KHAI THÁC kết hợp Đặc trưng THỊ GIÁC, âm thanh VÀ VĂN BẢN nhằm cải thiện chất lượng gom cụm kết quả tìm kiếm video

***ThS Nguyễn Quang Phúc, ThS Trần Duy Thanh***

**TÓM TẮT**

*Bài viết này nhằm mục đích mở rộng các nghiên cứu trước đây của chúng tôi về gom cụm kết quả tìm kiếm video trên web đã được trình bày trong [5], [6], [7]. Để tìm kiếm video, người dùng thường sử dụng các hệ thống tìm kiếm video trực tuyến như YouTube, Google Video. Tuy nhiên, kết quả tìm kiếm được trả về của các hệ thống này có thể bao gồm nhiều video thuộc các danh mục chủ đề khác nhau làm cho người dùng gặp khó khăn trong việc xác định video mà họ quan tâm. Do đó, gom cụm kết quả tìm kiếm video trên web là cần thiết để cải thiện hiệu quả tìm kiếm. Ý tưởng chính của bài viết này dựa trên việc phân tích và kết hợp các đặc trưng được trích xuất từ video để tìm ra bộ đặc trưng phù hợp nhằm cải thiện chất lượng gom cụm kết quả tìm kiếm video.*

**Từ khóa:** biểu diễn video, gom cụm web video, kết hợp đa đặc trưng.

1. **GIỚI THIỆU**

Thông thường, để tìm kiếm video người dùng 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. Số lượng video sẽ trải rộng trên nhiều chủ đề khác nhau tùy thuộc vào cách diễn đạt từ khóa tìm kiếm của người dùng. Điều này khiến người dùng mất thời gian trong việc xác định video cần tìm. 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.

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



**...**

**Đầu vào**  
(danh sách video trả về từ kết quả tìm kiếm trên Web)

**...**



**Cụm 1**



**Cụm 2**



**Cụm 3**

**Đầu ra**  
(các cụm video theo từng chủ đề)

*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”

Để gom cụm video, một vấn đề được đặt ra là bằng cách nào để *ước lượng độ tương tự giữa các video*. Kết quả gom cụm video được tạo ra bằng cách áp dụng các giải thuật 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. 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 (textual).

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 [8]. 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 [1], [4]. Trong [1], 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 Bounded Coordinate System (BCS) 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 [4], 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 Vector Space Model (VSM) 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 chú trọng 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 [5], [6], [7]. Trong bài báo này, chúng tôi tiếp tục phát triển hướng nghiên cứu gom cụm kết quả tìm kiếm video của chúng tôi 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*.

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.

1. PHƯƠNG PHÁP ĐỀ XUẤT
2. **Mô hình đề xuất**

Việc khai thác đặc trưng thị giác sẽ giúp gom các video có 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 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. Mô hình đề xuất được thể hiện ở Hình 2.

**Thuật toán gom cụm**

**Tập**

**dữ liệu**

**video**

**Đặc trưng thị giác**

**Đặc trưng âm thanh**

**Độ tương tự**

**Độ tương tự**

**Độ tương tự kết hợp**

**Kết quả gom cụm video**

**Thông tin văn bản**

**Độ tương tự**

*Hình 2.* Mô hình kết hợp đa đặc trưng cho bài toán gom cụm kết quả tìm kiếm video

1. **Biểu diễn và tính độ tương tự video theo đặc trưng thị giác**

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 3 bao gồm các bước sau:

* *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 keypoint được rút trích, một bộ mô tả 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.

**(iii) Gom cụm đặc trưng**

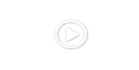
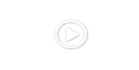
**(ii) Rút trích đặc trưng  
thị giác**

**(i) Rút trích frame**

**(iv) Biểu diễn video**

Tập các frame

Tập dữ liệu video



**...**

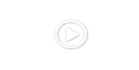
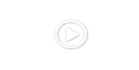
Tập các visual word (từ điển)



**...**

**...**

Tập các đặc trưng



*Hình 3.* Quá trình biểu diễn video theo đặc trưng thị giác

* *Gom cụm các keypoint, xây dựng “visual vocabulary[[1]](#footnote-1)”*: 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.

1. **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 4.

**...**

**Lấy mẫu**

**gom cụm**

**...**

**...**

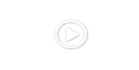
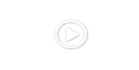
**...**

Tập các véc tơ đặc trưng

Từ điển

Tập dữ liệu video

**...**



**Trích xuất audio**

**đặc trưng**

**Rút trích**

**đặc trưng**

**...**

Tập audio



*Hình 4.* Sơ đồ mô tả quá trình tạo từ điển biểu diễn video dựa trên đặc trưng âm thanh

1. **Tính độ tương tự video dựa trên thông tin văn bản đi kèm**

Đối với thông tin văn bản, chúng tôi sử dụng bộ từ điển các từ đồng nghĩa WordNet[[2]](#footnote-2) 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 5.

Tiêu đề + Mô tả

(Title + Description)

Độ tương tự ngữ nghĩa

Thông tin văn bản

**WordNet**

Độ tương tự

Tập

dữ liệu

video

Các thẻ từ khóa

(Tags)

Độ tương tự ngữ nghĩa

*Hình 5.* Quá trình tính độ tương tự video dựa trên thông tin văn bản đi kèm sử dụng từ điển WordNet

Ở mô hình trên, 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. 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.

1. **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 gom cụm dữ liệu video**

**Input**: Danh sách 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 dựa trên đặc trưng thị giác theo [6].
2. Biểu diễn mỗi video thành vector dựa trên đặc trưng âm thanh.
3. Biểu diễn thông tin văn bản đi kèm mỗi video thành tập hợp các từ theo [6].

//Tính độ tương tự giữa các video

1. Tính độ tương tự giữa các vector theo công thức tính khoảng cách cosine.
2. Tính độ tương tự giữa các vector theo công thức tính khoảng cách cosine.
3. Tính độ tương tự ngữ nghĩa giữa các tập hợp từ dựa trên từ điển WordNet [6].
4. Tính độ tương tự kết hợp đa đặc trưng giữa video theo công thức:

với

//Gom cụm video

1. 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**

1. THỰC NGHIỆM
2. **Bộ dữ liệu video**

Chúng tôi tải dữ liệu video thực từ kết quả tìm kiếm video trên YouTube bởi phần mềm mã nguồn mở TubeKit[[3]](#footnote-3). 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 1580 video của 18 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. Jaguar | 86 | 4 | 5.1 |
| 5. Java | 87 | 4 | 7.2 |
| 6. Jupiter | 82 | 4 | 5.1 |
| 7. Leopard | 95 | 5 | 6.4 |
| 8. Lion | 89 | 4 | 6.2 |
| 9. Lotus | 91 | 6 | 5.5 |
| 10. Mustang | 83 | 5 | 5.6 |
| 11. Ocean | 90 | 7 | 5.5 |
| 12. Panda | 97 | 5 | 5.8 |
| 13. Pluto | 85 | 7 | 8.8 |
| 14. Python | 85 | 4 | 5.1 |
| 15. Tiger | 81 | 4 | 4.3 |
| 16. Venus | 89 | 7 | 6.9 |
| 17. Viper | 87 | 5 | 4.5 |
| 18. Zebra | 99 | 7 | 6.0 |

1. **Phương pháp đánh giá**

Kết quả 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 *Cj* với *j* = *1*,...,*k* và thuật toán gom cụm *n* video vào *k* cụm *Pi* với *i* = *1*,...,*k*. *Entropy* đánh giá chất lượng gom cụm được tính theo công thức sau:

(1)

trong đó, *ni* là số video trong cụm *Pi*, *nij* là số video trong cụm *Pi* thuộc chủ đề *Cj*, *n* là tổng số video trong tất cả các cụm.

Trường hợp lý tưởng là mỗi cụm chỉ chứa vi deo 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*:

(2)

1. **Cài đặt thực nghiệm**

Nhằm phân tích 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 các thí nghiệm sau:

* Gom cụm video theo từng đặc trưng riêng lẻ.
* *V (Visual)*: gom cụm video dựa trên đặc trưng thị giác.
* *A (Audio)*: gom cụm video dựa trên đặc trưng âm thanh.
* *T (Textual)*: gom cụm video dựa trên thông tin văn bản đi kèm.
* 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, ưu thế của các đặc trưng được cân bằng nhau.
* *VA (Visual + Audio)*: gom cụm video dựa trên đặc trưng thị giác và đặc trưng âm thanh.
* *VT (Visual + Textual)*: gom cụm video dựa trên đặc trưng thị giác và thông tin văn bản đi kèm video.
* *AT (Audio + Textual)*: gom cụm video dựa trên đặc trưng âm thanh và thông tin văn bản đi kèm video.
* *VAT (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 tuyến tính có trọng số theo công thức được trình bày trong giải thuật đề xuất.
* *VAT\* (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 phương pháp cài đặt biểu diễn video ứng với từng loại đặc trưng:

*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) [2], [3] 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*, tương tự như quá trình biểu diễn video theo đặc trưng thị giác, chúng tôi sử dụng Mel-Frequency Cepstral Coefficients (MFCC) [9] như là một loại đặc trưng âm thanh được trích xuất từ video. Sau khi đặc trưng âm thanh (biểu diễn dạng tập các véc tơ) được trích xuất từ tập dữ liệu video, 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.

*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].

Để thực hiện gom cụm video, chúng tôi thử nghiệm 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.

1. **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 VT [6] cho kết quả gom cụm video tốt hơn (đạt giá trị Entropy thấp hơn) phương pháp VT [1], VT [4] 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 [6] 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.

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truy vấn |  |  | |  | | Entropy | | | | | | | |  |
| V [6] | | A | | T [6] | | VA | VT [6] | VT [1] | VT [4] | AT | VAT | VAT\* | |
| 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. Jaguar | 0.4713 | | 0.4465 | | 0.5251 | | **0.3518** | 0.3681 | 0.4402 | 0.4240 | 0.4237 | 0.3723 | **0.2146** | |
| 5. Java | 0.2844 | | 0.3266 | | 0.5149 | | **0.1584** | 0.2083 | 0.3525 | 0.2322 | 0.3529 | 0.1187 | **0.0570** | |
| 6. Jupiter | 0.3300 | | 0.4182 | | 0.4875 | | **0.2538** | 0.2701 | 0.3992 | 0.3080 | 0.4467 | 0.2891 | **0.1883** | |
| 7. Leopard | 0.4160 | | 0.5057 | | 0.5610 | | **0.2252** | 0.2686 | 0.3767 | 0.3234 | 0.5320 | 0.2487 | **0.1029** | |
| 8. Lion | 0.5412 | | 0.5030 | | 0.5570 | | **0.4660** | 0.4828 | 0.5311 | 0.5113 | 0.4893 | 0.4880 | **0.3126** | |
| 9. Lotus | 0.5096 | | 0.5018 | | 0.6525 | | **0.3423** | 0.3751 | 0.4857 | 0.4426 | 0.5789 | 0.3894 | **0.1431** | |
| 10. Mustang | 0.5500 | | 0.5203 | | 0.5887 | | **0.4347** | 0.4828 | 0.5233 | 0.5111 | 0.5137 | 0.4662 | **0.1869** | |
| 11. Ocean | 0.5716 | | 0.5351 | | 0.6559 | | **0.4622** | 0.5207 | 0.5766 | 0.5421 | 0.5708 | 0.4971 | **0.3064** | |
| 12. Panda | 0.4066 | | 0.5106 | | 0.6058 | | **0.2693** | 0.2803 | 0.4181 | 0.3321 | 0.5396 | 0.3069 | **0.2082** | |
| 13. Pluto | 0.3546 | | 0.3166 | | 0.5026 | | **0.2887** | 0.3396 | 0.3715 | 0.3402 | 0.4191 | 0.3223 | **0.1773** | |
| 14. Python | 0.3320 | | 0.4048 | | 0.5246 | | **0.2023** | 0.2352 | 0.3685 | 0.2545 | 0.4521 | 0.2467 | **0.1068** | |
| 15. Tiger | 0.4181 | | 0.4147 | | 0.5460 | | **0.3301** | 0.3682 | 0.4120 | 0.3811 | 0.4237 | 0.3561 | **0.2185** | |
| 16. Venus | 0.5598 | | 0.5001 | | 0.6751 | | **0.4336** | 0.4813 | 0.5426 | 0.5069 | 0.4813 | 0.4112 | **0.2072** | |
| 17. Viper | 0.5415 | | 0.5018 | | 0.5927 | | **0.3729** | 0.4301 | 0.5560 | 0.4842 | 0.5356 | 0.4160 | **0.2527** | |
| 18. Zebra | 0.6405 | | 0.5963 | | 0.6863 | | **0.5156** | 0.5598 | 0.6302 | 0.6098 | 0.6532 | 0.4992 | **0.3094** | |
| **Trung bình** | 0.4741 | | 0.4691 | | 0.5713 | | **0.3552** | 0.3945 | 0.4749 | 0.4291 | 0.4900 | 0.3738 | **0.2174** | |

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 (VA) cho kết quả gom cụm tốt hơn so với các phương pháp kết hợp đặc trưng thị giác với thông tin văn bản (VT) hay đặc trưng âm thanh với thông tin văn bản (AT). 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.

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 (VAT), (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 (VAT\*).

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Truy vấn |  |  | |  | | Purity | | | | | | | |  |
| V [6] | | A | | T [6] | | VA | VT [6] | VT [1] | VT [4] | AT | VAT | VAT\* | |
| 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. Jaguar | 0.4419 | | 0.5349 | | 0.4651 | | 0.6512 | 0.6628 | 0.5698 | 0.5930 | 0.6279 | 0.6047 | **0.8333** | |
| 5. Java | 0.7126 | | 0.6897 | | 0.4483 | | **0.8621** | 0.8276 | 0.6552 | 0.7586 | 0.6437 | 0.9195 | **0.9529** | |
| 6. Jupiter | 0.6543 | | 0.5802 | | 0.4938 | | **0.7407** | 0.7037 | 0.6049 | 0.6790 | 0.5432 | 0.6790 | **0.8462** | |
| 7. Leopard | 0.6316 | | 0.5474 | | 0.4842 | | **0.8211** | 0.7474 | 0.6632 | 0.6947 | 0.4526 | 0.7895 | **0.9053** | |
| 8. Lion | 0.4270 | | 0.4944 | | 0.3820 | | **0.5169** | 0.4831 | 0.4157 | 0.4607 | 0.4494 | 0.5056 | **0.7528** | |
| 9. Lotus | 0.4835 | | 0.4835 | | 0.3626 | | **0.6703** | 0.6264 | 0.5275 | 0.5275 | 0.4176 | 0.6374 | **0.8681** | |
| 10. Mustang | 0.4578 | | 0.4940 | | 0.4096 | | **0.6386** | 0.5663 | 0.4819 | 0.5060 | 0.5060 | 0.5542 | **0.8675** | |
| 11. Ocean | 0.4556 | | 0.4778 | | 0.4000 | | **0.5667** | 0.5111 | 0.4667 | 0.5000 | 0.4222 | 0.5333 | **0.7444** | |
| 12. Panda | 0.5567 | | 0.4124 | | 0.3711 | | **0.7423** | 0.6804 | 0.4948 | 0.6289 | 0.4536 | 0.7010 | **0.8041** | |
| 13. Pluto | 0.6706 | | 0.6824 | | 0.5647 | | **0.7294** | 0.6941 | 0.6706 | 0.6824 | 0.6118 | 0.7059 | **0.8171** | |
| 14. Python | 0.6786 | | 0.6235 | | 0.4471 | | **0.7765** | 0.7294 | 0.6471 | 0.7059 | 0.5529 | 0.7176 | **0.9294** | |
| 15. Tiger | 0.5062 | | 0.5309 | | 0.3827 | | **0.6420** | 0.6049 | 0.5185 | 0.5556 | 0.5062 | 0.5926 | **0.7654** | |
| 16. Venus | 0.4607 | | 0.5393 | | 0.3483 | | **0.6404** | 0.5618 | 0.4494 | 0.5393 | 0.5281 | 0.5955 | **0.8315** | |
| 17. Viper | 0.4368 | | 0.4943 | | 0.3908 | | **0.6667** | 0.6092 | 0.4368 | 0.5057 | 0.4598 | 0.6092 | **0.7586** | |
| 18. Zebra | 0.3737 | | 0.4242 | | 0.3535 | | **0.5152** | 0.4949 | 0.4040 | 0.4040 | 0.3232 | 0.5051 | **0.7857** | |
| **Trung bình** | 0.5138 | | 0.5267 | | 0.4225 | | **0.6588** | 0.6147 | 0.5166 | 0.5652 | 0.5016 | 0.6349 | **0.8098** | |

Trong phương pháp VAT, 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 VA cho kết quả gom cụm tốt hơn so với phương pháp VAT khi vai trò của các đặc trưng được cân bằng.

Với phương pháp VAT\*, 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 (ứng với đặc trưng thị giác), (ứng với đặc trưng âm thanh), (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 VAT\* 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.

1. 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.

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**

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.

D. G. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.

D. G. Lowe, “Object Recognition from Local Scale-Invariant Features,” *International Conference on Computer Vision*, vol. 2, pp. 1150-1157, 1999.

H. Huang, Y. Lu, F. Zhang, S. Sun, “A Multi-modal Clustering Method for Web Videos,” *Trustworthy Computing and Services*, pp. 163-169, 2013.

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.

P. Q. Nguyen, A. T. Nguyen-Thi, T. D. Ngo, T. A. H Nguyen, “Using Textual Semantic Similarity to Improve Clustering Quality of Web Video Search Results,” in *Proceedings of the 7th International Conference on Knowledge and Systems Engineering (KSE)*, pp. 156-161, 2015.

P. Q. Nguyen, T. Do, A. T. Nguyen-Thi, T. D. Ngo, D. D. Le, T. A. H. Nguyen, “Clustering Web Video Search Results with Convolutional Neural Networks,” in *Proceedings of the 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS),* pp. 135-140, 2016.

S. Liu, M. Zhu, Q. Zheng, “Mining similarities for clustering web video clips,” *CSSE*, vol. 4, pp. 759-762, 2008.

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*, vol. 27, no. 1, pp. 105-141, 2005.

Y. H. Li, Z. Bandar, D. McLean, “An approach for measuring semantic similarity using multiple information sources,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, 2003.

1. 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”. [↑](#footnote-ref-1)
2. <http://wordnet.princeton.edu> [↑](#footnote-ref-2)
3. <www.tubekit.org> [↑](#footnote-ref-3)