Mô hình truy vấn không gian vector và mô hình truy vấn phân cụm

Nguyễn Đình Hóa
hoand@ptit.edu.vn 0942807711

Mô hình truy vấn không gian vector

- Giả thiết: luôn có một bộ hữu hạn các từ khóa (terms) để biểu diễn các văn bản và câu truy vấn.

$$D_i = [Ti_1, T_{i2}, \dots, T_{ik}, \dots, T_{iN}]$$
$$Q_j = [Q_{j1}, Q_{j2}, \dots, Q_{jk}, \dots, Q_{jN}]$$

  - T_{ik} là trọng số của từ khóa thứ k trong văn bản i,
  - Q_{jk} là trọng số của từ khóa thứ k trong câu truy vấn j,
  - N là tổng số từ khóa được sử dụng (N cố định).
  - T_{ik} và Q_{jk} có thể mang giá trị nhị phân {1, 0} hoặc mang các giá trị trọng số nào đó.

Mô hình truy vấn không gian vector

- Hỏi:
  - Trong trường hợp câu truy vấn có ít từ khóa thì làm thế nào so sánh với các văn bản?
  - Trong trường hợp các văn bản khác nhau có số từ khóa khác nhau thì làm thế nào để so sánh với nhau?

Mô hình truy vấn không gian vector

- Kết quả truy vấn phụ thuộc vào sự tương đồng giữa nội dung câu truy vấn và nội dung văn bản
- Sự tương đồng giữa văn bản D_i với câu truy vấn Q_j được thể hiện bởi công thức

$$S(D_i, Q_j) = \sum_{k=1}^{N} d_{ik} . q_{jk}$$

Mô hình truy vấn không gian vector

- Hỏi:
  - Các văn bản khác nhau với số các từ khóa khác nhau có ảnh hưởng đến kết quả truy vấn thế nào?

VD:

- D_1= máy bay, Malaysia, khủng bố, hạ cánh, mặt đất
- D_2 = máy bay, Indonesia, mất tích, biển
- Q = máy bay, Malaysia, mất tích.

Mô hình truy vấn không gian vector

- Công thức chuẩn hóa, đánh giá sự khác nhau về số từ khóa giữa các văn bản khác nhau, và giữa văn bản với câu truy vấn:

$$S(D_i, Q_j) = \frac{\sum_{k=1}^{N} d_{ik} \cdot q_{jk}}{\sqrt{\sum_{k=1}^{N} {d_{ik}}^2 \cdot \sum_{k=1}^{N} {q_{jk}}^2}}$$

- Kết quả truy vấn được sắp xếp theo thứ tự giảm dần về sự tương đồng.

Mô hình truy vấn không gian vector

- Ví dụ: Cho 4 văn bản với các vector trọng số như sau

$D_1 = [0.2, 0.3, 0.5, 0.0, 0.1]$

$D_2 = [0.3, 0.3, 0.2, 0.1, 0.2]$

$D_3 = [0.2, 0.2, 0.1, 0.2, 0.3]$

$D_4 = [0.1, 0.1, 0.3, 0.1, 0.0]$

Hãy tính sự tương đồng giữa câu truy vấn sau với các văn bản trên

$Q = [0.2, 0, 0.1, 0, 0.2]$

Mô hình truy vấn không gian vector

- Theo công thức chưa chuẩn hóa

  $S(D_1, Q) = ?$

  $S(D_2, Q) = ?$

  $S(D_3, Q) = ?$

  $S(D_4, Q) = ?$

- Theo công thức chuẩn hóa

  $S(D_1, Q) = ?$

  $S(D_2, Q) = ?$

  $S(D_3, Q) = ?$

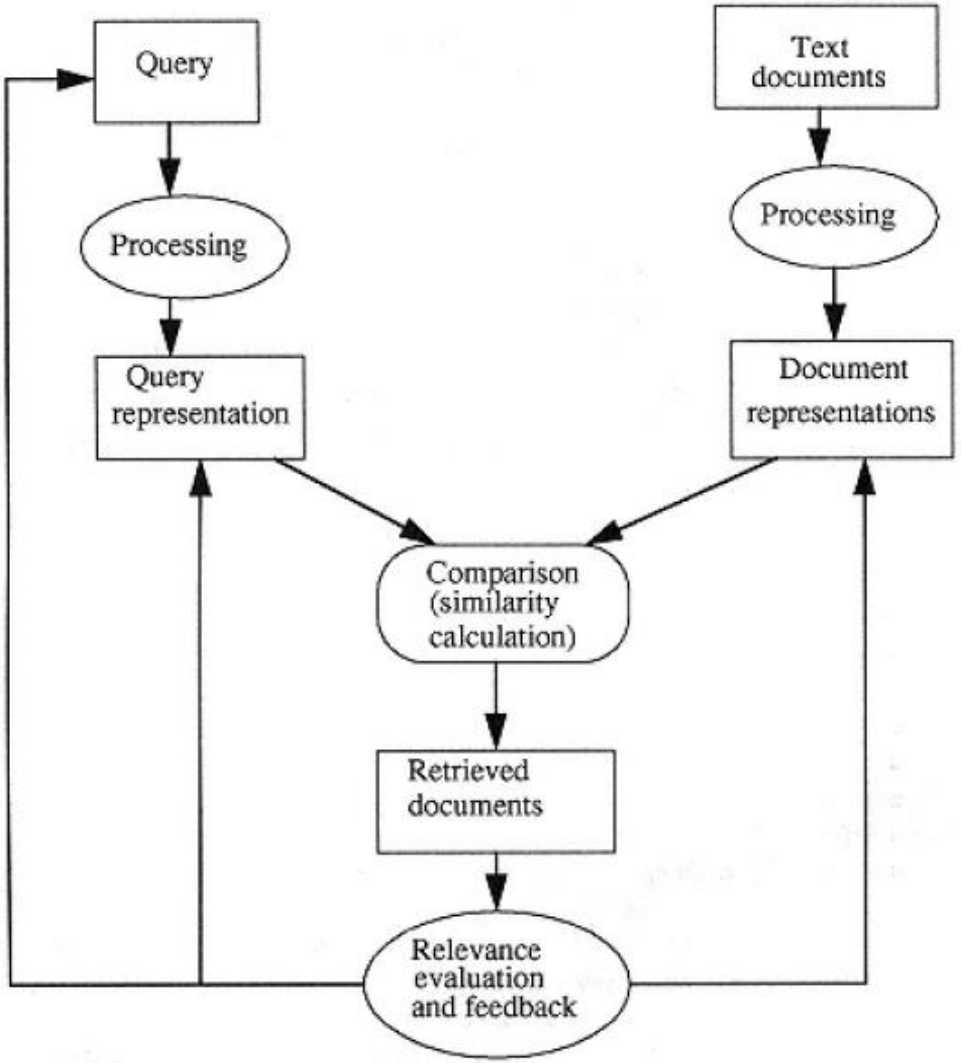
  $S(D_4, Q) = ?$

Mô hình truy vấn không gian vector

- Nhược điểm:
  - Mô hình truy vấn không gian vector không tính đến sự khác nhau về nghĩa giữa các từ khóa
  - Chỉ hoạt động tốt với các văn bản và câu truy vấn có độ dài thấp.

Chỉ số hóa dữ liệu văn bản

- Truy vấn có phản hồi



Mô hình truy vấn không gian vector

- Hệ thống có phản hồi
  - Thay đổi câu truy vấn nhằm cải thiện kết quả tìm kiếm
    - Các từ khóa có trong văn bản có liên quan nhưng không có trong câu truy vấn sẽ được thêm vào câu truy vấn với trọng số khởi tạo nào đó, hoặc tăng trọng số của chúng trong câu truy vấn nếu chúng đã ở trong câu truy vấn.
    - Các từ khóa có cả ở câu truy vấn cũng như của văn bản không liên quan sẽ được xóa khỏi câu truy vấn, hoặc giảm trọng số của chúng trong câu truy vấn
    - Thực hiện tìm kiếm theo câu truy vấn mới.

Mô hình truy vấn không gian vector

- Công thức thay đổi câu truy vấn:

$$Q_{j+1} = Q_j + \alpha . \sum_{D^k \in Rel} D^k - \beta . \sum_{D^t \in NonRel} D^t$$

Trong đó:

- Rel : tập các văn bản có liên quan đến nội dung truy vấn trong số các kết quả trả về
- $NonRel$: tập các văn bản không liên quan đến nội dung truy vấn trong số các kết quả trả về

Mô hình truy vấn không gian vector

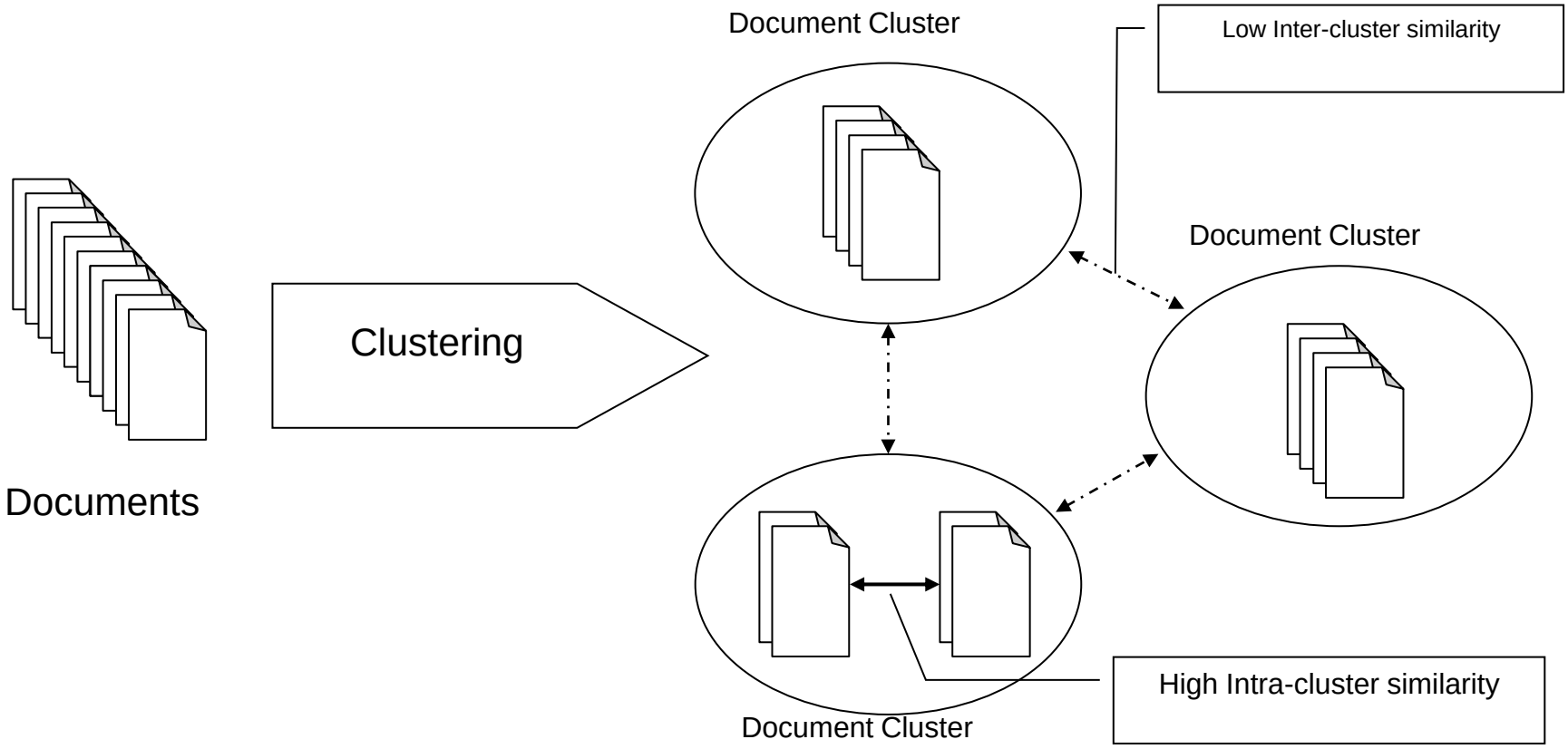
- Thay đổi cách hiển thị nội dung văn bản – nhằm tác động đến các câu truy vấn khác (của người khác).
  - Các từ khóa có trong câu truy vấn, nhưng không có trong văn bản được coi là có liên quan thì được thêm vào danh mục các từ khóa của văn bản đó với trọng số được khởi tạo bằng giá trị nào đó
  - Trọng số của các từ khóa có cả trong câu truy vấn lẫn văn bản có liên quan thì được tăng lên.
  - Trọng số của các từ khóa không có trong câu truy vấn nhưng có trong văn bản thì được giảm đi.
  - Thực hiện câu truy vấn mới với nội dung tương tự với câu trước.
- Phương pháp này sẽ không hữu ích nếu câu truy vấn mới quá khác so với câu truy vấn trước đó đã dùng để thay đổi văn bản.

Mô hình truy vấn dựa trên phân cụm

- Phân cụm: là quá trình gom nhóm các bản ghi giống nhau về nội dung vào thành từng cụm theo quy tắc:
  - Các bản ghi trong cùng một cụm phải giống nhau
  - Các bản ghi trong các cụm khác nhau thì khác nhau.

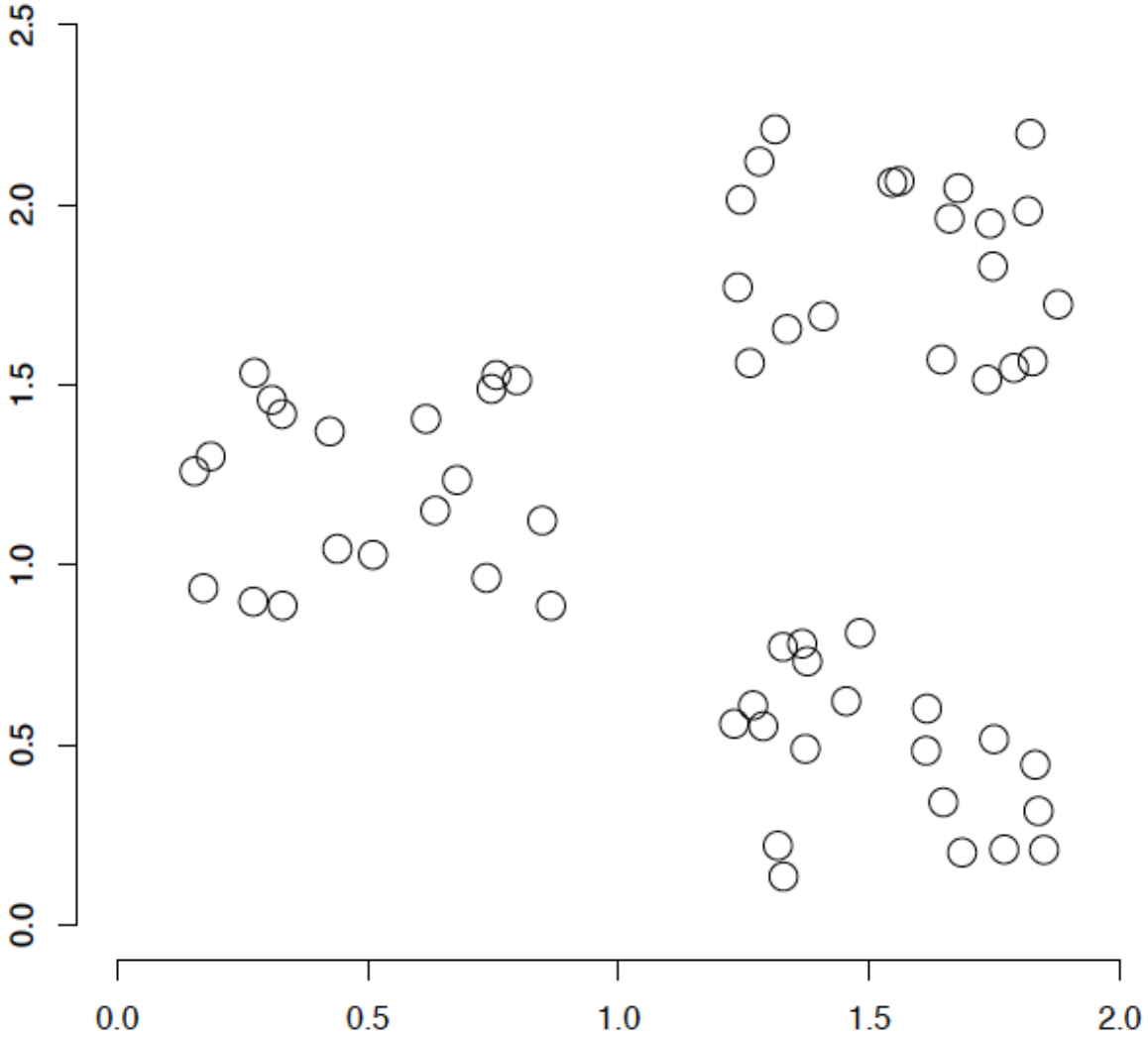
Mô hình truy vấn dựa trên phân cụm

- Phân cụm



Ví dụ về bộ dữ liệu có cấu trúc phân cụm

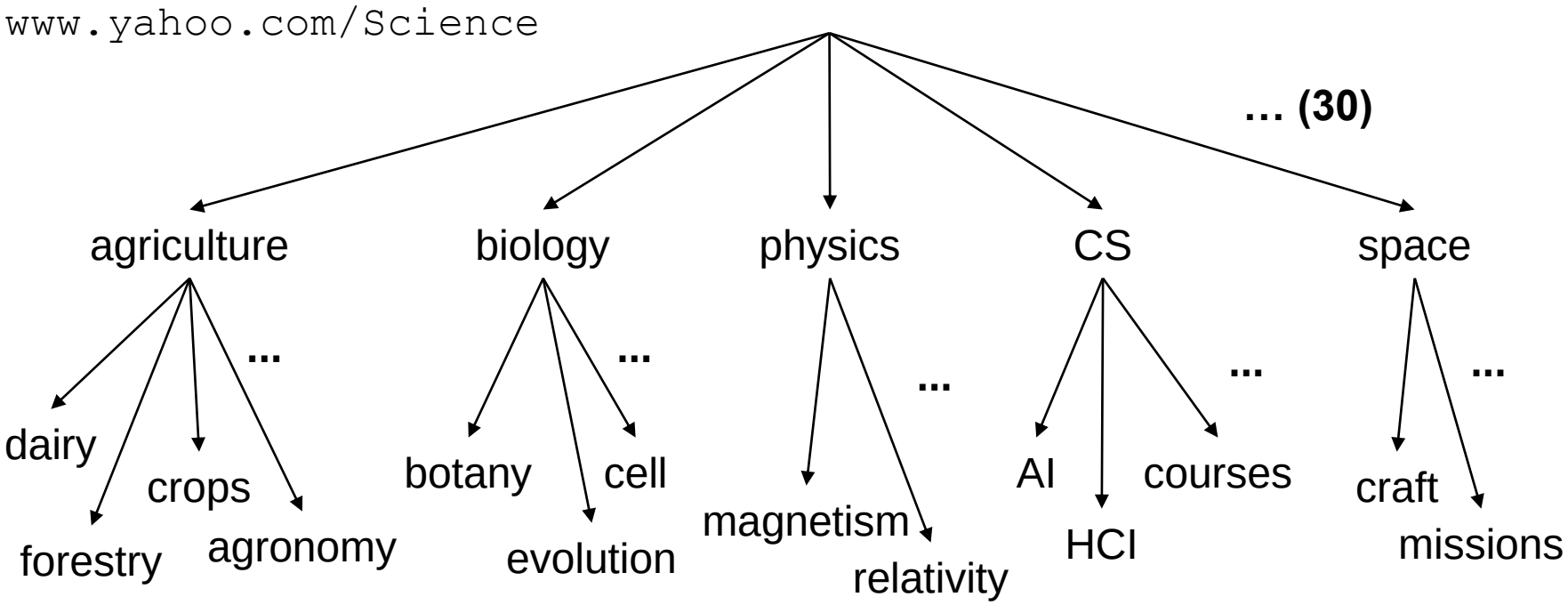
- Hãy nêu một thuật toán để chia bộ dữ liệu này thành 3 nhóm khác nhau.



Các thước đo độ tương đồng

- Độ tương tự giữa 2 bản ghi
- Khoảng cách giữa 2 bản ghi
  - L_p-norm: $L_p(D_i, D_j) = \sqrt[p]{\sum_{k=1}^{N} |d_{ik} - d_{jk}|^p}$
  - L_1-norm: $L_1(D_i, D_j) = \sum_{k=1}^{N} |d_{ik} - d_{jk}|$
  - L_2-norm: $L_2(D_i, D_j) = \sqrt{\sum_{k=1}^{N} |d_{ik} - d_{jk}|^2}$

Ví dụ về cấu trúc dữ liệu phân cụm của Yahoo



Mô hình truy vấn dựa trên phân cụm

- Các phương pháp phân cụm:
  - Phân cụm dựa trên so sánh theo từng cặp
  - Phân cụm dựa trên lựa chọn ngẫu nhiên
  - Phương pháp K-means

Mô hình truy vấn dựa trên phân cụm

- Phân cụm dựa trên so sánh theo cặp
  - Mỗi bản ghi được coi là một cụm chứa riêng nó
  - Hình thành các cặp cụm bản ghi giữa các cụm hiện có
  - Hai cụm bản ghi giống nhau nhất được gộp với nhau để tạo thành cụm mới
    - Tiêu chí so sánh hai cụm bản ghi với nhau:
      - Dựa trên sự giống nhau nhất của từng cặp bản ghi giữa hai cụm
      - Dựa trên sự khác nhau nhất của từng cặp bản ghi giữa hai cụm
      - Dựa trên trung bình sự giống nhau của tất cả các cặp bản ghi giữa hai cụm
  - Quá trình được lặp lại đến khi các bản ghi đều nằm ở một cụm nào đó

Mô hình truy vấn dựa trên phân cụm

- Tiêu chí dựa trên giống nhau nhất theo cặp

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

- Tiêu chí dựa trên khác nhau nhất theo cặp

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

Ví dụ 1:

- Cho các dữ liệu theo cấu trúc vector thuộc tính như sau

  $D_1 = [0.4, 0.5, 0.5]$

  $D_2 = [0.2, 0.3, 0.2]$

  $D_3 = [0.6, 0.5, 0.4]$

  $D_4 = [0.3, 0.5, 0.4]$

  $D_5 = [0.2, 0.1, 0.2]$

  Hãy phân cụm các dữ liệu trên bằng phương pháp so sánh theo cặp, áp dụng công thức tính khoảng cách L1-norm và sử dụng tiêu chí tính khoảng cách giữa các cụm dữ liệu:

  - Khoảng cách dài nhất
  - Khoảng cách trung bình

Mô hình truy vấn dựa trên phân cụm

- Phân cụm theo phương pháp lựa chọn ngẫu nhiên
  - Bản ghi đầu tiên được chọn làm một cụm của chính nó
  - Lựa chọn tiếp một bản ghi ngẫu nhiên khác, tính toán sự giống nhau giữa một bản ghi đó với các cụm bản ghi đã có.
  - Nếu bản ghi mới giống với một cụm nào đã có (dựa trên ngưỡng so sánh λ) thì được gộp vào đó, nếu không, nó trở thành một cụm mới
  - Chu trình được lặp lại đến khi toàn bộ bản ghi được xem xét.

Ví dụ 2:

- Cho các dữ liệu theo cấu trúc vector trọng số như sau

  D_1 = [0.1, 0.3, 0.3]

  D_2 = [0.2, 0.4, 0.2]

  D_3 = [0.4, 0.1, 0.5]

  D_4 = [0,2, 0.5, 0.3]

  D_5 = [0.5, 0.1, 0.4]

  Hãy phân cụm các dữ liệu trên theo phương pháp lựa chọn ngẫu nhiên sử dụng tiêu chí khoảng cách trung bình, áp dụng công thức tính khoảng cách L1-norm, và ngưỡng khoảng cách $\lambda = 0.4$

Mô hình truy vấn dựa trên phân cụm

- Thuật toán K-means
  - Các bản ghi được biểu diễn bởi các vector thuộc tính
  - Khởi tạo K điểm gốc trong không gian dữ liệu
  - Các bản ghi lần lượt được chọn vào một trong K nhóm gần nhất theo khoảng cách đến điểm gốc.
  - Vị trí mới của K điểm gốc được thiết lập là điểm trung bình của K nhóm dữ liệu mới hình thành
  - Chu trình được lắp lại cho đến khi
    - Không còn sự thay đổi về nhóm của các bản ghi
    - Điểm gốc của các nhóm không thay đổi
    - Vượt quá một số cố định vòng lặp cho trước.

Ví dụ 3

- Cho các dữ liệu theo cấu trúc vector trọng số như sau

  D_1 = [0.2, 0.2, 0.1]

  D_2 = [0.5, 0.3, 0.3]

  D_3 = [0.3, 0.1, 0.2]

  D_4 = [0.6, 0.3, 0.4]

  D_5 = [0.5, 0.2, 0.3]

  D_6 = [0.5, 0.4, 0.5]

  Hãy phân cụm các dữ liệu trên theo phương pháp K-means với K = 2, sử dụng công thức tính khoảng cách L1-norm.

Truy vấn dữ liệu trong mô hình phân cụm

- Cụm bản ghi gần với câu truy vấn nhất được xác định dựa trên khoảng cách giữa câu truy vấn với điểm trung bình của cụm
- Nếu cụm bản ghi chứa ít dữ liệu thì toàn bộ các bản ghi trong cụm sẽ là kết quả tra cứu
- Nếu cụm bản ghi nhiều dữ liệu: chỉ 1 hoặc K bản ghi giống nhất với câu truy vấn sẽ được làm kết quả.