TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC

CS336 – TRUY VẤN THÔNG TIN ĐA PHƯƠNG TIỆN

Giảng viên hướng dẫn: ThS. Nguyễn Trọng Chỉnh

Nhóm sinh viên:

- 19522143 – Trương Minh Sơn
- 19522274 – Hồ Thịnh
- 19522057 – Trần Hồ Thiên Phước

Mục lục

I. Giới thiệu truy vấn thông tin

1. Khái niệm:
   Information Retrieval là hoạt động tìm kiếm tài liệu có bản chất phi cấu trúc (unstructured) như văn bản, hình ảnh, video,.. sao cho phù hợp (relevant) với một nhu cầu thông tin (information need) nào đó, từ một tập hợp dữ liệu lớn (large collections). Các ứng dụng hệ thống truy xuất thông tin: Google, MS Bing, công cụ search của máy tính,…
2. Quá trình truy xuất thông tin:
   Người dùng đưa vào truy vấn vào công cụ tìm kiếm → hệ thống so khớp để tìm ra các thông tin liên quan đến truy vấn của người dùng → danh sách các tài liệu liên quan nhất.

II. VECTOR SPACE MODEL

1. Ý tưởng:
   Biểu diễn văn tài liệu và các câu truy vấn dưới dạng vector sau đó tính độ tương đồng của truy vấn ứng với từng tài liệu theo một công thức Δ để tìm ra các tài liệu phù hợp nhất với câu truy vấn. Vấn đề đặt ra ở đây là làm sao biểu diễn các document và query dưới dạng vector và làm thế nào để xác định được công thức tính độ tương đồng giữa câu truy vấn và tài liệu.
2. Biểu diễn các tài liệu và câu truy vấn bởi các Concept vector:
   Đầu tiên ta phân tách tài liệu của chúng ta thành các khái niệm/chủ đề, các khái niệm này gồm các từ ngữ liên quan. Mỗi khái niệm/ chủ đề sẽ là 1 chiều trong không gian → k concepts thì biểu diễn không gian k chiều. Bước tiếp theo ta cần làm là tìm trọng số cho các vector.
3. Trọng số của vector:
   Cách xác định và tính weights cho vector là hết sức quan trọng, ảnh hưởng đến độ chính xác của các thuật toán xếp hạng. Việc các từ có trọng số khác nhau là do không phải các từ đều có sự quan trọng giống nhau, sử dụng số lần xuất hiện của các từ làm vector không phải là một cách tối ưu. Ở phương diện các documents, một vài từ có thể mang nhiều thông tin hơn các từ còn lại.
   Các kỹ thuật tính trọng số: (các kỹ thuật này là các heuristic nên việc xác định cơ sở toán học trong các công thức tính toán có thể là từ thực nghiệm rút ra)
   a. TF(term frequency):
      Ý tưởng của kỹ thuật này là dựa theo tần suất xuất hiện của từ trong tài liệu tức là từ xuất hiện càng nhiều thì càng quan trọng.
      $tf(t,d) = \alpha + (1-\alpha)\frac{f(t,d)}{maxf(t,d)}$.

Trong đó maxf(t,d): số lần xuất hiện của từ xuất hiên nhiều nhất.
Hạn chế của kỹ thuật này là những từ xuất hiện nhiều/ rất ít (nằm ngoài một ngưỡng xác định) trong trong các tài liệu thì có thể ít quan trọng vì nó không giúp ta phân biệt các tài liệu với nhau.

b. IDF(Inverse Document Frequency):
Kỹ thuật này giúp ta đánh giá được độ quan trọng của từ trong tài liệu thay vì tần suất xuất hiện của chúng (khắc phục hạn chế của TF).

$$IDF(t) = 1 + log\frac{N}{df(t)}$$

Trong đó: N là số lượng tài liệu; df(t) là số tài liệu chứa t. Tuy nhiên vẫn có một số trường hợp các từ thuộc nhiều tài liệu khác nhau nhưng lại không mang nhiều thông tin cần thiết nên việc xác định trọng số không thực sự chính xác.

c. TF-IDF:
Đây là phép nhân của TF và IDF. Việc ta kết hợp 2 độ đo ở trên giúp ta xác định được những từ mà xuất hiện trong nhiều tài liệu nhưng ít phổ biến trong những tài liệu có nó giúp cho việc xác định trọng số hợp lý hơn.

4. Tính độ đo tương đồng:
   a. Euclidean distance:

dist(q,d) = $\sqrt{\sum_{t\in V}[tf(t,q)idf(t) - tf(t,d)idf(t)]^2}$

Hạn chế: khi tài liệu lớn/ dài thì sẽ có những trường hợp sai số.

b. Cosine:

Độ tương đồng Cosine là một phép đo để định lượng độ giống nhau giữa hai hoặc nhiều vectơ . Tính độ tương đồng cosine là tính cosine của góc giữa các vector.

$$\cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Độ tương tự có giá trị −1 có nghĩa là trái nghĩa hoàn toàn, với giá trị 1 nghĩa là giống nhau hoàn toàn, với 0 có nghĩa là trực giao hay tương quan (decorrelation), trong khi các giá trị ở giữa biểu thị sự giống nhau hoặc không giống nhau ở mức trung gian.

III. Lập chỉ mục:

Trước khi lập chỉ mục thì chúng ta sẽ xử lý dữ liệu trước qua các bước tiền xử lý: loại bỏ các kí tự không cần (. ; , ; :; …) vì các kí tự này không ảnh hưởng về mặt nghĩa của các từ → tách từ → loại bỏ các stopwords (các từ trong nhóm này thường không ảnh hưởng đến dữ liệu) → stemming (bước này đưa các từ về từ gốc giúp chúng ta giảm số lượng từ đáng kể).

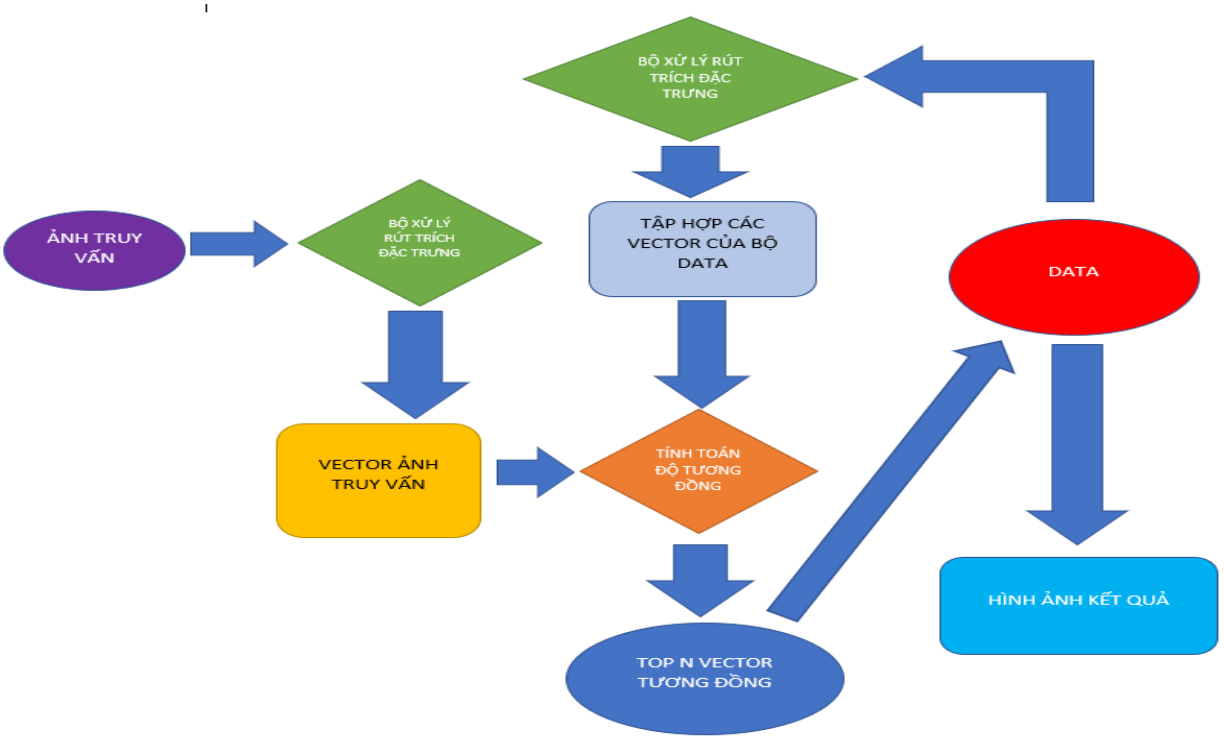
Các bước tiếp theo để lập chỉ mục: thống kê các từ theo docID của chúng (term, docID) → nối các danh sách từ của các document lại và sắp xếp chúng theo term → tập từ vựng → posting list.

IV. Task 1: Truy vấn ảnh dựa trên text:

Sau khi chạy thử đoạn chương trình do nhóm cài đặt, nhóm nhận thấy kết quả trả về đa số là kết quả mong muốn. Tuy nhiên vẫn còn một số trường hợp khi loại bỏ stopwords đã làm cho nghĩa của câu bị thay đổi ít nhiều, điều này dẫn đến một số kết quả truy vấn không thực sự liên quan tới câu truy vấn nhập vào. Cụ thể, trong cài đặt chạy thử với các câu truy vấn như ‘a cAt AnD A doG’ hay ‘CAR oN tHe roaD’ thì kết quả không thực sự khả quan lần lượt là 4/10 và 2/6 kết quả đúng trên 10 tài liệu liên quan đầu tiên trả về.

V. Task 2: Truy vấn ảnh dựa trên ảnh:

1. Mô hình chung:



2. Các đặc trưng ảnh:
   Cũng giống như văn bản, hình ảnh cũng có những đặc trưng để mô tả nội dung của chúng. Các đặc trưng của ảnh:
   a. Đặc trưng về màu sắc:
      Ảnh được mô tả từ tập hợp các pixel rất nhỏ, mỗi pixel là một màu trong một không gian màu. Trong không gian màu, mỗi màu được xác định bằng một giá trị độc lập. Trong không gian màu RGB, 1 màu được mô tả bằng ba màu cơ bản là đỏ (red), lục (green), lam (blue).
      Trong không gian màu HSV (Hue, Saturation, Value), mỗi màu được mô tả bằng các giá trị:
      - Hue là phần màu của mô hình màu, được biểu thị dưới dạng một số từ 0 đến 360 độ.
      - Độ bão hoà (Saturation) là lượng màu xám trong màu, từ 0 đến 100 phần trăm. Một hiệu ứng mờ nhạt có thể có được từ việc giảm độ bão hòa về không để giới thiệu nhiều màu xám hơn. Tuy nhiên, độ bão hòa đôi khi được xem trên phạm vi từ 0-1, trong đó 0 là màu xám và 1 là màu chính.
      - Giá trị (độ sáng) hoạt động kết hợp với độ bão hòa và mô tả độ sáng hoặc cường độ của màu sắc, từ 0-100 phần trăm, trong đó 0 là hoàn toàn đen và 100 là sáng nhất và cho thấy màu sắc nhất.

Histogram của màu sắc: là một độ thị cho ta biết về tần suất xuất hiện của một màu xác có trong một tập màu sắc (ảnh). Có thể dùng làm đặc trưng cho một hình ảnh.

b. Đặc trưng về hình dạng (Shape):
Đây là đặc trưng mô tả hình dạng của các vật thể trong bức ảnh. Đặc trưng về hình dạng được tìm bằng cách tìm các đường biên, là những nơi mà giá trị các pixel màu thay đổi đột ngột.
HOG là viết tắt của Histogram of Oriented Gradient - một loại "feature descriptor". Mục đích của "featura descriptor" là trừu tượng hóa đối tượng bằng cách trích xuất ra những đặc trưng của đối tượng đó và bỏ đi những thông tin không hữu ích. Vì vậy, HOG được sử dụng chủ yếu để mô tả hình dạng và sự xuất hiện của một đối tượng trong ảnh. Bản chất của phương pháp HOG là sử dụng thông tin về sự phân bố của các cường độ gradient (intensity gradient) hoặc của hướng biên (edge directins) để mô tả các đối tượng cục bộ trong ảnh. Các toán tử HOG được cài đặt bằng cách chia nhỏ một bức ảnh thành các vùng con, được gọi là "tế bào" (cells) và với mỗi cell, ta sẽ tính toán một histogram về các hướng của gradients cho các điểm nằm trong cell. Ghép các histogram lại với nhau ta sẽ có một biểu diễn cho bức ảnh ban đầu. Để tăng cường hiệu năng nhận dạng, các histogram cục bộ có thể được chuẩn hóa về độ tương phản bằng cách tính một ngưỡng cường độ trong một vùng lớn hơn cell, gọi là các khối (blocks) và sử dụng giá trị ngưỡng đó để chuẩn hóa tất cả các cell trong khối. Kết quả sau bước chuẩn hóa sẽ là một vector đặc trưng có tính bất biến cao hơn đối với các thay đổi về điều kiện ánh sáng.

3. Demo cài đặt truy vấn dựa trên ảnh:
Mô hình 1: sử dụng không gian màu RGB đưa tập hợp màu của không gian về 8 màu chính là: red,green,blue,black,white,yellow,cyan,purple. Các đặc trưng sẽ là histogram về sự phân bố của 8 màu.
Mô hình 2: sử dụng không gian màu HSV. Thu gọn các giá trị h,v,s (360,100,100) về tập giá trị nhỏ hơn (7,3,3). Tính histogram của 3 kênh màu và nối lại thành một vector đặc trưng
Mô hình 3: sử dụng đặc trưng Hog để tính đặc trưng.
4. Đánh giá kết quả đạt được:
Mô hình 1: không hiệu quả vì số lượng màu ít nên không mô tả được chính xác ảnh, gặp nhiễu.

Mô hình 2: ảnh kết quả chỉ tương đồng về mặt màu sắc. Chỉ hiệu quả trên các bức ảnh có số lượng màu ít. Đối với ảnh có màu phức tạp dễ gặp lỗi.
Mô hình 3: ảnh cho ra kết quả tương đồng về mặt hình dạng nhưng vẫn không chính xác. Gặp lỗi trên những bức ảnh có nhiều chi tiết nhỏ.