Gom cụm kết quả tìm kiếm video với hướng tiếp cận kết hợp đa đặc trưng

Clustering Web Video Search Results with a Multi-Feature Integration Approach

Nguyễn Quang Phúc

Abstract: *This paper aims to extend our previous researches on clustering web video search results, which reported in [1, 2, 3]. To search videos, users usually use online video search systems such as YouTube, Google Video. However, the returned search results of these systems may include many videos of different categories, and as a result, users find it difficult to locate video clips of interest. Therefore, clustering web video search results is necessary in order to improve the efficiency of searching. The main idea of paper based on analysing and combining the features extracted from video to find the set of appropriate features to improve the quality of video clusters.*

Keywords: *Clustering web video, video representation, multi-feature integration*

I. GIỚI THIỆU

Gom cụm kết quả tìm kiếm trên Web đã cho thấy tính hiệu quả, tiện lợi trong việc tìm kiếm qua các ứng dụng thực tế như ứng dụng gom cụm kết quả tìm kiếm đối với dữ liệu dạng văn bản như Clusty[1], Carrot2[2]; đối với dữ liệu hình ảnh như ứng dụng tìm kiếm ảnh của Google[3]. Với cùng ý tưởng gom cụm kết quả tìm kiếm đối với dữ liệu dạng văn bản và hình ảnh, hướng tiếp cận gom cụm kết quả tìm kiếm đối với dữ liệu video đã được đầu tư nghiên cứu trong những năm gần đây và đây là một hướng nghiên cứu còn khá mới mẻ với nhiều thách thức đặt ra. Để tìm kiếm video, người dùng thường sử dụng các công cụ tìm kiếm trực tuyến như YouTube, Google Video… thông qua các câu truy vấn. Với một câu truy vấn bất kỳ, người dùng sẽ nhận được một số lượng lớn kết quả trả về. Tùy thuộc vào khả năng diễn đạt từ khóa của người dùng mà số lượng video sẽ thay đổi và trải rộng trên nhiều chủ đề khác nhau. Điều này gây trở ngại cho người dùng vì phải tốn nhiều thời gian duyệt danh sách kết quả để tìm được video mong muốn. Đặc biệt, đối với các truy vấn quá ngắn hay mơ hồ do tính đa nghĩa của từ, hoặc trong trường hợp video của chủ đề quan tâm bị áp đảo bởi các chủ đề khác thì quá trình duyệt tìm video mong muốn của người dùng càng gặp nhiều khó khăn. Gom cụm kết quả tìm kiếm video là giải pháp khắc phục vấn đề này. Giải pháp này giúp người dùng có cái nhìn tổng quan hơn thông qua các chủ đề video cụ thể đã được gom cụm. Từ đó, người dùng có thể dễ dàng loại bỏ các cụm video không phù hợp và xác định được các video cần tìm trong thời gian ngắn thay vì phải duyệt toàn bộ danh sách kết quả video trả về.

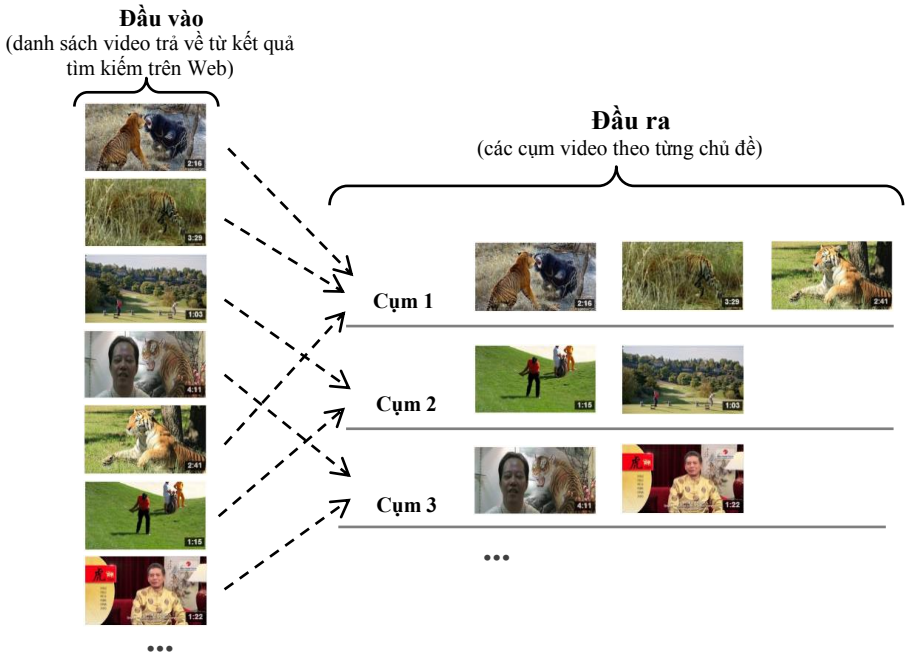
Dữ liệu đầu vào và đầu ra của bài toán gom cụm kết quả tìm kiếm video được minh họa trực quan ở Hình 1.

Một trong những thách thức lớn khi giải quyết bài toán gom cụm kết quả tìm kiếm video là *ước lượng độ tương tự giữa các video*. Danh sách video từ kết quả tìm kiếm video trên Web sẽ được gom thành từng cụm bằng cách áp dụng thuật toán gom cụm dựa trên độ tương tự giữa các video. Thông thường, độ tương tự giữa các video sẽ được tính toán dựa trên các *biểu diễn* của chúng.

[1] http://clusty.com

[2] http://carrot2.org

[3] https://images.google.com



Hình 1. Minh họa trực quan dữ liệu đầu vào và đầu ra của bài toán gom cụm kết quả tìm kiếm video ứng với truy vấn "Tiger" trên YouTube

Dữ liệu video là một dạng dữ liệu có cấu trúc phức tạp với nhiều loại đặc trưng như đặc trưng về thị giác (visual), âm thanh (audio) hay thông tin văn bản đi kèm. Để biểu diễn video, một cách đơn giản là chỉ sử dụng một loại đặc trưng cụ thể. Theo hướng tiếp cận này, Liu cùng các cộng sự đã khai thác thông tin từ đặc trưng thị giác để biểu diễn và so khớp video [4]. Tuy nhiên, để biểu diễn thông tin nội dung video một cách đầy đủ phù hợp cho việc so khớp hiệu quả thì việc chỉ sử dụng một đặc trưng riêng lẻ để biểu diễn video sẽ trở nên hạn chế.

Một hướng tiếp cận mới là sử dụng kết hợp đa đặc trưng nhằm khai thác ưu thế của từng loại đặc trưng giúp nâng cao hiệu quả so khớp và gom cụm video [5, 6]. Trong [5], Hindle cùng các cộng sự khai thác song song đặc trưng thị giác và thông tin văn bản đi kèm video. Tuy nhiên, các kỹ thuật được sử dụng để rút trích đặc trưng và biểu diễn video vẫn còn khá đơn giản chưa phát huy được ưu thế của từng loại đặc trưng. Đối với đặc trưng thị giác, tác giả đề xuất mô hình BCS (Bounded Coordinate System) để biểu diễn video, mô hình này chủ yếu khai thác thông tin màu sắc của video.

Mô hình này hiệu quả khi biểu diễn những video có màu sắc tương đối ổn định, đối với những video có nội dung đa dạng với các bối cảnh và màu sắc khác nhau thì mô hình này có phần hạn chế. Đối với thông tin văn bản đi kèm video, tác giả sử dụng hướng tiếp cận so sánh theo các cặp từ (word-by-word), hạn chế của phương pháp này là bỏ qua tính ngữ nghĩa của từ. Trong [6], Huang cùng các cộng sự cũng khai thác thông tin từ đặc trưng thị giác và thông tin văn bản đi kèm video.

Với đặc trưng thị giác, tác giả chú trọng vào tính bất biến của các đối tượng, hình ảnh trong video kết hợp với thông tin về màu sắc. Với thông tin văn bản đi kèm video, tác giả sử dụng mô hình VSM (Vector Space Model) để biểu diễn và so khớp thông tin văn bản. Mô hình này dựa vào tần suất xuất hiện của các từ trong văn bản để xác định độ tương đồng giữa các văn bản.

Tuy nhiên, do đặc điểm thông tin văn bản đi kèm video thường ở dạng văn bản ngắn và được mô tả bởi những người dùng khác nhau với các ngôn từ khác nhau nên tần suất xuất hiện của các từ giống nhau giữa

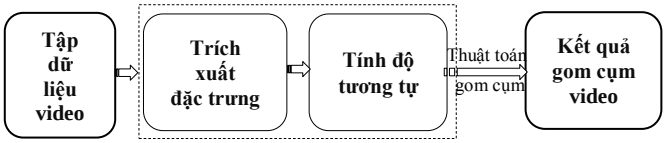
các văn bản là hiếm hoặc thậm chí là không có. Vì vậy, việc sử dụng mô hình VSM để biểu diễn và so khớp thông tin văn bản đi kèm video cũng chưa thật sự hiệu quả.

Nhìn chung, các công trình trước đây chú trọng vào việc khai thác các đặc trưng từ dữ liệu video và thiên về xử lý đặc trưng thị giác được trích xuất trực tiếp từ nội dung video hơn là các thông tin văn bản đi kèm.

Thông qua nghiên cứu các công trình liên quan trước đó, chúng tôi đã chọn hướng tiếp cận kết hợp đa đặc trưng để giải quyết bài toán gom cụm kết quả tìm kiếm video. Chúng tôi tập trung vào việc phân tích đặc điểm thông tin văn bản đi kèm video và chú trọng vào nội dung ngữ nghĩa kết hợp với đặc trưng thị giác để nâng cao chất lượng gom cụm video [1, 2]. Dựa trên việc phân tích đặc điểm các loại đặc trưng video, chúng tôi đã đề xuất sử dụng thêm đặc trưng âm thanh kết hợp với đặc trưng thị giác và thông tin văn bản đi kèm video để nâng cao chất lượng các cụm video [3].

Trong bài báo này, chúng tôi tiếp tục phát triển hướng nghiên cứu gom cụm kết quả tìm kiếm video của chúng tôi trong [1, 2, 3] dựa trên việc phân tích, kết hợp các đặc trưng dữ liệu video để tìm ra *bộ đặc trưng phù hợp* nhằm nâng cao chất lượng gom cụm video. Ý tưởng chính là kết hợp độ tương tự giữa các video theo từng loại đặc trưng. Cụ thể, chúng tôi tận dụng thông tin từ các loại đặc trưng như: *thị giác*, *âm thanh* và *thông tin văn bản đi kèm video* để *làm tăng khả năng khai thác độ tương đồng giữa các video từ đó nâng cao chất lượng gom cụm video*. Ngoài ra, một ứng dụng web được xây dựng minh họa chức năng gom cụm kết quả tìm kiếm video, với chức năng này các kết quả tìm kiếm video thay vì được trình bày như một danh sách phẳng thuộc nhiều chủ đề được trộn lẫn với nhau thì được tổ chức theo các cụm ứng với từng chủ đề cụ thể từ đó giúp người dùng xác định được video mà họ quan tâm một cách nhanh chóng.

Mô hình tổng quát cho bài toán gom cụm kết quả tìm kiếm video được thể hiện ở Hình 2 bao gồm các thành phần sau:



Hình 2. Mô hình tổng quát cho bài toán gom cụm kết quả tìm kiếm video

- **Dữ liệu video:** Dữ liệu video được thu thập từ kết quả tìm kiếm video trên các kênh video trực tuyến (ví dụ như YouTube, Google Video).
- **Trích xuất đặc trưng biểu diễn video:** Video được biểu diễn dựa trên các đặc trưng như: đặc trưng thị giác (visual), đặc trưng âm thanh (audio), thông tin văn bản đi kèm video. Kết quả giai đoạn này là mỗi video sẽ được đại diện bởi một véc tơ đặc trưng đa chiều ứng với từng đặc trưng.
- **Tính độ tương tự:** Độ tương tự được tính nhằm mục đích so khớp hai video có tương tự nhau về nội dung hay không. Độ tương tự càng lớn thì khả năng hai video có nội dung tương tự nhau càng cao. Độ tương tự giữa hai video sẽ được ước lượng dựa trên khoảng cách giữa hai véc tơ đặc trưng biểu diễn chúng với các độ đo phổ biến hiện nay như Cosine, L1 (Manhattan), L2 (Euclidean)...
- **Gom cụm video:** Áp dụng thuật toán gom cụm để thực hiện gom cụm video dựa trên các độ đo tương tự.

Trong bài báo này, chúng tôi tập trung vào hai thành phần chính là *trích xuất đặc trưng biểu diễn video* và *tính độ tương tự so khớp video*. Chúng tôi không đặt trọng tâm vào việc phân tích thuật toán gom cụm vì các thuật toán gom cụm hiện nay được xây dựng khá ổn định, mặt khác chất lượng kết quả gom cụm video phụ thuộc chủ yếu vào độ tương đồng giữa các video dựa trên các biểu diễn của chúng.

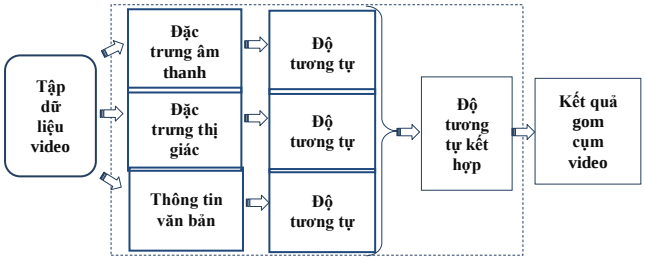
Các mục tiếp theo của bài báo được tổ chức như sau: mục 2 trình bày chi tiết về giải pháp đề xuất, mục 3 trình bày các kết quả thí nghiệm, mục 4 thảo luận về kết quả đạt được.

II. GIẢI PHÁP ĐỀ XUẤT

II.1 Mô hình đề xuất

Việc khai thác đặc trưng thị giác sẽ giúp gom các video có thể hiện thị giác (sự xuất hiện của những đối tượng, hình ảnh) giống nhau về cùng một cụm. Tuy nhiên, với sự đa dạng của dữ liệu video trên Web, những video có nội dung tương tự nhau (tức thuộc cùng một chủ đề) nhưng có thể có những đối tượng và hình ảnh không giống nhau. Khi đó, việc khai thác nội dung ngữ nghĩa từ thông tin văn bản đi kèm video (ví dụ như các thành phần tiêu đề, mô tả hay các thẻ từ khóa) sẽ giúp gom các video có nội dung tương đồng ngữ nghĩa về cùng một cụm. Do đó, đặc trưng thị giác và thông tin văn bản đi kèm video sẽ góp phần bổ sung cho nhau để biểu diễn nội dung video một cách "đầy đủ" làm tăng khả năng khai thác độ tương đồng cũng như chất lượng gom cụm video. Tuy nhiên, một vấn đề đặt ra là việc khai thác nội dung thông tin văn bản đi kèm video chỉ thực sự hiệu quả khi chúng được mô tả đúng với nội dung thực sự của video. Trong thực tế, các thông tin đi kèm video sẽ được người dùng khai báo khi chia sẻ trên các kênh video trực tuyến. Các thông tin này có thể không khớp với nội dung thực sự của video bởi nhiều lý do khác nhau như do cảm nhận chủ quan của người dùng, thu hút lượt xem.... Trong ngữ cảnh như vậy, chúng tôi tin rằng việc khai thác kết hợp đặc trưng âm thanh (ví dụ như những video về ca nhạc thường có các âm thanh như tiếng reo hò, tiếng vỗ tay; những video đua xe thì âm thanh đi kèm là tiếng động cơ xe...) sẽ góp phần cải thiện chất lượng gom cụm video.

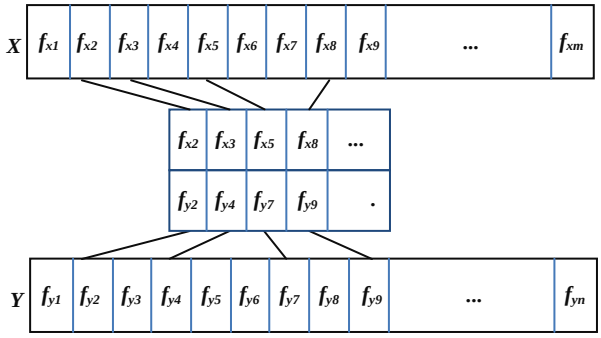
Từ những phân tích trên, chúng tôi xem xét mối kết hợp *đặc trưng thị giác*, *đặc trưng âm thanh* và *thông tin văn bản đi kèm video* để giải quyết bài toán gom cụm kết quả tìm kiếm video (xem Hình 3).



Hình 3. Mô hình kết hợp đa đặc trưng giải quyết bài toán gom cụm kết quả tìm kiếm video

II.2 Biểu diễn và tính độ tương tự video theo đặc trưng thị giác

Một video bao gồm một tập hợp tuần tự các frame. Đặc trưng thị giác được rút trích trực tiếp từ mỗi frame và được biểu diễn dưới dạng véc tơ đặc trưng. Mỗi video có thể được biểu diễn bằng một tập các véc tơ đặc trưng. Với cách biểu diễn này, độ tương tự giữa các video được tính thông qua việc so sánh độ tương tự từng frame của mỗi video (tức mỗi frame trong video này phải được so sánh với tất cả các frame trong video kia) (xem Hình 4). Phương pháp này không hiệu quả khi số lượng frame trong video cũng như số lượng video càng lớn.



Hình 4. Video X với m frame, video Y với n frame. Độ tương tự giữa hai video được tính thông qua việc so sánh từng cặp frame (frame-by-frame)

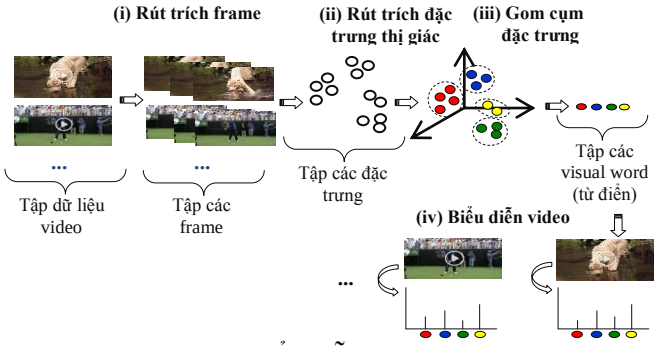
Mặt khác, dữ liệu video trên các kênh video trực tuyến có thể được tùy chỉnh và chia sẻ bởi nhiều người dùng. Điều này có thể dẫn đến số lượng frame khác nhau hoàn toàn trong các phiên bản của cùng một video. Trong những trường hợp này, nếu xem xét tính tương đồng giữa các video dựa trên việc ước lượng số frame tương tự của chúng thì phương pháp nêu trên không phản ánh hoàn toàn độ tương đồng giữa các

video. Cho video *X*, tạo video *Y* bằng cách chọn một frame của video *X* và lặp lại nhiều lần. Nếu số frame của video *Y* lớn hơn số frame của video *X* thì hai video *X* và *Y* được xem như là tương tự nhau mặc dù chúng chỉ có một frame tương tự.

Vấn đề trên có thể khắc phục bằng cách gom các frame tương tự trong cùng một video thành các cụm không giao nhau. Một cụm lý tưởng chỉ chứa các frame tương tự nhau và không có bất kỳ frame tương tự nào nằm ở cụm khác. Khi đó, độ tương tự giữa hai video *X* và *Y* được ước lượng thông qua việc xem xét số cụm được tạo ra từ *hợp hai tập frame* của video *X* và *Y* (*X* ∪ *Y*). Nếu trong một cụm mà có chứa các frame thuộc hai video thì các frame này được xem như là tương tự nhau theo đặc trưng thị giác. Tỷ lệ giữa *số cụm cùng chứa các frame của hai video* và *tổng số cụm được tạo ra* được xem như là độ tương tự giữa hai video. Độ tượng tự này có thể được xem là lý tưởng.Tuy nhiên, chi phí thực hiện tính toán cao. Giả sử cần tính độ tương tự giữa hai video có *l* frame, yêu cầu đầu tiên là phải thực hiện tính toán khoảng cách tương đồng của l^2 cặp frame trước khi chạy thuật toán gom cụm các frame và tính độ tương tự giữa hai video. Hơn nữa, các tính toán này đòi hỏi phải lưu trữ toàn bộ dữ liệu video. Điều này là không phù hợp cho những ứng dụng có cơ sở dữ liệu lớn.

Trong nhiều ứng dụng thực tế như đánh chỉ mục, tìm kiếm video hay xác định các video trùng lặp thì độ tương tự giữa các video được *ước lượng xấp xỉ* nhằm giảm chi phí tính toán thay vì phải biểu diễn toàn bộ thông tin dữ liệu video để tìm ra một độ tương tự lý tưởng với chi phí tính toán và không gian lưu trữ lớn. Trong bài báo này, thay vì phải ước lượng tỷ lệ các frame tương tự nhau để tính độ tương tự giữa các video, chúng tôi chọn hướng tiếp cận biểu diễn dữ liệu video với một đại diện có kích thước cố định như *véc tơ đặc trưng đa chiều*. *Độ tương tự giữa các video được ước lượng thông qua việc tính toán khoảng cách giữa các véc tơ đặc trưng đại diện chúng*.

Quá trình biểu diễn video theo đặc trưng thị giác được thể hiện ở Hình 5 bao gồm các bước chính sau:



Hình 5. Quá trình biểu diễn video theo đặc trưng thị giác

- **Rút trích frame:** các frame được rút trích từ tập dữ liệu video.
- **Rút trích các keypoint từ mỗi frame và mô tả các keypoint (keypoint descriptor):** rút trích keypoint (hay interest point) là xác định vị trí (điểm ảnh) "hấp dẫn" trên mỗi frame. "Hấp dẫn" ở đây có nghĩa là điểm đó có thể có các đặc trưng bất biến khi thay đổi cường độ chiếu sáng, co giãn hay xoay ảnh.... Sau khi các key-point được rút trích, một bộ mô tả (descriptor) được sử dụng để mô tả các keypoint dưới dạng các véc tơ đặc trưng đa chiều phục vụ cho việc tính toán khoảng cách, gom cụm các keypoint được thực hiện ở bước kế tiếp.
- **Gom cụm các keypoint, xây dựng "visual vocabulary[4]":** thuật toán gom cụm được áp dụng để thực hiện gom cụm các keypoint, mỗi cụm được xem như một "visual word" trong từ điển "visual vocabulary".
- **Biểu diễn video:** Tính tần suất xuất hiện trong video của mỗi "visual word" trong "visual vocabulary". Kết thúc bước này, video được biểu diễn bởi một histogram (tạm dịch là biểu đồ tần suất) với các cột mô tả số lần xuất hiện của các "visual word" trong video. Histogram này có thể ánh xạ thành véc tơ đặc trưng có số chiều tương ứng với số "visual word" có trong từ điển.

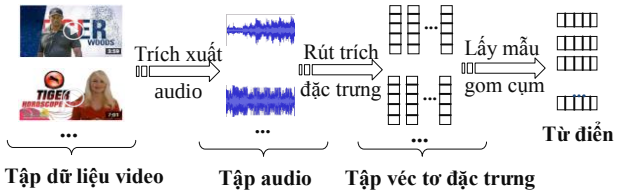
[4] Trong biểu diễn dữ liệu dạng văn bản (text), các từ được định nghĩa là "word". Trong xử lý video, khái niệm "visual word" được hiểu tương tự như "word" trong xử lý văn bản, "visual vocabulary" được xem như một bộ từ điển chứa các "visual word".

II.3. Biểu diễn và tính độ tương tự video theo đặc trưng âm thanh

Như phân tích trước đó, đặc trưng âm thanh đóng một vai trò quan trọng trong việc thể hiện nội dung video giúp làm tăng khả năng khai thác sự tương đồng giữa các video.

Tương tự như quá trình biểu diễn video dựa trên đặc trưng thị giác, sau khi đặc trưng âm thanh được trích xuất từ tập dữ liệu video và được biểu diễn dạng tập các véc tơ đặc trưng, quá trình gom cụm các đặc trưng tạo từ điển được tiến hành. Cuối cùng, mỗi video sẽ được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số từ trong từ điển. Độ tương tự giữa các video được tính là khoảng cách giữa các véc tơ đại diện chúng.

Quá trình tạo từ điển biểu diễn video theo đặc trưng âm thanh được thể hiện ở sơ đồ Hình 6.



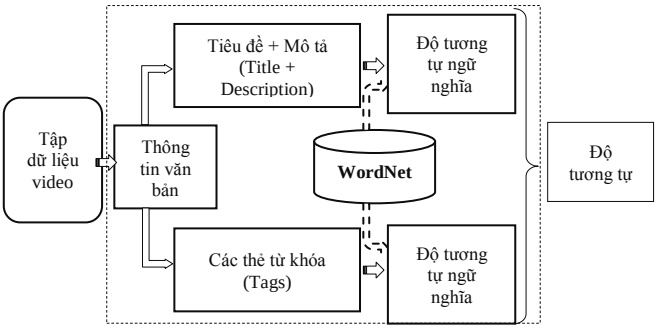
Hình 6. Sơ đồ mô tả quá trình tạo từ điển biểu diễn video dựa trên đặc trưng âm thanh

II.4. Tính độ tương tự video dựa trên thông tin văn bản đi kèm

Thông tin văn bản đi kèm video (ví dụ như tiêu đề (title), mô tả (description), các thẻ từ khóa (tags)) thể hiện nội dung ngữ nghĩa video giúp cải thiện chất lượng gom cụm video. Thông tin văn bản đi kèm video thường được người dùng mô tả dưới dạng cụm từ, câu hay đoạn văn bản ngắn. Độ tương đồng video được ước tính dựa trên độ tương đồng ngữ nghĩa của các mô tả này.

Các phương pháp truyền thống tính độ tương đồng văn bản (ví dụ như Bag-of-Words hay Vector Space Model) chủ yếu tập trung phân tích các từ ngữ dùng chung (sự giống nhau giữa các từ) trong các văn bản. Các phương pháp này hiệu quả khi áp dụng cho các văn bản dài bởi vì trong các văn bản dài có nội dung tương tự nhau thường chứa đựng các từ ngữ giống nhau. Tuy nhiên, trong các văn bản ngắn thì tần suất xuất hiện các từ giống nhau là rất hiếm hay thậm chí là không có từ ngữ nào giống nhau. Điều này chủ yếu là do tính linh hoạt vốn có của ngôn ngữ tự nhiên cho phép người dùng thể hiện cùng một nội dung nhưng với các ngôn từ khác nhau.

Trong bài báo này, chúng tôi đề xuất sử dụng bộ từ điển các từ đồng nghĩa WordNet[5] để tính độ tương tự ngữ nghĩa giữa các từ thể hiện trong thông tin văn bản đi kèm video. Mô hình tính độ tương tự giữa các video dựa trên thông tin văn bản đi kèm sử dụng từ điển WordNet được thể hiện ở Hình 7.



Hình 7. Quá trình tính độ tương tự video dựa trên thông tin văn bản đi kèm sử dụng từ điển WordNet [2, 3]

Ở mô hình thể hiện ở Hình 7, chúng tôi kết hợp *tiêu đề* và *mô tả* của video chung trong một thành phần vì đối với các loại video được chia sẻ trên Web như YouTube thì việc mô tả thông tin cho video tại các thành phần trong thông tin văn bản là không bị ràng buộc theo bất kỳ quy tắc nào, tức các thông tin mang tính giới thiệu, mô tả nội dung video có thể được diễn đạt chi tiết ở thành phần *tiêu đề (title)* hoặc cũng có thể được diễn đạt chi tiết ở thành phần *mô tả (description)* của video. Do đó, để tận dụng tất cả các thông tin có thể, chúng tôi kết hợp *tiêu đề* và *mô tả* của video chung trong một thành phần và xem chúng như là các văn bản ngắn, chúng tôi cũng xem xét *các thẻ từ khóa* của video như là các văn bản ngắn khác.

[5] http://wordnet.princeton.edu

Khi đó, độ tương tự giữa các video sẽ được ước lượng dựa trên độ tương tự ngữ nghĩa giữa các văn bản ngắn trong hai thành phần *tiêu đề + mô tả*, *các thẻ từ khóa* mô tả thông tin văn bản của video.

II.5 Gom cụm video dựa trên độ tương tự kết hợp đa đặc trưng

Mỗi video 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 giữa hai video bất kỳ X và Y được tính theo công thức sau:

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

Trong đó:

- $Sim(X,Y)$ là độ tương tự giữa hai video X và Y.
- $Sim_{vis}(X,Y)$ là độ tương tự giữa hai video X và Y theo đặc trưng thị giác.
- $Sim_{aud}(X,Y)$ là độ tương tự giữa hai video X và Y theo đặc trưng âm thanh.
- $Sim_{tex}(X,Y)$ là độ tương tự giữa hai video 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 video được tính, thuật toán gom cụm dữ liệu được áp dụng để thực hiện gom cụm video với đầu vào là ma trận lưu độ tương tự giữa các video.

II.6 Giải thuật tổng quát cho giải pháp đề xuất

Các bước thực hiện gom cụm kết quả tìm kiếm video của giải pháp đề xuất được thể hiện ở giải thuật sau:

Giải thuật tổng quát cho giải pháp đề xuất

Input: Danh sách n video trả về của 1 truy vấn bất kỳ trên bộ máy tìm kiếm, số cụm k (ứng với số chủ đề của truy vấn)

Output: Các cụm video.

Begin

//Biểu diễn các video thành các vector đặc trưng

1. Biểu diễn mỗi video thành vector $v_i \in V, i = \overline{1,n}$ dựa trên đặc trưng thị giác theo [2].
2. Biểu diễn mỗi video thành vector $a_i \in A, i = \overline{1,n}$ dựa trên đặc trưng âm thanh theo [3].
3. Biểu diễn thông tin văn bản đi kèm mỗi video thành tập hợp các từ $w_i \in W, i = \overline{1,n}$ theo [2].

//Tính độ tương tự giữa các video

4. Tính độ tương tự $VSim_j, j = \overline{1,\frac{n(n-1)}{2}}$ giữa các vector $v_i \in V, i = \overline{1,n}$ theo công thức tính khoảng cách cosine.
5. Tính độ tương tự $ASim_j, j = \overline{1,\frac{n(n-1)}{2}}$ giữa các vector $a_i \in A, i = \overline{1,n}$ theo công thức tính khoảng cách cosine.
6. Tính độ tương tự ngữ nghĩa $WSim_j, j = \overline{1,\frac{n(n-1)}{2}}$ giữa các tập hợp từ $w_i \in W, i = \overline{1,n}$ dựa trên từ điển WordNet [2].
7. Tính độ tương tự kết hợp đa đặc trưng giữa n video theo công thức:
   $Sim = \alpha.VSim_j + \beta.ASim_j + (1-\alpha-\beta).WSim_j$
   với $j = \overline{1,\frac{n(n-1)}{2}}$

//Gom cụm video

8. Áp dụng thuật toán gom cụm K-Medoids để thực hiện gom cụm video dựa trên độ đo tương tự kết hợp đa đặc trưng giữa các video được thực hiện tính trước đó.

End

Vấn đề cốt lõi để giải quyết bài toán gom cụm kết quả tìm kiếm video là ước lượng độ tương đồng giữa các video dựa trên các *biểu diễn* của chúng. Quá trình *trích xuất đặc trưng biểu diễn video* được xử lý offline (quá trình này được xử lý tại máy chủ của công cụ tìm kiếm video tại cùng một thời điểm khi video được lập chỉ mục). *Quá trình được thực hiện trực tuyến (online) trong thời gian thực là gom cụm video*. Quá trình này không mất nhiều thời gian tính toán (độ phức tạp tính toán được ước tính theo thuật toán gom cụm K-Medoids cho mỗi lần lặp là $O(kn^2)$ với k là số cụm, n là số video). Điều này là phù hợp với một hệ thống tìm kiếm video trong thực tế bởi vì người dùng luôn kỳ vọng rằng kết quả tìm kiếm video cần được trả về một cách nhanh chóng sau khi họ nhập truy vấn.

III. THỰC NGHIỆM

Trong phần này, chúng tôi trình bày về các thực nghiệm đánh giá chất lượng gom cụm kết quả tìm kiếm video dựa trên cách tiếp cận kết hợp đa đặc trưng. Thứ nhất, chúng tôi mô tả về bộ dữ liệu video. Thứ hai, chúng tôi trình bày về phương pháp đánh giá chất lượng gom cụm video. Thứ ba, chúng tôi trình bày về các cài đặt thực nghiệm. Cuối cùng, chúng tôi trình bày chi tiết về kết quả thực nghiệm và các thảo luận.

III.1. Bộ dữ liệu video

Dữ liệu video thực được tải từ kết quả tìm kiếm video trên YouTube bởi phần mềm mã nguồn mở TubeKit[6]. Với mỗi truy vấn, chúng tôi tải về khoảng 80 đến 100 video và thực hiện loại bỏ một số video biệt lập, ít liên quan đến truy vấn tìm kiếm. Sự loại bỏ này là hợp lý bởi vì chúng tôi đang thử nghiệm tính năng hậu xử lý gom cụm kết quả tìm kiếm video chứ không phải là tìm kiếm chính xác của một công cụ tìm kiếm video. Các video sau khi tải về sẽ được gán nhãn thủ công theo từng chủ đề cụ thể để làm cơ sở đánh giá kết quả gom cụm video. Các thí nghiệm được tiến hành trên bộ dữ liệu gồm 1752 video của 20 truy vấn với các từ khóa khác nhau. Thông tin chi tiết về bộ dữ liệu video được mô tả ở Bảng 1.

Bảng 1. Bộ dữ liệu video thực nghiệm

Truy vấn	Số video	Số chủ đề	Tổng số giờ video
1. Apple	80	4	7.5
2. Aston	82	4	5.3
3. Cobra	92	5	5.0
4. Dragon	82	6	5.6
5. Jaguar	86	4	5.1
6. Java	87	4	7.2
7. Jupiter	82	4	5.1
8. Leopard	95	5	6.4
9. Lion	89	4	6.2
10. Lotus	91	6	5.5
11. Mustang	83	5	5.6
12. Ocean	90	7	5.5
13. Panda	97	5	5.8
14. Pluto	85	7	8.8
15. Python	85	4	5.1
16. Scorpion	90	6	6.7
17. Tiger	81	4	4.3
18. Venus	89	7	6.9
19. Viper	87	5	4.5
20. Zebra	99	7	6.0

III.2. Phương pháp đánh giá

Chất lượng gom cụm video được đánh giá bởi hai độ đo phổ biến là *Entropy* và *Purity*.

Giả sử có một tập gồm n video thuộc k chủ đề được gán nhãn thủ công ký hiệu là C_j với $j = 1,...,k$ và thuật toán gom cụm n video vào k cụm P_i với $i = 1,...,k$. *Entropy* đánh giá chất lượng gom cụm được tính theo công thức sau:

$$Entropy = -\sum_i \frac{n_i}{n} \sum_j \frac{n_{ij}}{n_i} log \frac{n_{ij}}{n_i} \quad (2)$$

Trong đó: n_i là số video trong cụm P_i, n_{ij} là số video trong cụm P_i thuộc chủ đề C_j, n là tổng số video trong tất cả các cụm.

Trường hợp lý tưởng là mỗi cụm chỉ chứa video thuộc cùng một chủ đề duy nhất. Khi đó, giá trị *Entropy* bằng không. Nói một cách tổng quát, giá trị *Entropy* càng nhỏ thì cho chất lượng gom cụm càng tốt.

Ngược lại với *Entropy*, *Purity* phản ánh độ tinh khiết của các cụm, giá trị *Purity* lớn thì cho kết quả gom cụm tốt hơn. *Purity* đánh giá chất lượng gom cụm được tính theo công thức sau với các ký hiệu có ý nghĩa tương tự như trong công thức tính *Entropy*:

$$Purity = \sum_i \frac{n_i}{n} (max_j \frac{n_{ij}}{n_i}) \quad (3)$$

III.3. Các cài đặt thực nghiệm

[6] www.tubekit.org

Với mục đích so sánh và đánh giá hiệu quả của giải pháp đề xuất, chúng tôi tiến hành cài đặt các phương pháp cơ sở trong [2, 3, 5, 6]. Mặt khác, để làm cơ sở phân tích đánh giá ưu thế của từng loại đặc trưng và xác định bộ đặc trưng phù hợp nhằm nâng cao chất lượng kết quả gom cụm video, chúng tôi tiến hành cài đặt bổ sung các thí nghiệm *kết hợp các bộ đặc trưng khác nhau*. Cụ thể các phương pháp cài đặt của chúng tôi bao gồm:

- Gom cụm video theo từng đặc trưng riêng lẻ.
  - *V (Visual)*: gom cụm video dựa trên đặc trưng thị giác [2, 3].
  - *A (Audio)*: gom cụm video dựa trên đặc trưng âm thanh [3].
  - *T (Textual)*: gom cụm video dựa trên thông tin văn bản đi kèm [2, 3].
- Gom cụm video dựa trên cách kết hợp các bộ đặc trưng khác nhau với cách kết hợp tuyến tính không có trọng số. Với cách kết hợp này, vai trò của các đặc trưng được đánh giá tương đương nhau.
  - *V-A (Visual – Audio)*: gom cụm video dựa trên đặc trưng thị giác và đặc trưng âm thanh.
  - *V-T (Visual – Textual)*: gom cụm video dựa trên đặc trưng thị giác và thông tin văn bản đi kèm video. Ở kịch bản thử nghiệm này, nhằm mục đích đánh giá hiệu quả của phương pháp mà chúng tôi đề suất sử dụng trong [2] với các phương pháp được sử dụng trong [5, 6], chúng tôi thực hiện các cài đặt sau:
    - V-T [2]: Rút trích và biểu diễn đặc trưng thị giác với SIFT (Scale-Invariant Feature Transform) + so khớp thông tin văn bản đi kèm video sử dụng từ điển WordNet.
    - V-T [5]: Rút trích và biểu diễn đặc trưng thị giác với mô hình BCS + biểu diễn và so khớp thông tin văn bản đi kèm video sử dụng mô hình Bag-of-Words nguyên thủy.
    - V-T [6]: Rút trích và biểu diễn đặc trưng thị giác với SIFT + biểu diễn và so khớp thông tin văn bản đi kèm video sử dụng mô hình VSM.
  - *A-T (Audio – Textual)*: gom cụm video dựa trên đặc trưng âm thanh và thông tin văn bản đi kèm video.
  - *V-A-T (Visual – Audio – Textual)*: gom cụm video dựa trên đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản đi kèm video.
- Gom cụm video dựa trên cách kết hợp đa đặc trưng với cách kết hợp có trọng số theo công thức (1).
  - *V^*-A^*-T^* (Visual – Audio – Textual)*: gom cụm video dựa trên đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản đi kèm video có sử dụng trọng số cho mỗi đặc trưng.

Sau đây là chi tiết về các phương pháp cài đặt biểu diễn video, lựa chọn trọng số cho mỗi đặc trưng và quá trình thực hiện gom cụm video:

Biểu diễn video:

Với đặc trưng thị giác, một trong những yếu tố quan trọng để tăng độ chính xác so khớp video là các điểm đặc trưng cục bộ (local keypoint features) được rút trích từ các frame phải bất biến với những biến đổi về độ sáng, tỉ lệ co giãn, phép xoay.... Một trong những phương pháp rút trích và mô tả các đặc trưng cục bộ đáp ứng yêu cầu trên được sử dụng phổ biến nhất hiện nay là Scale-Invariant Feature Transform (SIFT) [7, 8] 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 video. Với mỗi đặc trưng, một véc tơ 128 chiều được tạo ra từ bộ mô tả SIFT.

Như vậy, mỗi frame của video sẽ được biểu diễn bao gồm một tập các véc tơ đặc trưng 128 chiều. Video được biểu diễn bằng tập hợp tập các véc tơ đặc trưng biểu diễn cho từng frame. Từ tập các véc tơ đặc trưng biểu diễn cho các video, chúng tôi sử dụng thuật toán gom cụm Approximate K-Means để tạo từ điển gồm 1000 từ (ứng với các visual word) với 10 lần lặp. Sau cùng, theo mô hình Bag-of-Words, mỗi video sẽ được biểu diễn thành một véc tơ đặc trưng với 1000 chiều. Độ tương tự giữa các video được tính là khoảng cách giữa các véc tơ đại diện chúng.

Với đặc trưng âm thanh, chúng tôi sử dụng Mel-Frequency Cepstral Coefficients (MFCC) [9] để biểu diễn đặc trưng âm thanh được trích xuất từ video. 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 video) 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 khoản 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. Sau đó, quá trình gom cụm các véc tơ đặc trưng tạo từ điển được tiến hành.

Dựa trên mô hình Bag-of-Words, đặc trưng âm thanh được biểu diễn dưới dạng tập các véc tơ được trích xuất từ tập dữ liệu video 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 video 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 video sẽ được biểu diễn bởi một véc tơ đặc trưng với số chiều tương ứng với số cluster (audio word) có trong từ điển. Độ tương tự giữa các video được tính dựa trên khoảng cách giữa các véc tơ đại diện chúng.

Với các thông tin văn bản đi kèm video, sau khi nghiên cứu rộng rãi một số phương pháp, chúng tôi đề xuất sử dụng phương pháp của tác giả Li khai thác từ điển các từ đồng nghĩa WordNet để tính độ tương tự ngữ nghĩa giữa các từ, phương pháp này có sự tương quan tốt nhất với sự đánh giá của con người về mức độ tương tự ngữ nghĩa giữa các từ như được trình bày trong [10].

Lựa chọn trọng số:

Đối với sự đa dạng của dữ liệu video trên web thì đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản đi kèm *đều có vai trò nhất định* trong việc thể hiện nội dung video. Trong từng trường hợp cụ thể thì *vai trò của các đặc trưng thể hiện không giống nhau*. Việc sử dụng trọng số alpha, beta cho từng loại đặc trưng trong công thức (1) *nhằm tối ưu hóa chất lượng kết quả gom cụm video*. Với các trọng số $\alpha, \beta \in (0,1)$ trong công thức (1), chúng tôi tiến hành chạy thực nghiệm bằng cách thay đổi lần lượt giá trị các trọng số với bước nhảy 0.1 để tìm ra bộ trọng số phù hợp. Cụ thể, $(\alpha = i, \beta = j)$ với $i = \overline{0.1,0.9}$ và $j = \overline{0.1,(1-i)}$. Ví dụ: $(\alpha = 0.1, \beta = \overline{0.1,0.9})$, $(\alpha = 0.2, \beta = \overline{0.1,0.8})$, ..., $(\alpha = 0.9, \beta = 0.1)$. Qua thực nghiệm, chúng tôi nhận thấy với bộ trọng số $\alpha = 0.4$ (ứng với đặc trưng thị giác), $\beta = 0.5$ (ứng với đặc trưng âm thanh), $1 - \alpha - \beta = 0.1$ (ứng với thông tin văn bản đi kèm video) cho kết quả tốt hơn các trường hợp còn lại.

Gom cụm video:

Có nhiều thuật toán gom cụm phổ biến như: K-Means, K-Medoids. Tuy nhiên, chúng tôi thử nghiệm gom cụm video với thuật toán K-Medoids 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 video được xử lý tính toán trước đó.

Đối với bài toán gom cụm tổng quát thì số cụm được khai báo linh động bởi người dùng. Số cụm càng ít thì tỷ lệ các đối tượng khác nhau được gom về cùng một cụm càng cao, số cụm càng nhiều thì tỷ lệ các đối tượng giống nhau được gom vào các cụm khác nhau càng lớn. Trong bài báo này, để công bằng trong việc đánh giá giữa các phương pháp thực nghiệm, chúng tôi tiến hành thử nghiệm thuật toán gom cụm với số cụm đầu vào tương ứng với số chủ đề của mỗi truy vấn.

III.4. Kết quả thí nghiệm

Kết quả gom cụm trên các bộ dữ liệu video ứng với các truy vấn khác nhau được đánh giá qua hai chuẩn

độ đo *Entropy* và *Purity* được thể hiện ở Bảng 2 và Bảng 3.

Kết quả thể hiện ở Bảng 2 cho thấy phương pháp V-T [2] cho kết quả gom cụm video tốt hơn (đạt giá trị Entropy thấp hơn) phương pháp V-T [5], V-T [6] trên toàn bộ dữ liệu video của các truy vấn. Điều này chứng tỏ rằng phương pháp rút trích và biểu diễn đặc trưng thị giác với SIFT kết hợp với phương pháp so khớp thông tin văn bản đi kèm video sử dụng từ điển WordNet mà chúng tôi đề xuất sử dụng trong [2] cho chất lượng gom cụm video tốt hơn so với các phương pháp được sử dụng trước đó. Vì thế, trong các thực nghiệm tiếp theo, chúng tôi sẽ sử dụng SIFT để biểu diễn đặc trưng thị giác và từ điển WordNet trong việc so khớp thông tin văn bản đi kèm video.

Sau đây, chúng tôi tiếp tục đánh giá vai trò của từng loại đặc trưng cụ thể ảnh hưởng đến chất lượng gom cụm video. Dựa vào kết quả thực nghiệm ở Bảng 2, chúng tôi thấy rằng trên đa số các truy vấn thì phương pháp sử dụng đặc trưng thị giác (V) và đặc trưng âm thanh (A) cho kết quả gom cụm video tốt hơn (đạt giá trị Entropy thấp hơn) so với thông tin văn bản đi kèm (T). Điều này cho thấy đặc trưng thị giác và đặc trưng âm thanh chiếm ưu thế hơn so với thông tin văn bản đi kèm video khi thực hiện gom cụm video dựa trên từng loại đặc trưng riêng lẻ.

Ngoài ra, kết quả gom cụm video bằng việc kết hợp các cặp đặc trưng khác nhau cũng cho thấy phương pháp kết hợp đặc trưng thị giác và đặc trưng âm thanh (V-A) cho kết quả gom cụm tốt hơn so với các phương pháp kết hợp đặc trưng thị giác với thông tin văn bản (V-T) hay đặc trưng âm thanh với thông tin văn bản (A-T). Điều này cho thấy xu hướng những video có nội dung tương tự nhau (tức thuộc cùng chủ đề) thường có những đối tượng hình ảnh, âm thanh giống nhau.

Bảng 2. Kết quả gom cụm video được đánh giá theo Entropy

Truy vấn	**Entropy**									
	V [2, 3]	**A [3]**	**T [2, 3]**	**V-A**	**V-T [2]**	**V-T [5]**	**V-T [6]**	**A-T**	**V-A-T**	**V^*-A^*-T^***
1. Apple	0.5414	0.5004	0.5122	**0.4442**	0.4586	0.5141	0.5001	0.4895	0.4378	**0.2884**
2. Aston	0.5130	0.4277	0.5111	**0.3896**	0.4465	0.4918	0.4861	0.4299	0.3953	**0.3276**
3. Cobra	0.5523	0.5145	0.5837	**0.4545**	0.5258	0.5593	0.5341	0.4883	0.4675	**0.3048**
4. Dragon	0.5317	0.4649	0.6410	**0.3454**	0.4403	0.5312	0.4929	0.5382	0.3892	**0.2817**
5. Jaguar	0.4713	0.4465	0.5251	**0.3518**	0.3681	0.4402	0.4240	0.4237	0.3723	**0.2146**
6. Java	0.2844	0.3266	0.5149	**0.1584**	0.2083	0.3525	0.2322	0.3529	0.1187	**0.0570**
7. Jupiter	0.3300	0.4182	0.4875	**0.2538**	0.2701	0.3992	0.3080	0.4467	0.2891	**0.1883**
8. Leopard	0.4160	0.5057	0.5610	**0.2252**	0.2686	0.3767	0.3234	0.5320	0.2487	**0.1029**
9. Lion	0.5412	0.5030	0.5570	**0.4660**	0.4828	0.5311	0.5113	0.4893	0.4880	**0.3126**
10. Lotus	0.5096	0.5018	0.6525	**0.3423**	0.3751	0.4857	0.4426	0.5789	0.3894	**0.1431**
11. Mustang	0.5500	0.5203	0.5887	**0.4347**	0.4828	0.5233	0.5111	0.5137	0.4662	**0.1869**
12. Ocean	0.5716	0.5351	0.6559	**0.4622**	0.5207	0.5766	0.5421	0.5708	0.4971	**0.3064**
13. Panda	0.4066	0.5106	0.6058	**0.2693**	0.2803	0.4181	0.3321	0.5396	0.3069	**0.2082**
14. Pluto	0.3546	0.3166	0.5026	**0.2887**	0.3396	0.3715	0.3402	0.4191	0.3223	**0.1773**
15. Python	0.3320	0.4048	0.5246	**0.2023**	0.2352	0.3685	0.2545	0.4521	0.2467	**0.1068**
16. Scorpion	0.4294	0.3707	0.6082	**0.3099**	0.3735	0.4445	0.3987	0.4044	0.3331	**0.2454**
17. Tiger	0.4181	0.4147	0.5460	**0.3301**	0.3682	0.4120	0.3811	0.4237	0.3561	**0.2185**
18. Venus	0.5598	0.5001	0.6751	**0.4336**	0.4813	0.5426	0.5069	0.4813	0.4112	**0.2072**
19. Viper	0.5415	0.5018	0.5927	**0.3729**	0.4301	0.5560	0.4842	0.5356	0.4160	**0.2527**
20. Zebra	0.6405	0.5963	0.6863	**0.5156**	0.5598	0.6302	0.6098	0.6532	0.4992	**0.3094**
Trung bình	0.4748	0.4640	0.5766	**0.3525**	0.3958	0.4763	0.4308	0.4881	0.3725	**0.2220**

Với sự phong phú, đa dạng của dữ liệu video trên web thì những video thuộc cùng một chủ đề nhưng có thể có những đối tượng hình ảnh và âm thanh khác nhau. Khi đó, chúng tôi tin rằng việc khai thác thông tin văn bản đi kèm video sẽ giúp cải thiện chất lượng gom cụm. Như vậy, các thông tin được trích xuất từ đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản đi kèm video sẽ bổ trợ cho nhau làm tăng khả năng khai thác sự tương đồng giữa các video từ đó nâng cao chất lượng kết quả gom cụm.

Tuy nhiên, vấn đề đặt ra là kết hợp như thế nào để có thể tận dụng được ưu thế của từng loại đặc trưng. Để xem xét vấn đề này, chúng tôi tiến hành hai thí nghiệm sau: (i) kết hợp tuyến tính không sử dụng trọng số giữa đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản (V-A-T), (ii) kết hợp đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản với các trọng số khác nhau cho mỗi đặc trưng (V^*-A^*-T^*).

Trong phương pháp V-A-T, ưu thế của các đặc trưng được xem như cân bằng nhau. Kết quả thực nghiệm cho thấy phương pháp này cũng cho kết quả tốt hơn so với việc sử dụng từng loại đặc trưng riêng lẻ trên hầu hết các bộ dữ liệu video của các truy vấn. Điều này một lần nữa minh chứng cho tính hiệu quả của việc kết hợp đa đặc trưng. Tuy nhiên, với dữ liệu video thực tế thì mỗi loại đặc trưng đóng một vai trò khác nhau trong việc thể hiện nội dung video dẫn tới việc kết hợp nhiều loại đặc trưng với sự cân bằng về vai trò chưa hẳn sẽ cho một kết quả gom cụm tốt nhất. Giả định rằng một trong các đặc trưng không thể hiện tốt nội dung video thì việc kết hợp với sự cân bằng về ưu thế sẽ làm hạn chế vai trò của các đặc trưng còn lại.

Ví dụ như trong trường hợp thông tin văn bản đi kèm video được người dùng mô tả không sát với nội dung thực sự của video thì việc kết hợp thêm thông tin văn bản với sự cân bằng về vai trò sẽ làm hạn chế ưu thế của đặc trưng thị giác và đặc trưng âm thanh. Kết quả Bảng 2 cho thấy phương pháp V-A cho kết quả gom cụm tốt hơn so với phương pháp V-A-T khi vai trò của các đặc trưng được cân bằng.

Với phương pháp V*-A*-T*, mỗi đặc trưng được gán trọng số khác nhau thể hiện vai trò khác nhau. Kết quả Bảng 2 cho thấy phương pháp này cho kết quả gom cụm video tốt nhất (đạt giá trị Entropy thấp nhất chứng minh xác suất phân bố các video thuộc cùng một chủ đề vào các cụm khác nhau là thấp nhất) trên hầu hết các bộ dữ liệu video thực nghiệm. Bằng thực nghiệm, chúng tôi thấy rằng với bộ trọng số $\alpha = 0.4$ (ứng với đặc trưng thị giác), $\beta = 0.5$ (ứng với đặc trưng âm thanh), $1 - \alpha - \beta = 0.1$ (ứng với thông tin văn bản đi kèm video) cho kết quả tốt hơn các trường hợp còn lại.

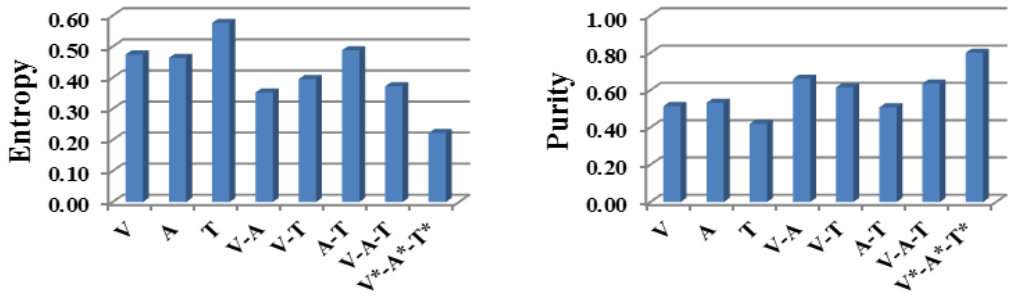
Kết quả gom cụm video thể hiện ở Bảng 3 cho thấy phương pháp V*-A*-T* cũng cho kết quả gom cụm video tốt nhất (đạt giá trị Purity cao nhất chứng minh tỉ lệ phân bố những video thuộc cùng một chủ đề vào cùng một cụm là cao nhất) so với các phương pháp thực nghiệm khác.

Tóm lại, đối với dữ liệu video trên web thì đặc trưng thị giác, đặc trưng âm thanh và thông tin văn bản đi kèm đều có vai trò nhất định trong việc thể hiện nội dung video. Trong từng trường hợp cụ thể thì vai trò của các đặc trưng thể hiện không giống nhau. Kết quả thực nghiệm của chúng tôi cho thấy rằng việc kết hợp đặc trưng thị giác, âm thanh và thông tin văn bản đi kèm video với các trọng số phù hợp sẽ mang đến hiệu quả cải thiện đáng kể chất lượng gom cụm video. Hình 8 thể hiện chất lượng gom cụm video được đánh giá trên toàn bộ các truy vấn qua các phương pháp thực nghiệm.

Với kết quả thực nghiệm đạt được, chúng tôi xây dựng một ứng dụng web minh họa cho chức năng tổ chức kết quả tìm kiếm video trả về theo các cụm. Với chức năng này, người dùng có thể duyệt qua kết quả tìm kiếm video một cách dễ dàng hơn thay vì phải xem xét một danh sách phẳng với nhiều video thuộc nhiều chủ đề trộn lẫn vào nhau. Song song với chức năng hiển thị kết quả tìm kiếm video theo dạng danh sách như các công cụ tìm kiếm video trước đây, ứng dụng hỗ trợ chức năng hiển thị kết quả tìm kiếm video theo các cụm giúp người dùng có cái nhìn trực quan hơn đối với những video mà họ quan tâm (xem Hình 9).

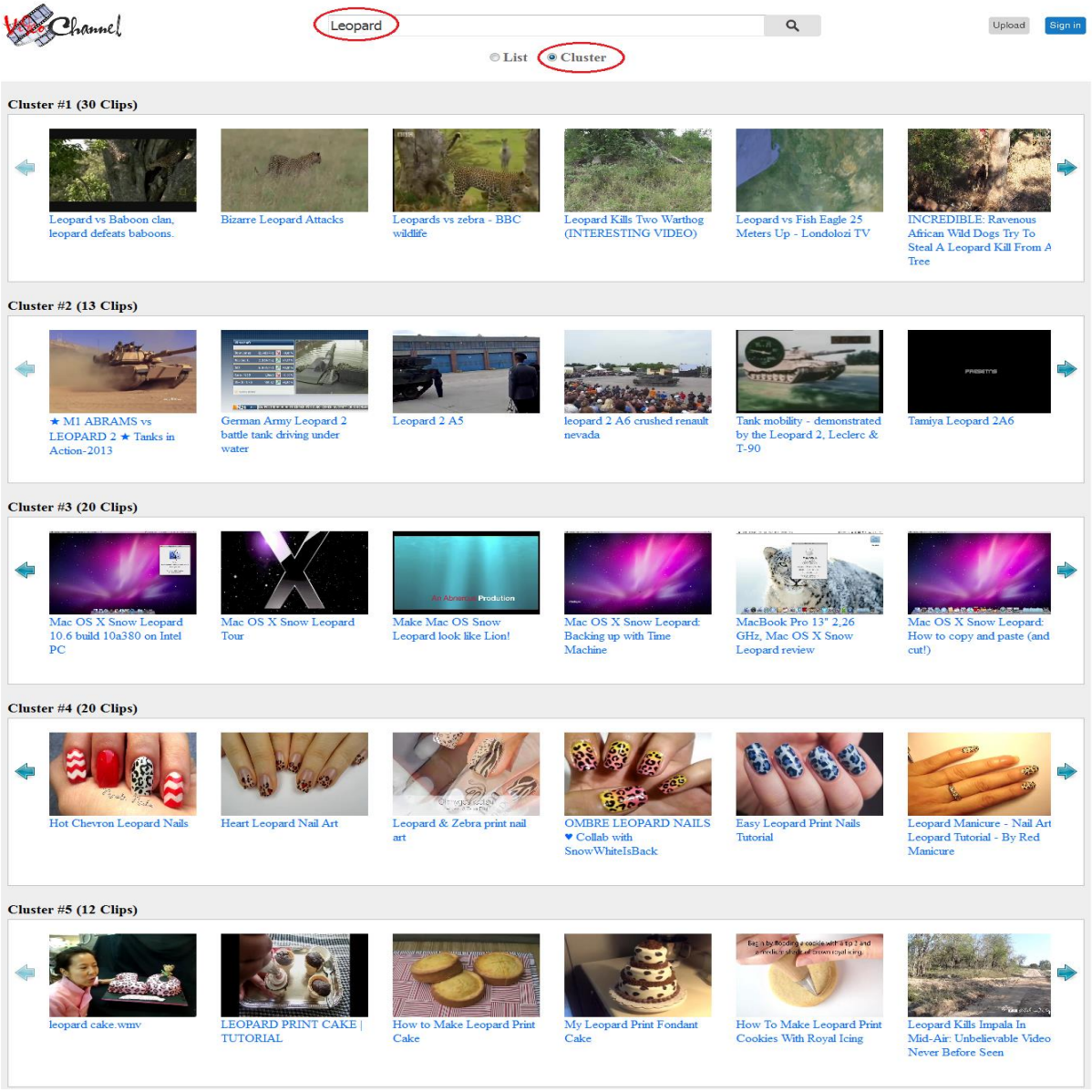
Bảng 3. Kết quả gom cụm video được đánh giá theo Purity

Truy vấn	Purity									
	V [2, 3]	A [3]	T [2, 3]	V-A	V-T [2]	V-T [5]	V-T [6]	A-T	V-A-T	V*-A*-T*
1. Apple	0.4625	0.4875	0.4375	0.5375	0.5500	0.4375	0.5000	0.4500	0.6000	**0.7250**
2. Aston	0.4268	0.5610	0.4512	**0.5976**	0.5122	0.4634	0.4756	0.5488	0.6341	**0.6585**
3. Cobra	0.4130	0.4239	0.4130	**0.5435**	0.5000	0.3913	0.4565	0.5326	0.5435	**0.7303**
4. Dragon	0.4390	0.5122	0.3780	**0.6341**	0.5976	0.5000	0.5122	0.4756	0.5854	**0.6829**
5. Jaguar	0.4419	0.5349	0.4651	0.6512	0.6628	0.5698	0.5930	0.6279	0.6047	**0.8333**
6. Java	0.7126	0.6897	0.4483	**0.8621**	0.8276	0.6552	0.7586	0.6437	0.9195	**0.9529**
7. Jupiter	0.6543	0.5802	0.4938	**0.7407**	0.7037	0.6049	0.6790	0.5432	0.6790	**0.8462**
8. Leopard	0.6316	0.5474	0.4842	**0.8211**	0.7474	0.6632	0.6947	0.4526	0.7895	**0.9053**
9. Lion	0.4270	0.4944	0.3820	**0.5169**	0.4831	0.4157	0.4607	0.4494	0.5056	**0.7528**
10. Lotus	0.4835	0.4835	0.3626	**0.6703**	0.6264	0.5275	0.5275	0.4176	0.6374	**0.8681**
11. Mustang	0.4578	0.4940	0.4096	**0.6386**	0.5663	0.4819	0.5060	0.5060	0.5542	**0.8675**
12. Ocean	0.4556	0.4778	0.4000	**0.5667**	0.5111	0.4667	0.5000	0.4222	0.5333	**0.7444**
13. Panda	0.5567	0.4124	0.3711	**0.7423**	0.6804	0.4948	0.6289	0.4536	0.7010	**0.8041**
14. Pluto	0.6706	0.6824	0.5647	**0.7294**	0.6941	0.6706	0.6824	0.6118	0.7059	**0.8171**
15. Python	0.6786	0.6235	0.4471	**0.7765**	0.7294	0.6471	0.7059	0.5529	0.7176	**0.9294**
16. Scorpion	0.6000	0.6444	0.4111	**0.7111**	0.6556	0.5778	0.6222	0.6333	0.6778	**0.7556**
17. Tiger	0.5062	0.5309	0.3827	**0.6420**	0.6049	0.5185	0.5556	0.5062	0.5926	**0.7654**
18. Venus	0.4607	0.5393	0.3483	**0.6404**	0.5618	0.4494	0.5393	0.5281	0.5955	**0.8315**
19. Viper	0.4368	0.4943	0.3908	**0.6667**	0.6092	0.4368	0.5057	0.4598	0.6092	**0.7586**
20. Zebra	0.3737	0.4242	0.3535	**0.5152**	0.4949	0.4040	0.4040	0.3232	0.5051	**0.7857**
Trung bình	0.5144	0.5319	0.4197	**0.6602**	0.6159	0.5188	0.5654	0.5069	0.6345	**0.8007**



Hình 8. Chất lượng gom cụm video được đánh giá trên toàn bộ các truy vấn qua các phương pháp thực nghiệm

Kết quả thí nghiệm thể hiện ở Hình 9 bao gồm 5 cụm video liên quan đến truy vấn "Leopard". Cụm 1 bao gồm những video liên quan đến động vật (con báo). Cụm 2 bao gồm những video liên quan đến xe tăng. Cụm 3 bao gồm những video liên quan đến hệ điều hành máy tính (hệ điều hành Snow Leopard của hãng Apple). Cụm 4 bao gồm những video liên quan đến nghệ thuật vẽ móng tay và cụm 5 bao gồm những video liên quan đến bánh ngọt. Thông qua kết quả gom cụm video trực quan, người dùng có thể xác định được những video mà họ quan tâm một cách dễ dàng hơn.



Hình 9. Ứng dụng web gom cụm kết quả tìm kiếm video ứng với truy vấn "Leopard"

Giả định rằng với truy vấn "Leopard", người dùng muốn tìm kiếm những video liên quan đến xe tăng nhưng hầu hết các kết quả tìm kiếm video trả về liên quan đến động vật, hệ điều hành máy tính và những chủ đề khác. Khi đó, việc gom cụm kết quả tìm kiếm video theo các chủ đề riêng biệt sẽ giúp người dùng định hướng tìm kiếm một cách nhanh chóng.

IV. KẾT LUẬN

Trên cơ sở phân tích đặc điểm các đặc trưng của dữ liệu video, chúng tôi đã đề xuất các giải pháp kết hợp nhằm tìm ra bộ đặc trưng phù hợp giúp nâng cao chất lượng gom cụm kết quả tìm kiếm video trên các kênh video trực tuyến. Kết quả thực nghiệm cho thấy rằng việc sử dụng bộ đặc trưng bao gồm đặc trưng thị giác, âm thanh và thông tin văn bản đi kèm video đã làm tăng hiệu quả cải thiện chất lượng gom cụm video. Bằng thực nghiệm chúng tôi đã đề xuất được bộ trọng số phù hợp cho các đặc trưng.

Về mặt thực tiễn, chúng tôi bước đầu xây dựng một ứng dụng web thử nghiệm tìm kiếm video với chức năng gom cụm kết quả trả về. Với chức năng này, danh sách video trả về sẽ được gom theo từng cụm với từng chủ đề nhằm giúp người dùng có thể xác định video cần tìm một cách nhanh chóng thay vì phải quệt qua một danh sách phẳng các video thuộc nhiều chủ đề được trộn lẫn với nhau.

Trong tương lai, bằng cách dịch và so sánh các thông tin văn bản đi kèm video với các ngôn ngữ khác nhau, chúng tôi hy vọng có thể gom cụm các video có nội dung tương tự mặc dù thông tin văn bản đi kèm có thể được thể hiện bởi một ngôn ngữ khác với truy vấn.

TÀI LIỆU THAM KHẢO

[1] NGUYỄN QUANG PHÚC, NGUYỄN HOÀNG TÚ ANH, NGÔ ĐỨC THÀNH, LÊ ĐÌNH DUY, "*Gom cụm dữ liệu web video theo hướng tiếp cận early fusion cho đặc trưng văn bản*", Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ 7 về Nghiên cứu cơ bản & ứng dụng Công nghệ thông tin (FAIR), tr. 145-152, 2014.

[2] PHUC QUANG NGUYEN, ANH-THU NGUYEN-THI, THANH DUC NGO, TU-ANH HOANG NGUYEN, "*Using Textual Semantic Similarity to Improve Clustering Quality of Web Video Search Results*", Proceedings of the 7th International Conference on Knowledge and Systems Engineering (KSE), pp. 156-161, 2015.

[3] NGUYỄN QUANG PHÚC, NGUYỄN THỊ ANH THƯ, NGÔ ĐỨC THÀNH, LÊ ĐÌNH DUY, NGUYỄN HOÀNG TÚ ANH, "*Nâng cao chất lượng gom cụm kết quả tìm kiếm video sử dụng kết hợp đặc trưng âm thanh, đặc trưng thị giác và thông tin văn bản*", Kỷ yếu Hội thảo Quốc gia về Điện tử, Truyền thông và Công nghệ thông tin (REV-ECIT), tr. 130-135, 2015.

[4] S. LIU, M. ZHU, Q. ZHENG, "*Mining similarities for clustering web video clips*", CSSE (4), pp. 759-762, 2008.

[5] A. HINDLE, J. SHAO, D. LIN, J. LU, R. ZHANG, "*Clustering Web Video Search Results Based on Integration of Multiple Features*", WWW, pp. 53-73, 2011.

[6] H. HUANG, Y. LU, F. ZHANG, S. SUN, "*A Multi-modal Clustering Method for Web Videos*", Trustworthy Computing and Services, pp. 163-169, 2013.

[7] D. G. LOWE, "*Distinctive Image Features from Scale-Invariant Keypoints*", International Journal of Computer Vision, 60(2), pp. 91-110, 2004.

[8] D. G. LOWE, "*Object Recognition from Local Scale-Invariant Features*", International Conference on Computer Vision, vol. 2, pp. 1150-1157, 1999.

[9] U. SRINIVASAN, S. PFEIFFER, S. NEPAL, M. LEE, L. GU, S. BARRASS, "*A Survey of Mpeg-1 Audio, Video and Semantic Analysis Techniques*", Multimedia Tools and Applications, 27(1), pp. 105-141, 2005.

[10] Y. H. LI, Z. BANDAR, D. MCLEAN, "*An approach for measuring semantic similarity using multiple information sources*", IEEE Transactions on Knowledge and Data Engineering, 15(4), pp. 871-882, 2003.

Nhận bài ngày: 16/03/2016

SƠ LƯỢC VỀ TÁC GIẢ

NGUYỄN QUANG PHÚC

Tốt nghiệp cử nhân tại Trường ĐH Sư phạm TP. HCM, chuyên ngành Sư phạm Tin học năm 2012.

Hiện đang là học viên cao học của Trường ĐH Công nghệ thông tin, ĐH Quốc gia TP. HCM chuyên ngành Khoa học máy tính.

Hướng nghiên cứu: khai thác dữ liệu đa phương tiện, thị giác máy tính và máy học.

Email: phucnq@uit.edu.vn