(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **2-0001864**

(51)[7] **G06F 17/30**

(13) **Y**

(21) 2-2016-00146
(22) 29.04.2016
(45) 25.10.2018 367
(43) 25.10.2016 343
(73) ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH (VN)
Khu phố 6, phường Linh Trung, quận Thủ Đức, thành phố Hồ Chí Minh
(72) Nguyễn Hoàng Tú Anh (VN), Ngô Đức Thành (VN), Nguyễn Quang Phúc (VN), Lê Đình Duy (VN)

(54) PHƯƠNG PHÁP GOM CỤM KẾT QUẢ TÌM KIẾM VIĐEO TRÊN CÁC KÊNH VIĐEO TRỰC TUYẾN

(57) Giải pháp hữu ích đề xuất phương pháp gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến thông qua việc thu thập dữ liệu viđeo và hai quá trình chính: (1) trích xuất đặc trưng biểu diễn viđeo và tính độ tương tự giữa các viđeo theo từng loại đặc trưng như âm thanh, thị giác, thông tin văn bản đi kèm; (2) áp dụng thuật toán gom cụm dữ liệu để thực hiện gom cụm viđeo dựa trên độ tương tự kết hợp đa đặc trưng. Giải pháp đề xuất nhằm nâng cao độ chính xác gom cụm kết quả tìm kiếm viđeo giúp người dùng có thể dễ dàng xác định được viđeo mà họ quan tâm một cách nhanh chóng thông qua các cụm viđeo trực quan thay vì phải duyệt qua một danh sách phẳng bao gồm nhiều viđeo thuộc nhiều chủ đề trộn lẫn với nhau.

Lĩnh vực kỹ thuật được đề cập

Lĩnh vực công nghệ thông tin, cụ thể là phương pháp gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến nhằm tổ chức kết quả tìm kiếm viđeo trả về theo các cụm chủ đề giúp người dùng tìm kiếm những viđeo mà họ mong muốn một cách dễ dàng hơn.

Tình trạng kỹ thuật của giải pháp hữu ích

Vấn đề cốt lõi để thực hiện gom cụm viđeo là xác định độ tương đồng giữa các viđeo thông qua các biểu diễn dựa trên các đặc trưng của viđeo (ví dụ đặc trưng âm thanh, đặc trưng thị giác hay các thông tin văn bản đi kèm viđeo). Đối với bài toán gom cụm kết quả tìm kiếm viđeo, một thách thức lớn đặt ra là làm sao biểu diễn viđeo để thực hiện tính toán so khớp hiệu quả. Tùy theo từng mục đích cụ thể thì sẽ có những cách biểu diễn khác nhau nhưng với tiêu chí chung là làm sao phải thể hiện được nội dung viđeo một cách "đầy đủ" nhất. Hướng tiếp cận chủ yếu để biểu diễn viđeo là dựa trên các đặc trưng của chúng.

Ở Việt Nam những nghiên cứu cùng loại thì chưa thấy.

Thế giới: Đây là bài toán gom cụm dữ liệu viđeo (video clustering). Một nhánh của bài toán này là ứng dụng gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến. Thách thức đặt ra cho bài toán này là làm sao gom các viđeo thuộc cùng một chủ đề về cùng một cụm thông qua việc tính độ tương tự giữa các viđeo dựa trên các biểu diễn theo các đặc trưng của chúng.

Một số công trình nghiên cứu trước đây chủ yếu khai thác nội dung viđeo dựa trên một loại đặc trưng cụ thể, cách làm này đơn giản nhưng không thể hiện đầy đủ nội dung viđeo. Do đó, hiệu quả so khớp viđeo không cao làm giảm độ chính xác gom cụm viđeo. Một số công trình nghiên cứu khác theo hướng tiếp cận kết hợp đa đặc trưng. Tuy nhiên, ưu điểm của từng loại đặc trưng chưa được phân tích làm rõ qua quá trình rút trích đặc trưng biểu diễn viđeo. Vì vậy, ưu thế của từng loại đặc trưng chưa được khai thác nhiều trong quá trình xử lý so khớp và gom cụm viđeo. Dữ liệu viđeo có cấu trúc phức tạp, nội dung của viđeo chứa đựng đồng thời các loại đặc trưng thị giác (visual), âm thanh (audio) hay văn bản (textual). Các nghiên cứu trước đây chủ yếu khai thác độ tương đồng giữa các viđeo dựa trên đặc trưng thị giác. Cách làm này sẽ giúp gom các viđeo có thể hiện thị giác (sự xuất hiện của các đối tượng,

hình ảnh, ...) gần giống nhau về cùng một cụm. Tuy nhiên, với sự đa dạng của dữ liệu viđeo trên web, các viđeo thuộc cùng một chủ đề (tức sẽ thuộc cùng một cụm sau khi thực hiện gom cụm) nhưng có thể có thể hiện thị giác khác nhau. Khi đó, hướng tiếp cận này sẽ không thật sự hiệu quả.

Song song với việc sử dụng đặc trưng thị giác, gần đây có một số ít công trình nghiên cứu tận dụng thông tin từ các văn bản đi kèm viđeo (ví dụ như tiêu đề, mô tả hay các thẻ từ khóa) nhằm làm tăng khả năng khai thác sự tương đồng giữa các viđeo. Tuy nhiên, các kỹ thuật để thực hiện so khớp các thông tin văn bản đi kèm viđeo còn khá đơn giản và chưa khai thác nội dung ngữ nghĩa thật sự của chúng. Việc tận dụng thông tin văn bản đi kèm viđeo sẽ thật sự hiệu quả khi chúng được mô tả đúng với nội dung thật sự của viđeo. Tuy nhiên, dữ liệu viđeo trên các kênh viđeo trực tuyến thường được tải lên bởi nhiều người dùng, các thông tin văn bản đi kèm viđeo cũng được người dùng khai báo. Trong thực tế, có thể vì những mục đích riêng (ví dụ như thu hút lượt xem) hoặc do cảm nhận chủ quan, người dùng có thể mô tả các thông tin văn bản đi kèm không đúng với nội dung thật sự của viđeo. Trong những trường hợp tương tự như vậy, việc khai thác nội dung từ thông tin văn bản đi kèm viđeo sẽ không thật sự hiệu quả trong quá trình so khớp và gom cụm viđeo.

Bản chất kỹ thuật của giải pháp hữu ích

Nhằm tận dụng thông tin hữu ích từ các loại đặc trưng khác nhau của dữ liệu viđeo, giải pháp đã nghiên cứu và phân tích đặc điểm của từng loại đặc trưng, từ đó đề xuất sử dụng các phương pháp phù hợp để rút trích thông tin trên từng loại đặc trưng nhằm nâng cao hiệu quả so khớp cũng như chất lượng gom cụm kết quả tìm kiếm viđeo.

Mục tiêu của giải pháp gom cụm kết quả tìm kiếm viđeo là gom các viđeo thuộc cùng một chủ đề vào trong cùng một cụm, nhằm giúp người dùng có thể xác định được những viđeo mà họ quan tâm một cách dễ dàng hơn thay vì phải "tốn công" duyệt qua một danh sách phẳng bao gồm nhiều viđeo thuộc nhiều chủ đề được trộn lẫn với nhau như đầu ra của các công cụ tìm kiếm viđeo hiện nay (ví dụ như YouTube, Google Video, ...). Thông qua kết quả gom cụm viđeo trực quan, người dùng có thể định hướng tìm kiếm một cách nhanh chóng và bỏ qua các cụm viđeo không phù hợp (đặc biệt trong những trường hợp người dùng gửi một truy vấn quá ngắn hoặc một truy vấn mơ hồ do tính đa nghĩa của từ khóa truy vấn).

Từ những vấn đề còn hạn chế của các công trình trước đó, giải pháp đã phân tích ưu điểm của từng loại đặc trưng cụ thể để làm cơ sở cho việc kết hợp đa đặc trưng. Giải pháp đề xuất kết hợp đặc trưng âm thanh, đặc trưng thị giác và thông tin

văn bản đi kèm viđeo để biểu diễn nội dung viđeo một cách đầy đủ nhất nhằm làm tăng khả năng khai thác sự tương đồng giữa các viđeo từ đó nâng cao chất lượng gom cụm viđeo.

Phương pháp gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến bao gồm các bước sau:

- Thu thập dữ liệu viđeo: tập dữ liệu viđeo được thu thập từ kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến (ví dụ YouTube, Google Video). Cụ thể, chúng tôi sử dụng phần mềm mã nguồn mở TubeKit để tải dữ liệu viđeo thực từ YouTube thông qua YouTube API. Khi yêu cầu một truy vấn (query), TubeKit sẽ gửi yêu cầu đó đến YouTube và nhận về một danh sách viđeo tương ứng cùng với các thông tin văn bản đi kèm như tiêu đề (title), mô tả (description) và các thẻ từ khóa (tags).
- Trích xuất đặc trưng biểu diễn viđeo: Biểu diễn nội dung viđeo thông qua các đặc trưng như âm thanh (audio), thị giác (visual) hay thông tin văn bản đi kèm viđeo (textual). Tùy theo đặc điểm của từng loại đặc trưng, viđeo được biểu diễn dưới dạng véc tơ đặc trưng đa chiều.
- Tính độ tương tự: Độ tương tự giữa các viđeo được tính bằng khoảng cách giữa các véc tơ đặc trưng đại diện chúng. Cho hai viđeo X và Y được đại diện bởi hai véc tơ đặc trưng $V_X = (x_1, x_2, ..., x_n)$ và $V_Y = (y_1, y_2, ..., y_n)$, khoảng cách giữa chúng có thể được xác định bằng một trong các độ đo phổ biến như L1 (Manhattan), L2 (Euclidean), Cosine.

  L1: là độ đo khoảng cách cơ bản giữa hai véc tơ. Khoảng cách d giữa hai véc tơ $V_X = (x_1, x_2, ..., x_n)$ và $V_Y = (y_1, y_2, ..., y_n)$ được xác định như sau:

$$d(V_X, V_Y) = \sum_{i=1}^{n} |x_i - y_i|$$

  L2: là độ đo được sử dụng khá phổ biến để tính khoảng cách giữa hai véc tơ. Khoảng cách d giữa hai véc tơ $V_X = (x_1, x_2, ..., x_n)$ và $V_Y = (y_1, y_2, ..., y_n)$ được xác định như sau:

$$d(V_X, V_Y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

  Cosine: là độ đo khoảng cách giữa hai véc tơ được xác định dựa trên góc tạo bởi giữa chúng. Khoảng cách d giữa hai véc tơ $V_X = (x_1, x_2, ..., x_n)$ và $V_Y = (y_1, y_2, ..., y_n)$ được xác định như sau:

$$d(V_X, V_Y) = \frac{V_X \cdot V_Y}{\|V_X\| \, \|V_Y\|} = \frac{\sum_{i=1}^{n} x_i \times y_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}$$

Giá trị Cosine luôn thuộc đoạn [0,1]. Không giống L1 hay L2, giá trị Cosine đánh giá tốt độ tương đồng giữa hai véc tơ khi chúng chưa được chuẩn hóa giá trị. Do đó, Cosine phù hợp khi so khớp viđeo được biểu diễn bằng cách kết hợp nhiều véc tơ đặc trưng với nhau (chấp nhận sự chênh lệch về giá trị giữa các véc tơ).

- Gom cụm: Áp dụng thuật toán gom cụm dữ liệu để thực hiện gom cụm viđeo dựa trên độ tương tự kết hợp đa đặc trưng. Cụ thể với đầu vào là ma trận lưu độ đo tương tự giữa các viđeo được tính trước đó, đầu ra là các cụm viđeo.

Mô tả vắn tắt hình vẽ

Hình 1: Minh họa trực quan dữ liệu đầu vào và đầu ra cho bài toán gom cụm kết quả tìm kiếm viđeo ứng với từ khóa truy vấn "Aston".

Hình 2: Mô hình kết hợp đặc trưng âm thanh, đặc trưng thị giác và thông tin văn bản giải quyết bài toán gom cụm kết quả tìm kiếm viđeo.

Hình 3: Minh họa quá trình tính độ tương tự viđeo dựa trên đặc trưng âm thanh (hệ số phổ tần số Mel - MFCC) được biểu diễn theo mô hình túi đựng từ (BoW – Bag of Words).

Hình 4: Minh họa quá trình tính độ tương tự viđeo dựa trên đặc trưng thị giác biến đổi các thuộc tính bất biến theo tỷ lệ (SIFT – Scale Invariant Feature Transform) được biểu diễn theo mô hình BoW.

Hình 5: Minh họa quá trình tính độ tương tự viđeo dựa trên thông tin văn bản đi kèm sử dụng từ điển WordNet.

Mô tả chi tiết giải pháp hữu ích

Giải pháp bắt đầu từ việc thu thập dữ liệu viđeo. Tập dữ liệu viđeo được thu thập từ kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến (ví dụ YouTube, Google Video). Giải pháp sử dụng phần mềm mã nguồn mở TubeKit để tải dữ liệu viđeo thực từ YouTube thông qua giao diện ứng dụng YouTube (Youtube API). Khi yêu cầu một truy vấn (query), TubeKit sẽ gửi yêu cầu đó đến YouTube và nhận về một danh sách viđeo tương ứng cùng với các thông tin văn bản đi kèm như tiêu đề (title), mô tả (description) và các thẻ từ khóa (tags). Sau đó, giải pháp tiến hành gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến thông qua 2 giai đoạn chính:

Giai đoạn 1: Trích xuất đặc trưng biểu diễn viđeo và tính độ tương tự viđeo theo từng loại đặc trưng.

- Đặc trưng âm thanh (audio):

  i) Trích xuất audio từ viđeo: viđeo ở định dạng mp4 sẽ được chuyển đổi sang

tập tin âm thanh dạng wmv để xử lý cho các bước sau.

ii) Rút trích đặc trưng âm thanh: giải pháp sử dụng phương pháp dựa vào hệ số phổ tần số Mel (MFCC - Mel-Frequency Cepstral Coefficients) như là một loại đặc trưng âm thanh được trích xuất từ viđeo. Kỹ thuật rút trích đặc trưng âm thanh dựa trên việc thực hiện biến đổi để chuyển dữ liệu âm thanh đầu vào (tập tin âm thanh ứng với mỗi viđeo) về thang đo tần số Mel, kỹ thuật trích chọn này bao gồm các bước biến đổi liên tiếp, trong đó dữ liệu đầu ra của phép biến đổi này sẽ làm dữ liệu đầu vào cho bước biến đổi tiếp theo. Tín hiệu âm thanh được rời rạc hóa, bao gồm các mẫu liên tiếp nhau khi biểu diễn trên máy tính. Chúng tôi thực hiện lấy mẫu với tần số trong khoảng 300Hz-3700Hz, chia tín hiệu âm thanh thành các đoạn nhỏ với 25ms cho mỗi khung hình. Rút trích đặc trưng MFCC cho ta tập đặc trưng (biểu diễn dạng véc tơ) cho mỗi khung hình. Như vậy, mỗi tập tin âm thanh sẽ được biểu diễn bởi một tập hợp tập các véc tơ đặc trưng biểu diễn cho từng khung hình được chia.

iii) Xây dựng từ điển và biểu diễn viđeo: dựa trên mô hình BoW (Bag of Words), đặc trưng âm thanh được biểu diễn dưới dạng tập các véc tơ được trích xuất từ tập dữ liệu viđeo sẽ được gom cụm vào các nhóm (cluster), mỗi cluster ứng với một audio word (về ý nghĩa tương tự như word (từ) trong xử lý văn bản). Tập các cluster này tạo thành một từ điển. Sau khi rút trích đặc trưng âm thanh ở bước trước thì mỗi viđeo được biểu diễn bởi một tập các véc tơ đặc trưng, ở bước này mỗi véc tơ đặc trưng sẽ được gán vào cluster gần nhất trong từ điển (dựa vào khoảng cách mỗi véc tơ đến các tâm của các cluster đại diện). Sau cùng, mỗi viđeo sẽ được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số cluster (audio word) có trong từ điển.

iv) Độ tương tự giữa các viđeo được tính dựa trên khoảng cách giữa các véc tơ đại diện chúng.

- Đặc trưng thị giác (visual):

i) Trích xuất khung hình (frame): trung bình mỗi viđeo có 24-25 khung hình /giây. Để tránh sự lặp lại của nhiều khung hình tương tự nhau nên cứ mỗi 2 giây chúng tôi sẽ trích xuất một khung hình.

ii) Rút trích đặc trưng thị giác: Để tăng độ chính xác so khớp viđeo thì một trong những yêu cầu quan trọng là các điểm đặc trưng cục bộ (local keypoint features) được rút trích từ các khung hình phải bất biến với những biến đổi về độ sáng, tỉ lệ co giãn, phép xoay, ... Giải pháp sử dụng mô hình

biến đổi các thuộc tính bất biến theo tỷ lệ (SIFT - Scale Invariant Feature Transform) bao gồm các bước chính là phát hiện và mô tả các điểm đặc trưng. Các điểm đặc trưng sẽ được phát hiện và mô tả trên từng frame của mỗi viđeo. Với mỗi đặc trưng, một véc tơ 128 chiều được tạo ra từ bộ mô tả SIFT. Như vậy, mỗi frame của viđeo sẽ được biểu diễn thành một tập các véc tơ đặc trưng 128 chiều. Viđeo được biểu diễn bằng tập hợp tập các véc tơ đặc trưng biểu diễn cho từng khung hình.

iii) Xây dựng từ điển và biểu diễn viđeo: sử dụng mô hình BoW để xây dựng từ điển và biểu diễn viđeo tương tự như quá trình biểu diễn viđeo theo đặc trưng âm thanh. Mỗi viđeo sẽ được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số từ trong từ điển.

iv) Độ tương tự giữa các viđeo được tính dựa trên khoảng cách giữa các véc tơ đại diện chúng.

- Thông tin văn bản (textual):

  - Từ tập dữ liệu viđeo các thông tin văn bản như thành phần tiêu đề, mô tả hay các thẻ từ khóa được rút trích để xem xét nội dung ngữ nghĩa nhằm làm tăng khả năng khai thác sự tương đồng giữa các viđeo.

  - Tính độ tương tự: thông tin văn bản có nội dung ý nghĩa tương tự nhau có thể được diễn đạt với nhiều ngôn từ khác nhau. Do đó, giải pháp đề xuất sử dụng từ điển WordNet để tính độ tương tự ngữ nghĩa giữa các từ thể hiện trong thông tin văn bản đi kèm viđeo. WordNet là một cơ sở dữ liệu từ vựng lớn tiếng Anh. Trong WordNet, các từ được gom lại theo các bộ từ đồng nghĩa (synsets) thông qua mối quan hệ ngữ nghĩa giữa chúng và được tổ chức ở dạng cây phân cấp. Độ tương tự giữa các thông tin văn bản đi kèm viđeo sẽ được ước lượng dựa trên độ tương tự ngữ nghĩa giữa các từ. Độ tương tự ngữ nghĩa giữa các từ được tính theo khoảng cách chiều dài và độ sâu của chúng được tổ chức trong từ điển WordNet dạng cây phân cấp.

Giai đoạn 2: Kết hợp độ tương tự giữa các viđeo theo từng loại đặc trưng được tính toán ở giai đoạn 1 và áp dụng thuật toán gom cụm để tiến hành gom cụm viđeo dựa trên độ tương tự kết hợp đa đặc trưng.

Mỗi viđeo được biểu diễn với các đặc trưng về thị giác, âm thanh và văn bản được xem như một đối tượng cụ thể. Độ tương tự kết hợp đa đặc trưng giữa hai viđeo bất kỳ X và Y được tính theo công thức sau:

$$Sim(X,Y) = \alpha * Sim_{vis}(X,Y) + \beta * Sim_{aud}(X,Y) + (1-\alpha-\beta) * Sim_{tex}(X,Y)$$

Trong đó:

$Sim(X, Y)$ là độ tương tự giữa hai viđeo X và Y.

$Sim_{vis}(X, Y)$ là độ tương tự giữa hai viđeo X và Y theo đặc trưng thị giác.

$Sim_{aud}(X, Y)$ là độ tương tự giữa hai viđeo X và Y theo đặc trưng âm thanh.

$Sim_{tex}(X, Y)$ là độ tương tự giữa hai viđeo X và Y theo thông tin văn bản đi kèm.

$\alpha, \beta \in (0, 1)$ là các trọng số của các đặc trưng. Trọng số này nhằm nhấn mạnh ưu thế của từng đặc trưng cụ thể. Chẳng hạn như $\alpha = 0.5$, $\beta = 0.3$, $1 - \alpha - \beta = 0.2$, trọng số α lớn hơn cho thấy đặc trưng thị giác được nhấn mạnh.

Sau khi độ tương tự giữa các viđeo được tính, thuật toán gom cụm dữ liệu được áp dụng để thực hiện gom cụm viđeo với đầu vào là ma trận lưu độ tương tự giữa các viđeo. Giải pháp sử dụng thuật toán K-Medoids (một thuật toán gom cụm phổ biến) vì đặc điểm của thuật toán này là chọn các đối tượng cụ thể để làm trọng tâm của các cụm và độ đo khoảng cách giữa các đối tượng chỉ cần tính một lần. Điều này là phù hợp với đầu vào là độ đo tương tự kết hợp đa đặc trưng giữa các viđeo được xử lý tính toán trước đó.

Ví dụ thực hiện giải pháp hữu ích

Ví dụ như với truy vấn "Aston" được thực hiện tìm kiếm trên kênh YouTube thì kết quả trả về bao gồm nhiều viđeo thuộc các chủ đề khác nhau như: âm nhạc, bóng đá, xe hơi, ... được trộn lẫn với nhau (xem Hình 1). Điều này gây khó khăn cho người dùng khi xác định viđeo cần tìm. Người dùng phải "tốn công" duyệt qua một danh sách kết quả trả về để tìm được viđeo phù hợp. Trường hợp xấu hơn xảy ra khi kết quả của các chủ đề khác áp đảo chủ đề mà người dùng quan tâm. Trong kịch bản như vậy, việc gom cụm kết quả tìm kiếm viđeo là cần thiết nhằm giúp người dùng dễ dàng xác định viđeo cần tìm. Giả sử với truy vấn "Aston", người dùng muốn tìm kiếm những viđeo liên quan đến chủ đề xe hơi nhưng hầu hết các kết quả tìm kiếm viđeo trả về liên quan đến âm nhạc, bóng đá và những chủ đề khác. Khi đó, việc gom cụm kết quả tìm kiếm viđeo theo các chủ đề riêng biệt sẽ giúp người dùng định hướng tìm kiếm một cách dễ dàng.

Hiệu quả/ lợi ích có thể đạt được

Giải pháp hữu ích này có thể được triển khai trên các kênh viđeo trực tuyến phổ biến hiện nay như YouTube, Google Video. Ứng dụng của nó là gom cụm các kết quả tìm kiếm viđeo trải rộng ở nhiều mức độ ngữ nghĩa theo từng chủ đề cụ thể từ đó giúp người dùng xác định được viđeo mà họ quan tâm một cách nhanh chóng hơn.

Ngoài ra, giải pháp hữu ích này cũng được ứng dụng nhiều trong thực tế, đặc biệt

là trong lĩnh vực an ninh như tổ chức gom cụm các viđeo có chứa các đối tượng tình nghi từ tập dữ liệu viđeo của các camera giám sát, hay làm bước tiền xử lý cho các hệ thống tìm kiếm viđeo (đối với cơ sở dữ liệu lớn, viđeo được tổ chức theo các cụm, nhóm phục vụ cho mục trích truy xuất nhanh kết quả tìm kiếm).

YÊU CẦU BẢO HỘ

1. Phương pháp gom cụm kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến bao gồm các bước sau:

- thu thập dữ liệu viđeo: tập dữ liệu viđeo được thu thập từ kết quả tìm kiếm viđeo trên các kênh viđeo trực tuyến;

- trích xuất đặc trưng biểu diễn viđeo: biểu diễn nội dung viđeo thông qua các đặc trưng như âm thanh (audio), thị giác (visual) hay thông tin văn bản đi kèm viđeo (textual); tùy theo đặc điểm của từng loại đặc trưng, viđeo được biểu diễn dưới dạng véc tơ đặc trưng đa chiều;

- tính độ tương tự: dựa trên véc tơ đặc trưng được tạo ra từ quá trình trích xuất đặc trưng biểu diễn viđeo, tính độ tương tự giữa các viđeo dựa trên khoảng cách giữa các véc tơ đặc trưng biểu diễn chúng nhằm ước lượng độ tương đồng hay so khớp hai viđeo có tương tự nhau về nội dung hay không;

- gom cụm: áp dụng thuật toán gom cụm dữ liệu để thực hiện gom cụm viđeo dựa trên độ tương tự kết hợp đa đặc trưng;

trong đó khác biệt ở chỗ:

bước trích xuất đặc trưng biểu diễn viđeo và tính độ tương tự theo đặc trưng âm thanh bao gồm các bước:

i) trích xuất audio từ viđeo: viđeo ở định dạng mp4 sẽ được chuyển đổi sang tập tin âm thanh dạng wmv để xử lý cho các bước sau;

ii) rút trích đặc trưng âm thanh: giải pháp sử dụng phương pháp dựa vào hệ số phổ tần số Mel (MFCC - Mel-Frequency Cepstral Coefficients) như là một loại đặc trưng âm thanh được trích xuất từ viđeo dựa trên việc thực hiện biến đổi để chuyển dữ liệu âm thanh đầu vào về thang đo tần số Mel; tín hiệu âm thanh được rời rạc hóa, bao gồm các mẫu liên tiếp nhau khi biểu diễn trên máy tính; thực hiện lấy mẫu với tần số trong khoảng 300Hz-3700Hz, chia tín hiệu âm thanh thành các đoạn nhỏ với 25ms cho mỗi khung hình; rút trích đặc trưng MFCC tạo ra tập đặc trưng (biểu diễn dạng véc tơ) cho mỗi khung hình; mỗi tập tin âm thanh sẽ được biểu diễn bởi một tập hợp tập các véc tơ đặc trưng biểu diễn cho từng khung hình được chia;

iii) xây dựng từ điển và biểu diễn viđeo: dựa trên mô hình túi đựng từ (BoW), đặc trưng âm thanh được biểu diễn dưới dạng tập các véc tơ được trích xuất từ tập dữ liệu viđeo sẽ được gom cụm vào các cluster (nhóm), mỗi cluster ứng với một từ audio; tập các cluster này tạo thành một từ điển; mỗi véc tơ đặc trưng sẽ được gán vào cluster gần nhất trong từ điển (dựa vào khoảng

cách mỗi véc tơ đến các tâm của các cluster đại diện); sau cùng, mỗi viđeo sẽ được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số cluster (từ audio) có trong từ điển;

iv) độ tương tự giữa các viđeo được tính dựa trên khoảng cách giữa các véc tơ đại diện chúng;

bước trích xuất đặc trưng biểu diễn viđeo và tính độ tương tự theo đặc trưng thị giác bao gồm các bước :

i) trích xuất khung hình (frame): để tránh sự lặp lại của nhiều khung hình tương tự nhau nên cứ mỗi 2 giây giải pháp sẽ trích xuất một khung hình;

ii) rút trích đặc trưng thị giác: sử dụng mô hình biến đổi các thuộc tính bất biến theo tỷ lệ (SIFT – Scale Invariant Feature Transform) để phát hiện và mô tả các điểm đặc trưng trên từng khung hình của mỗi viđeo; mỗi khung hình của viđeo sẽ được biểu diễn thành một tập các véc tơ đặc trưng 128 chiều; viđeo được biểu diễn bằng tập hợp tập các véc tơ đặc trưng biểu diễn cho từng khung hình;

iii) xây dựng từ điển và biểu diễn viđeo: sử dụng mô hình BoW để xây dựng từ điển và biểu diễn viđeo; mỗi viđeo được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số từ trong từ điển;

iv) độ tương tự giữa các viđeo được tính dựa trên khoảng cách giữa các véc tơ đại diện chúng;

bước trích xuất đặc trưng biểu diễn viđeo và tính độ tương tự theo thông tin văn bản bao gồm các bước:

i) từ tập dữ liệu viđeo các thông tin văn bản như thành phần tiêu đề, mô tả hay các thẻ từ khóa được rút trích để xem xét nội dung ngữ nghĩa nhằm làm tăng khả năng khai thác sự tương đồng giữa các viđeo;

ii) tính độ tương tự: sử dụng từ điển WordNet để tính độ tương tự ngữ nghĩa giữa các từ thể hiện trong thông tin văn bản đi kèm viđeo; các từ được gom lại theo các bộ từ đồng nghĩa (synsets) thông qua mối quan hệ ngữ nghĩa giữa chúng và được tổ chức ở dạng cây phân cấp; độ tương tự giữa các thông tin văn bản đi kèm viđeo sẽ được ước lượng dựa trên độ tương tự ngữ nghĩa giữa các từ được tính theo khoảng cách chiều dài và độ sâu của trong từ điển WordNet;

trong đó bước gom cụm thực hiện gom cụm viđeo dựa trên độ tương tự kết hợp đa đặc trưng gồm các bước :

i) tính độ tương tự kết hợp đa đặc trưng giữa hai viđeo bất kỳ X và Y theo công thức sau :

$$Sim(X,Y) = \alpha * Sim_{vis}(X,Y) + \beta * Sim_{aud}(X,Y) + (1-\alpha-\beta) * Sim_{tex}(X,Y)$$

trong đó:

$Sim(X,Y)$ là độ tương tự giữa hai viđeo X và Y;

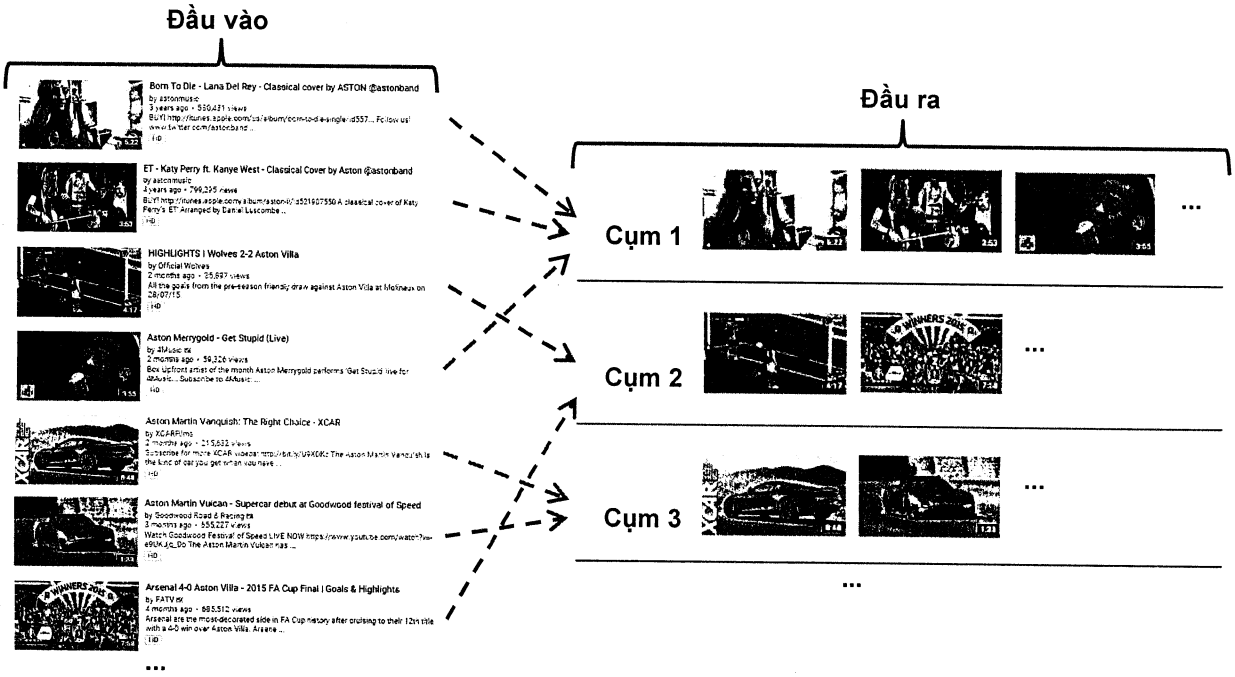
$Sim_{vis}(X,Y)$ là độ tương tự giữa hai viđeo X và Y theo đặc trưng thị giác;

$Sim_{aud}(X,Y)$ là độ tương tự giữa hai viđeo X và Y theo đặc trưng âm thanh;

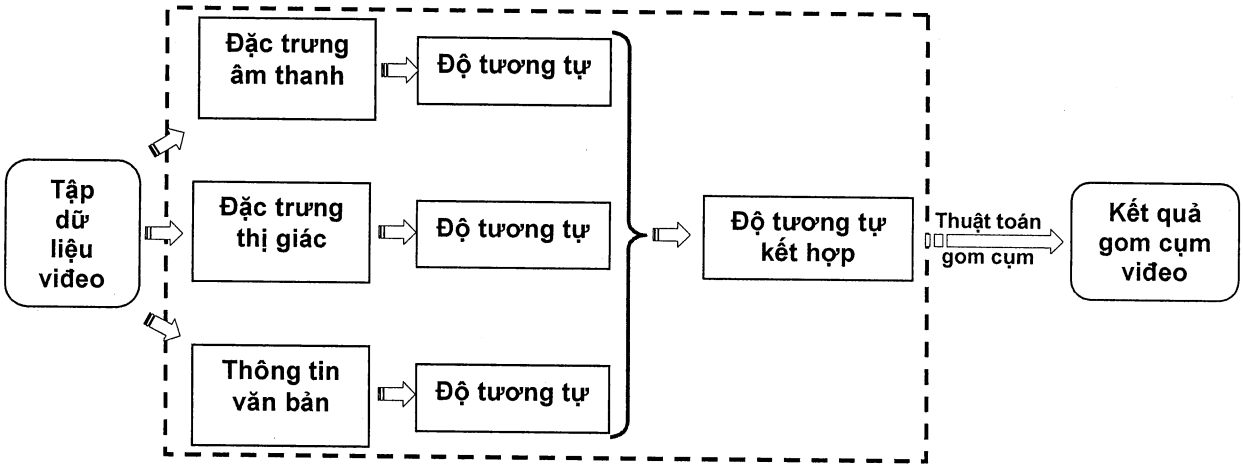
$Sim_{tex}(X,Y)$ là độ tương tự giữa hai viđeo X và Y theo thông tin văn bản đi kèm;

$\alpha, \beta \in (0,1)$ là các trọng số của các đặc trưng nhằm nhấn mạnh ưu thế của từng đặc trưng cụ thể;

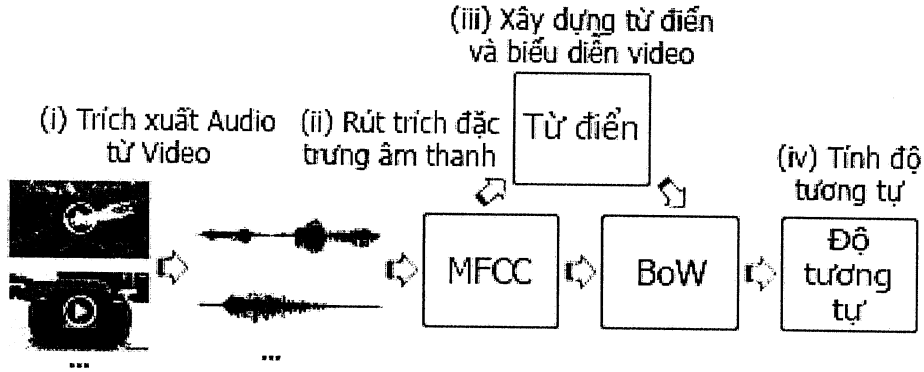
ii) sử dụng thuật toán K-Medoids để thực hiện gom cụm viđeo với đầu vào là ma trận lưu độ tương tự giữa các viđeo.


Đầu vào
Born To Die - Lana Del Rey - Classical cover by ASTON @astonband
ET - Katy Perry ft. Kanye West - Classical Cover by Aston @astonband
HIGHLIGHTS | Wolves 2-2 Aston Villa
Aston Merrygold - Get Stupid (Live)
Aston Martin Vanquish: The Right Choice - XCAR
Aston Martin Vulcan - Supercar debut at Goodwood festival of Speed
Arsenal 4-0 Aston Villa - 2015 FA Cup Final | Goals & Highlights
...
Đầu ra
Cụm 1
Cụm 2
Cụm 3
...

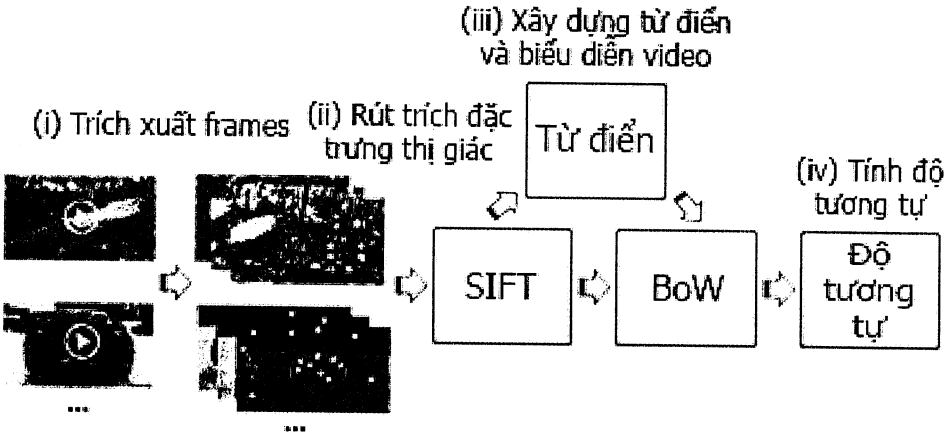
Hình 1


Tập dữ liệu viđeo
Đặc trưng âm thanh
Đặc trưng thị giác
Thông tin văn bản
Độ tương tự
Độ tương tự
Độ tương tự
Độ tương tự kết hợp
Thuật toán gom cụm
Kết quả gom cụm viđeo

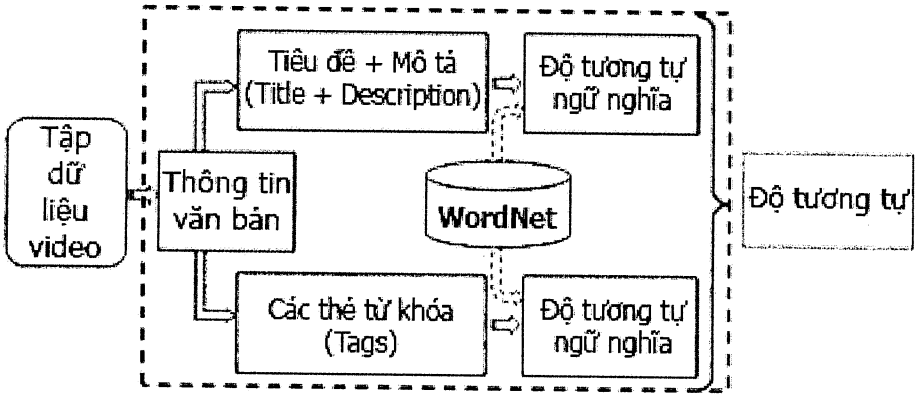
Hình 2



Hình 3



Hình 4



Hình 5