VIỆN TRÍ TUỆ NHÂN TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

K-MEANS VÀ LẬP TRÌNH MAPREDUCE TRONG PHÂN CỤM ẢNH

Nguyễn Phú Lộc - 22022547
Phạm Chiến - 22022634
Bùi Ngọc Khánh - 22022551
Nguyễn Quang Huy - 22022582

VIỆN TRÍ TUỆ NHÂN TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Table of Contents

Khái niệm về dữ liệu lớn

Theo Wikipedia, dữ liệu lớn là bộ dữ liệu có kích thước hoặc độ phức tạp lớn đến mức các phương pháp xử lý truyền thống không đáp ứng được. Theo Gartner, dữ liệu lớn được đặc trưng bởi 3Vs: Khối lượng (Volume), Tốc độ (Velocity) và Đa dạng (Variety).

Volume

Lượng dữ liệu khổng lồ, tăng trưởng nhanh qua thời gian.

Velocity

Tốc độ tạo ra và xử lý dữ liệu rất nhanh.

Variety

Dữ liệu đến từ nhiều nguồn và có nhiều định dạng khác nhau.

Đặc trưng của dữ liệu lớn

Ngoài 3Vs, dữ liệu lớn còn có những đặc trưng khác như Độ chính xác (Veracity) và Giá trị (Value). Độ chính xác đề cập đến chất lượng và độ tin cậy của dữ liệu, trong khi giá trị thể hiện lợi ích thu được từ dữ liệu.

1. **Volume**
   Khối lượng dữ liệu lớn, tăng trưởng nhanh.
2. **Velocity**
   Tốc độ tạo ra và xử lý dữ liệu nhanh.
3. **Variety**
   Đa dạng nguồn và định dạng dữ liệu.
4. **Veracity**
   Độ chính xác và tin cậy của dữ liệu.
5. **Value**
   Giá trị và lợi ích thu được từ dữ liệu.

Công nghệ xử lý dữ liệu lớn

Để xử lý dữ liệu lớn một cách hiệu quả, các công nghệ chính được sử dụng là tính toán phân tán, tính toán song song, song song hóa bằng CPU đa nhân và GPU, xử lý phân tán với hệ thống cluster và xử lý phân tán trên Cloud.

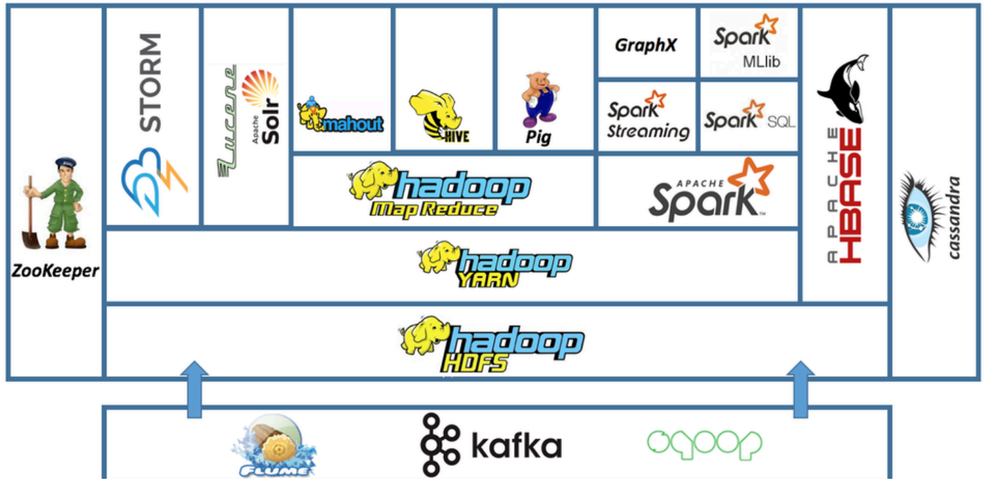
Tính toán phân tán

Tính toán song song

Xử lý phân tán trên Cloud

Tổng quan Hadoop

Hadoop là một công nghệ phân tán và mã nguồn mở được sử dụng phổ biến để xử lý và lưu trữ khối dữ liệu lớn. Được tạo ra bởi Doug Cutting và Mike Cafarella năm 2005, Hadoop được phát triển bởi Apache Software Foundation dựa trên công nghệ Google File System và MapReduce.



2005
Hadoop được tạo ra bởi Doug Cutting và Mike Cafarella.

1

2

2006
Hadoop được phát hành dưới dạng mã nguồn mở.

2008
Apache Software Foundation đảm nhận việc phát triển Hadoop.

3

4

Hiện nay
Hadoop được sử dụng rộng rãi trong các hệ thống xử lý dữ liệu lớn.

Thành phần của Hadoop

Hadoop bao gồm nhiều module, nhưng hai thành phần quan trọng nhất là HDFS (Hadoop Distributed File System) và MapReduce. HDFS cung cấp khả năng lưu trữ và truy vấn dữ liệu song song, trong khi MapReduce là một khung làm việc để xử lý dữ liệu theo lô.

HDFS
Hệ thống file phân tán cung cấp khả năng truy vấn song song.

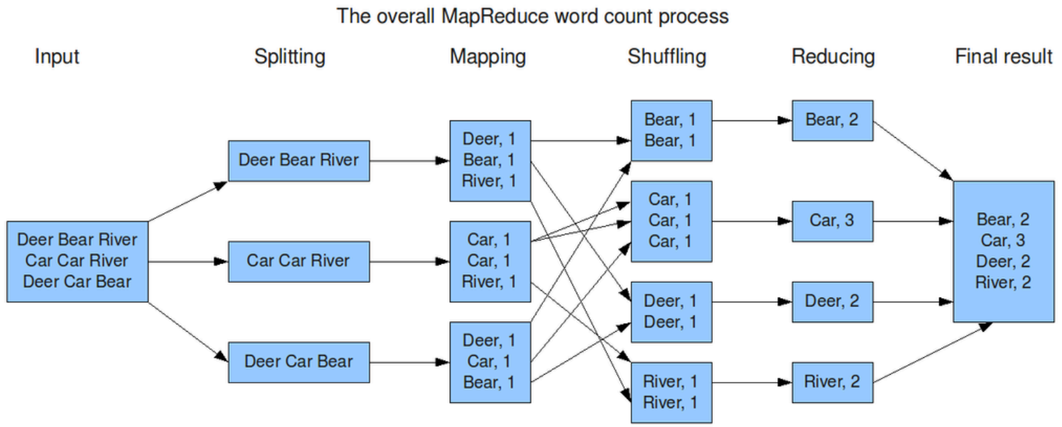
MapReduce
Khung làm việc xử lý dữ liệu theo lô trên cụm máy tính.

YARN
Framework quản lý lập lịch tác vụ và tài nguyên.

Tổng quan MapReduce

MapReduce là một khung làm việc xử lý dữ liệu phân tán được sử dụng để xử lý dữ liệu lớn trên Hadoop. Được phát triển bởi Google và sau đó được Apache Software Foundation phát hành dưới dạng một phần của hệ sinh thái Hadoop.

1. **Bước Map**
   Dữ liệu được chia thành các phần nhỏ và xử lý độc lập trên từng nút trong cụm Hadoop.
2. **Bước Shuffle**
   Dữ liệu đầu ra từ bước Map được sắp xếp và gom nhóm.
3. **Bước Reduce**
   Dữ liệu được xử lý lại và kết quả cuối cùng được trả về.



Định Nghĩa K-Means

Phân cụm không giám sát

K-Means là thuật toán học máy không giám sát, nhóm dữ liệu không có nhãn thành các cụm khác nhau.

Phân chia dữ liệu

Mục tiêu là phân chia n quan sát thành k cụm, mỗi quan sát thuộc cụm có giá trị trung bình gần nhất.

Tối ưu hóa

Thuật toán tìm cách giảm tổng bình phương khoảng cách giữa các điểm dữ liệu và tâm cụm.

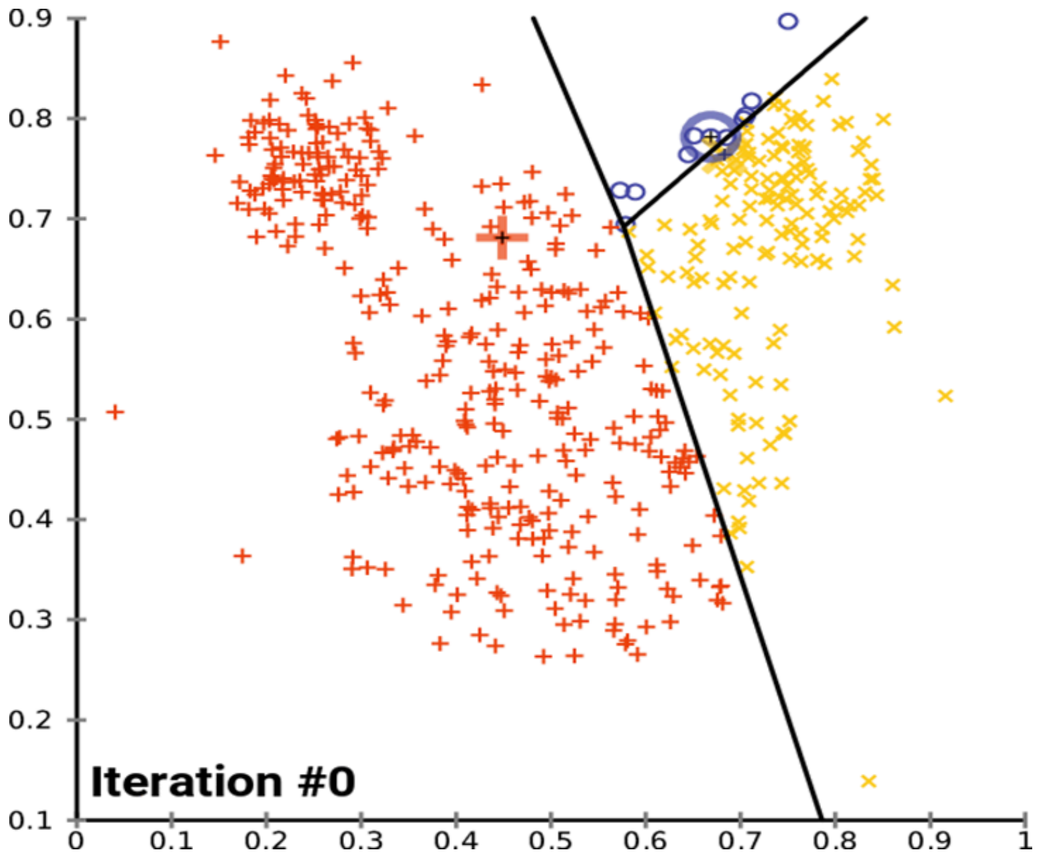
Ưu và Nhược Điểm

Ưu điểm

- Hoạt động tốt với dữ liệu được phân tách rõ ràng
- Nhanh hơn so với các kỹ thuật phân cụm khác
- Cung cấp sự kết hợp mạnh mẽ giữa các điểm dữ liệu

Nhược điểm

- Khó xác định số lượng K tối ưu
- Không phù hợp khi các điểm dữ liệu chồng chéo
- Nhạy cảm với nhiễu và có thể bị kẹt ở cực tiểu cục bộ



Lịch Sử Phát Triển

1

1956
Hugo Steinhaus đề xuất ý tưởng ban đầu về phân cụm K-Means.

2

1957
Stuart Lloyd từ Bell Labs đề xuất thuật toán chuẩn đầu tiên.

3

1965
Edward W. Forgy công bố phương pháp tương tự, dẫn đến tên gọi thuật toán Lloyd-Forgy.

4

1967
James MacQueen sử dụng thuật ngữ "k-means" lần đầu tiên.

Các Bước Thực Hiện Thuật Toán

1

Bước 1: Khởi tạo
Tạo K trung tâm ngẫu nhiên trong không gian dữ liệu.

2

Bước 2: Gán cụm
Gán mỗi điểm dữ liệu vào cụm có trung tâm gần nhất.

3

Bước 3: Cập nhật trung tâm
Tính toán lại trung tâm mới cho mỗi cụm dựa trên các điểm đã được gán.

4

Bước 4: Lặp lại
Lặp lại bước 2 và 3 cho đến khi hội tụ hoặc đạt số lần lặp tối đa.

Ứng Dụng trong Phân Cụm Ảnh

Phân đoạn hình ảnh

K-Means được sử dụng để phân đoạn hình ảnh thành các vùng khác nhau dựa trên màu sắc hoặc kết cấu, hữu ích trong nhận dạng và theo dõi đối tượng.

Nén hình ảnh

Thuật toán có thể được sử dụng để giảm số lượng màu trong hình ảnh, giúp nén dữ liệu mà vẫn duy trì chất lượng hình ảnh.

Phát hiện đối tượng

K-Means giúp phân biệt đối tượng với nền, hỗ trợ các ứng dụng thị giác máy tính và xử lý hình ảnh nâng cao.

Ứng Dụng trong Kinh Doanh

Phân khúc khách hàng

Nhóm khách hàng dựa trên hành vi mua sắm và dữ liệu nhân khẩu học.

Hệ thống đề xuất

Đề xuất sản phẩm dựa trên lịch sử mua hàng và sở thích của người dùng.

Phát hiện gian lận

Xác định các giao dịch bất thường hoặc đáng ngờ trong hệ thống tài chính.

Ứng Dụng trong Công Nghệ

K-Means có nhiều ứng dụng trong lĩnh vực công nghệ, bao gồm phát hiện xâm nhập mạng, phân cụm tài liệu, phát hiện bất thường trong dữ liệu chuỗi thời gian và nén hình ảnh.

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.2 Các phép đo khoảng cách

2.2.1 Khoảng cách Euclidean

- Khoảng cách Euclid (Euclidean Distance) là một khái niệm trong toán học dùng để đo khoảng cách "trực tiếp" giữa hai điểm trong không gian Euclid. Đây là một dạng mở rộng của định lý Pythagoras trong không gian nhiều chiều.
- Khoảng cách Euclid giữa hai điểm $x = (x_1, x_2, \dots, x_n)$ và $y = (y_1, y_2, \dots, y_n)$ trong không gian n chiều được tính bằng công thức:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.2 Các phép đo khoảng cách

2.2.1 Khoảng cách Euclidean

- Khoảng cách Euclid (Euclidean Distance) là một khái niệm trong toán học dùng để đo khoảng cách "trực tiếp" giữa hai điểm trong không gian Euclid. Đây là một dạng mở rộng của định lý Pythagoras trong không gian nhiều chiều.
- Khoảng cách Euclid giữa hai điểm $x = (x_1, x_2, \dots, x_n)$ và $y = (y_1, y_2, \dots, y_n)$ trong không gian n chiều được tính bằng công thức:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.2 Các phép đo khoảng cách

2.2.2 Khoảng cách Manhattan

- Khoảng cách Manhattan (Manhattan Distance) hay còn gọi là khoảng cách L1, là một phương pháp đo khoảng cách giữa hai điểm trong không gian bằng cách cộng tổng giá trị tuyệt đối của sự khác biệt giữa các tọa độ tương ứng của chúng.
- Khoảng cách Manhattan giữa hai điểm $x = (x_1, x_2, \ldots, x_n)$ và $y = (y_1, y_2, \ldots, y_n)$ trong không gian n chiều được tính bằng công thức:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.2 Các phép đo khoảng cách

2.2.3 Khoảng cách Cosin

- Khoảng cách Cosine (Cosine Distance) là một phương pháp đo lường sự khác biệt giữa hai vector trong không gian đa chiều. Thay vì tập trung vào độ lớn của vector, khoảng cách Cosine chủ yếu quan tâm đến góc giữa các vector, thể hiện độ tương đồng hoặc không tương đồng về hướng.
- Khoảng cách Cosine được tính dựa trên **Cosine Similarity**. Nếu hai vector u và v có góc giữa chúng là θ, thì Cosine Similarity được định nghĩa là:

$$\text{Cosine Similarity} = \cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

$$\text{Cosine Distance} = 1 - \cos(\theta)$$

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.2 Các phép đo khoảng cách

2.2.3 Khoảng cách Minkowski

- **Khoảng cách Minkowski** là một cách tổng quát hóa của các khoảng cách phổ biến như **khoảng cách Euclid** và **khoảng cách Manhattan**. Nó được sử dụng trong không gian đa chiều để đo lường khoảng cách giữa hai điểm dựa trên một tham số p, cho phép điều chỉnh mức độ ảnh hưởng của từng thành phần.
- Khoảng cách Minkowski giữa hai điểm $u = (u_1, u_2, \dots, u_n)$và $v = (v_1, v, \dots, v_n)$ trong không gian n chiều được tính bằng công thức:

$$d(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{n} |u_i - v_i|^p \right)^{\frac{1}{p}}$$

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.3 Nguyên lí hoạt động của K-means trong phân cụm ảnh

2.3.1 K-Means trong phân cụm các điểm ảnh

- **Nguyên tắc:** Dựa vào khoảng cách giữa các điểm ảnh

Bước 1: Khởi tạo

- Xác định số lượng k cụm
- Chọn ngẫu nhiên k tâm cụm ban đầu
- **Nhược điểm của cách chọn cụm của Kmeans**: nó nhạy cảm với việc khởi tạo các tâm hoặc các điểm trung bình. Nếu một centroid được khởi tạo tại một điểm "xa xôi", nó có thể không có điểm nào liên kết với nó. Tương tự, nhiều centroid có thể được khởi tạo trong cùng một cụm, dẫn đến việc phân cụm kém hiệu quả.
- Để khắc phục nhược điểm trên ta có thể sử dụng cách phân cụm của thuật toán **Kmeans++**

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

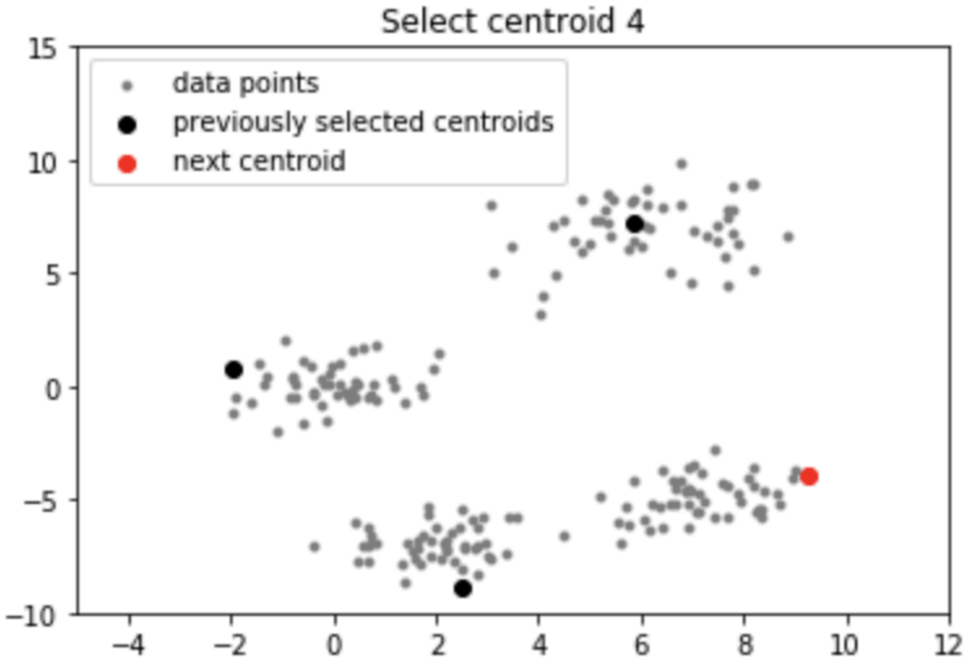
2.3 Nguyên lí hoạt động của K-means trong phân cụm ảnh

2.3.1 K-Means trong phân cụm các điểm ảnh

- **Nguyên tắc:** Dựa vào khoảng cách giữa các điểm ảnh

Bước 1: Khởi tạo bằng Kmeans++

- Chọn ngẫu nhiên centroid đầu tiên từ các điểm dữ liệu.
- Tính khoảng cách từ từng điểm dữ liệu đến centroid gần nhất đã được chọn trước đó.
- Chọn centroid tiếp theo từ các điểm dữ liệu sao cho xác suất chọn một điểm làm centroid tỷ lệ thuận với khoảng cách của điểm đó đến centroid gần nhất đã chọn.
- Lặp lại cho đến khi chọn đủ k centroid.
- Chúng ta sẽ chọn các centroid cách xa nhau. Điều này làm tăng khả năng ban đầu chọn được các centroid nằm ở các cụm khác nhau.



II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.3 Nguyên lí hoạt động của K-means trong phân cụm ảnh

2.3.1 K-Means trong phân cụm các điểm ảnh

- **Nguyên tắc:** Dựa vào khoảng cách giữa các điểm ảnh

2

Bước 2: Phân cụm các điểm ảnh

- Tính khoảng cách Euclid (Manhattan, Cosine, Minkowski) giữa các điểm ảnh và tâm cụm.
- Gán pixel vào cụm gần nhất.

3

Bước 3: Cập nhật tâm cụm

- Tính trung bình toạ độ cho các pixel trong cùng cụm và đặt làm tâm cụm mới.

4

Bước 4: Lặp lại

- Tiếp tục phân nhóm và cập nhật tâm cụm cho đến khi hội tụ.

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.3 Nguyên lí hoạt động của K-means trong phân cụm ảnh

2.3.2 K-Means trong phân cụm các ảnh có cùng đặc điểm

- **Nguyên tắc:** Dựa vào độ tương đồng giữa các vector Embedding của các ảnh

1

Bước 1: Khởi tạo

- Xác định số lượng k cụm. Chọn ngẫu nhiên k tâm cụm ban đầu
- Hoặc sử dụng thuật toán Kmean++

2

Bước 2: Phân cụm các điểm ảnh

- Tính khoảng cách Euclid (Manhattan, Cosine, Minkowski) giữa các điểm ảnh và tâm cụm.
- Gán ảnh vào cụm gần nhất.

3

Bước 3: Cập nhật tâm cụm

- Cập nhật tâm cụm bằng trung bình của các vector Embedding được gán cho cụm ngay tại lần lặp đó

4

Bước 4: Lặp lại

- Tiếp tục phân nhóm và cập nhật tâm cụm cho đến khi hội tụ.

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.3 Ưu điểm và nhược điểm của K-means

▼Ưu điểm

- Tính đơn giản, dễ hiểu và triển khai nhanh.
- Khả năng xử lý lượng dữ liệu lớn (ảnh có độ phân giải lớn, số ảnh lớn)

▼Nhược điểm:

- Nhạy cảm với việc khởi tạo tâm cụm.
- Có thể hội tụ vào cực tiểu cục bộ nếu không khởi tạo tốt.
- Hiệu quả phụ thuộc vào việc chọn số cụm k.
- Độ chính xác không cao bằng các phương pháp hiện đại

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

2.4 So sánh với các phương pháp phân vùng khác

Tiêu chí	K-means	Fuzzy C-means	Threhold-base
Tính đơn giản	Cao	Trung bình	Cao
Tính linh hoạt	Thấp (đối với biến đổi cạnh)	Cao	Cao
Độ nhạy với	Cao	Thấp	Trung bình

II | Thuật toán K-means

2. Ứng dụng K-means trong phân cụm ảnh

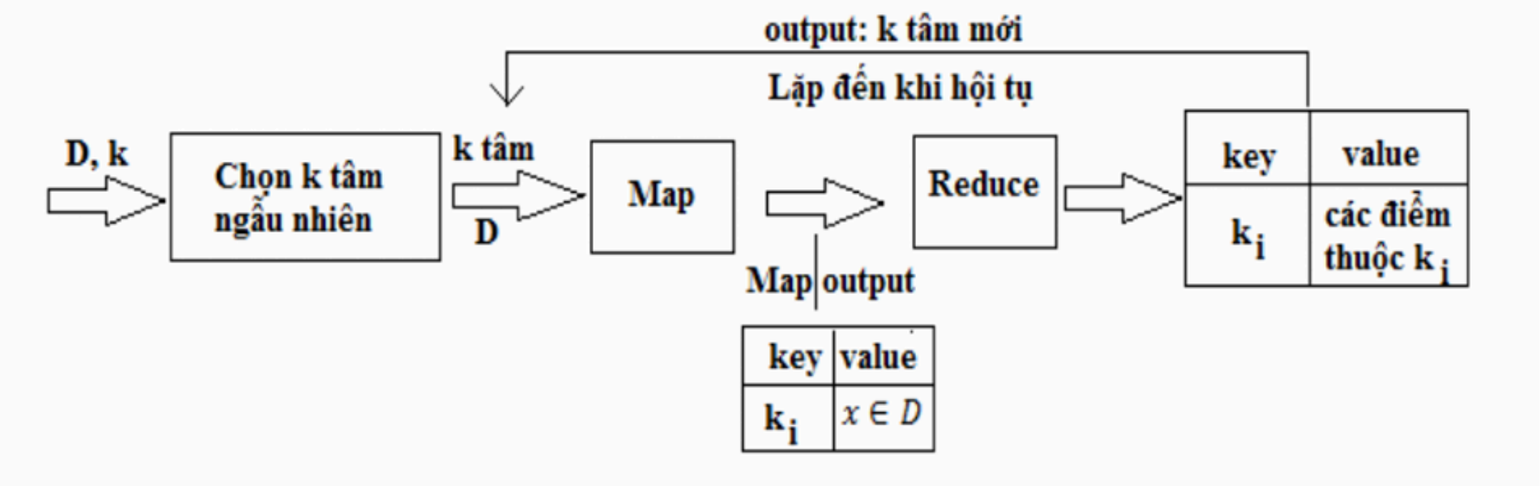
2.5 Kết luận

- Thuật toán K-Means là công cụ hữu hiệu cho phân vùng ảnh nhờ tính đơn giản và hiệu quả cao. Tuy nhiên, cần khắc phục nhược điểm về việc khởi tạo và chọn tham số k để đạt được kết quả tối ưu.
- Trong tương lai, có thể kết hợp với các phương pháp khác như subtractive clustering để cải thiện độ chính xác và hiệu năng.

III | ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Ý tưởng KMeans dựa trên MapReduce:

- MapReduce tận dụng sức mạnh của các hệ thống phân tán để xử lý dữ liệu đồng thời, làm giảm thời gian tính toán của KMeans.
- Việc chia nhỏ dữ liệu và xử lý từng phần giúp giảm áp lực lên bộ nhớ máy chủ, phù hợp hơn với dữ liệu lớn.
- Sơ đồ thuật toán:



ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Ý tưởng KMeans dựa trên MapReduce:

- **Khởi tạo tâm cụm (centroids):**
  - Tâm cụm ban đầu được khởi tạo ngẫu nhiên (hoặc sử dụng phương pháp KMeans++). Những tâm này sẽ được chia sẻ giữa các node trong hệ thống phân tán.
- **Giai đoạn Map:**
  - Mỗi mapper nhận một phần dữ liệu từ bộ dữ liệu lớn.
  - Với mỗi điểm dữ liệu trong phần đó, mapper xác định tâm cụm gần nhất (dựa trên khoảng cách Euclidean hoặc các phương pháp tính khoảng cách khác).
  - Mapper tạo ra cặp giá trị dưới dạng: (tâm cụm, điểm dữ liệu).

ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Ý tưởng KMeans dựa trên MapReduce:

- **Giai đoạn Shuffle and Sort:**
  - Hệ thống MapReduce sẽ nhóm tất cả các điểm dữ liệu thuộc cùng một tâm cụm lại với nhau. Điều này tạo ra các nhóm điểm dữ liệu tương ứng với mỗi tâm cụm.
- **Giai đoạn Reduce:**
  - Mỗi reducer nhận nhóm các điểm dữ liệu thuộc về một tâm cụm.
  - Reducer tính toán tâm cụm mới bằng cách lấy trung bình tọa độ của tất cả các điểm dữ liệu trong nhóm.
  - Tâm cụm mới sẽ được gửi lại để khởi chạy bước lặp tiếp theo.

ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

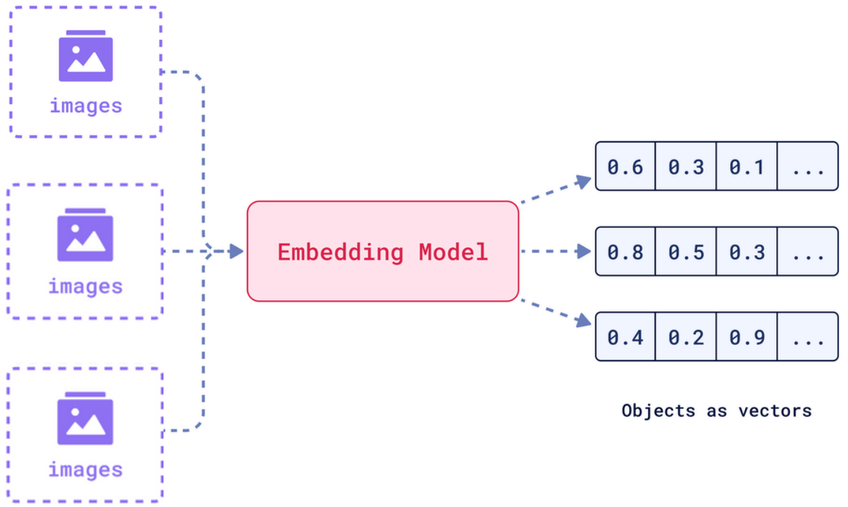
Ý tưởng KMeans dựa trên MapReduce:

- **Lặp lại (Iterative Process):**
  - Các tâm cụm mới được cập nhật và phân phối đến các mapper để tiếp tục xử lý.
  - Quy trình này được lặp lại cho đến khi các tâm cụm hội tụ (tức là không thay đổi đáng kể giữa các bước lặp) hoặc đạt đến số lần lặp tối đa.
- **Kết quả:**
  - Sau khi hội tụ, các tâm cụm cuối cùng được lưu trữ, và mỗi điểm dữ liệu sẽ được gán nhãn dựa trên tâm cụm gần nhất.

ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Lưu đồ thuật toán:

- **Dữ liệu đầu vào:**
  - Dữ liệu đầu vào là toạ độ điểm ảnh hoặc tập hợp ảnh được biểu diễn dưới dạng vector đặc trưng.
  - Ví dụ $x = [f, f, ..., f]$ trong đó f là gí trị đặc trưng thứ k của ảnh.



ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Lưu đồ thuật toán:

- **Kmeans Mapper**
  - Input:
    - keyIn: Số thứ tự ảnh
    - valIn: Vector đặc trưng của ảnh
  - Output:
    - keyInt: Chỉ số cụm j mà ảnh thuộc về.
    - valInt:Vector đặc trưng của ảnh.

ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Lưu đồ thuật toán:

- **KMeans Reducer**
  - Input:
    - keyIn: Chỉ số cụm j
    - valIn: Danh sách các vector đặc trưng của cụm j
  - Output:
    - keyOut: Chỉ số cụm j
    - valOut:Tâm cụm mới C_j^{t+1}

III | ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Lưu đồ thuật toán:

- **Điều kiện hội tụ**
  - Thuật toán ngừng lặp khi đạt đến điều kiện hội tụ

$$||c_j^{t+1} - c_j^t||_2 \leqslant \delta$$

III | ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Phân cụm các điểm ảnh (Image Compression)

- Dữ liệu: Một bức ảnh bất kì được biểu diễn dưới dạng các điểm ảnh
- Thử nghiệm: Thực hiện phân cụm các điểm ảnh với nhiều tham số khác nhau bao gồm số cluster, phép đo khoảng cách khác nhau (Euclid, Cosine, Manhattan, ...)
- Kết quả: Thuật toán hội tự sau 7 lần chạy

ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

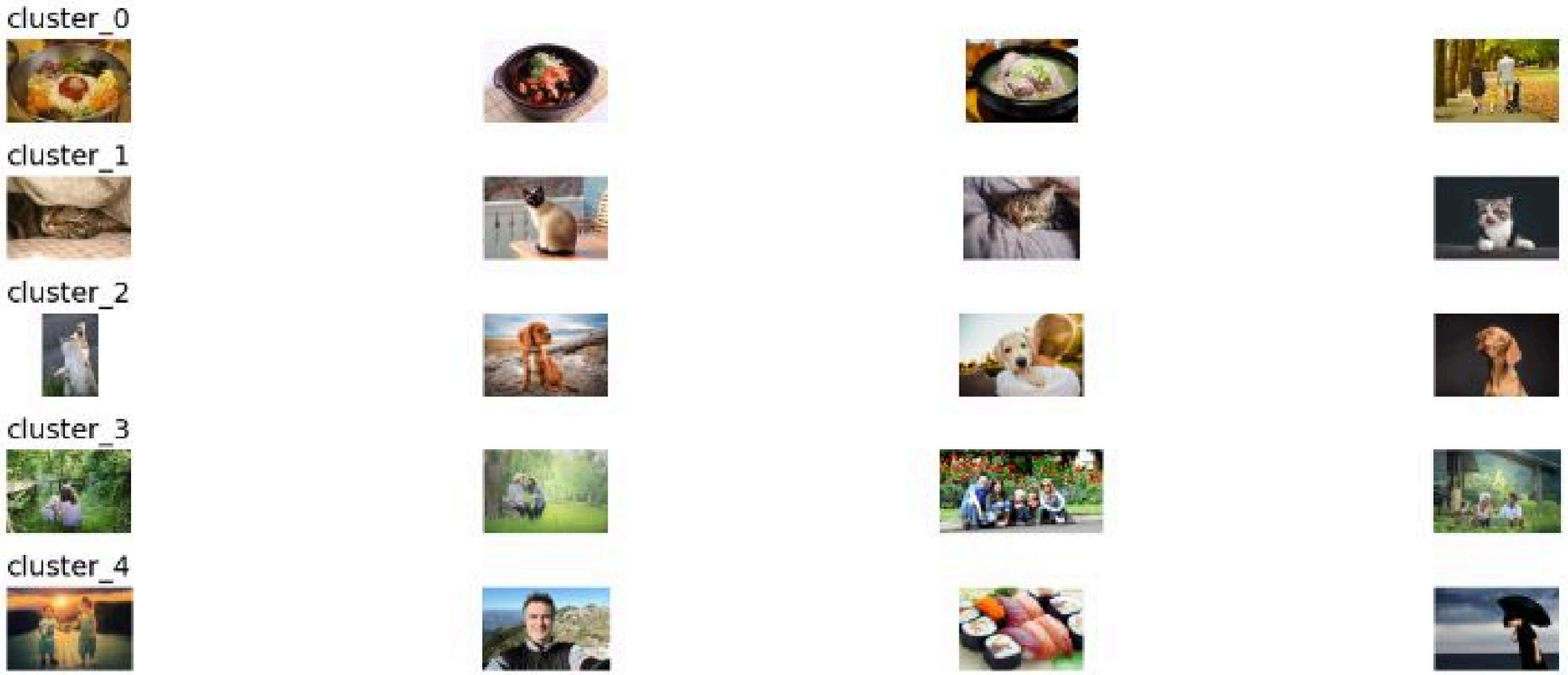
Phân cụm các ảnh (Image Clustering)

- Dữ liệu: Sử dụng bộ dữ liệu nhỏ trên Kaggle dùng cho bài toán phân cụm "Small image dataset for unsupervised clustering", bao gồm 80 ảnh thuộc 5 class khác nhau (dog, cat, family, alone, food)
- Thử nghiệm: Trước khi đưa vào mô hình, các ảnh được biểu diễn dưới dạng vector đặc trưng. Dữ liệu được huấn luyện với nhiều phép đo khoảng cách khác nhau
- Kết quả: Thuật toán hội tụ sau 10 lần chạy

III | ỨNG DỤNG MAPREDUCE TRONG PHÂN CỤM ẢNH BẰNG K-MEANS

Phân cụm các ảnh (Image Clustering)

- Kết quả



IV | KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Kết luận:

- Hiểu rõ khái niệm, vai trò của Big Data, cùng các công cụ xử lý dữ liệu lớn như Hadoop và mô hình MapReduce, đặc biệt là cách tối ưu hóa xử lý dữ liệu phân tán.
- Nắm bắt chi tiết thuật toán K-Means, các bước thực hiện, ưu nhược điểm, và các ứng dụng thực tế trong phân cụm dữ liệu.
- Triển khai thành công chương trình kết hợp K-Means với MapReduce để thực hiện phân cụm ảnh, qua đó khai thác được tiềm năng của việc xử lý dữ liệu lớn trong việc phân tích hình ảnh .
- Thử nghiệm chương trình với tập dữ liệu ảnh cụ thể, thu được kết quả khả quan và tiến hành đánh giá hiệu năng của chương trình.

IV | KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Một số hạn chế cần khắc phục:

- Quy mô dữ liệu thử nghiệm còn nhỏ, chưa đủ để đánh giá toàn diện khả năng mở rộng và hiệu quả của hệ thống trong các bài toán thực tế với dữ liệu lớn.
- Chương trình demo còn hạn chế trong việc hỗ trợ các tập dữ liệu có tính đa dạng cao hoặc các định dạng phức tạp khác.
- Các chiến lược khởi tạo cụm và lựa chọn số cụm k chưa được tối ưu, dẫn đến việc phân cụm đôi khi không đạt hiệu quả cao nhất.

IV | KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hướng phát triển:

- **Mở rộng quy mô xử lý**: Phát triển chương trình để xử lý dữ liệu lớn hơn, ứng dụng vào các lĩnh vực thực tế
- **Nâng cao độ chính xác**: Kết hợp thuật toán K-Means với các phương pháp hiện đại như Gaussian Mixture Models (GMM), hoặc tích hợp với các mô hình học sâu để tăng cường khả năng nhận diện và phân cụm chính xác hơn.
- **Cải thiện hiệu suất**: Nghiên cứu và triển khai các phép đo khoảng cách đa dạng như Cosine, Mahalanobis, hoặc Minkowski để tăng tính linh hoạt của thuật toán khi làm việc với các tập dữ liệu có cấu trúc khác nhau.

IV | KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hướng phát triển:

- **Tích hợp công nghệ mới:** Áp dụng các công cụ và framework hiện đại như Apache Spark, TensorFlow hoặc PyTorch để xử lý dữ liệu nhanh hơn và hiệu quả hơn.
- **Phát triển ứng dụng đa nền tảng**: Tích hợp chương trình vào các hệ thống thực tế như ứng dụng di động hoặc nền tảng web, từ đó nâng cao tính khả dụng và phổ biến của giải pháp.
- **Nghiên cứu thêm thuật toán mới**: Khám phá và thử nghiệm các thuật toán phân cụm khác như DBSCAN, Spectral Clustering để đánh giá hiệu năng so với K-Means.

VIỆN TRÍ TUỆ NHÂN TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

THANKS FOR LISTENING