



IT4853

# Tìm kiếm và trình diễn thông tin

---

Bài 18. Chia cụm văn bản

IIR.C16. Flat clustering

TS. Nguyễn Bá Ngọc, *Bộ môn Hệ thống thông tin,  
Viện CNTT & TT*  
*ngocnb@soict.hust.edu.vn*

Hà Nội, 2016



# Nội dung chính

---

- Bài toán chia cụm
- Ứng dụng chia cụm trong tìm kiếm
- Giải thuật K-means

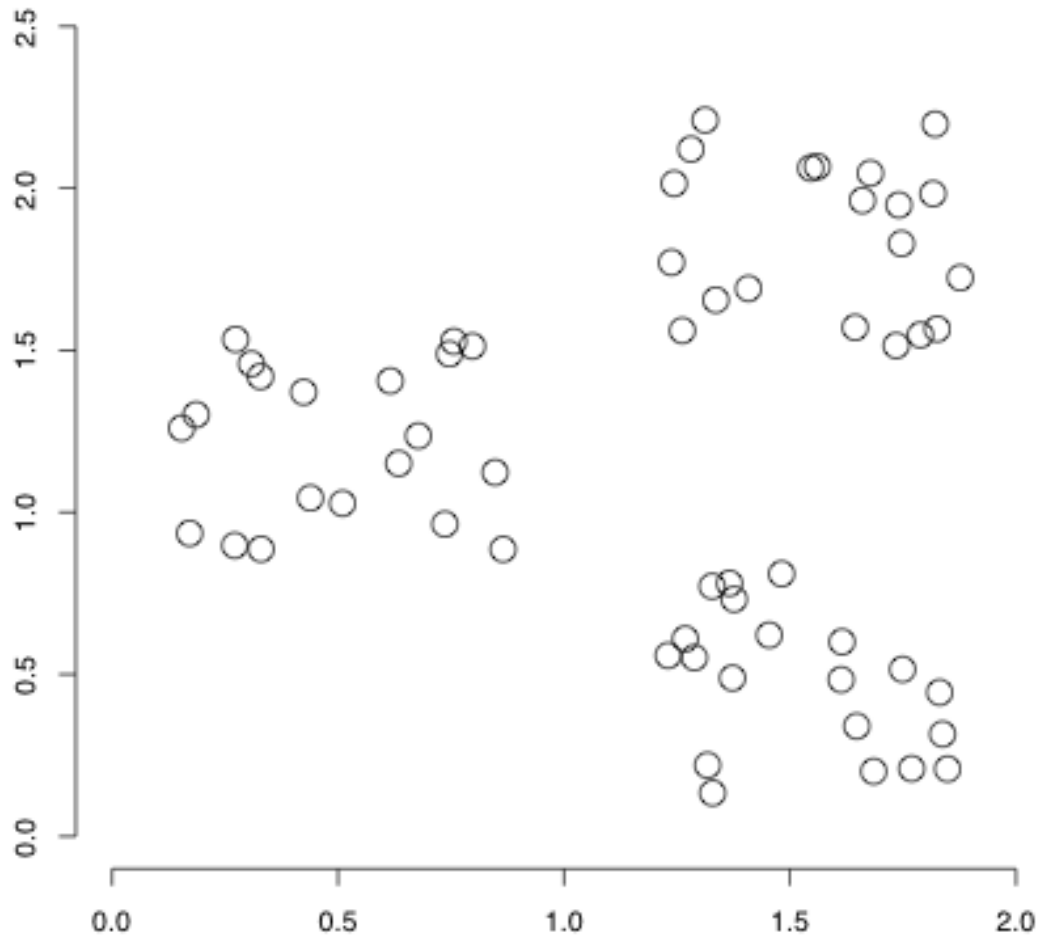


# Bài toán chia cụm

---

- Chia cụm là chia một tập văn bản lớn thành nhiều tập nhỏ với nội dung tương tự. Mỗi tập văn bản nhỏ là một cụm:
  - Các văn bản trong cùng một cụm phải giống nhau;
  - Các văn bản khác cụm phải khác nhau;
  - Số lượng cụm phải phù hợp với bộ dữ liệu:
    - Có thể được xác định bằng phương pháp bán tự động.
- Mục tiêu phụ:
  - Kích thước cụm không quá lớn hoặc quá nhỏ;
  - Các cụm phản ánh một chủ đề tường minh, cụ thể;
  - V.V.

## Bài toán chia cụm (2)



Làm cách nào để  
chia cụm như trong  
hình vẽ?



# Phân lớp vs. chia cụm

---

- Phân lớp: Học có giám sát
  - Sử dụng dữ liệu luyện;
  - Phân lớp mẫu được thực hiện thủ công.
- Chia cụm: Học không giám sát
  - Cụm được suy diễn trực tiếp từ dữ liệu;
  - Không sử dụng dữ liệu luyện;
  - Có thể tùy chỉnh giải thuật bằng các tham số: số cụm, độ tương đồng, biểu diễn văn bản v.v.



# Cụm phẳng vs. cụm phân cấp

---

- Giải thuật chia cụm phẳng:
  - Thường bắt đầu với một cách chia ngẫu nhiên;
  - Sau đó lặp quá trình xác định lại cụm;
  - Giải thuật tiêu biểu: K-means.
- Chia cụm phân cấp:
  - Tổ chức cụm theo cấu trúc cây;
  - Bottom-up, agglomerative;
  - Top-down, divisive.



# Đường biên cứng vs. mềm

---

- Đường biên cứng: Mỗi văn bản chỉ thuộc một cụm duy nhất.
  - Đơn giản hơn so với chia cụm mềm;
- Đường biên mềm: Mỗi văn bản có thể thuộc nhiều cụm.

K-Means là phương pháp chia cụm phẳng, đường biên cứng.



# Nội dung chính

---

- Bài toán chia cụm
- Ứng dụng chia cụm trong tìm kiếm
- Giải thuật K-means





## Giả thuyết chia cụm

---

- Các văn bản trong cùng một cụm có xu hướng cùng phù hợp với một nhu cầu thông tin.
- *"Closely associated documents tend to be relevant to the same requests".*

[ *Van Rijbergen* ]



# Ứng dụng chia cụm trong tìm kiếm

Ứng dụng	Tập văn bản chia cụm?	Lợi ích
Chia cụm kết quả	Tập kết quả	Dễ tìm kết quả phù hợp hơn
Chia cụm – gom nhóm (Scatter-Gather)	Bộ văn bản	Giao diện duyệt tập văn bản (search without typing)
Lọc văn bản theo cụm	Bộ văn bản	Xử lý truy vấn nhanh hơn
...	...	...

# Chia cụm kết quả tìm kiếm



**Vivísimo®**   [Search](#) [Advanced Search](#) [Help](#)

**Clustered Results** Top 208 results of at least 20,373,974 retrieved for the query **jaguar** ([Details](#))

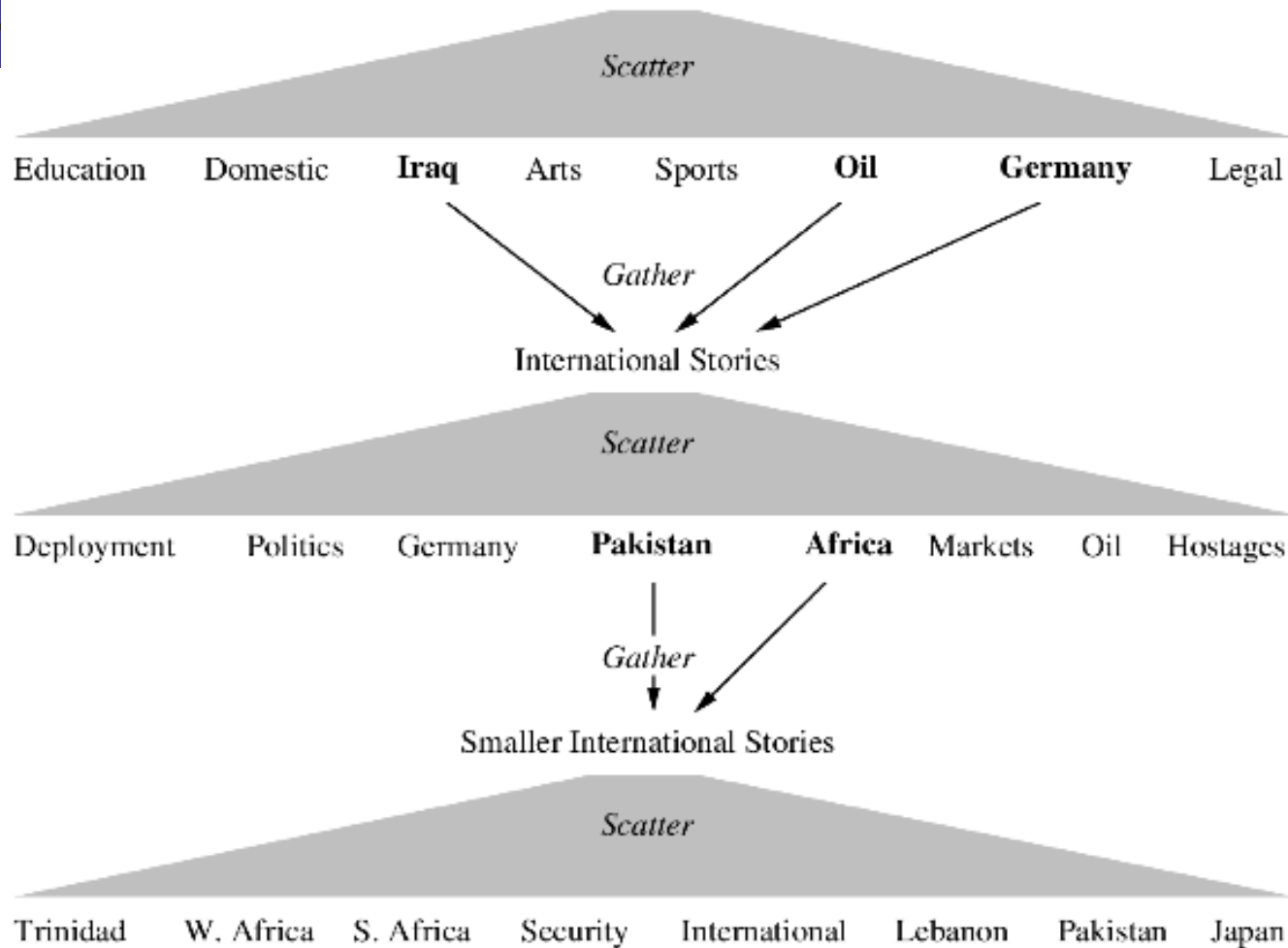
[jaguar](#) (208)

- [Cars](#) (74)
- [Club](#) (34)
- [Cat](#) (23)
- [Animal](#) (13)
- [Restoration](#) (10)
- [Mac OS X](#) (8)
- [Jaguar Model](#) (8)
- [Request](#) (5)
- [Mark Webber](#) (5)
- [Maya](#) (5)
- ▼ [More](#)

Find in clusters:  
 [Go](#)

- [Jag-lovers - THE source for all Jaguar information](#) [new window] [frame] [cache] [preview] [clusters]  
... Internet! Serving Enthusiasts since 1993 The Jag-lovers Web Currently with 40661 members The Premier **Jaguar** Cars web resource for all enthusiasts Lists and Forums Jag-lovers originally evolved around its ...  
[www.jag-lovers.org](#) - Open Directory 2, Wisenut 8, Ask Jeeves 8, MSN 9, Looksmart 12, MSN Search 18
- [Jaguar Cars](#) [new window] [frame] [cache] [preview] [clusters]  
[...] redirected to [www.jaguar.com](#)  
[www.jaguarcars.com](#) - Looksmart 1, MSN 2, Lycos 3, Wisenut 5, MSN Search 9, MSN 29
- [http://www.jaguar.com/](#) [new window] [frame] [cache] [preview] [clusters]  
[www.jaguar.com](#) - MSN 1, Ask Jeeves 1, MSN Search 3, Lycos 9
- [Apple - Mac OS X](#) [new window] [frame] [preview] [clusters]  
Learn about the new OS X Server, designed for the Internet, digital media and workgroup management. Download a technical factsheet.  
[www.apple.com/macosx](#) - Wisenut 1, MSN 3, Looksmart 25

# Chia cụm-gom nhóm





# Tăng độ đầy đủ

---

- Mở rộng tập kết quả tìm kiếm:
  - Chia cụm văn bản trong bộ dữ liệu;
  - Trả về các văn bản trong cùng cụm với những văn bản phù hợp (mở rộng tập kết quả);

Mong đợi trả về các văn bản chứa từ automobile cho truy vấn car.



# Nội dung chính

---

- Bài toán chia cụm
- Ứng dụng chia cụm trong tìm kiếm
- Giải thuật K-means



# Giải thuật K-means

---

- Biểu diễn văn bản dưới dạng vec-tơ
  - tương tự như trong VSM;
- Sử dụng khoảng cách Euclide để đánh giá độ khác biệt giữa các văn bản.



## Giải thuật K-means (2)

---

- Trọng tâm (centroid) của cụm  $\omega$  là:

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





## Giải thuật K-means (3)

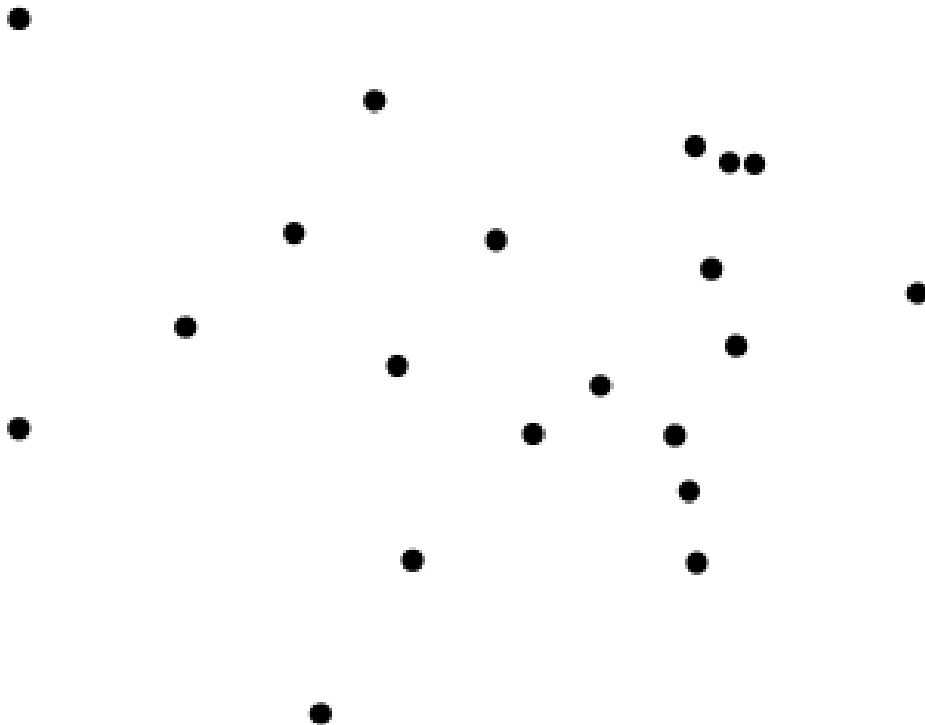
---

- Khởi tạo tâm cụm:
  - Có thể lựa chọn ngẫu nhiên K văn bản.
- Lặp:
  - 1. Gắn mỗi vec-tơ với trọng tâm gần nhất;
  - 2. Xác định lại trọng tâm sau mỗi lần chia cụm;
  - 3. Nếu thỏa mãn điều kiện dừng thì kết thúc, nếu ngược lại thì quay lại bước 1.
- Hàm mục tiêu: Tổng bình phương khoảng cách giữa các văn bản và tâm cụm của văn bản đó.



# Ví dụ chia cụm theo K-means

---



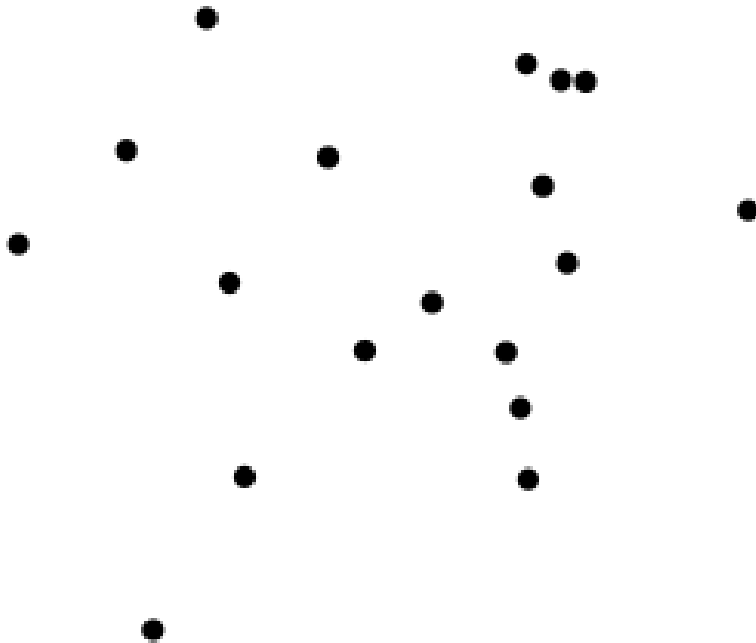


Ví dụ (2),  
khởi tạo ngẫu nhiên 2 trọng tâm

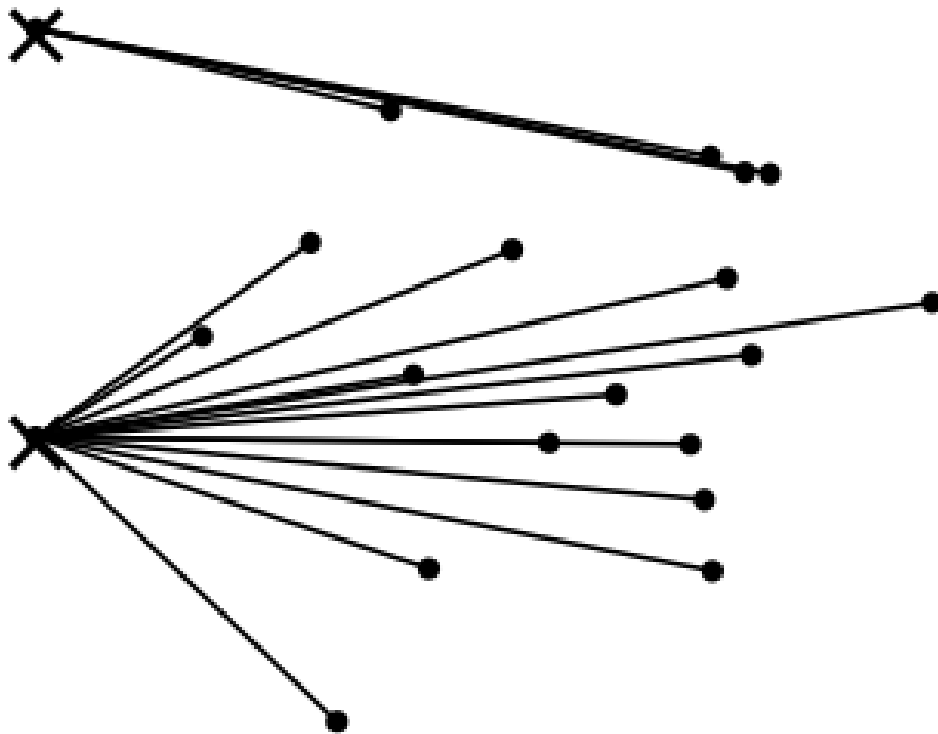
---

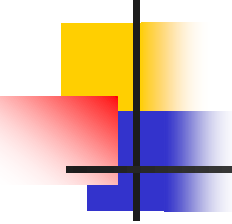
x

x



# Ví dụ (3), gắn văn bản với trọng tâm gần nhất





## Ví dụ (4), kết quả chia cụm

---

✕

2

222

1

1

1

1

1

1

1

1

1

1

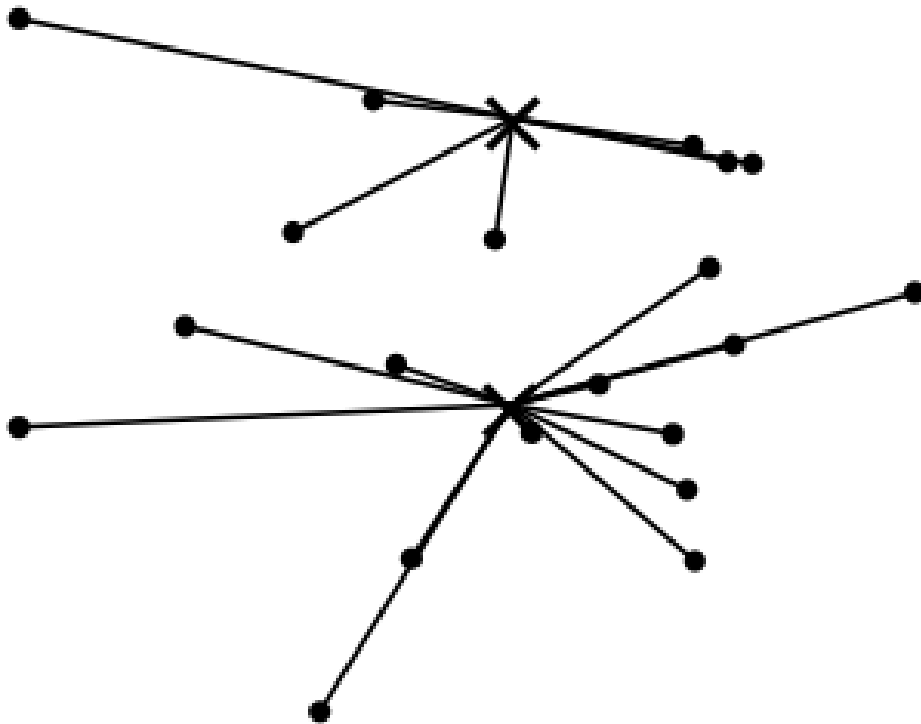
1

✕

1



## Ví dụ (6), chia lại cụm





## Ví dụ (7), kết quả chia cụm mới

---

2

2 X 22

2 2 1 1

1 1 1 1 1

1 1 1 1 1

1



1

2

2

2

1

1

1

22

1

1

1

1

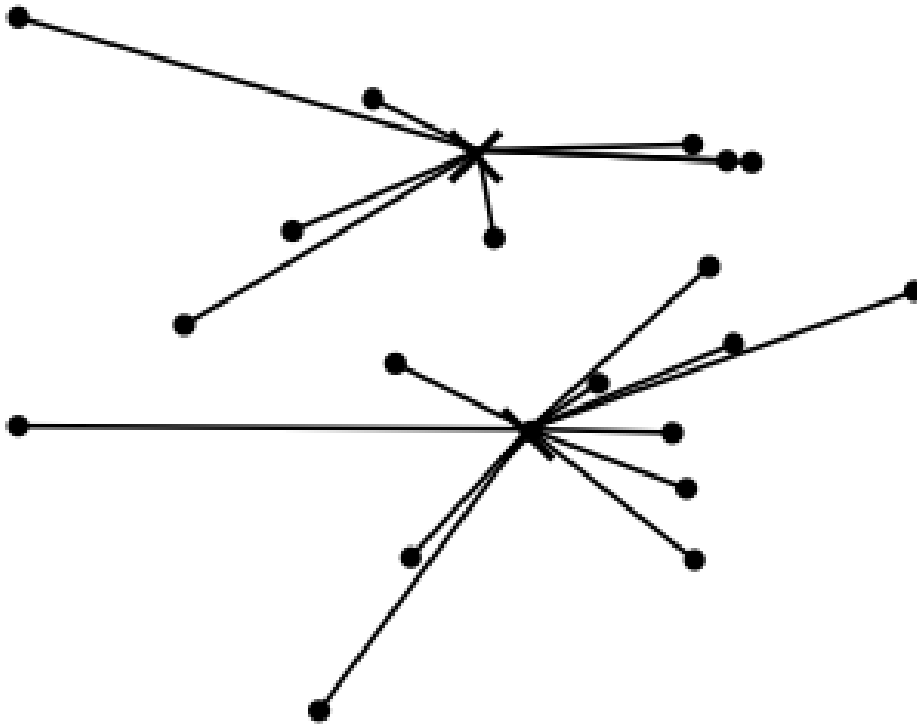
1

1



1

## Ví dụ (9), chia lại cụm



2

2

22

2

2

1

1

1

1

1

**X**

1

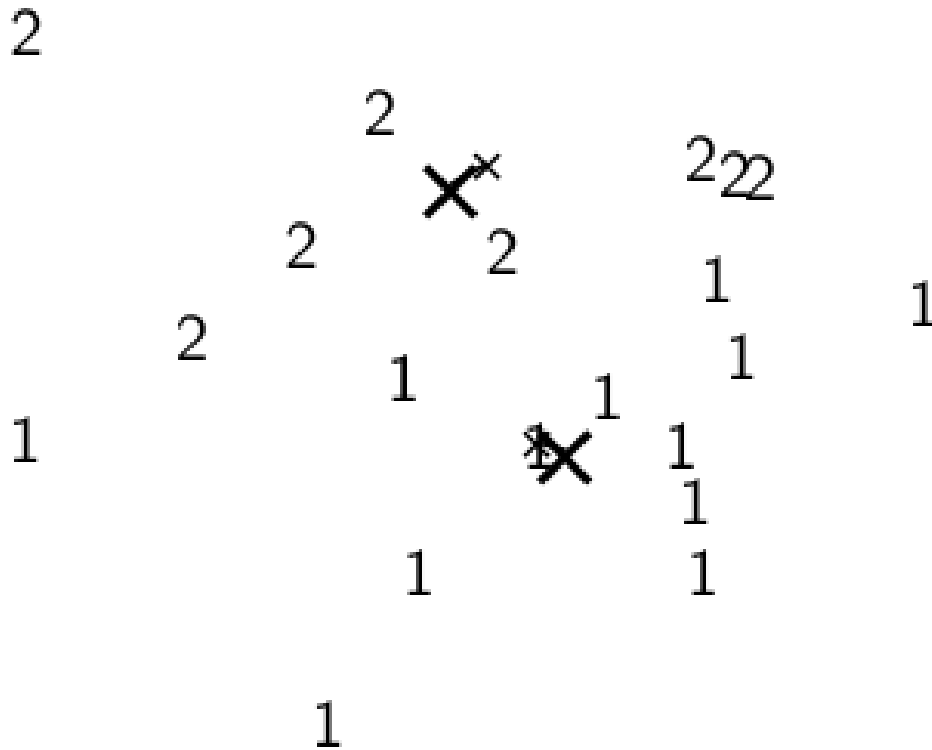
1

1

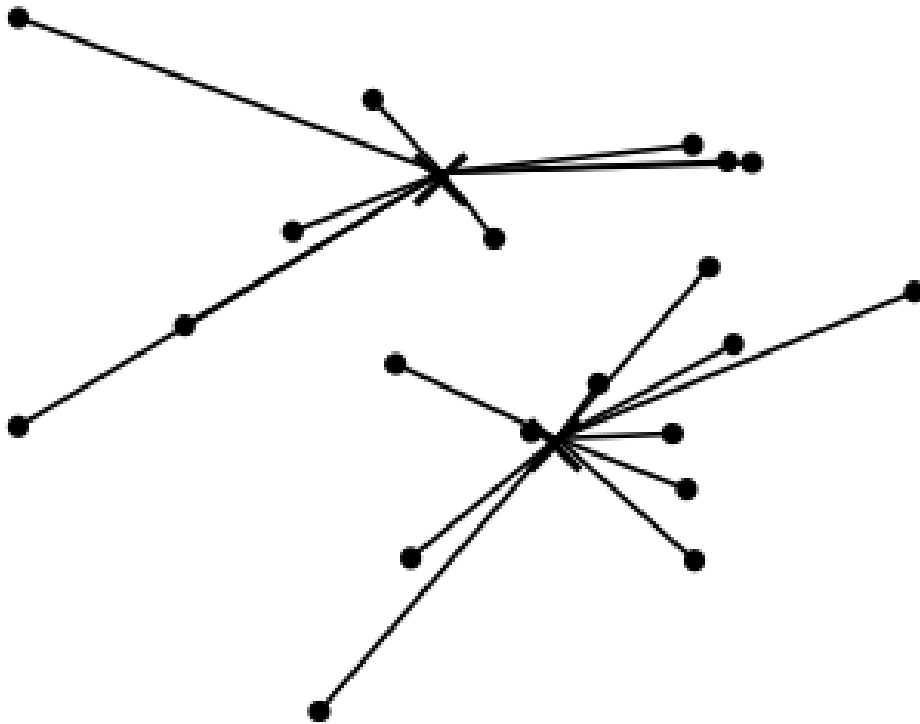
1

1

# Ví dụ (11), xác định lại trọng tâm

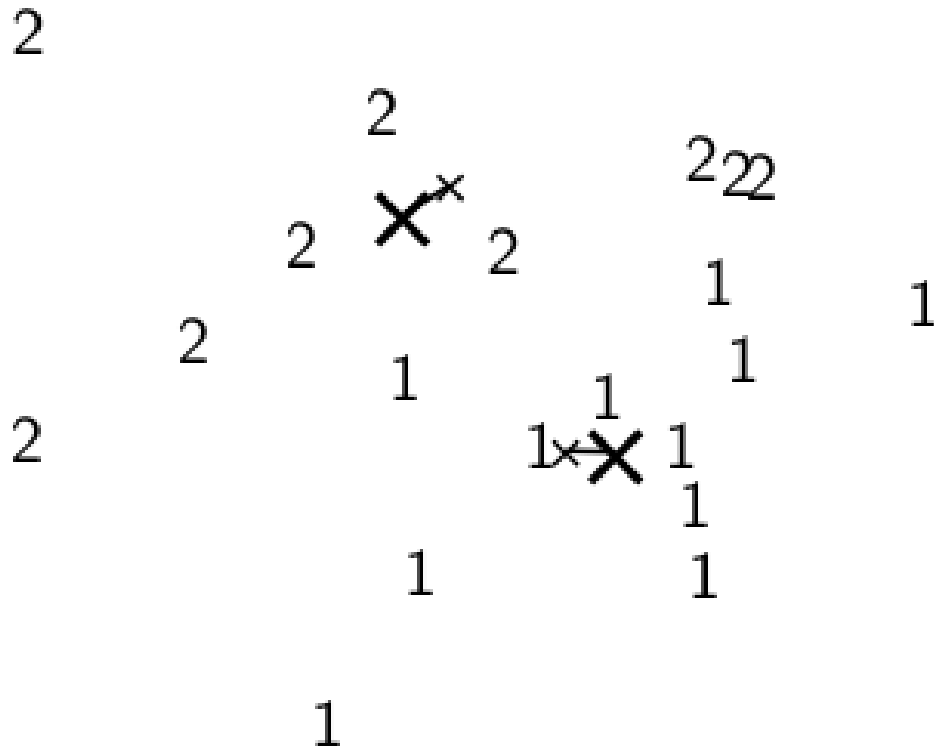


## Ví dụ (12), chia lại cụm

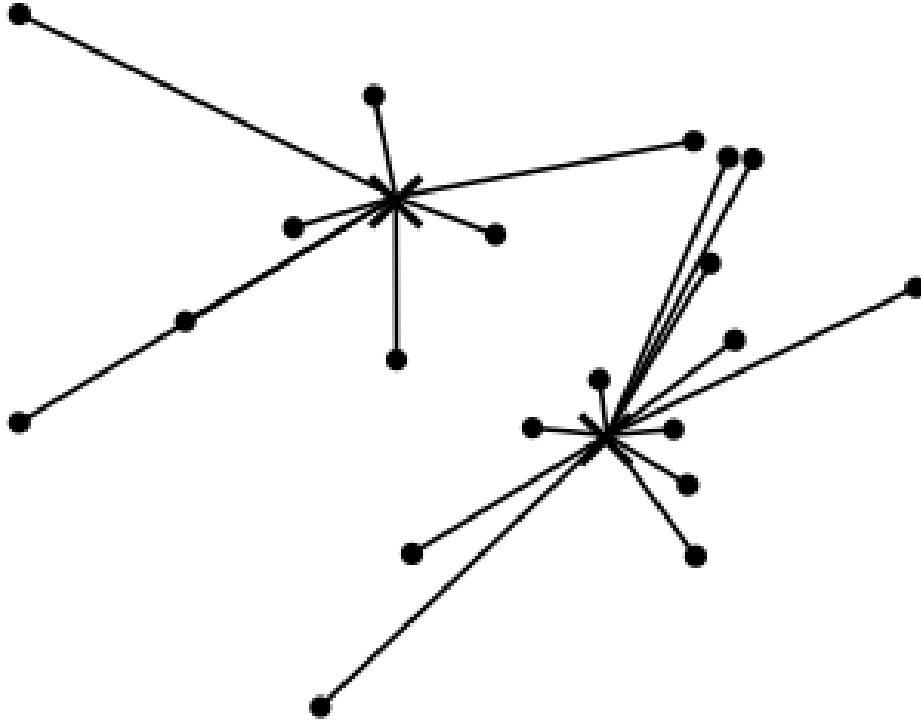


A diagram of a Coxeter-Dynkin diagram for a rank 4 root system. It consists of 10 nodes arranged in a branching structure. The top node is labeled 2. It has two children, both labeled 2. The left child has two children, both labeled 2. The right child has one child labeled 1. The leftmost node has two children, both labeled 1. The node below the leftmost node has one child labeled 1. The node below the rightmost node has two children, both labeled 1. There are two nodes marked with an 'X': the top node and the node below the rightmost node.

# Ví dụ (14), xác định lại trọng tâm



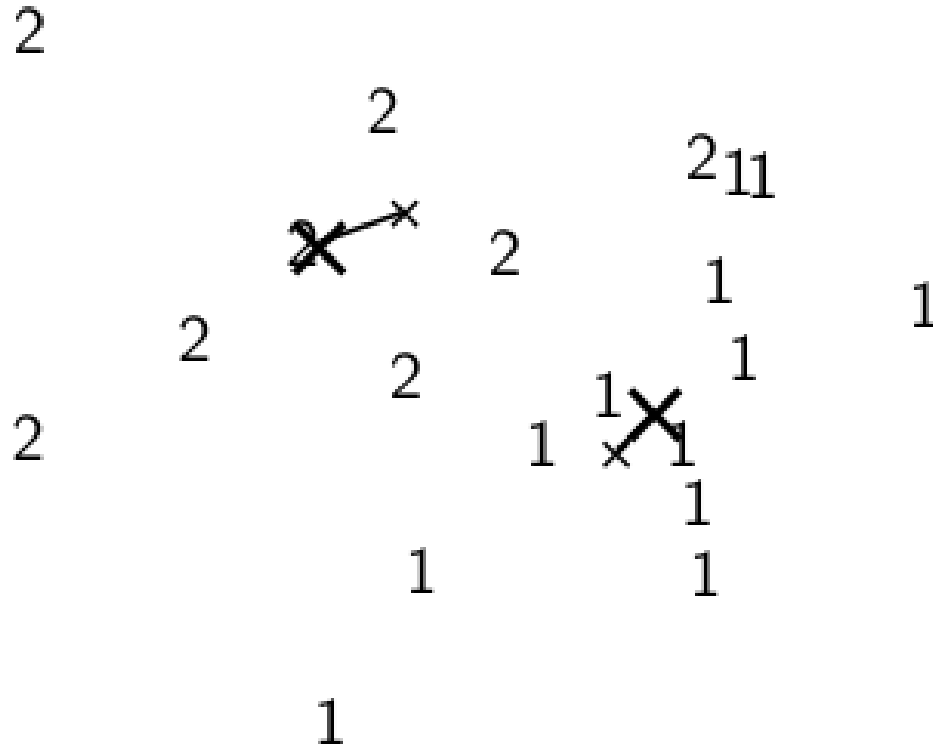
## Ví dụ (15), chia lại cụm



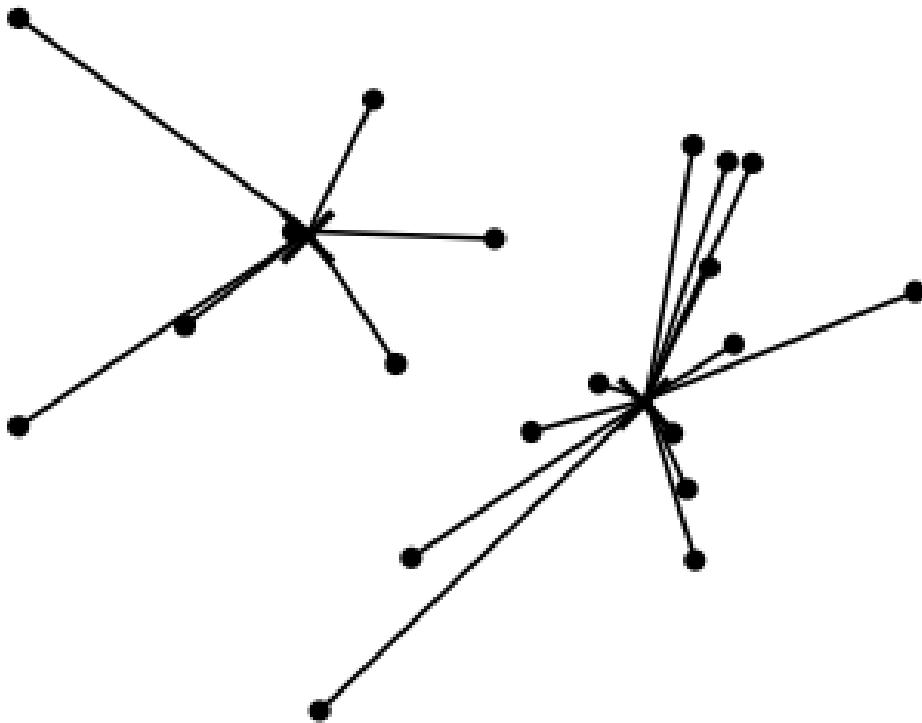


33

# Ví dụ (17), xác định lại trọng tâm



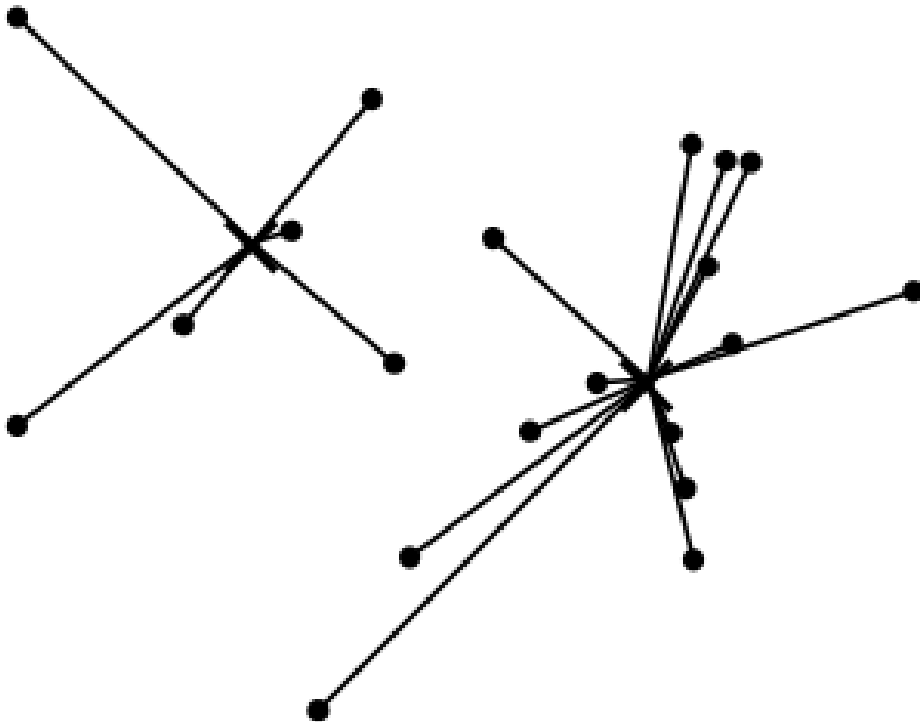
## Ví dụ (18), chia lại cụm



2

[illegible]

## Ví dụ (21), chia lại cụm



A diagram of a 2D hexagonal lattice with 19 nodes. Two nodes are marked with a large 'X' and labeled  $x^2$  and  $x$ . The node  $x^2$  is at the top-left, and  $x$  is at the center-right. Other nodes are labeled with integers: 2, 1, 11, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1.

40





## Ví dụ (24), kết quả chia cụm ổn định

---

2

2

1<sub>11</sub>

× 2 1 1 1

2 2 1 × 1 1

2 1 1 1 1

1



## Bài tập 18.1

---

Giả sử nếu hai văn bản bất kỳ có 2 từ chung thì tương đồng. Hãy thử lấy hai văn bản bất kỳ và một câu truy vấn cùng với nhu cầu thông tin để minh họa một tình huống sai của giả thuyết chia cụm.



## Bài tập 18.2

---

Hãy lấy một ví dụ đơn giản trên không gian một chiều (điểm trên trục số) để minh họa cho trường hợp kém hiệu quả của phương pháp tìm kiếm trên cơ sở chia cụm.

Trong ví dụ, kết quả tìm kiếm trong cụm gần với câu truy vấn phải kém hơn kết quả tìm kiếm những láng giềng gần nhất.

