}
 E_{11} &= N \times P(t) \times P(c) = N \times \frac{N_{11} + N_{10}}{N} \times \frac{N_{11} + N_{01}}{N} \\
 &= N \times \frac{49 + 141}{N} \times \frac{49 + 27652}{N} \approx 6.6
 \end{aligned}$$

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \approx 284$$

Phân bố χ^2 với một bậc tự do

p	χ^2 critical value
0.1	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

- Với $X^2 \geq 6.63$ có thể khẳng định chắc chắn 99% về sự phụ thuộc của 2 sự kiện.
- X^2 lớn chứng tỏ từ là đặc trưng tốt để phân lớp.

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
7. Đánh giá kết quả chia cụm

Đánh giá kết quả phân lớp

- Phải được thực hiện trên dữ liệu khác với dữ liệu huấn luyện
- Các độ đo thông dụng: Độ chính xác (P), độ đầy đủ (R), trung bình điều hòa của P và R (F1)

	Thuộc lớp	Không thuộc lớp
Dự đoán thuộc lớp	A (TP)	B (FP)
Dự đoán không thuộc lớp	C (FN)	D (TN)

$$P = \frac{|A|}{|A \cup B|} = \frac{TP}{TP + FP}$$

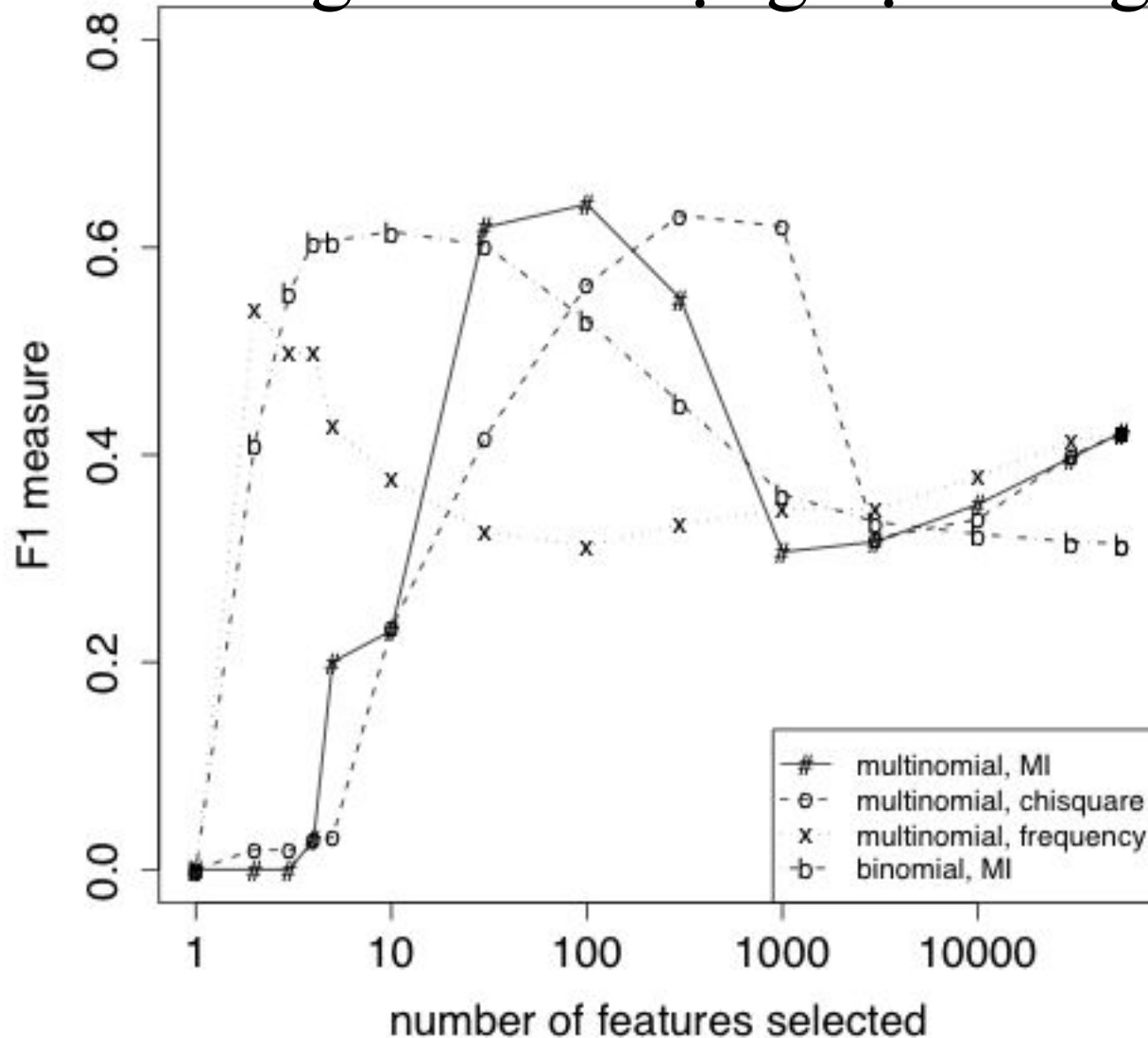
$$R = \frac{|A|}{|A \cup C|} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2PR}{P + R}$$

Lấy trung bình

- Tính P, R, và F_1 cho từng lớp
- Nhưng chúng ta muốn có một chỉ số duy nhất cho tất cả các lớp trên bộ dữ liệu kiểm thử.
- Lấy trung bình Macro
 - Tính F_1 cho từng lớp
 - Lấy trung bình các giá trị F_1
- Lấy trung bình Micro
 - Thống kê TP, TN, FP, FN cho từng lớp
 - Lấy tổng các đại lượng thống kê này
 - Tính F_1 trên các giá trị tổng hợp

Ảnh hưởng của số lượng đặc trưng



(multinomial = Naïve Bayes đa thức, binomial = Naïve Bayes nhị thức) 72

So sánh các phương pháp phân lớp

F1 trên Reuters-21578. a [Li & Yang, 2003]; b, kNN[Dumais et al., 1998];

(a)		NB	Rocchio	kNN	SVM
	micro-avg-L (90 classes)	80	85	86	89
	macro-avg (90 classes)	47	59	60	60

(b)		NB	Rocchio	kNN	trees	SVM
	earn	96	93	97	98	98
	acq	88	65	92	90	94
	money-fx	57	47	78	66	75
	grain	79	68	82	85	95
	crude	80	70	86	85	89
	trade	64	65	77	73	76
	interest	65	63	74	67	78
	ship	85	49	79	74	86
	wheat	70	69	77	93	92
	corn	65	48	78	92	90
	micro-avg (top 10)	82	65	82	88	92
	micro-avg-D (118 classes)	75	62	n/a	n/a	87

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
7. Đánh giá kết quả chia cụm

Bài toán chia cụm văn bản

Phát biểu bài toán chia cụm ở dạng sơ lược nhất:

- Chia cụm văn bản là bài toán phân chia 1 tập văn bản thành nhiều tập con/cụm văn bản sao cho:
 - Các văn bản trong cùng cụm phải giống nhau;
 - Các văn bản trong các cụm khác nhau phải khác nhau.
 - *(Giống kết quả phân lớp)*
- Chia cụm là 1 hình thức học không giám sát
 - Học không giám sát = Học từ chính dữ liệu cần được xử lý, không có dữ liệu huấn luyện như trong phân lớp.
 - Có thể có các tham số điều khiển, tiêu biểu như số lượng cụm.
 - *(Tuy nhiên số lượng cụm cũng thể được xác định tự động bằng các phương pháp tối ưu hóa.)*
- Chia cụm văn bản có nhiều ứng dụng quan trọng trong TKTT và các lĩnh vực khác

Một số vấn đề trong bài toán chia cụm

- Biểu diễn cụm
 - Biểu diễn văn bản: Không gian vec-tơ, chuẩn hóa
 - Đo độ tương đồng/khoảng cách
- Số lượng cụm
 - Được cung cấp như tham số cho giải thuật.
 - hoặc được suy diễn từ dữ liệu
 - Hạn chế các cụm đơn điệu - quá lớn hoặc quá nhỏ
 - Nếu cụm quá lớn trong trường hợp biểu diễn kết quả có thể bỏ qua nhiều kết quả chất lượng.
 - v.v..

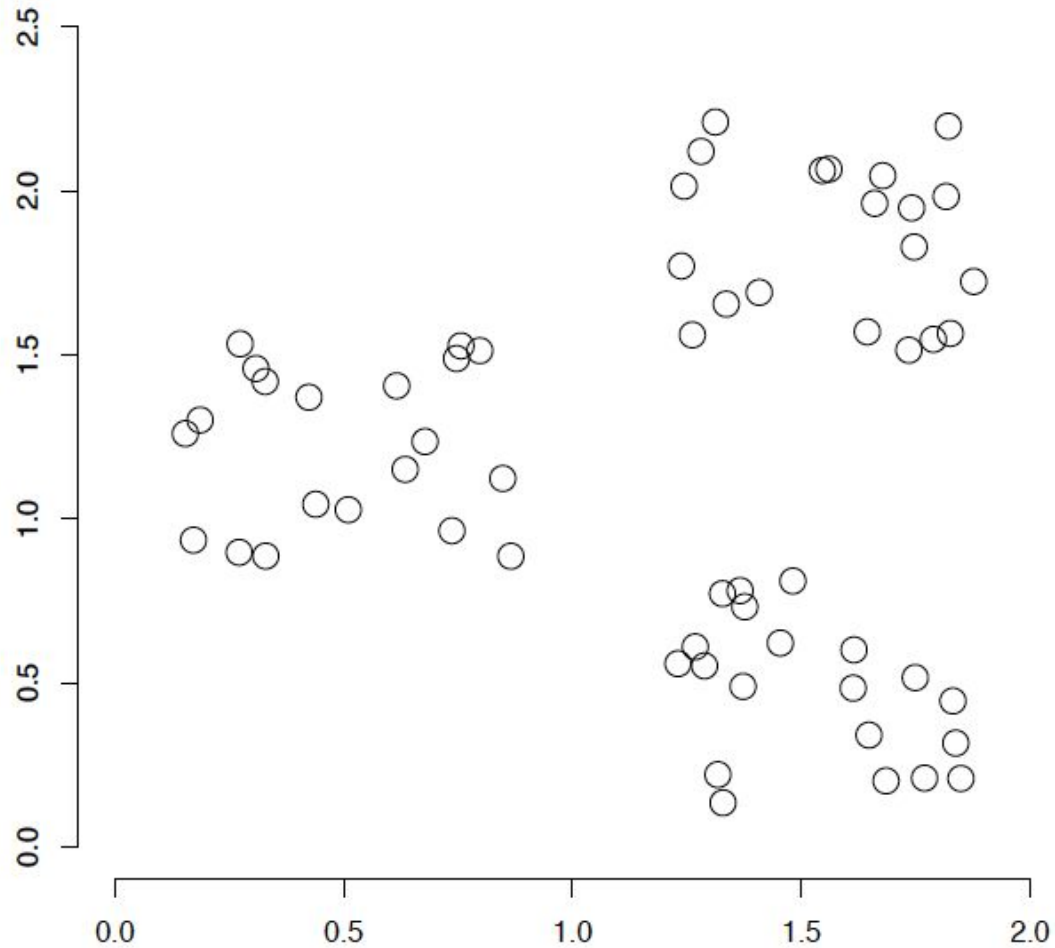
Độ đo độ tương đồng/khoảng cách

Trong không gian vec-tơ:

- Độ tương đồng cosine
- tf.idf
- Khoảng cách

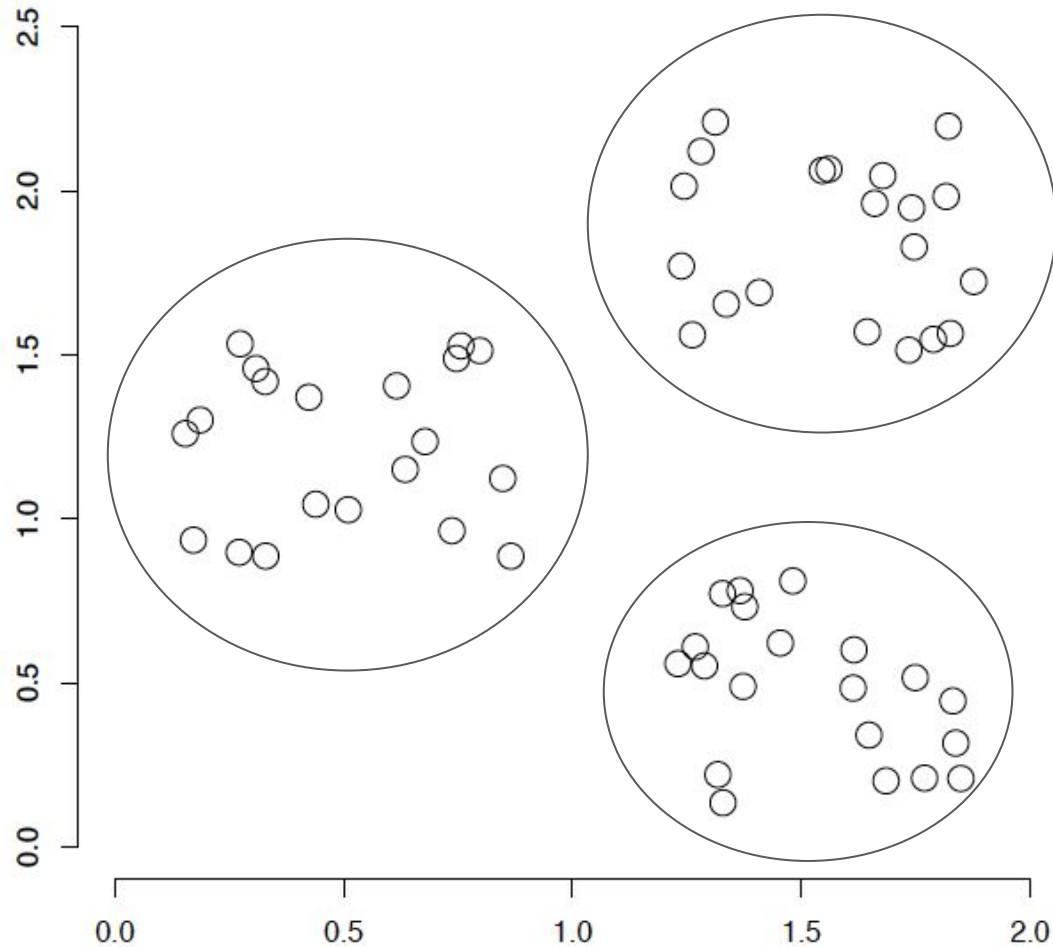
Các phương pháp chia cụm trong bài giảng này chủ yếu sử dụng khoảng cách Euclide. Trong thực tế độ tương đồng cosine và các đại lượng khác có thể được sử dụng.

Ví dụ 8.11. Chia cụm



Thử dự đoán kết quả chia cụm?

Ví dụ 8.11. Chia cụm₍₂₎



Làm sao để tìm 3 cụm như trong ví dụ này?

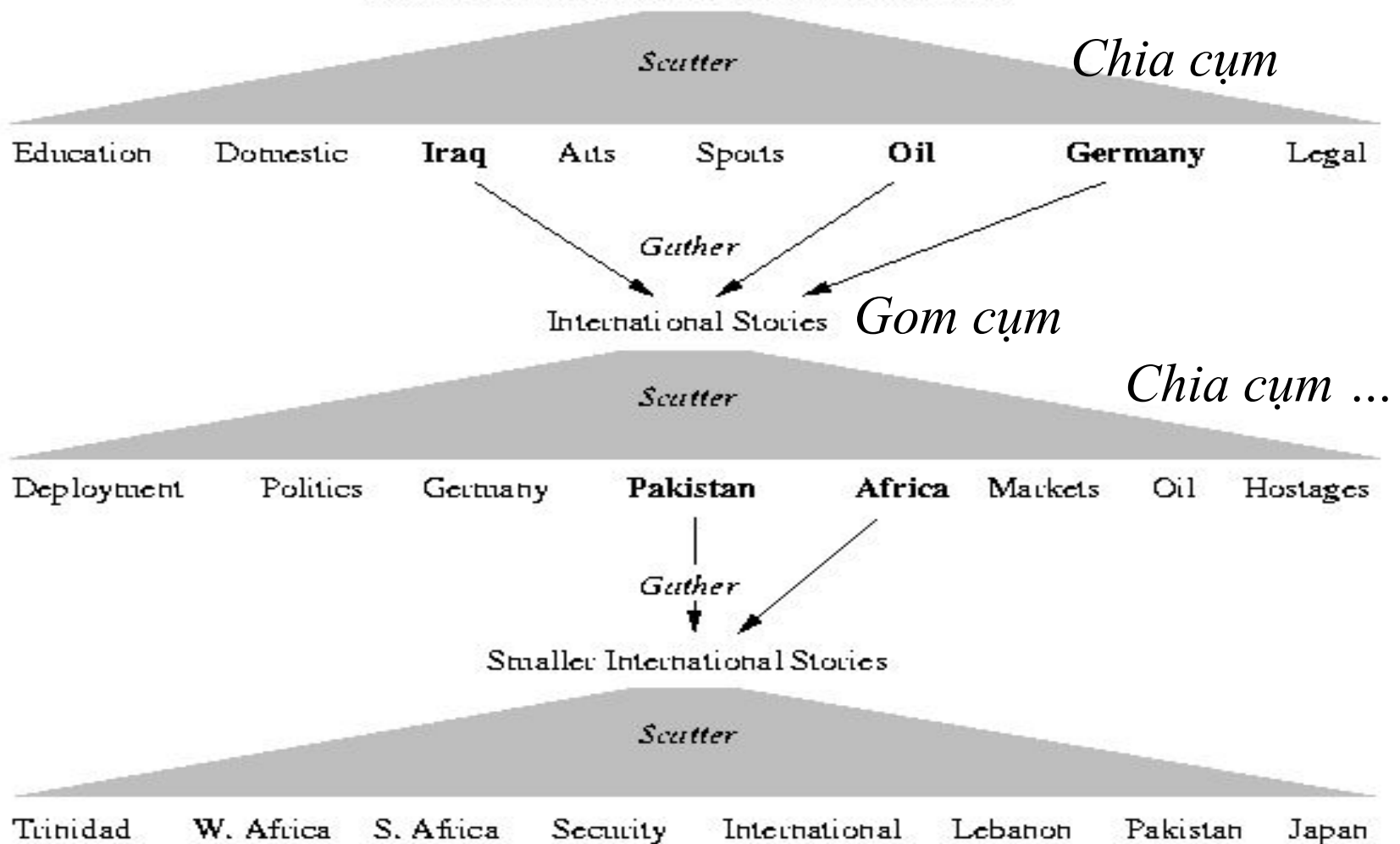
Một số ứng dụng của chia cụm trong TKTT

- Duyệt nội dung theo chủ đề
- Mở rộng phạm vi kết quả với các văn bản cùng cụm
- Phân cụm danh sách kết quả tìm kiếm
- Giới hạn phạm vi tìm kiếm theo cụm
 - *(Xử lý truy vấn trong VSM)*

Ý tưởng duyệt theo chủ đề

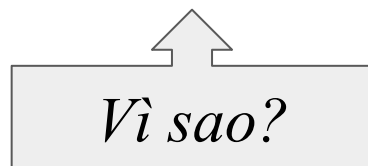
- Scatter/Gather [Cutting, Karger, and Pedersen]

New York Times News Service, August 1990



Mở rộng phạm vi kết quả

- Giả thuyết chia cụm: Các văn bản trong cùng cụm có nội dung tương tự, vì vậy các văn bản cùng cụm với kết quả tìm kiếm phù hợp có khả năng cao cũng là kết quả phù hợp.
- Để tăng số lượng kết quả phù hợp được trả về:
 - Phân cụm tập văn bản;
 - Với mỗi văn bản d trong kết quả tìm kiếm, hệ thống trả về thêm 1 số văn bản khác trong cùng cụm với d
- Hệ quả của mở rộng phạm vi kết quả:
 - Truy vấn "ô tô" có thể trả về cả các văn bản chứa từ "xe hơi"
 - *(Giả sử các văn bản chứa từ "ô tô" và các văn bản chứa từ "xe hơi" được đưa vào cùng cụm.)*



Phân cụm kết quả

← → http://search.yippy.com/search?v%3aproject=clusty&v%3afile=viv_Xpo6AV&v%3arecluster=&

Most Visited Getting Started Latest Headlines Fridge Filters

Y! Yahoo! Search SEARCH

web news images wikipedia jobs more »

clustering Search advanced preferences

clouds sources sites remix

All Results (185)

- Analysis (23)
- Method (22)
- Computing (15)
- Search, Engine (13)
- Hierarchical (16)
- Definition (11)
- High availability (13)
- Linux (11)
- Windows, Microsoft (9)
- Papers (8)

more | all clouds

find in clouds: Find

Font size: A A A A

Yippy Approved Shakespeare Searched

Top 179 results retrieved for the query **clustering** (definition) (details)

Clustering
Lower Latency In Your Data Center w/ Intel's **Cluster** Ready Solutions!
www.intel.com

Load Balancing 101
Learn the 'Nuts & Bolts' of Load Balancing with F5's White Paper
www.f5.com/load_balancing

Affordable Load Balancers
High Performance Load Balancing Solutions From KEMP- See Demo Today
kemptechnologies.com

Computer cluster - Wikipedia, the free encyclopedia
Middleware such as MPI (Message Passing Interface) or PVM (Parallel Virtual Machine) permits compute **clustering** programs to be portable to a /Computer_cluster
en.wikipedia.org/wiki/Computer_cluster - [cache] - Bing, Yahoo!

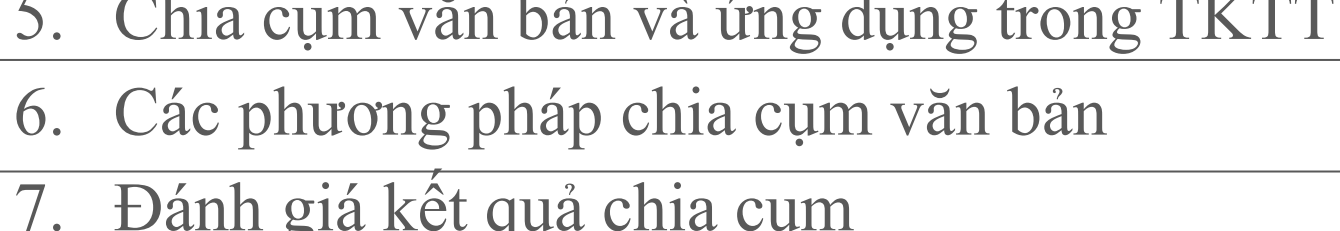
Writer's Web: Prewriting: Clustering
Prewriting: **Clustering** Melanie Dawson & Joe Essid (printable version here) **Clustering** is a type of prewriting that allows you to explore many ideas
writing2.richmond.edu/writing/wwwweb/cluster.html - [cache] - Bing, Yahoo!

Getting Started: Clustering Ideas - CT Community Colleges
Clustering. **Clustering** is similar to another process called Brainstorming. **Clustering** is something that you can do on your own or with friends or grammar.ccc.commnet.edu/grammar/composition/brainstorm_cluster.htm
grammar.ccc.commnet.edu/grammar/composition/brainstorm_clustering.htm - [cache] - Bing, Yahoo!

Advanced Clustering | Home

Ít gặp trong thực tế

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
 2. Các phương pháp phân lớp văn bản
 3. Các phương pháp trích chọn đặc trưng
 4. Đánh giá kết quả phân lớp
 5. Chia cụm văn bản và ứng dụng trong TKTT
 6. Các phương pháp chia cụm văn bản
 7. Đánh giá kết quả chia cụm
- 

Phân loại theo đường biên cụm

- Chia cụm cứng: Mỗi văn bản thuộc về đúng một cụm
- Chia cụm mềm: Một văn bản có thể thuộc nhiều hơn 1 cụm
 - Thích hợp hơn để tổ chức văn bản theo cấu trúc cây
 - Bạn có thể xếp đôi giày thể thao vào 2 cụm: (i) Thể thao và (ii) giày dép. => chia cụm mềm
- Chia cụm cứng dễ thực hiện hơn và phổ biến hơn.

Các phương pháp chia cụm được phân tích trong bài giảng này đều sử dụng đường biên cứng

Phân loại theo cấu trúc cụm

- Chia cụm phẳng

- Các cụm ngang hàng, không có các cụm lồng nhau;
- Thường bắt đầu với một phân cụm ngẫu nhiên;
- Sau đó tinh chỉnh dần theo vòng lặp;
- Ví dụ, chia cụm K-means

- Chia cụm phân cấp

- Các cụm được tổ chức theo cấu trúc cây, một cụm có thể được chia thành nhiều cụm nhỏ hơn.
- Chia cụm có thể được thực hiện theo hướng:
 - Từ đáy lên, các cụm nhỏ được kết hợp lại (agglomerative)
 - Từ đỉnh xuống, các cụm lớn được chia nhỏ (divisive)

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
 2. Các phương pháp phân lớp văn bản
 3. Các phương pháp trích chọn đặc trưng
 4. Đánh giá kết quả phân lớp
 5. Chia cụm văn bản và ứng dụng trong TKTT
 6. Các phương pháp chia cụm văn bản
- 6.1. Giải thuật K-means
- 6.2. Chia cụm phân cấp
7. Đánh giá kết quả chia cụm

Giải thuật K-Means

- Các văn bản được biểu diễn như các vec-tơ
- Cụm được xác định dựa trên các trọng tâm (tâm trọng lực) của các điểm trong một cụm, với cụm c :

$$\vec{\mu}(c) = \frac{1}{|c|} \sum_{x \in c} \vec{x}$$

- Văn bản được đưa vào cụm của trọng trọng tâm gần nhất
 - Hoặc chỉ số tương đương khác, ví dụ chỉ số tương đồng

Trọng tâm được xác định tương tự các giải thuật Rocchio

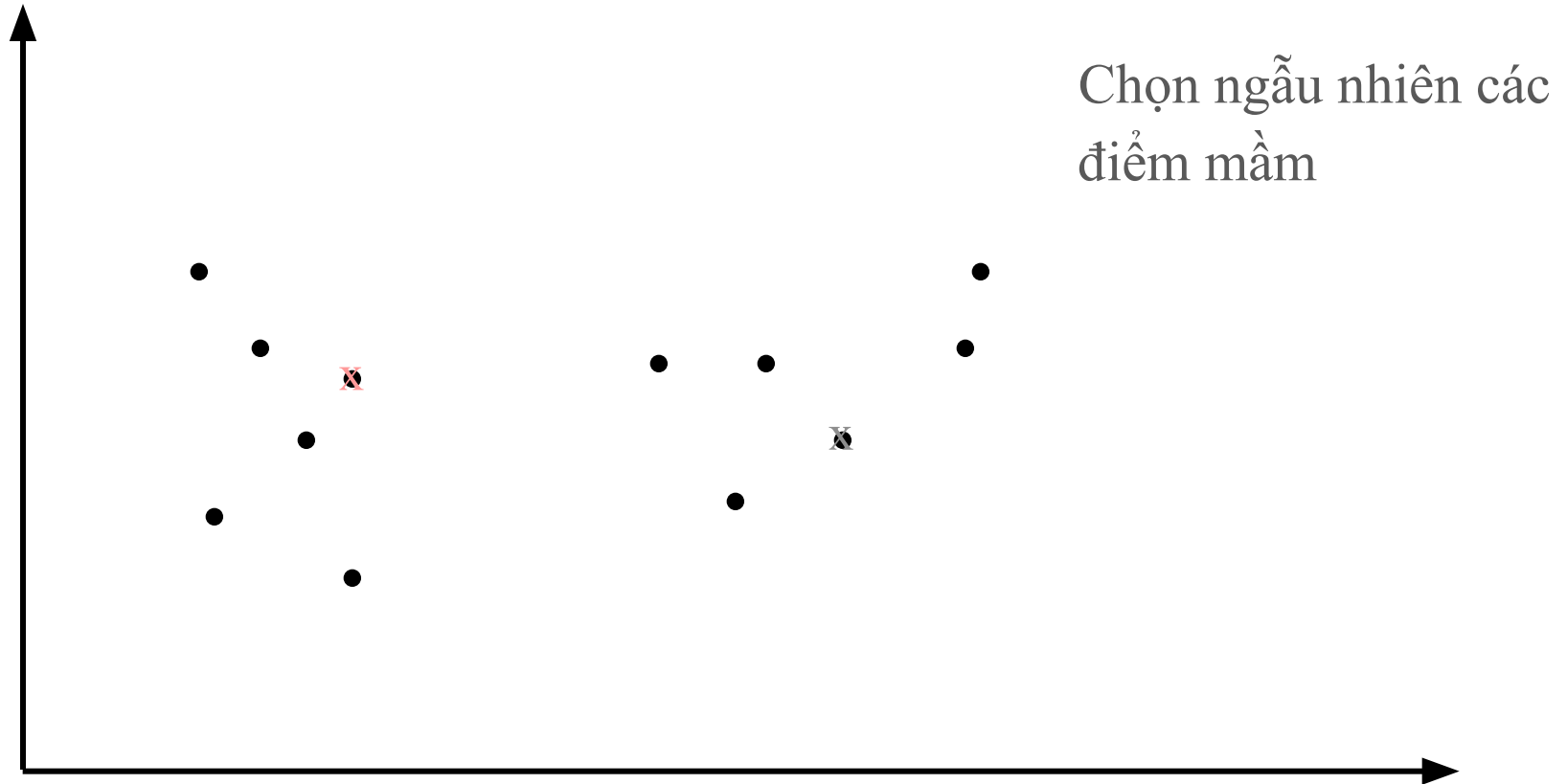
Giải thuật K-Means

- Lựa chọn K điểm mầm: Tập mầm $\{s_1, s_2, \dots, s_K\}$:
 - Lựa chọn ngẫu nhiên K văn bản;
 - K tâm cụm trong lượt chia cụm trước.
 - v.v..
- Lặp chuỗi thao tác sau cho tới khi hội tụ:
 - Gán các văn bản d_i vào cụm c_j sao cho khoảng cách $\text{dist}(x_i, s_j)$ là cực tiểu
 - Cập nhật điểm mầm bằng tâm cụm.

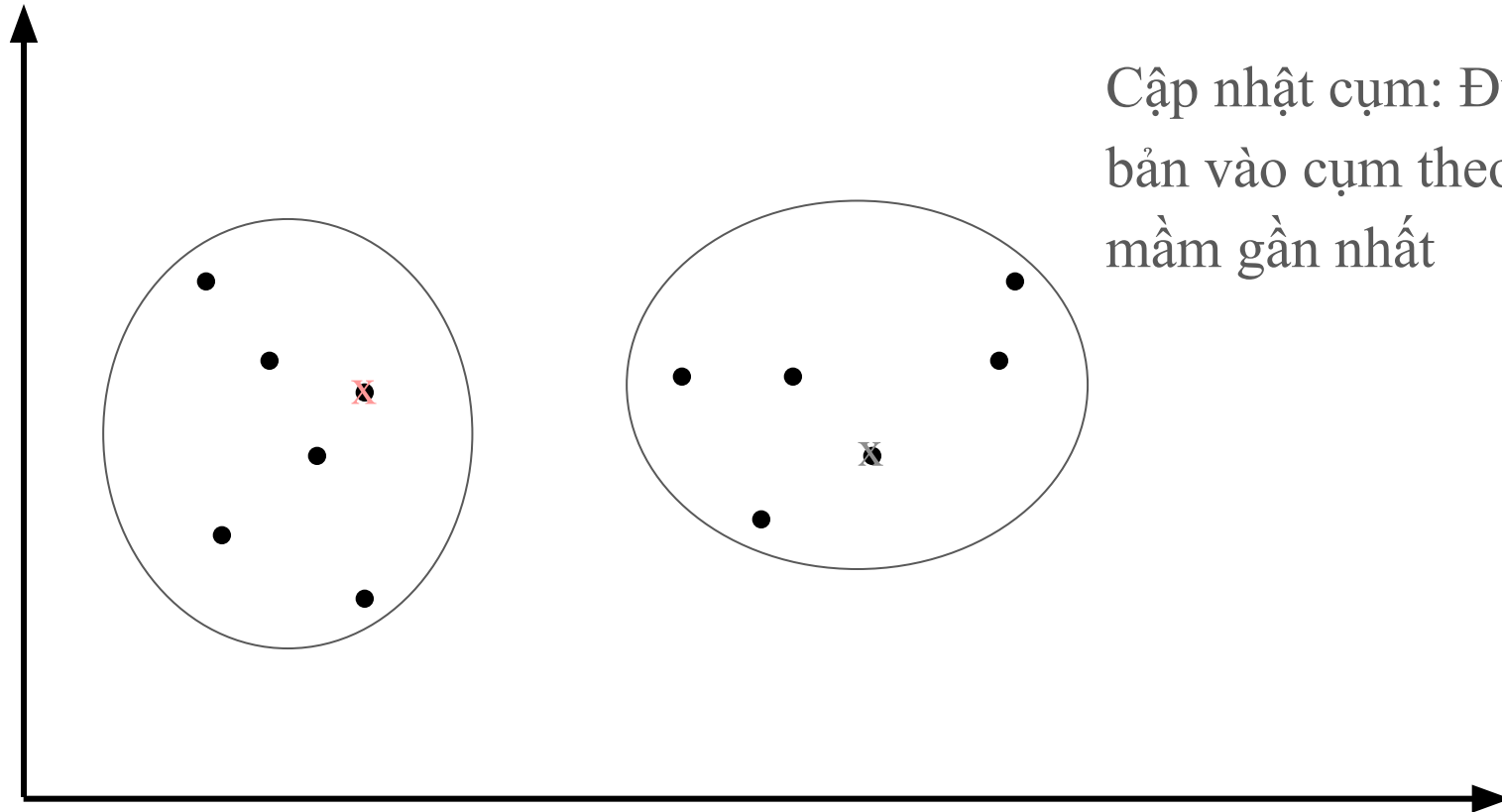
for each c_j

$$\vec{s}_j = \vec{\mu}(c_j)$$

Ví dụ 8.12. K-Means ($K = 2$)

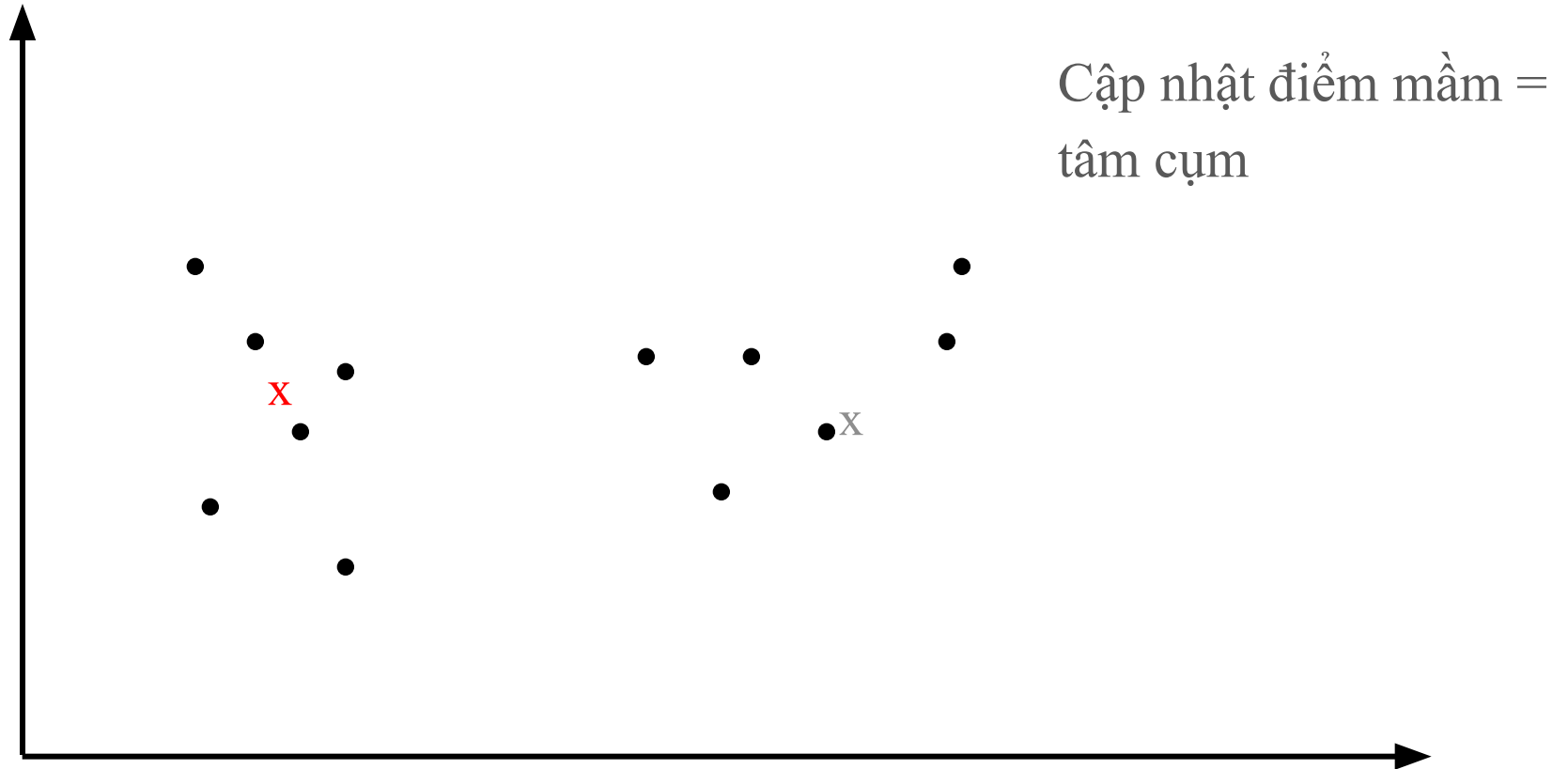


Ví dụ 8.12. K-Means ($K = 2$)₍₂₎

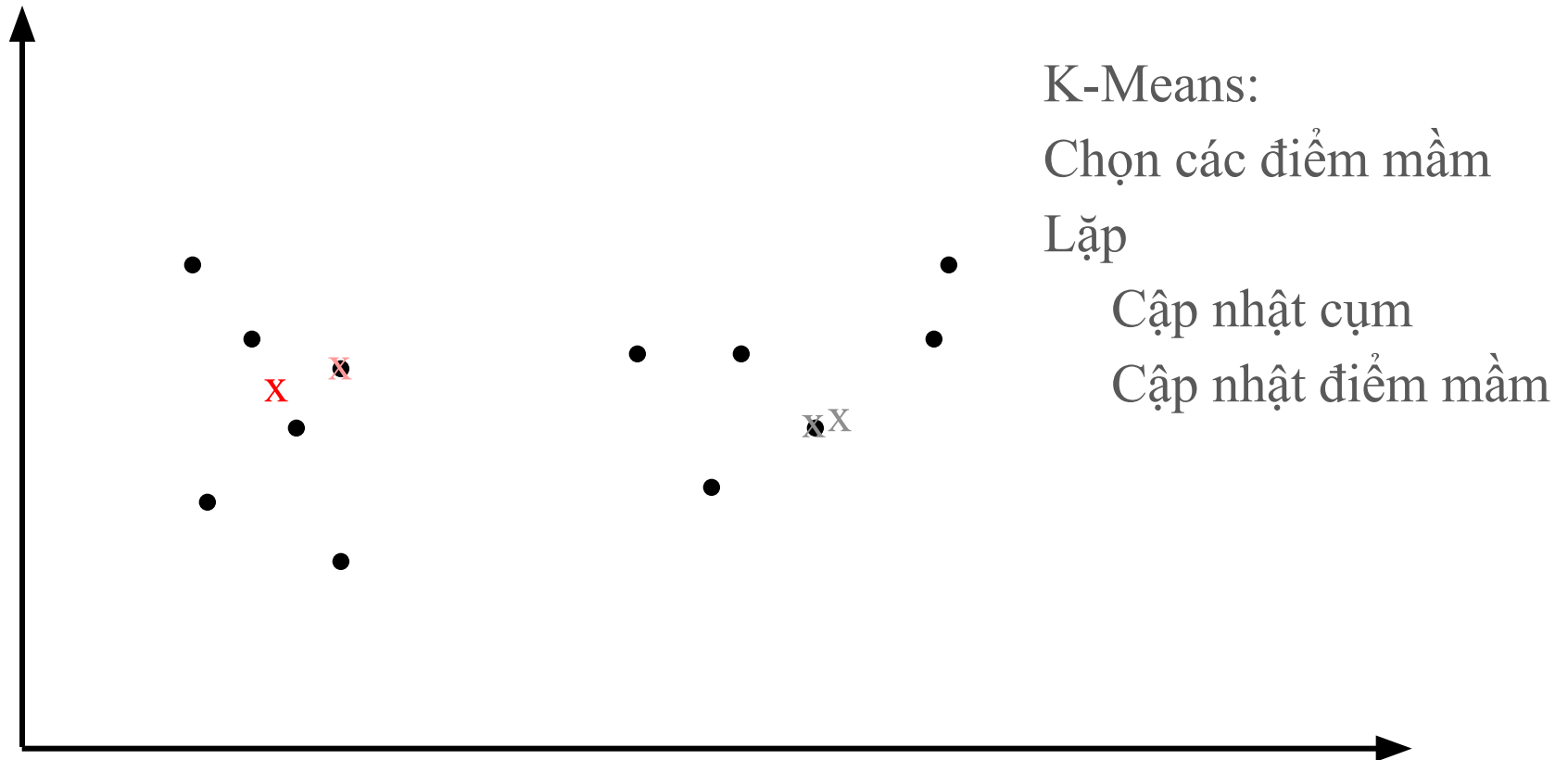


Cập nhật cụm: Đưa văn bản vào cụm theo điểm gần nhất

Ví dụ 8.12. K-Means ($K = 2$)₍₃₎



Ví dụ 8.12. K-Means ($K = 2$)₍₄₎



K-means: Các điều kiện dừng

- Giới hạn số lượng vòng lặp: Dừng sau n vòng lặp.
 - Các cụm không thay đổi
 - Vị trí tâm cụm không thay đổi
- } *Tương đương?*

Tính hội tụ của K-Means

- Vì sao giải thuật K-means có thể tiến tới trạng thái ổn định?
 - Trạng thái mà các cụm không thay đổi sau vòng lặp.
- K-means là một trường hợp đặc biệt của giải thuật cực đại hóa kỳ vọng (EM/Expectation Maximization)
 - EM luôn hội tụ
 - Số lượng vòng lặp có thể lớn
 - Nhưng ít khi xảy ra trong thực tế.

Tính hội tụ của K-Means₍₂₎

- Tổng bình phương khoảng cách tới tâm (Residual Sum of Squares/RSS):
 - Chia cụm hướng tới RSS nhỏ nhất.
- $RSS_j = \sum_i |d_i - c_j|^2$ (cộng tổng với tất cả d_i thuộc cụm c_j)
- $RSS = \sum_j RSS_j$
- Các thao tác cập nhật trong vòng lặp K-means liên tục làm giảm RSS
- Bởi vì ...

Các bước cập nhật trong K-means

- $RSS = \sum_{k=1..K} RSS_k$
- $RSS_k(\vec{\mu}) = \sum_{\vec{x} \in \omega_k} \|\vec{\mu} - \vec{x}\|^2$
- $RSS_k(\vec{\mu}) = \sum_{\vec{x} \in \omega_k} \sum_{i=1..M} (\mu_i - x_i)^2$
- $\frac{\partial RSS_k(\vec{\mu})}{\partial \mu_i} = \sum_{\vec{x} \in \omega_k} 2(\mu_i - x_m)$ → nếu =0 thì RSS_k đạt cực tiểu.
- $\mu_i = \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} x_i$ ←

Cả 2 thao tác cập nhật đều làm giảm RSS (nếu có thay đổi):

- RSS có xu hướng giảm sau khi cập nhật tâm cụm.
- RSS có xu hướng giảm sau khi đưa văn bản vào điểm mầm của cụm gần nhất.

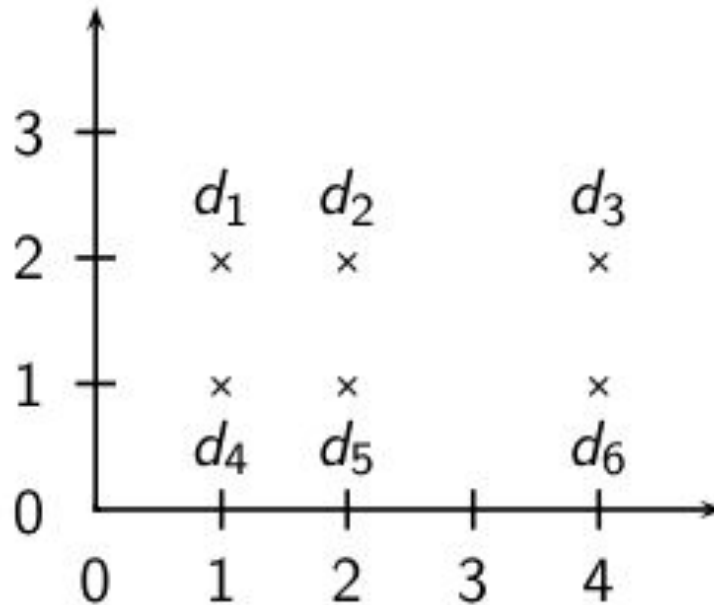
Độ phức tạp

- Tính khoảng cách giữa 2 văn bản: $O(M)$, trong đó M là số chiều của không gian vec-tơ = số lượng từ duy nhất trong tập văn bản.
- Gán lại cụm: $O(KN)$ thao tác tính khoảng cách = $O(KNM)$ thao tác cơ bản.
- Cập nhật trọng tâm: Mỗi văn bản được xử lý 1 lần: $O(NM)$
- Giả sử giải thuật hội tụ sau I vòng lặp: $O(IKNM)$

Cực tiểu địa phương của RSS

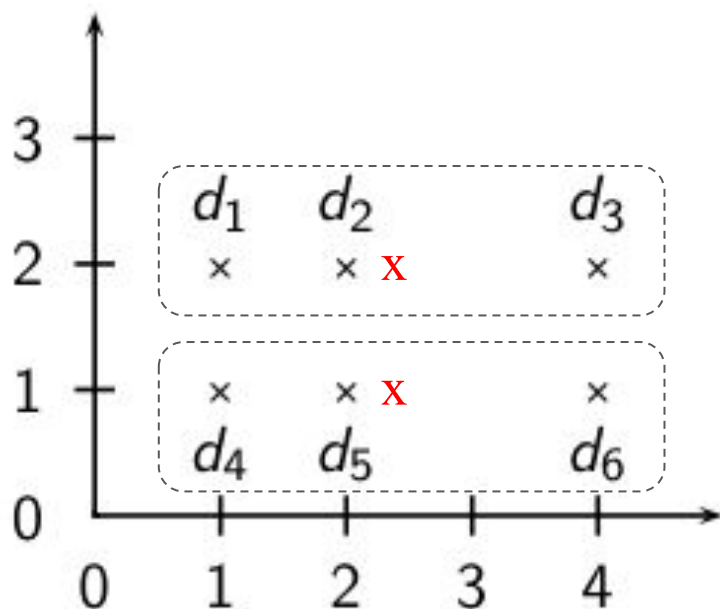
- Có thể có nhiều trạng thái hội tụ khác nhau phụ thuộc vào cách lựa chọn tập mẫu
- Giải thuật chia cụm có thể hội tụ chậm, hoặc hội tụ ở trạng thái tối ưu địa phương:
 - Làm sao để lựa chọn tập mẫu tốt?
 - Khởi tạo với các văn bản được lựa chọn ngẫu nhiên?
 - Thử nhiều phương án khởi tạo?
 - Khởi tạo với các kết quả của phương pháp khác?
 - v.v..

Ví dụ 8.13. Cực tiểu địa phương

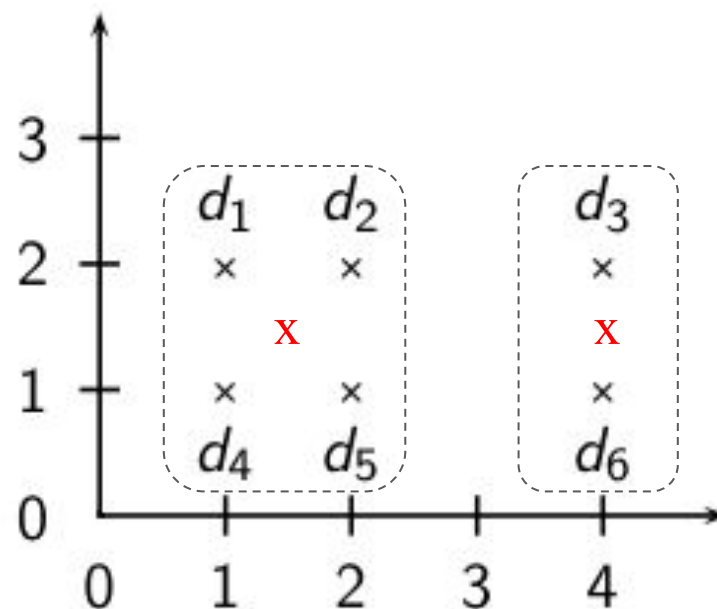


Thử chia thành 2 cụm với các tập mẫu $\{d_2, d_5\}$ và $\{d_2, d_3\}$?

Ví dụ 8.13. Cực tiểu địa phương²₍₂₎



$\{d_2, d_5\} - (7/3, 1) (7/3, 2)$
 $RSS1 = ?$



và $\{d_2, d_3\} - (1.5, 1.5), (4, 1.5)$
 $RSS2 = ?$

Thử tính đại lượng RSS cho các trạng thái hội tụ?

Một số đặc điểm của K-means

- Tính lại trọng tâm khi thay đổi cụm của 1 văn bản (thay vì sau khi xử lý tất cả văn bản) có thể tăng tốc độ hội tụ của K-means.
- Các cụm được cho là các hình cầu trong không gian vec-tơ
 - Kết quả dễ thay đổi theo thay đổi tọa độ, cách đánh giá trọng số
- Các cụm không giao nhau, có thể hợp nhất lại thành tập văn bản ban đầu
 - Không có khái niệm ngoại lệ
 - Nhưng có thể bổ xung các bộ lọc.

Vấn đề lựa chọn số lượng cụm tối ưu

- Trong 1 số trường hợp như với các kết quả tìm kiếm cho 1 truy vấn, số lượng cụm thường không được biết trước.
- Lợi ích của kết quả chia cụm:
 - Coi độ tương đồng cosine với tâm cụm là lợi ích của mỗi văn bản
 - Lợi ích của kết quả chia cụm là tổng lợi ích của tất cả các văn bản.
- Hạn chế số lượng cụm:
 - Sử dụng số lượng cụm phù hợp: Không quá nhiều và cũng không quá ít;
 - Trừ điểm nếu có quá nhiều cụm

Thử tìm cách chia cụm N văn bản với lợi ích cực đại ($=N$) và $RSS = 0$? Chia cụm như vậy có hữu ích cho các ứng dụng?

Vấn đề lựa chọn số lượng cụm tối ưu₍₂₎

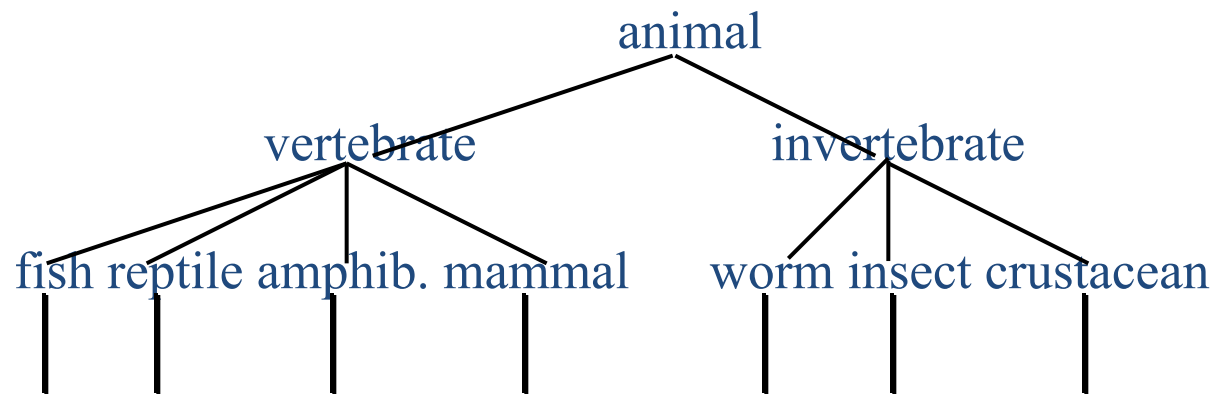
- Tính chi phí C cho mỗi cụm.
- Chi phí chia K cụm $= KC$
- Giá trị của 1 kết quả chia cụm $= \text{Lợi ích} - \text{Chi phí}$
- Mục đích: Tìm số lượng cụm K theo kết quả chia cụm với giá trị cao nhất:
 - Lợi ích lớn nhất của kết quả chia cụm tăng khi tăng K .
 - ... Nhưng chi phí cũng tăng theo K .
 - Giá trị sẽ ngừng tăng khi lợi ích tăng chậm hơn chi phí.

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
2. Các phương pháp phân lớp văn bản
3. Các phương pháp trích chọn đặc trưng
4. Đánh giá kết quả phân lớp
5. Chia cụm văn bản và ứng dụng trong TKTT
6. Các phương pháp chia cụm văn bản
 - 6.1. Giải thuật K-means
 - 6.2. Chia cụm phân cấp
7. Đánh giá kết quả chia cụm

Chia cụm phân cấp

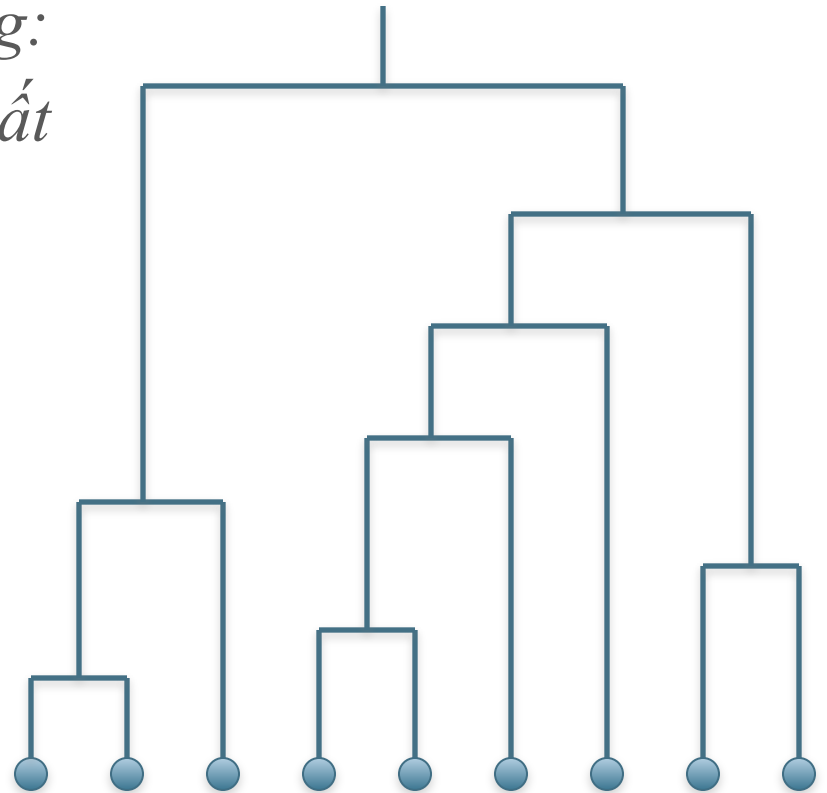
- Tổ chức tập văn bản theo một cấu trúc phân cấp dạng cây/cây phân cấp (dendrogram)



- Các hướng tiếp cận:
 - Từ đỉnh đến đáy/Phân rã: Áp dụng giải thuật chia cụm nhiều lần theo hình thức đệ quy;
 - Từ đáy tới đỉnh/Kết hợp: Lặp hợp nhất các cụm nhỏ thành cụm lớn hơn, bắt đầu với trạng thái mỗi văn bản là 1 cụm.

Suy diễn kết quả từ cây phân cấp

*Kết quả chia cụm thu được bằng
cách cắt cây phân cấp ở một tầng:
Các tầng phía dưới được hợp nhất
thành cụm.*



Hướng tiếp cận từ đáy tới đỉnh/Kết hợp

Hierarchical Agglomerative Clustering (HAC) - Chia cụm phân cấp theo hướng kết hợp.

- Bắt đầu với trạng thái mỗi văn bản là 1 cụm
 - Sau đó lặp kết hợp cặp các cụm gần nhau nhất, cho tới khi chỉ còn 1 cụm duy nhất.
- Tiến trình hợp nhất tạo thành một cây nhị phân/cây phân cấp
- Cặp cụm gần nhau nhất có thể được xác định theo:
 - Độ tương đồng của cặp văn bản có độ tương đồng lớn nhất: Liên kết đơn (Single-link)
 - Độ tương đồng của cặp văn bản xa nhất, có độ tương đồng nhỏ nhất: Liên kết đủ (Complete-link)
 - Độ tương đồng của các trọng tâm: Trọng tâm (Centroid)
 - Trung bình độ tương đồng giữa tất cả các cặp: Liên kết trung bình (Average-link)

Chia cụm kết hợp liên kết đơn

- Độ tương đồng của 2 cụm là độ tương đồng cực đại của các cặp văn bản trong các cụm.

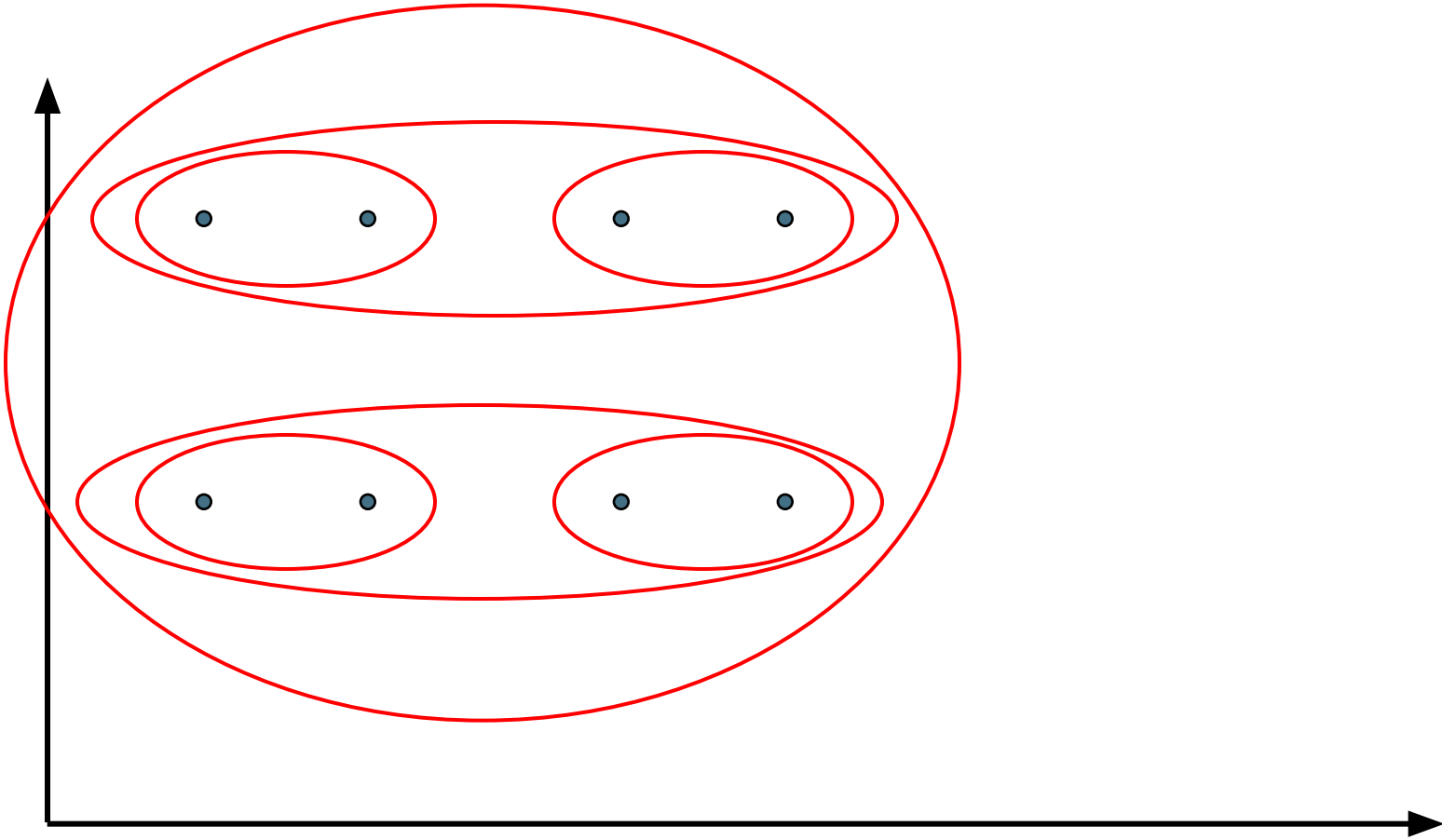
$$\text{sim}(c_i, c_j) = \max_{x \in c_i, y \in c_j} \text{sim}(x, y)$$

- Sau khi hợp nhất c_i và c_j , độ tương đồng của cụm kết quả so với cụm c_k khác là:

$$\text{sim}((c_i \cup c_j), c_k) = \max(\text{sim}(c_i, c_k), \text{sim}(c_j, c_k))$$

Sai số có thể dẫn đến các cụm mỏng và dài (trong không gian)

Ví dụ 8.14. Liên kết đơn



Liên kết đủ

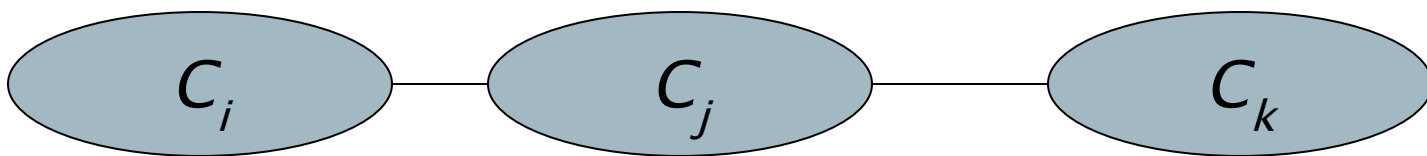
- Độ tương đồng của 2 cụm là độ tương đồng cực tiểu của các cặp văn bản

$$\text{sim}(c_i, c_j) = \min_{x \in c_i, y \in c_j} \text{sim}(x, y)$$

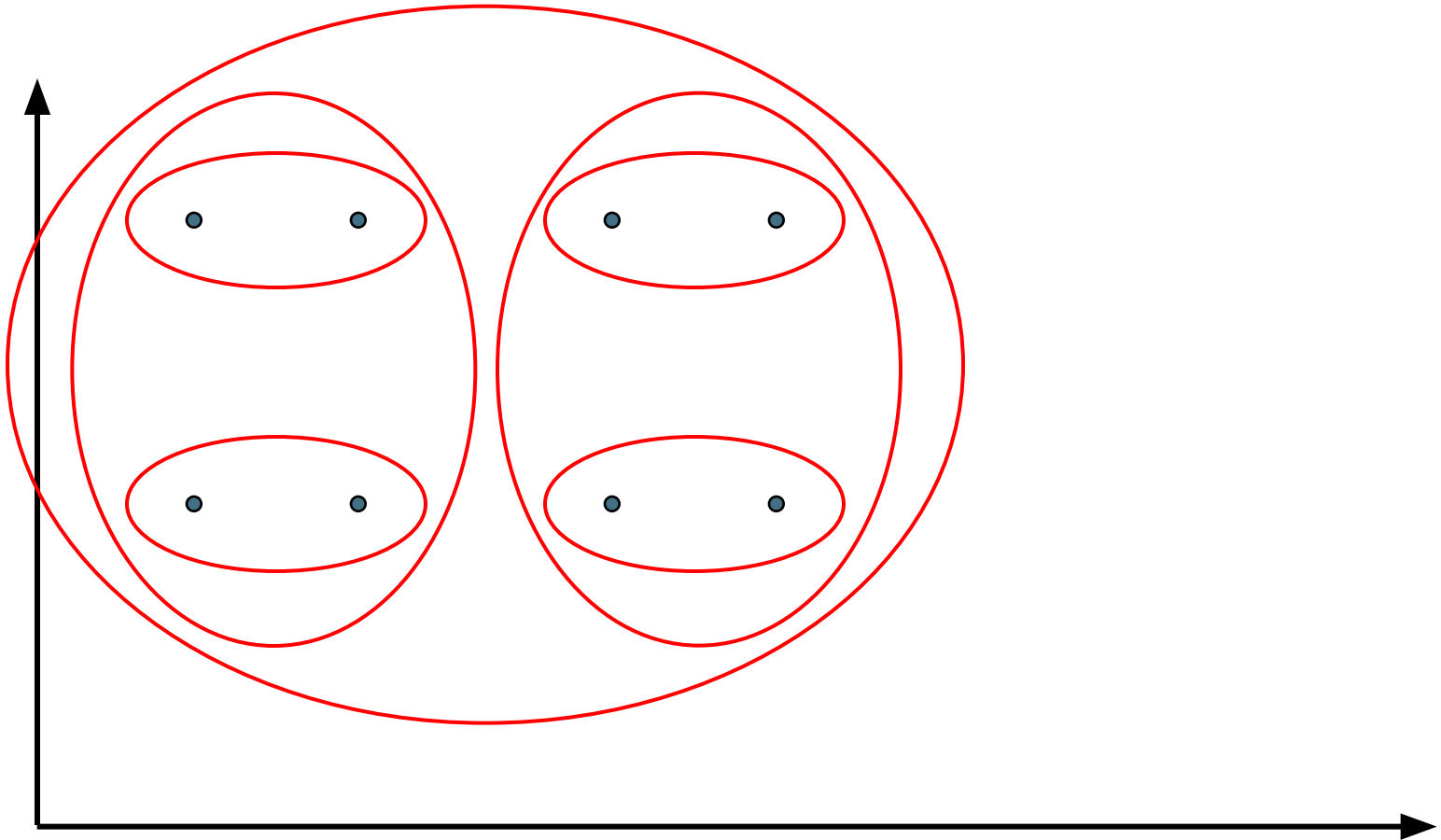
- Sau khi hợp nhất c_i và c_j , độ tương đồng của cụm kết quả với một cụm c_k khác là:

$$\text{sim}((c_i \cup c_j), c_k) = \min(\text{sim}(c_i, c_k), \text{sim}(c_j, c_k))$$

Tạo thành các cụm mật độ dày đặc hơn, giống hình cầu hơn (thường được ưa thích hơn).



Ví dụ 8.15. Liên kết đủ



Giải thuật HAC khái quát và độ phức tạp

1. Tính độ tương đồng giữa tất cả các cặp văn bản


 $O(N^2)$

2. Thực hiện N - 1 lần:

1. Tìm cặp văn bản/cụm gần nhau nhất để hợp nhất

 Naïve: $O(N^2)$  Priority Queue: $O(N)$  Single link: $O(N)$

2. Cập nhật độ tương đồng của các văn bản/cụm với cụm mới

 Naïve: $O(N)$  Priority Queue: $O(N \log N)$  Single link: $O(N)$

Nhanh nhất

Liên kết trung bình

- Độ tương đồng của 2 cụm = Trung bình độ tương đồng của tất cả các cặp có trong cụm được hợp nhất

$$sim(c_i, c_j) = \frac{1}{|c_i \cup c_j|(|c_i \cup c_j| - 1)} \sum_{\vec{x} \in (c_i \cup c_j)} \sum_{\vec{y} \in (c_i \cup c_j): \vec{y} \neq \vec{x}} sim(\vec{x}, \vec{y})$$

- Cân đối giữa liên kết đơn và liên kết đủ
- Lựa chọn khác:
 - Trung bình trên tất cả các cặp trong 2 cụm ban đầu
 - Không có sự khác biệt đáng kể về tính hiệu quả

Tính trung bình độ tương đồng

- Luôn duy trì tổng các vec-tơ trong mỗi cụm

$$\vec{s}(c_j) = \sum_{\vec{x} \in c_j} \vec{x}$$


- Tính độ tương đồng của các cụm

$$\text{sim}(c_i, c_j) = \frac{(\vec{s}(c_i) + \vec{s}(c_j)) \bullet (\vec{s}(c_i) + \vec{s}(c_j)) - (|c_i| + |c_j|)}{(|c_i| + |c_j|)(|c_i| + |c_j| - 1)}$$

Chia cụm thể nào là tốt?

- Các chỉ số bên trong: Chia cụm hướng tới giá trị tối ưu của các chỉ số:
 - Độ tương đồng trong-cụm (trong lớp) cao
 - Độ tương đồng giữa các cụm thấp
 - Chất lượng đo được của một kết quả chia cụm phụ thuộc vào biểu diễn văn bản và độ đo tương đồng được sử dụng
- Các chỉ số bên ngoài: Được biểu diễn bằng khả năng phát hiện thêm các mẫu, các lớp ẩn
 - Được đánh giá dựa trên dữ liệu kiểm thử

Nội dung

1. Phân lớp văn bản và ứng dụng trong TKTT
 2. Các phương pháp phân lớp văn bản
 3. Các phương pháp trích chọn đặc trưng
 4. Đánh giá kết quả phân lớp
 5. Chia cụm văn bản và ứng dụng trong TKTT
 6. Các phương pháp chia cụm văn bản
 7. Đánh giá kết quả chia cụm
- 

Đánh giá kết quả chia cụm

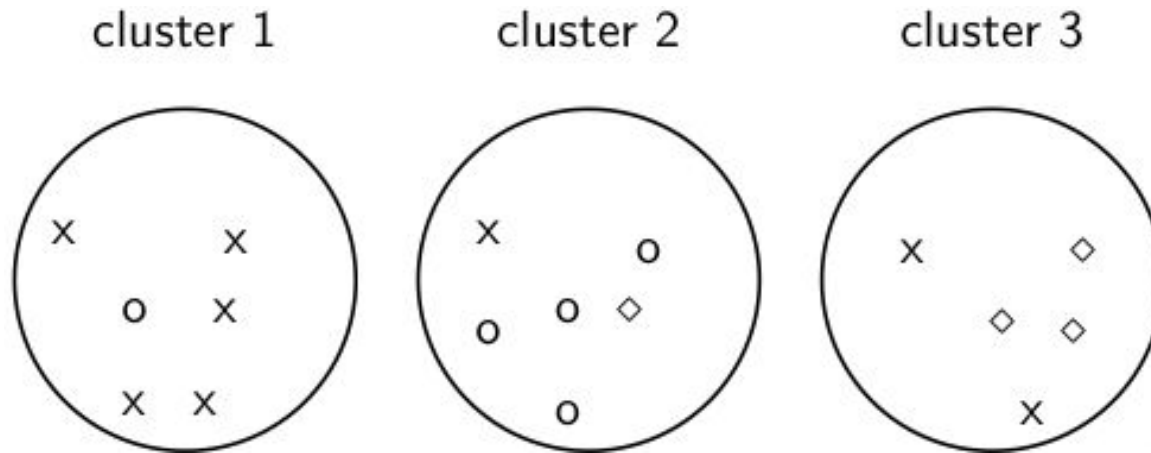
- Dữ liệu kiểm thử: Có thể coi một tập văn bản đã được gán nhãn/phân lớp như một phương án chia cụm mẫu.
 - Đánh giá kết quả chia cụm bằng cách so sánh với các lớp.
- Các độ đo thông dụng:
 - Độ thuần khiết: Purity
 - Chỉ số ngẫu nhiên: Rand Index

Độ đo Purity

$$\text{purity}(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$$

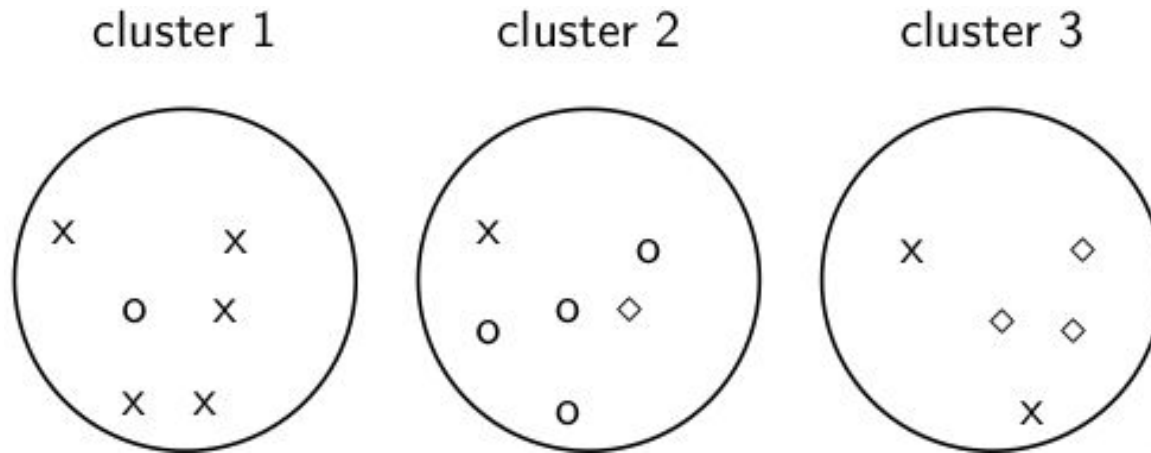
- $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$ là tập cụm,
- $C = \{c_1, c_2, \dots, c_J\}$ là tập lớp.
- Độ thuần khiết là tỉ lệ giữa số lượng văn bản thuộc lớp chiếm đa số trong cụm và số lượng mẫu trong cụm.
- Lệch về phía nhiều cụm: Độ thuần khiết của phương án chia n cụm $= n$.
- Chỉ số khác là Entropy của lớp trong cụm (hoặc hàm lượng thông tin tương hỗ giữa lớp và cụm)

Ví dụ 8.16. Tính Purity



Thử tính Purity?

Ví dụ 8.16. Tính Purity₍₂₎



- $\max_j |\omega_1 \cap c_j| = 5;$
 $\max_j |\omega_2 \cap c_j| = 4;$
 $\max_j |\omega_3 \cap c_j| = 3.$
- $\text{Purity} = (1/17) \times (5 + 4 + 3) \approx 0.71.$

Chỉ số ngẫu nhiên (Rand Index)

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

	Cùng lớp	Khác lớp
Cùng cụm	TP	FP
Khác cụm	FN	TN

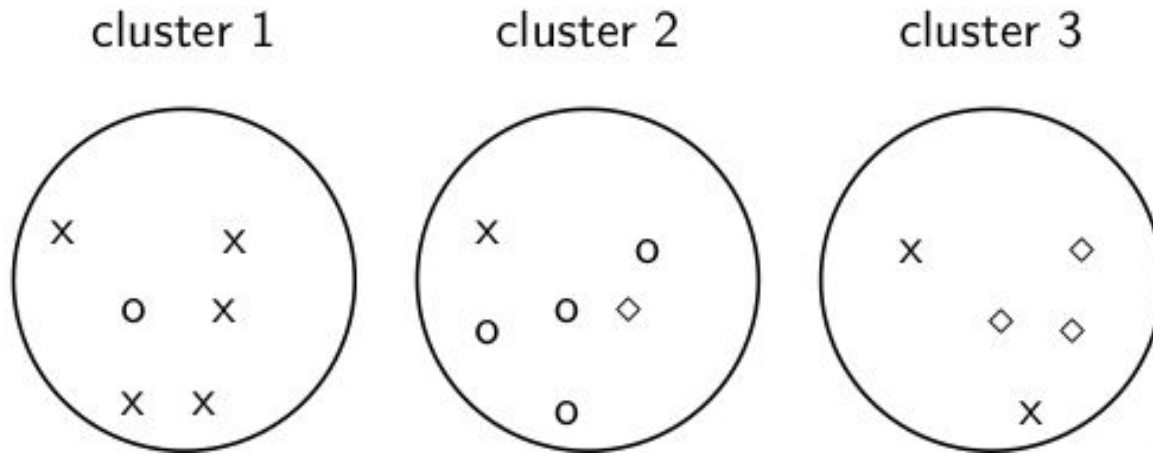
- So sánh với độ chính xác và độ đầy đủ:

$$P = TP/(TP + FP)$$

$$R = TP/(TP + FN)$$

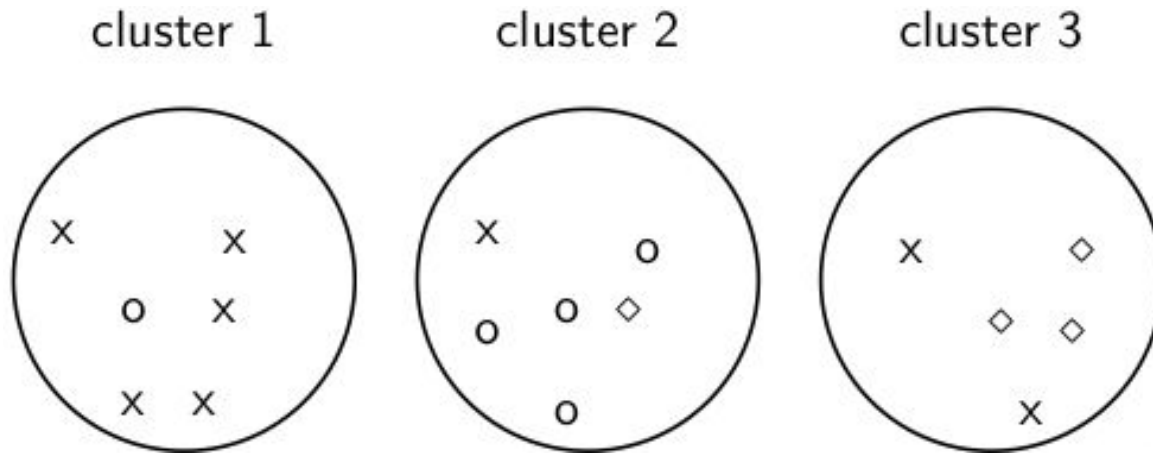
- Trong một số nghiên cứu còn định nghĩa và sử dụng độ đo F cho kết quả chia cụm.
 - Độ đo F cũng là một chỉ số tin cậy để đánh giá kết quả chia cụm

Ví dụ 8.17. Rand Index



Thử tính RI?

Ví dụ 8.17. Rand Index₍₂₎



$$TP = \binom{5}{2} + \binom{4}{2} + \binom{3}{2} + \binom{2}{2} = 20$$

$$TP + FP = \binom{6}{2} + \binom{6}{2} + \binom{5}{2} = 40$$

$FP = 40 - 20 = 20$, FN và TN được xác định tương tự.

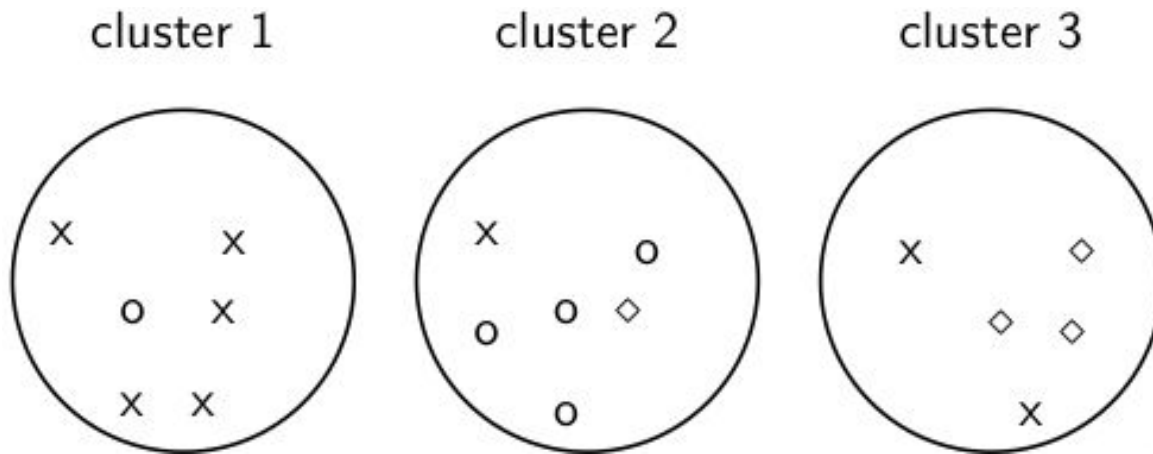
Ví dụ 8.17. Rand Index₍₃₎

	Cùng lớp	Khác lớp
Cùng cụm	TP = 20	FP = 20
Khác cụm	FN = 24	TN = 72

$$RI = \frac{TP+TN}{TP+FP+FN+TN}$$

$$RI = (20 + 72)/136$$

Tổng hợp các độ đo



	purity		RI
lower bound	0.0		0.0
maximum	1.0		1.0
value for example	0.71		0.68

Bài tập 8.1. Phân lớp Rocchio

	docID	words in document	in $c = \textit{China}$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Phân lớp Rocchio sử dụng cấu hình: Biểu diễn vec-tơ bằng tf, đánh giá độ tương đồng bằng tích vô hướng.

Bài tập 8.2. Phân lớp kNN

	docID	Các từ trong văn bản	Lớp c = China?
Tập huấn luyện	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
	5	Tokyo Kyoto Japan Japan	no
	6	Japan Japan Tokyo Shanghai	no
Tập kiểm thử	7	Chinese Chinese Japan Tokyo	??

Phân lớp kNN sử dụng cấu hình: Biểu diễn vec-tơ theo tf, tính độ tương đồng theo tích vô hướng, $k = 3$.

Bài tập 8.3

Hãy so sánh các biểu diễn văn bản theo mô hình đa thức và mô hình Bernoulli

- 1) He moved from London, Ontario, to London, England.
- 2) He moved from London, England, to London, Ontario.
- 3) He moved from England to London, Ontario.

Bài tập 8.4. MI

Hãy tính MI cho Kyoto/Japan

	docID	Từ trong văn bản	c = Japan?
Tập huấn luyện	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no

Bài tập 8.5

Hãy lựa chọn 2 từ theo χ^2 và MI dựa trên các thống kê sau cho lớp Coffee

Từ	N₀₀	N₀₁	N₁₀	N₁₁
brazil	98012	102	1835	51
council	96322	133	3525	20
producers	98524	119	1118	34
roasted	99824	143	23	10

Bài tập 8.6

Đối với các đại lượng trong công thức tính χ^2 hãy chứng minh:

$$|N_{11} - E_{11}| = |N_{10} - E_{10}| = |N_{01} - E_{01}| = |N_{00} - E_{00}|$$

Bài tập 8.7

Hãy phân lớp theo Naive Bayes đa thức và Naive Bayes Bernoulli, có thể giản lược các tham số không sử dụng tới.

	docID	Từ trong văn bản	c = China?
Tập huấn luyện	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
Tập kiểm thử	5	Taiwan Taiwan Sapporo	?

Bài tập 8.8

Hai điều kiện dừng của giải thuật K-means:

- (i) Kết quả phân cụm không thay đổi;
- (ii) Tâm cụm không thay đổi.

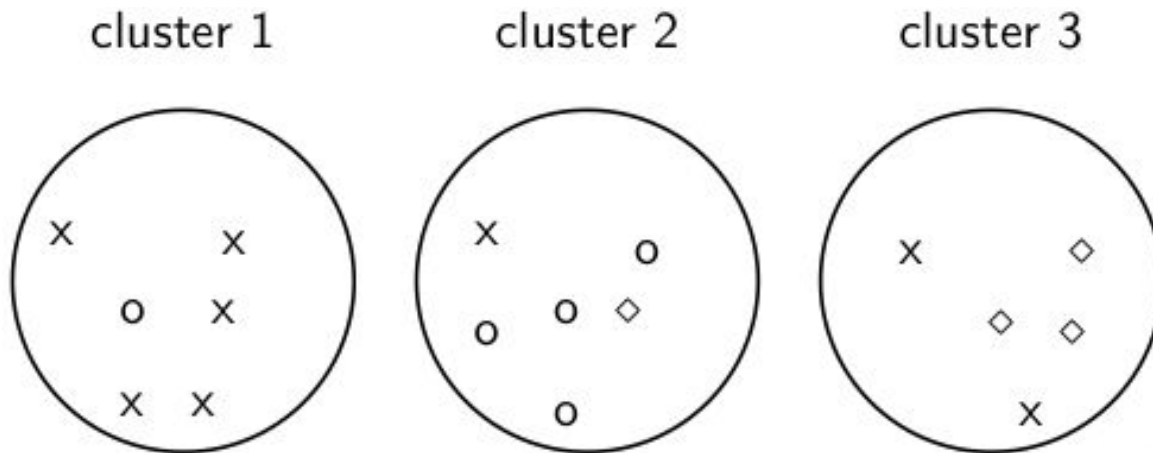
Từ điều kiện (i) có suy ra được điều kiện (ii) hay không?

Từ điều kiện (ii) có suy ra được điều kiện (i) hay không?

Bài tập 8.9

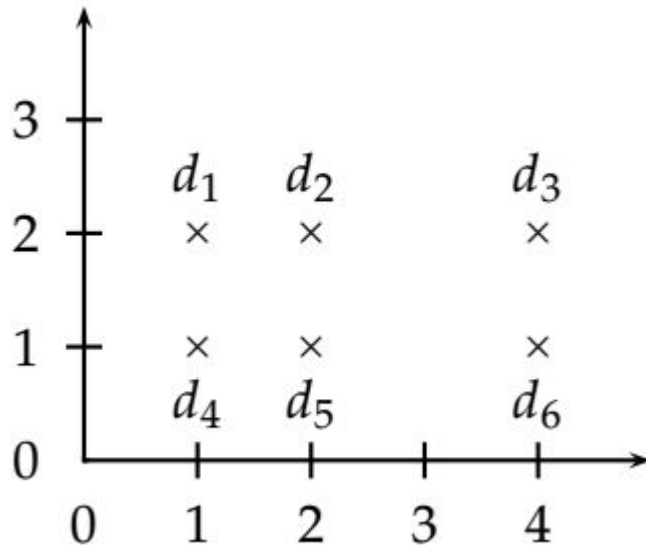
Thay thế mỗi văn bản trên hình vẽ bằng hai văn bản. Sau đó hãy tính Purity và RI.

Thêm các văn bản trùng lặp có làm quá trình chia cụm khó hơn không? Đại lượng nào thay đổi/không thay đổi?



Bài tập 8.10

Hãy tính RSS cho kết quả chia cụm trong cả hai trường hợp.



$$\{\{d_1, d_2, d_3\}, \{d_4, d_5, d_6\}\}.$$

$$\{\{d_1, d_2, d_4, d_5\}, \{d_3, d_6\}\},$$

Bài tập 8.11

Hãy lấy ví dụ một tập điểm và 3 trọng tâm ban đầu sao cho kết quả phân cụm 3-means hội tụ với cụm rỗng. (ii) Kết quả chia cụm với cụm rỗng có thể là kết quả tối ưu toàn cục theo RSS?

Bài tập 8.12

Chứng minh $RSS_{\min}(K)$ là hàm đơn điệu giảm đối với biến K .

