TRƯỜNG ĐẠI HỌC CÔNG NGHỆ - ĐHQGHN

VIỆN TRÍ TUỆ NHÂN TẠO

-----*****-----

BÁO CÁO MÔN KỸ THUẬT VÀ CÔNG NGHỆ DỮ LIỆU LỚN

ĐỀ TÀI

THUẬT TOÁN K-MEANS & LẬP TRÌNH MAPREDUCE HÓA TRONG PHÂN CỤM ẢNH

Nhóm sinh viên thực hiện:

1. Trần Quốc Sáng - 22022671
2. Nguyễn Quang Trung - 22022665
3. Thái Thị Thùy Linh - 22022631

Giảng viên hướng dẫn: TS. Trần Hồng Việt

LỜI MỞ ĐẦU

Trong thời đại dữ liệu lớn, việc xử lý và phân tích lượng dữ liệu khổng lồ với tốc độ nhanh và độ chính xác cao đã trở thành một trong những thách thức hàng đầu. Các kỹ thuật và công nghệ dữ liệu lớn, như Hadoop và Spark, không chỉ cung cấp khả năng lưu trữ phân tán mà còn hỗ trợ xử lý dữ liệu ở quy mô lớn thông qua các mô hình lập trình song song như MapReduce. Trong lĩnh vực xử lý ảnh, phân cụm đóng vai trò quan trọng trong nhận diện mẫu, phân đoạn hình ảnh và giảm nhiễu. Thuật toán K-Means, khi được kết hợp với MapReduce, tận dụng sức mạnh của các hệ thống dữ liệu lớn để tối ưu hóa hiệu suất tính toán và khả năng mở rộng.

Báo cáo này sẽ tập trung phân tích cách thức áp dụng hai công nghệ trên trong phân cụm ảnh, minh họa trực quan qua việc xử lí phân cụm ảnh chụp CT não để xác định khối u.

Báo cáo bao gồm 5 chương:

Chương 1: Tổng quan về dữ liệu lớn

Chương 2: Giải thuật K-means

Chương 3: Phân cụm ảnh dùng giải thuật K-means song song MapReduce

Chương 4: Kết quả và đánh giá

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ DỮ LIỆU LỚN

1.1. Định nghĩa

Dữ liệu lớn (Big data) là một thuật ngữ cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Dữ liệu lớn bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan, truy vấn và tính riêng tư.

Đặc trưng cơ bản của dữ liệu lớn:

(1) Khối lượng lớn (Volume): Khối lượng dữ liệu rất lớn và đang ngày càng tăng lên, tính đến 2014 thì có thể trong khoảng vài trăm terabyte.

(2) Tốc độ (Velocity): Khối lượng dữ liệu gia tăng rất nhanh.

(3) Đa dạng (Variety): Ngày nay hơn 80% dữ liệu được sinh ra là phi cấu trúc(tài liệu, blog, hình ảnh,...).

(4) Độ tin cậy/chính xác(Veracity): Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu đang là tính chất quan trọng của bigdata.

(5) Giá trị(Value): Giá trị thông tin mang lại.

1.2. Tổng quan về MapReduce

MapReduce là mô hình lập trình song song bắt nguồn từ lập trình chức năng và được Google đề xuất để xử lý lượng dữ liệu lớn trong môi trường phân tán.Dự án Hadoop cung cấp hệ thống file phân tán (HDFS) và hỗ trợ mô hình MapReduce. Hadoop cho phép các ứng dụng làm việc với hàng ngàn nodes với hàng petabyte dữ liệu.

Một tiến trình xử lý MapReduce cơ bản có thể tính toán đến terabytes hoặc petabyte dữ liệu trên hệ thống được kết nối thành cụm các nodes. Dữ liệu được chia thành các mảnh nhỏ rồi đưa vào các nodes độc lập, vì vậy số lượng và kích thước của các mảnh phụ thuộc vào số nodes được kết nối trong mạng

Quy trình thực hiện công việc trên MapReduce:

- B1: Chia dữ liệu đầu vào thành các mảnh dữ liệu

- B2: Thực hiện công việc Map trên từng mảnh dữ liệu đầu vào (Xử lý song song các mảnh dữ liệu trên nhiều máy tính trong cụm).

Bộ ánh xạ (Mapper): xử lý tập dữ liệu đầu vào dưới dạng (key, value) và tạo ra tập dữ liệu trung gian là cặp (key, value).

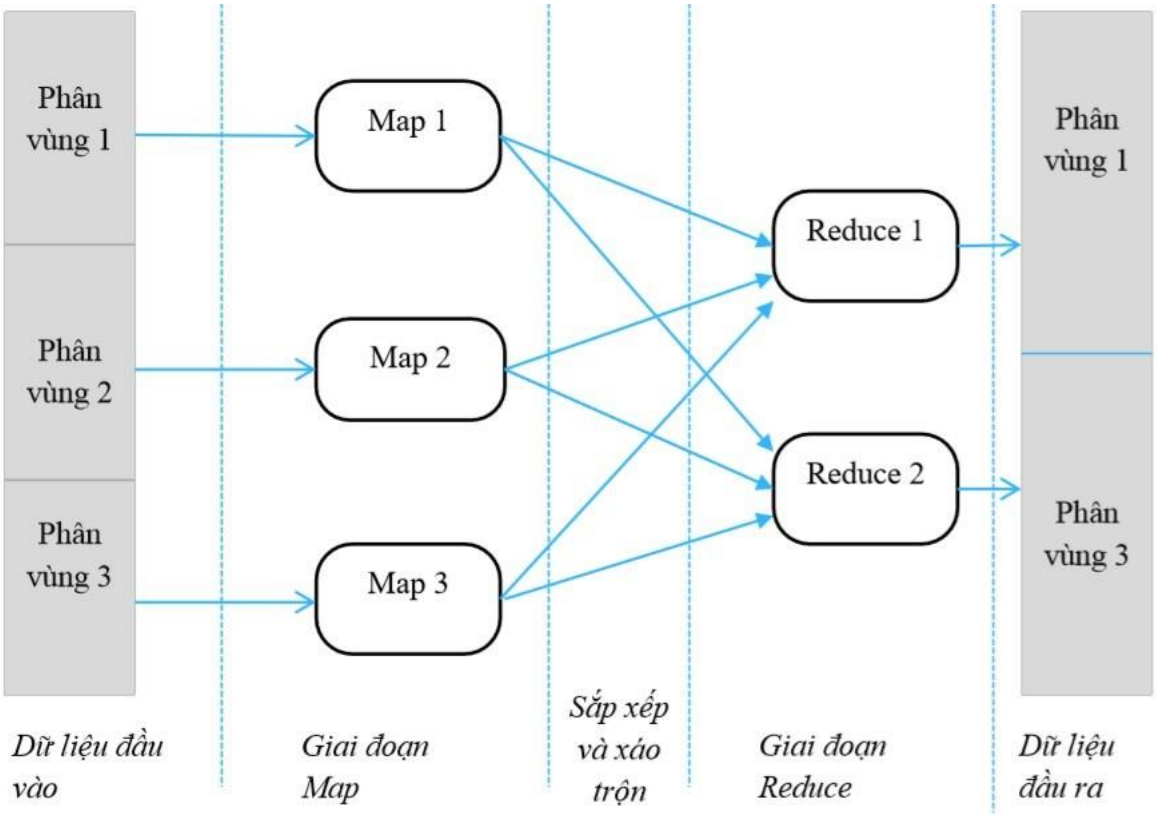
1: Ánh xạ cho mỗi nhóm dữ liệu đầu vào dưới dạng (key, value).

2: Thực thi việc Map, xử lý cặp (key,value) để tạo (key, value) mới, công việc này được gọi là chia nhóm.

3: Đầu ra của bộ ánh xạ được lưu trữ và định vị cho mỗi bộ Reducer.

- B3: Tổng hợp kết quả trung gian (Sắp xếp, trộn).
- B4: Sau khi tất cả công việc Map hoàn thành, thực hiện công việc Reduce trên từng mảnh dữ liệu trung gian

  => Thực hiện song song các mảnh dữ liệu trung gian trên nhiều máy tính trong cụm
- B5: Tổng hợp kết quả hàm Reduce để cho kết quả cuối cùng



CHƯƠNG 2: GIẢI THUẬT K-MEANS

2.1. Tổng quan

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các quy trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác định trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.

2.2. Triển khai

1. Khởi tạo các centroids ban đầu

Chọn k điểm bất kì làm điểm trung tâm (k < n).

2. Nhóm dữ liệu

Với mỗi điểm dữ liệu, tính khoảng cách của nó đến tất cả các centroids và gán điểm đó vào cụm có centroid gần nhất (thường sử dụng khoảng cách Euclid).

3. Cập nhật các centroids

Sau khi gán tất cả các điểm dữ liệu vào các cụm, tính lại tâm cụm mới bằng cách lấy trung bình các điểm dữ liệu trong cụm đó:

$$C_j = \frac{1}{|S_j|} \sum_{x \in S_j} x$$

Với S_j là tập hợp các điểm thuộc cụm j.

4. Lặp lại

Lặp lại các bước 2 và 3 cho đến khi:

- Các centroids không thay đổi (hoặc thay đổi rất ít).
- Số lần lặp đạt ngưỡng tối đa.
- Tổng lỗi (tổng bình phương khoảng cách từ các điểm dữ liệu đến centroid của cụm) không giảm đáng kể.

❖ **Hạn chế:**

- Trong trường hợp xấu nhất, độ phức tạp trở thành superpolynomial
- Việc khởi tạo centroids ngẫu nhiên có thể khiến việc clustering trở nên khó khăn hơn và kết quả sẽ có sự sai khác sau mỗi lần chạy.

CHƯƠNG 3: PHÂN CỤM ẢNH DÙNG GIẢI THUẬT K-MEANS SONG SONG MAPREDUCE

3.1. Bài toán

Dataset:

- Bộ dữ liệu gồm 59 ảnh chụp CT não bộ có khối u.
- Bộ dữ liệu gồm 59 ảnh mask là ảnh nhị phân thể hiện vị trí của khối u sau khi đã xử lí (được coi là đáp án đích).

➢ **Mục tiêu**:

- Phân cụm các pixel trong ảnh chụp CT não bằng K-means để tìm ra vùng bất thường (khối u). Ảnh ouput là ảnh nhị phân với vùng trắng là khối u.
- Sử dụng MapReduce để xử lý lượng lớn dữ liệu ảnh trong môi trường phân tán, giúp tăng tốc độ và khả năng mở rộng.

➢ **K-means**:

Chia tập dữ liệu (ở đây là các pixel ảnh) thành **k cụm** dựa trên khoảng cách của pixel đến các centroid (tâm cụm).

➢ **MapReduce**:

- **Mapper**: Gắn mỗi pixel với cụm gần nhất.
- **Reducer**: Cập nhật centroid của mỗi cụm và kiểm tra điều kiện hội tụ.

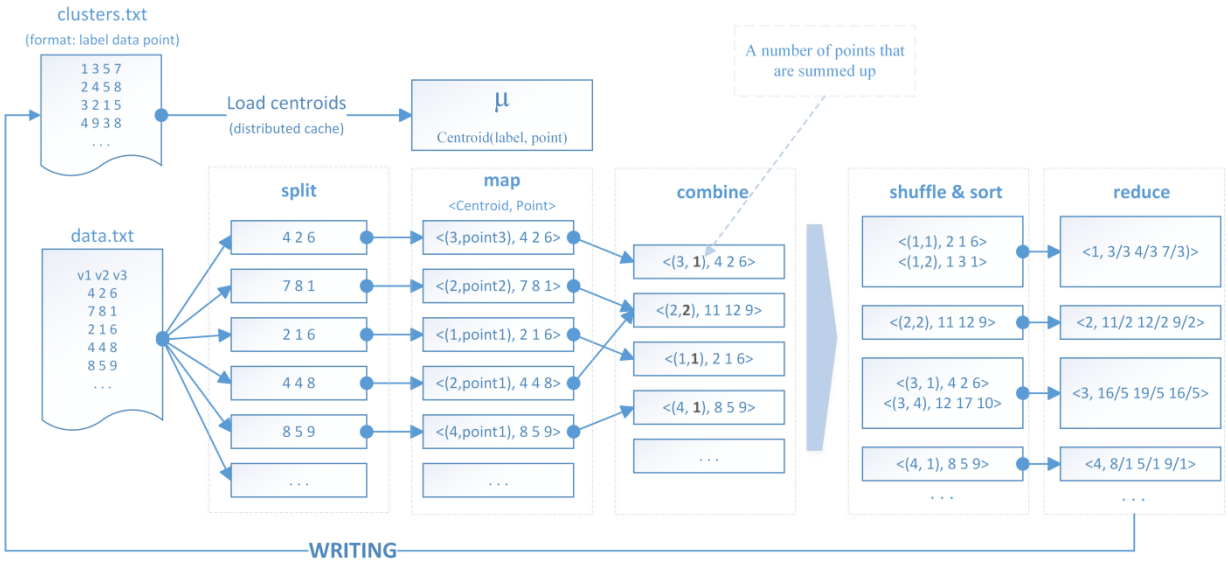
3.2. Triển khai

1. Đầu tiên, Centroid và Context (Configuration) được tải vào Distributed Cache. Điều này được thực hiện bằng cách ghi đè hàm thiết lập trong lớp Mapper và Reducer.

2. Sau đó, tệp dữ liệu đầu vào được chia nhỏ và mỗi điểm dữ liệu được xử lý bởi một trong các hàm map (trong Map process). Hàm này ghi các cặp key-value <Centroid, Point>, trong đó Centroid là điểm gần nhất với Point.

3. Tiếp theo, Combiner được sử dụng để giảm số lượng ghi cục bộ. Trong giai đoạn này, các điểm dữ liệu trên cùng một máy được cộng lại và số lượng các điểm dữ liệu đó được ghi lại, biến Point.number.

4. Bây giờ, vì lý do tối ưu hóa, các giá trị đầu ra được tự động xáo trộn và sắp xếp theo Centroid. Reducer thực hiện cùng một quy trình như Combiner, nhưng nó cũng kiểm tra xem các centroid có hội tụ hay không; so sánh sự khác biệt giữa các centroid cũ và mới với tham số đầu vào delta. Nếu một centroid hội tụ, thì Counter toàn cục sẽ không thay đổi, nếu không, nó sẽ được tăng lên.



Sau khi hoàn thành một lần lặp, các tâm mới được lưu và chương trình kiểm tra hai điều kiện, nếu chương trình đã đạt đến số lần lặp tối đa hoặc nếu giá trị Counter không thay đổi. Nếu một trong hai điều kiện này được đáp ứng, thì

chương trình đã hoàn tất, nếu không, toàn bộ quy trình MapReduce sẽ được chạy lại với các tâm đã cập nhật.

5. Xử lý sau phân cụm

Các pixel trong cụm liên quan đến khối u được làm sáng, các vùng khác được làm tối.

- Đầu ra bài toán: Ảnh nhị phân với khối u là màu trắng, phần còn lại là màu đen.

CHƯƠNG 4: KẾT QUẢ VÀ ĐÁNH GIÁ

4.1. Kết quả so sánh với data mẫu

Sau xử lí ta thu được 59 ảnh nhị phân với vùng trắng là khối u não. Tiến hành so sánh với data mẫu (mask), sử dụng thuật toán Dice Similarity Coefficient (DSC) để đánh giá mức độ chồng lấp giữa hai hình ảnh.

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

Trong đó:

- A: Tập hợp hoặc vùng của hình ảnh thứ nhất.
- B: Tập hợp hoặc vùng của hình ảnh thứ hai.
- $|A \cap B|$: Số phần tử chung (điểm ảnh chung) giữa A và B.
- $|A|$,$|B|$: Tổng số phần tử (điểm ảnh) trong A và B.

Ý nghĩa: Hệ số Dice dao động từ 0 đến 1:

- Dice = 1: Hai tập hoàn toàn giống nhau.
- Dice = 0: Hai tập hoàn toàn không có phần chung.

- **Kết quả**: file comparison_results.txt chứa kết quả so sánh của 59 cặp ảnh. Đạt được độ giống nhau cao nhất là 89.12%, độ giống nhau trung bình đạt 47.13%.

4.2. Đánh giá

1. Ưu điểm

- **Phân tán:** Kết hợp MapReduce cho phép xử lý dữ liệu lớn trên cụm máy tính, rất phù hợp khi cần phân cụm trên lượng lớn ảnh CT hoặc ảnh có độ phân giải cao.
- **Tốc độ:** MapReduce chia nhỏ công việc thành các tác vụ nhỏ hơn (map), đồng thời thực hiện giai đoạn giảm (reduce) để tổng hợp kết quả, giúp tối ưu thời gian xử lý.

- Quá trình lặp lại trong thuật toán K-means (tính tâm cụm và phân cụm lại) có thể được tự động hóa qua các vòng lặp MapReduce.

- Dễ triển khai.

2. Hạn chế

- Ảnh CT chứa nhiều chi tiết phức tạp, như nhiễu hoặc kết cấu không đồng nhất, K-means khó phát hiện chính xác ranh giới của khối u mà không có bước tiền xử lý kỹ càng.
- K-means chỉ xem xét các giá trị pixel (độ xám) mà không tận dụng ngữ cảnh không gian hoặc đặc điểm hình học, dễ dẫn đến lỗi phân cụm.
- Khi xử lý lượng dữ liệu vừa phải (số lượng ảnh nhỏ), MapReduce có thể gây ra chi phí lớn hơn lợi ích do quá trình khởi tạo và quản lý tác vụ.

3. Đề xuất cải tiến

- Tăng cường chất lượng ảnh đầu vào: Lọc nhiễu (Gaussian, Median), cân bằng độ sáng.
- Tối ưu hóa K-means: Sử dụng thuật toán như K-means++ , chỉ cần khởi tạo centroid đầu ngẫu nhiên thay vì k centroids.