**Video Search with CLIP and Interactive Text Query Reformulation**

**Abs :** Ngày nay, các mô hình dựa trên học sâu như CLIP cho phép thiết kế đơn giản các hệ thống tìm kiếm video đa phương thức có khả năng giải quyết nhiều nhiệm vụ được coi là có tính thách thức cao vài năm trước. Trong bài viết này, chúng tôi phân tích cách tiếp cận tìm kiếm dựa trên CLIP tập trung vào các tình huống mà người dùng không thể tìm thấy các truy vấn văn bản phù hợp để mô tả các phân đoạn video được tìm kiếm. Cách tiếp cận dựa trên các đề xuất của các lớp cho các tập kết quả trung gian được hiển thị và do đó cho phép người dùng nhận ra các từ và ý tưởng còn thiếu để mô tả các khung hình video. Cách tiếp cận này được hỗ trợ với một nghiên cứu sơ bộ cho thấy tiềm năng của phương pháp. Dựa trên kết quả, chúng tôi mở rộng hệ thống tìm kiếm mục đã biết được tôn trọng cho Cuộc trình diễn trình duyệt video, nơi các nhiệm vụ tìm kiếm mục đã biết bằng hình ảnh thách thức hơn được lên kế hoạch

1/ Intro:

Cuộc thi Video Browser Showdown (VBS) [6,9] đại diện cho một sự kiện quốc tế được tôn trọng, trong đó các nhóm khác nhau cố gắng giải quyết các nhiệm vụ đầy thách thức một cách tương tác trên một bộ sưu tập video không tầm thường (hiện đang sử dụng V3C [11] và máy chủ DRES [10]) . Cuộc thi tập trung vào các nhiệm vụ tìm kiếm mục đã biết và tìm kiếm đặc biệt [4,5] không thể tự động giải quyết bằng một truy vấn văn bản đơn nhất định (ít nhất là chưa). Vài năm gần đây đã chỉ ra rằng chìa khóa thành công là thiết kế các giao diện tìm kiếm tương tác trên các mô hình học sâu mới nhất. Ví dụ: ba hệ thống tìm kiếm mục đã biết hàng đầu tại VBS 2022 [1–3] đã sử dụng các mô hình dựa trên CLIP [8] cho phép tìm kiếm văn bản dạng tự do hiệu quả và thành phần kiểm tra tập hợp kết quả cung cấp các tùy chọn truy vấn bổ sung.

Sự kiện VBS 2023 giới thiệu một số thay đổi. Đầu tiên, các phân đoạn mục tiêu tìm kiếm vật phẩm đã biết bằng hình ảnh có thể ngắn hơn và do đó, các nhóm sẽ có ít tùy chọn hơn để chọn các đối tượng đặc biệt cao để truy vấn. Thật vậy, với CLIP và kết hợp truy vấn thời gian, một số nhóm đã đạt được thời gian ấn tượng để giải quyết các nhiệm vụ KIS trực quan tại VBS 2022. Thứ hai, một bộ sưu tập dành riêng cho miền sẽ được sử dụng cho các nhiệm vụ tìm kiếm trong một cụm đồng nhất. Cụ thể, một bộ sưu tập video hàng hải [12] sẽ được sử dụng khi các từ khóa thường được biết đến như cá, rạn san hô hoặc bơi lội không thể lọc đầy đủ tập kết quả. Trong bài báo này, chúng tôi trình bày một cách tiếp cận giải quyết các thách thức mới được đưa ra, nhưng vẫn giữ cho hệ thống đơn giản cho người dùng. Hãy nhấn mạnh rằng các phân đoạn mục tiêu KIS trực quan ngắn hơn và các video cụ thể theo miền sẽ làm tăng nhu cầu về đặc tả truy vấn văn bản. Do đó, chúng tôi tập trung vào việc xây dựng lại truy vấn tương tác với một hệ thống cung cấp các gợi ý về nhãn lớp. Cách tiếp cận này đã được chứng minh là có triển vọng [7] với các cách tiếp cận phân loại dựa trên mạng sâu “cổ điển” như GoogleNet. Đối với VBS 2023, chúng tôi dự định thử nghiệm phương pháp này bằng cách sử dụng mô hình CLIP thể hiện công nghệ tiên tiến nhất trong việc nhúng chung cũng như phân loại bằng không. Giả định của chúng tôi là các lớp được đề xuất bởi phân loại zero-shot CLIP có thể giúp nâng cao các truy vấn văn bản cho CLIP.

2/ Interactive Text Query Reformulation with CLIP:

Định dạng lại truy vấn văn bản là một chiến lược tìm kiếm tiêu chuẩn, trong đó người dùng thay đổi truy vấn văn bản nếu tập kết quả không hiển thị các mục có triển vọng. Người dùng dựa vào trực giác, lựa chọn khái niệm cũng như kỹ năng ngôn ngữ để chọn một mục mới hoặc mô tả cùng một mục bằng các từ khác nhau. Tuy nhiên, có thể có những tình huống mà người dùng không thể tìm thấy từ thích hợp. Do đó, các đề xuất có sự trợ giúp của máy tính có thể đại diện cho phản hồi quan trọng đối với người dùng đang cố gắng mô tả các mục đã tìm kiếm. Đồng thời, người dùng có thể hiểu rõ hơn (và tìm hiểu) cách mạng liên kết dữ liệu trực quan với văn bản. Ý tưởng cơ bản là hiển thị kết quả nâng cao của truy vấn văn bản ban đầu, trong đó đối với mỗi hình ảnh được hiển thị, hệ thống cũng cung cấp các từ khóa được gán cho hình ảnh. Hệ thống cũng có thể hiển thị tổng hợp các từ khóa thường xuyên nhất được chỉ định cho các hình ảnh top-k trong tập hợp kết quả. Xin lưu ý rằng phương pháp này là một giải pháp thay thế cho trình duyệt dựa trên kNN. Thay vì sử dụng tính năng nhúng của hình ảnh, chỉ (các) nhãn thực thể đã chọn được lấy từ hình ảnh mẫu.

Để cung cấp ước tính hiệu suất định lượng sơ bộ, chúng tôi đã thực hiện một nghiên cứu (với bốn người dùng) cho thấy tiềm năng của phương pháp này. Đầu tiên, để có được từ khóa cho từng hình ảnh cơ sở dữ liệu (nghĩa là khung hình video đã chọn), mô hình CLIP đã được sử dụng làm bộ phân loại zero-shot cho danh sách các lớp được chọn trước (các bộ khác nhau cho V3C và bộ dữ liệu hàng hải). Dựa trên kết quả phân loại, các lớp hàng đầu được lưu trữ cho mỗi hình ảnh. Sau đó, nghiên cứu được thực hiện với một công cụ dựa trên web đơn giản, nơi người dùng nhìn thấy một hình ảnh mục tiêu đã biết được chọn ngẫu nhiên (tức là mục tiêu tìm kiếm, ví dụ cần thông tin) và có thể viết một truy vấn văn bản để tìm thấy nó. Tập hợp kết quả của mỗi truy vấn được nâng cao với các từ khóa (lớp) được đề xuất thu được trong bước phân loại zero-shot ngoại tuyến. Cải tiến này được minh họa trong Hình 1 cho bộ dữ liệu video hàng hải. Người dùng có thể nhấp vào các từ khóa được đề xuất và do đó cập nhật truy vấn văn bản một cách tương tác (cũng cho phép chỉnh sửa thủ công, xem Bảng 1). Sau khi truy vấn được cập nhật, bạn có thể nhận được một tập kết quả mới. Hệ thống theo dõi vị trí của hình ảnh mục tiêu được tìm kiếm và ghi lại vị trí/thứ hạng của nó cho mỗi truy vấn. Đối với mục đích của nghiên cứu, chỉ cho phép một lần lặp lại các công thức tái tạo và tập hợp các hình ảnh mục tiêu đã biết đã chọn được giới hạn ở các hình ảnh có cá hoặc rạn san hô. Nghiên cứu đã sử dụng tập dữ liệu hình ảnh với 22K khung hình được trích xuất.

A screenshot of a sea creature

Description automatically generated

Hình 1: Minh họa tập hợp kết quả nâng cao với các lớp từ phân loại CLIP không ảnh hưởng. Màu sắc biểu thị tần suất xuất hiện của các loại (trong 10 loại hàng đầu cho mỗi hình ảnh), trong đó màu xanh lục cho biết loại này rất hiếm. (Hình màu trên mạng).

Bảng 1: Hai ví dụ, trong đó việc định dạng lại truy vấn văn bản tương tác đã giúp cải thiện đáng kể thứ hạng của hình ảnh mục tiêu được tìm kiếm. Chúng tôi có thể quan sát thấy sự thay đổi từ các truy vấn chung sang các mô tả giống như chuyên gia.

A white box with black text

Description automatically generated

Dựa trên vị trí/xếp hạng đã ghi của khung mục tiêu trong mỗi lần lặp lại, Hình 2 cho thấy hiệu quả của tìm kiếm văn bản và định dạng lại truy vấn đối với các nhiệm vụ khó khăn hơn. Chúng tôi có thể quan sát thấy rằng một số mục tiêu đã dễ dàng tìm thấy với truy vấn đầu tiên, trong khi trung bình có sự cải thiện rõ ràng trong bước cải tổ đầu tiên trong các tìm kiếm với hai lần lặp lại. Trong một số trường hợp, việc xây dựng lại có thể phản tác dụng (xem một số giá trị âm trong boxplot RI1–RI2). Tuy nhiên, các đề xuất thường giúp người dùng cải thiện truy vấn văn bản cho cùng một mục tiêu và cũng dần dần “dạy” cho người dùng kiến ​​thức miền quan trọng (nghĩa là sự ăn khớp tương ứng giữa các mẫu trực quan và nhãn lớp).

Xin lưu ý rằng không đảm bảo rằng CLIP gán nhãn lớp chính xác, đặc biệt là trong các bộ sưu tập cụ thể của miền. Tuy nhiên, miễn là CLIP gán nhãn sai (ví dụ: loài cá) một cách nhất quán cho các đặc điểm trực quan, việc sử dụng chúng để tìm kiếm trong không gian nối vẫn có thể hiệu quả

A graph with numbers and a line

Description automatically generated

Hình 2 : Thứ hạng của hình ảnh mục tiêu được tìm kiếm cho 66 tìm kiếm chỉ với một lần lặp RI1/1 và 122 tìm kiếm với hai lần lặp RI1/2, RI2/2. Sự khác biệt giữa thứ hạng của lần lặp đầu tiên và lần thứ hai (trong các tìm kiếm có hai lần lặp lại) được hiển thị dưới dạng RI1–RI2

**3/ VBS participation:**

Phương pháp nói trên đủ chung để được tích hợp vào bất kỳ hệ thống tìm kiếm video tương tác tìm kiếm mục đã biết nào hỗ trợ tìm kiếm dựa trên CLIP. Năm nay, chúng tôi dự định tiếp tục với CVHunter [3] triển khai các tính năng duyệt hữu ích tiêu chuẩn (cuộn tập hợp kết quả, kiểm tra video, xem trước video, v.v.) trên một số phương pháp xếp hạng dựa trên CLIP. Do các cảnh được trình bày từ V3C sẽ ngắn hơn và bộ dữ liệu video hàng hải chứa các video quay một lần, nên chúng tôi không có kế hoạch tích hợp trình xếp hạng nhận biết ngữ cảnh chung1 (để giữ thiết kế tối thiểu). Tuy nhiên, một trường hợp đặc biệt của xếp hạng nhận biết ngữ cảnh, truy vấn thời gian, vẫn có thể hữu ích trong trường hợp phân đoạn ngắn chứa hai cảnh riêng biệt. Nói cách khác, CVHunter sẽ vẫn cho phép xem xét ngữ cảnh của các khung được tìm kiếm. Công cụ này cũng sẽ giữ phản hồi về mức độ phù hợp của Bayesian, bao gồm cả biến thể theo thời gian của nó.

Tính năng gợi ý lớp học mới sẽ được tích hợp vào hệ thống theo hai cách. Đầu tiên, một bản tóm tắt các nhãn lớp được đề xuất từ ​​các kết quả hàng đầu của truy vấn có thể được trình bày trong bảng văn bản có sẵn (ví dụ: chú thích biểu mẫu ứng dụng). Thứ hai, các nhãn hàng đầu của mỗi hình ảnh có thể được hiển thị bên dưới hình ảnh trong bảng thông tin, cho phép sử dụng văn bản đơn giản bằng cách nhấp chuột. Hơn nữa, tất cả các nhãn lớp có sẵn cho một hình ảnh có thể được hiển thị khi cửa sổ chi tiết được mở

**4/ Conclusions:**

Chúng tôi đã phân tích một phương pháp tìm kiếm tương tác cho VBS cho phép người dùng chọn các nhãn lớp có triển vọng từ các đề xuất được cung cấp. Theo nghiên cứu được thực hiện, phương pháp này mang lại hiệu suất đầy hứa hẹn trong các tình huống mà người dùng gặp khó khăn hơn trong việc tìm các truy vấn văn bản phù hợp. Cách tiếp cận này được tích hợp vào hệ thống được thiết kế cho cuộc thi Video Browser Showdown, cung cấp một tùy chọn tìm kiếm tương tác thú vị khác

**Perfect Match in Video Retrieval**

Abs: Bài viết này trình bày phiên bản đầu tiên của hệ thống tìm kiếm video Perfect Match của chúng tôi cho cuộc thi Video Browser Showdown 2023. Hệ thống lập chỉ mục các video từ bộ dữ liệu video V3C lớn và tự động lấy các bộ mô tả nội dung trực quan. Hơn nữa, nó cung cấp giao diện người dùng (UI) tìm kiếm video tương tác, thực hiện các phương pháp tiếp cận từ miền đề xuất dựa trên phê bình, để cho phép người dùng tìm thấy phân đoạn video mong muốn nhanh nhất có thể

1/ Intro:

Việc xác định video nào phù hợp với yêu cầu đang trở thành một nhiệm vụ ngày càng phức tạp với số lượng video có sẵn ngày càng tăng. Do đó, việc truy xuất hiệu quả các video từ danh mục lớn là một thách thức quan trọng.

Video Browser Showdown (VBS) là một cuộc thi thường niên nhằm khuyến khích các nhà nghiên cứu trong độ tuổi phát triển các phương pháp sáng tạo để giải quyết thách thức này. Trong sự kiện, hai loại nhiệm vụ được xác định [9]. Đầu tiên, trong nhiệm vụ Ad-hoc Video search (AVS), mô tả bằng văn bản về nội dung video được đưa ra, với mục đích tìm càng nhiều cảnh càng tốt, trong một khoảng thời gian giới hạn, phù hợp với mô tả. Thứ hai, trong nhiệm vụ Known item search (KIS), đoạn video được yêu cầu đã được biết và người tham gia cần tìm đoạn video đó càng nhanh càng tốt.

Đối với VBS 2023, hai bộ dữ liệu, V3C1 và V3C2, từ Bộ sưu tập Vimeo Creative Commons được sử dụng [11]. Chúng bao gồm 17235 video với tổng thời lượng khoảng 2300 giờ. Ngoài ra, một bộ dữ liệu video hàng hải đặc biệt với các video dưới nước và lặn biển sẽ được sử dụng trong một thử thách bổ sung [13].

Để cho phép tìm kiếm bên trong bộ sưu tập video lớn này, chúng tôi đã phát triển một hệ thống tự động lập chỉ mục các video từ tập dữ liệu và suy ra các bộ mô tả nội dung trực quan từ các khung hình chính của video. Ứng dụng tìm kiếm video Perfect Match của chúng tôi nhận thông tin đầu vào của người dùng theo kiểu tương tác thông qua UI và sử dụng thông tin đó để soạn truy vấn cho các video được lập chỉ mục, bao gồm cả siêu dữ liệu được suy luận, nhằm tìm kết quả phù hợp hoàn hảo, tức là phân đoạn video giải quyết nhiệm vụ .

Trong ứng dụng của chúng tôi, các yêu cầu được thu thập theo cách tương tác bằng cách áp dụng các phương pháp tiếp cận dựa trên phê bình, tạo thành một loại hệ thống đề xuất dựa trên tri thức đặc biệt [5], thay vì đăng một truy vấn tìm kiếm đơn lẻ. Phần còn lại của bài viết này được tổ chức như sau. Đầu tiên, chúng tôi đưa ra phần giới thiệu về đề xuất dựa trên phê bình trong Phần. 2. Sau đó, hệ thống được trình bày chi tiết trong Phần. 3, tiếp theo là phần kết luận trong Phần. 4

2/ Critiquing

Knowledge-Based Recommender Systems (Hệ thống đề xuất dựa trên tri thức) tạo thành một loại đề xuất đặc biệt gợi ý các mặt hàng dựa trên kiến ​​thức rõ ràng về danh mục mặt hàng và sở thích của người dùng. Trái ngược với các hệ thống đề xuất khác dựa trên nội dung hoặc tương tác của người dùng trong quá khứ, các tùy chọn không được thu thập trước mà được xác định trong quy trình đề xuất đàm thoại [5].

Critiquing-Based Recommendations( Đề xuất dựa trên phê bình) [3,4] tạo thành một loại phụ của hệ thống đề xuất dựa trên tri thức, sử dụng các biện pháp tương tự khác nhau để truy xuất các mục tương tự nhất với yêu cầu của người dùng nhất định. Một đặc điểm của các hệ thống đó là, thay vì chỉ định trước tất cả các yêu cầu theo cách giống như truy vấn, các yêu cầu được chỉ định theo kiểu dựa trên điều hướng. Điều hướng (cũng là cách tiếp cận dựa trên trình duyệt) hỗ trợ các tình huống mà người dùng có thể không biết trước chính xác những gì họ đang tìm kiếm. Họ được hướng dẫn qua quy trình để cải thiện quy trình tìm kiếm mặt hàng họ cần.

Thông thường, người dùng chỉ định một số yêu cầu ban đầu thô và nhận được đề xuất đầu tiên về các mục đáp ứng yêu cầu. Người dùng xem xét mục được đề xuất và chỉ định một yêu cầu bổ sung chưa được đáp ứng, tức là phê bình. Trong phần sau, hệ thống đề xuất một giải pháp thay thế mới mà người dùng nên xem xét. Các bước này được lặp lại cho đến khi người dùng hài lòng với đề xuất hoặc không còn lựa chọn thay thế nào khác. Việc tích hợp các khía cạnh dựa trên phê bình đó cho mục đích tìm kiếm video được giải thích trong phần sau

3/ Perfect Match – System overview:

Hệ thống Kết hợp hoàn hảo của chúng tôi bao gồm hai giai đoạn xử lý với các tư thế mục đích khác nhau. Đầu tiên, giai đoạn ngoại tuyến được sử dụng để lập chỉ mục các phân đoạn video của tập dữ liệu V3C và tự động trích xuất siêu dữ liệu nội dung video. Thứ hai, trong giai đoạn trực tuyến, các truy vấn tìm kiếm của người dùng được tổng hợp và gửi tới API, API này sẽ xếp hạng các khung dựa trên truy vấn và siêu dữ liệu video được lập chỉ mục. Việc tách quá trình xử lý thành hai giai đoạn là một quyết định thiết kế cần thiết để đạt được hiệu suất tốt khi sử dụng giao diện người dùng để tìm kiếm trong giai đoạn trực tuyến. Lập chỉ mục video và trích xuất siêu dữ liệu là một nhiệm vụ rất tốn thời gian và phần cứng. Thực hiện nó theo yêu cầu trong giai đoạn trực tuyến sẽ không khả thi. Do đó, giai đoạn trực tuyến chỉ khai thác kết quả của quy trình này bằng cách truy vấn siêu dữ liệu liên tục, đủ nhanh để mang lại trải nghiệm tốt cho người dùng. Trong các phần sau, các chi tiết của các giai đoạn được giải thích thêm.

3.1/ Video Indexing:

Bước đầu tiên trong phần lập chỉ mục video là tạo tham chiếu cơ sở dữ liệu cho từng video trong bộ dữ liệu V3C. Một id duy nhất được chỉ định cho mỗi video. Sau đó, FFmpeg1 được sử dụng để trích xuất và lập chỉ mục các khung hình đã chọn của video với một id duy nhất.

Vì lý do hiệu suất và dự phòng, không thể trích xuất các bộ mô tả nội dung hình ảnh cho mọi khung hình trong video. Do đó, chúng tôi quyết định lấy mẫu các khung hình từ mỗi video theo định kỳ và sử dụng các khung hình đó để phân tích. Mỗi khung được phân tích cho nội dung trực quan và văn bản in. Kết quả phân tích này được lưu trữ trong cơ sở dữ liệu. Đối với mỗi khung, chúng tôi lưu trữ thông tin mà các thực thể đã được nhận dạng, tần suất chúng được nhận dạng và độ tin cậy để nhận dạng được cung cấp bởi các mô hình

Để phát hiện khái niệm, AlexNet [7] được sử dụng, đây là mạng (CNN) được đào tạo trước, để xác định nội dung của khung. Nó được train trên bộ dữ liệu ILSVRC2010 [12]. Model này phân loại hình ảnh với 1000 lớp khác nhau mô tả nội dung của chúng. Hơn nữa, chúng tôi áp dụng thuật toán phát hiện đối tượng YOLOv5 [6] để xác định các đối tượng bên trong khung. Nó được đào tạo để phân loại 80 đối tượng khác nhau của bộ dữ liệu COCO.

Ngoài ra, CLIP [10], một mạng lưới thần kinh được huấn luyện để học các khái niệm trực quan với sự giám sát bằng ngôn ngữ tự nhiên cũng được sử dụng. Nó được sử dụng để dự đoán chú thích thích hợp nhất của nămn hình ảnh đầu vào. Chúng tôi kết hợp nó với bộ dữ liệu Food101 [2] và SUN [14], cho phép phân loại thực phẩm được công nhận và phân loại cảnh tương ứng.

Cuối cùng, nhận dạng ký tự quang học (OCR) được áp dụng để nhận dạng các ký tự văn bản in trong hình ảnh. Sự kết hợp của các bộ mô tả nội dung trực quan đó sau đó được sử dụng để truy vấn cơ sở dữ liệu nhằm tìm khung hình video mong muốn

3.2/ Search Queries and Ranking:

Việc truy xuất các khung phù hợp được chia thành giai đoạn truy vấn và xếp hạng. Trong giai đoạn đầu tiên, các khung ứng cử viên phù hợp với các yêu cầu do người dùng chỉ định được xác định. Trong giai đoạn thứ hai, những ứng viên đó được xếp hạng theo điểm xác định mức độ phù hợp của họ với các yêu cầu.

Ứng dụng của chúng tôi cung cấp khả năng lọc danh mục video bằng cách chỉ định các khái niệm, đối tượng và giá trị OCR sẽ được đưa vào phân đoạn video mong muốn. Đối với những giá trị đã chọn đó, người dùng có thể tinh chỉnh thêm bằng cách bày tỏ những lời phê bình liên quan đến tầm quan trọng của các giá trị đơn lẻ trên thang điểm 3 với các giá trị “ít quan trọng”, “trung lập” hoặc “rất quan trọng”. Ví dụ: điều này hữu ích nếu một đối tượng cụ thể chủ yếu xuất hiện trong phân đoạn video được tìm kiếm. Những trọng số đó sẽ ảnh hưởng đến điểm số trong giai đoạn xếp hạng.

Ngoài ra, trong quá trình tương tác, người dùng có thể phản biện đề xuất đang hiển thị với ba tùy chọn khác nhau. Đầu tiên, người dùng có thể tuyên bố rằng khung hình hiện tại không thuộc về phân đoạn video được tìm kiếm, điều này sẽ loại trừ khung hình đó khỏi các ứng cử viên. Thứ hai, người dùng có thể đề cập rằng video lấy khung hiện tại không chứa khung được tìm kiếm. Điều này sẽ làm giảm các ứng cử viên theo tất cả các khung tiếp theo của video này. Thứ ba, trong trường hợp khung hình hiện tại không phải là một phần của phân đoạn video được tìm kiếm mà là một phần của cùng một video mà phân khúc được tìm kiếm, người dùng có thể khắc phục yêu cầu chỉ xem xét các khung hình của cùng một video. Một truy vấn cơ sở dữ liệu để truy xuất tập hợp ứng viên được biên dịch từ bộ sưu tập tùy chọn lọc Boolean cụ thể của các bộ mô tả nội dung được bao gồm và các video bị loại trừ hoặc cố định. Tập ứng viên được truy xuất bao gồm số lần xuất hiện của các giá trị bên trong khung.

Ý tưởng xếp hạng dựa trên giả định rằng các khung có nhiều phiên bản hơn của một giá trị siêu dữ liệu được chỉ định, có nhiều khả năng là khung chính xác hơn. Do đó, đối với mỗi khung trong tập hợp ứng cử viên, điểm số được tính bằng tổng số lần xuất hiện có trọng số của các giá trị siêu dữ liệu được chỉ định. Trọng số phụ thuộc vào việc người dùng có bày tỏ những lời phê bình có trọng số đối với một số giá trị hay không. Là một tiện ích mở rộng, chúng tôi áp dụng Tần số nghịch đảo tài liệu thuật ngữ (TF-IDF) [1] cho các lần xuất hiện để chuẩn hóa kết quả trước khi đưa chúng vào tính điểm. Áp dụng kỹ thuật này cho phép giảm tầm quan trọng của các bộ mô tả nội dung xuất hiện trong nhiều khung hình và nhấn mạnh các bộ mô tả cụ thể cho một lượng nhỏ khung hình. Khung có số điểm cao nhất được trả về dưới dạng gợi ý.

**3.3 User Interface:**

Khái niệm cốt lõi về giao diện người dùng ứng dụng Perfect Match của chúng tôi được lấy cảm hứng từ quy trình công việc được cung cấp trong các ứng dụng hẹn hò như Tinder 2. Những ứng dụng đó đề xuất hồ sơ người dùng được đại diện bởi ảnh hồ sơ cho người dùng, có thể nhanh chóng quyết định xem đề xuất đó có thú vị hay không cho anh ấy / cô ấy bằng cách vuốt sang trái hoặc phải.

Vì tương tác nhanh cũng phù hợp với cuộc thi VBS nên chúng tôi đã áp dụng cách tương tác này cho nhiệm vụ tìm kiếm video. Như được hiển thị trong Hình 1, đề xuất khung hiện tại được trình bày nổi bật ở phía bên trái của màn hình. Người dùng có khả năng sử dụng các nút bên dưới để quyết định điều gì sẽ xảy ra với khung. Nếu khung phù hợp (với nhiệm vụ tìm kiếm hiện tại), nó có thể được gửi. Nếu khung hình không khớp hoàn hảo nhưng người dùng mong đợi phần video chính xác từ cùng một video, thì họ có thể chỉ ra điều này. Mặt khác, nếu khung hình hoàn toàn không khớp với yêu cầu, người dùng có thể quyết định bỏ qua khung hình hiện tại hoặc tất cả các khung hình tiếp theo của video này. Cái sau hữu ích nếu họ cho rằng phần video được tìm kiếm thuộc về một video khác.

Để chỉ định các yêu cầu của tác vụ tìm kiếm hiện tại, phía bên phải của giao diện người dùng được sử dụng để xác định tiêu chí cho truy vấn. Người dùng có thể chọn các khái niệm và đối tượng từ tất cả các tùy chọn có thể bằng cách sử dụng hộp thả xuống có tự động hoàn thành, cũng như chỉ định nội dung OCR bằng kiểu nhập văn bản miễn phí. Sau khi chọn các yêu cầu, người dùng có thể tinh chỉnh thêm bằng cách gán trọng số

Truy vấn tìm kiếm được cập nhật mỗi khi người dùng thay đổi các yêu cầu hoặc giá trị, do đó, mọi tương tác đều có khả năng dẫn đến một đề xuất mới được xếp hạng cao nhất hiển thị ở bên trái của giao diện người dùng.

A person sitting at a desk

Description automatically generated

4/ Conclusion:

Trong bài báo này, chúng tôi giới thiệu phiên bản đầu tiên của ứng dụng tìm kiếm video Perfect Match có thể được sử dụng để tìm kiếm video trong bộ dữ liệu V3C và gửi giải pháp cho các nhiệm vụ của cuộc thi VBS. Hệ thống lập chỉ mục các video và tự động suy ra các mô tả nội dung trực quan của các khung hình đã chọn. Hơn nữa, nó cho phép người dùng tinh chỉnh các yêu cầu tìm kiếm một cách tương tác bằng cách thể hiện các lời phê bình để tìm đoạn video được tìm kiếm nhanh nhất có thể.

**Exploring Effective Interactive Text-Based Video Search in vitrivr**

Abs : vitrivr là một hệ thống truy xuất mục đích chung hỗ trợ nhiều phương thức truy vấn. Trong bài báo này, chúng tôi giới thiệu ngắn gọn về hệ thống và mô tả các thay đổi cũng như điều chỉnh được thực hiện cho lần lặp lại năm 2023 của cuộc thách đấu trình duyệt video. Chúng tập trung chủ yếu vào các sơ đồ truy xuất dựa trên văn bản và cơ chế phản hồi của người dùng tương ứng.

KeyWord: Video browser showdown · Interactive video retrieval · Content-based retrieval

1/ Intro:

Video Browser Showdown (VBS) [9,14,19] là một chiến dịch đánh giá dài hạn để truy xuất đa phương tiện tương tác và tìm kiếm video lấy người dùng làm trung tâm. Kể từ năm 2012 [27], VBS đã cung cấp một thiết lập có tính cạnh tranh cao, trong đó các hệ thống và người điều hành của chúng được giao nhiệm vụ tìm các phân đoạn video trong một bộ sưu tập lớn. Bộ sưu tập hiện đang được sử dụng là một tập hợp con của Bộ sưu tập Vimeo Creative Commons (V3C) [26], bao gồm 2300 giờ tài liệu video có tổng dung lượng lên tới 1,6 TB. Ngoài V3C1 [3] và V3C2 [25], phiên bản VBS năm 2023 sẽ có bộ dữ liệu lặn dưới nước/lặn bình khí đồng nhất có tên Marine Video Kit [30], có dung lượng khoảng 230 GB và thời lượng khoảng 11,5 giờ

VBS bao gồm hai loại nhiệm vụ: Tìm kiếm mục đã biết (KIS) và Tìm kiếm video đặc biệt (AVS) [12]. Cái trước liên quan đến việc tìm một phân đoạn video cụ thể dựa trên bản xem trước trực quan hoặc mô tả bằng văn bản. Loại thứ hai yêu cầu tìm các mục quan tâm phù hợp với mô tả chung hơn (độ chính xác của chúng được đánh giá thủ công trong cuộc thi).

Trong bài báo này, chúng tôi trình bày vitrivr – một hệ thống truy xuất đa phương tiện đa phương thức dựa trên nội dung mã nguồn mở – và những cải tiến được thực hiện cho nó so với các phiên bản trước đó. Kỳ năm 2023 đánh dấu lần thứ 91 vitrivr tham gia VBS liên tiếp [7], với hai lần giành chiến thắng trong bốn năm qua [8,24]. A diagram of a video search engine

Description automatically generated

Hình 1: Tổng quan về hệ thống cho vitrivr và ba thành phần chính của nó: vitrivr-ng, Cineast và Cottontail DB. Phiên bản sửa đổi đôi chút của [8, Hình 1]

Phần còn lại của bài báo này được cấu trúc như sau: Phần. 2 cung cấp một cái nhìn tổng quan về vitrivr stack, Sect. 3 làm nổi bật các bổ sung khác nhau cho ngăn xếp và chúng tôi kết luận bài báo trong Phần. 4

2/ **Vitrivr:**

vitrivr [23] là ngăn xếp truy xuất đa phương tiện nguồn mở, có khả năng hỗ trợ nhiều loại phương tiện và truy vấn bao gồm nhưng không giới hạn ở video (tìm kiếm). Tổng quan về kiến ​​trúc của vitrivr được cung cấp trong Hình 1 [8, Hình 1]. Ngăn xếp bao gồm ba thành phần chính:

**Cottontail DB:** là lớp cơ sở dữ liệu của vitrivr và có thể được sử dụng để lưu trữ, quản lý và truy vấn siêu dữ liệu vô hướng cũng như các vectơ đặc trưng chiều cao. Cottontail DB cho phép tương tự hiệu quả và truy xuất Boolean.

**Cineast :** là công cụ trích xuất và truy xuất tính năng của ngăn xếp. Nó tạo ra các biểu diễn tính năng khác nhau từ dữ liệu đầu vào (video, truy vấn do người dùng cung cấp), sắp xếp việc thực thi truy vấn và thực hiện tổng hợp kết quả và hợp nhất điểm sốc.

**Vitrivr-ng:** là giao diện người dùng dựa trên web của vitrivr và tạo điều kiện thuận lợi cho việc xây dựng truy vấn, trình bày kết quả và cho phép khám phá hiệu quả. Ngoài ra, nó cũng cho phép lọc và hợp nhất giai đoạn cuối.

Tất cả các thành phần của vitrivr đều có sẵn miễn phí trên trang web của dự án.Một số thành phần được đặt tên ở trên cũng đóng vai trò là cơ sở cho các hệ thống truy xuất đa phương tiện khác như vitrivr-vr [28,29] và Lifegraph [18]

3/ **Novelties for VBS 2023:**

Vì lần lặp lại trước của VBS đã chỉ ra [9] rằng hầu hết các hệ thống hoạt động hàng đầu đều dựa vào đồng nhúng video-văn bản và hỗ trợ một số dạng truy vấn tạm thời, chúng tôi đã tập trung vào việc tinh chỉnh và mở rộng các phương tiện cho phép truy vấn dựa trên văn bản tìm kiếm.

3.1/ **Improved Visual Text Co-embedding:**

Chúng tôi đã giới thiệu phiên bản đầu tiên của tính năng đồng nhúng văn bản trực quan trong [29], bao gồm một mạng nông dự đoán đầu ra của hai xương sống được đào tạo trước đơn phương thức vào một không gian chung, được căn chỉnh về mặt ngữ nghĩa. Để xử lý nhiều khung hình của một video thay vì chỉ một hình ảnh duy nhất, tất cả các khung hình được chuyển qua xương sống trực quan của chúng và đầu ra của nó được gộp lại trước khi chiếu. Đối với phiên bản VBS năm nay, chúng tôi đã cập nhật một số khía cạnh của phương pháp này. Hai xương sống đã được thay thế và sơ đồ tổng hợp và phương pháp chiếu đã được tinh chỉnh. Để nhúng văn bản, chúng tôi hiện sử dụng xương sống đa ngôn ngữ [31] để tăng khả năng tiếp cận đối với những người không nói tiếng Anh bản ngữ. Đối với các nhúng cấp khung trực quan, chúng tôi sử dụng kiến ​​trúc tích chập mới hơn [11] và loại bỏ cả phân loại cuối cùng cũng như các lớp tổng hợp không gian. Lấy cảm hứng từ [2], chúng tôi mở rộng sơ đồ tổng hợp để không tập hợp các phần nhúng hình ảnh một cách bừa bãi mà chú ý đến nguồn gốc không gian và nhịp độ của các phần nhúng hình ảnh.

**3.2/ CLIP and Vitrir:**

Việc OpenAI phát hành mô hình CLIP [16] vào năm 2021 đã đánh dấu một bước thay đổi về chất lượng có thể đạt được khi tìm kiếm hình ảnh bằng cách sử dụng văn bản mô tả nội dung ngữ nghĩa của chúng. Trong phiên bản 2022 của VBS, một số nhóm đạt điểm cao nhất ít nhất một phần dựa vào các biểu diễn tính năng do CLIP tạo ra [1,10,13]. Để chứng minh tính hiệu quả mang tính quyết định này, chúng tôi đã thêm một trình trích xuất tính năng dựa trên CLIP vào vitrivr cho phiên bản 2023 của VBS. Các tính năng được trích xuất chỉ dựa trên một khung hình đại diện cho mỗi lần chụp, như được cung cấp bởi bộ dữ liệu [26]. Trong thời gian chạy, chúng tôi cung cấp cho người dùng phương tiện để chọn đồng nhúng hoặc CLIP của chúng tôi làm trình xử lý truy vấn.

Vì kỹ thuật nhắc nhở dường như là một yếu tố có liên quan trong việc thực hiện hiệu quả các mô hình ngôn ngữ hình ảnh chung hiện đại, vì những thay đổi nhỏ trong đầu vào có thể dẫn đến những thay đổi khá lớn trong kết quả trả về trong một số trường hợp, chúng tôi cũng sử dụng các phương pháp chú thích hình ảnh do CLIP hướng dẫn, cụ thể là [4] và [15], để tạo một (mỗi) chú thích cho mỗi khung hình đại diện của mỗi cảnh quay. Những chú thích này không nhằm mục đích sử dụng trực tiếp cho tìm kiếm (mặc dù chức năng như vậy cũng được hỗ trợ) mà là để cung cấp phản hồi cho người điều hành, truy vấn văn bản hợp lý sẽ là gì để truy xuất bất kỳ kết quả nào. Người vận hành có thể sử dụng cơ chế phản hồi này để tự làm quen với sự phức tạp của tính năng nhằm giúp họ xây dựng các lời nhắc tìm kiếm hiệu quả hơn

3.3 **SIMD Support for Cottontail DB:**

Tốc độ thực hiện truy vấn là điều cốt yếu đối với việc truy xuất video tương tác, đặc biệt là trong các cài đặt cạnh tranh như VBS. Trong phiên bản mới nhất của Cottontail DB [5]—lớp cơ sở dữ liệu đa phương tiện được sử dụng bởi vitrivr—do đó, chúng tôi đã bắt đầu khai thác việc sử dụng các lệnh SIMD để tăng tốc thực thi truy vấn cho tìm kiếm brute-force. Việc sử dụng rõ ràng SIMD đã được kích hoạt bằng cách tạo JAVA VECTOR API mới được đề xuất gần đây trong JEP 338, 414 và 417.3 Mặc dù việc triển khai hiện tại khá đơn giản và mặc dù tính năng này vẫn đang ở giai đoạn thử nghiệm sớm, chúng tôi có thể báo cáo tăng tốc từ 20–30%, đặc biệt đối với các vectơ chiều cao (d > 1024). Chúng tôi hy vọng sẽ đạt được khả năng tăng tốc hơn nữa bằng cách chuyển đổi công cụ thực thi truy vấn cơ bản từ trình vòng lặp sang mô hình xử lý theo đợt trong tương lai (không quá xa)

3.4 **Humman in the Loop:**

Một thành phần quan trọng trong tìm kiếm video lấy người dùng làm trung tâm là người vận hành hệ thống (truy xuất) của con người. Nhóm vitrivr sử dụng các lần chạy khô kiểu VBS thông thường từ khá lâu với thiết lập đánh giá chuyên dụng. Trong phần năm nay, chúng tôi sử dụng triển khai hệ thống DRES [20] của riêng mình, với các tác vụ được tạo riêng cho mục đích đó. Chúng tôi phân tích từng lần chạy khô và đặc biệt là phân tích những nhiệm vụ mà chúng tôi—người vận hành hệ thống và hệ thống—không thể tìm thấy mục tiêu. Như đã nêu trong Mục. 3.2, chúng tôi sử dụng, trong số những phương tiện khác, phương tiện trình bày đặc điểm văn bản của mục tiêu để tìm hiểu những cụm từ tìm kiếm nào sẽ có mục đích. Cụ thể đối với phần cài đặt năm 2023, chúng tôi cũng sẽ chạy thử nghiệm vào tháng 12 với các đồng nghiệp không làm việc trên vitrivr để mô phỏng các phiên làm việc mới và chúng tôi có thể cần điều chỉnh một số chức năng giao diện người dùng dựa trên phản hồi của họ.

**4/Conclusion:**

Trong bài báo này, chúng tôi đã trình bày phiên bản của vitrivr mà chúng tôi dự định tham gia tại VBS 2023. Như phân tích gần đây đã chỉ ra, xu hướng hiện tại trong tìm kiếm video lấy người dùng làm trung tâm hướng tới việc nhúng đồng văn bản video hỗ trợ học sâu, chẳng hạn như KẸP. Vì vậy, chúng tôi tập trung vào các cải tiến trong lĩnh vực này bằng cách mở rộng tính năng đồng nhúng văn bản trực quan—trong số những thứ khác—với một đường trục đa ngôn ngữ và cũng giới thiệu mô hình CLIP trong hệ thống của chúng tôi. Hơn nữa, nhiều phần khác nhau của hệ thống truy xuất mã nguồn mở vitrivr đã được cải thiện và chúng tôi sẽ đào tạo hệ thống một cách nghiêm túc cho những người vận hành hệ thống của mình với hệ thống cũng như mô phỏng các phiên làm việc mới với các đồng nghiệp địa phương của chúng tôi

**QIVISE: A Quantum-Inspired Interactive Video Search Engine in VBS2023**

**ABS**: Trong bài báo này, chúng tôi trình bày một công cụ tìm kiếm video tương tác lấy cảm hứng từ lượng tử (QIVISE), sẽ được thử nghiệm trong VBS2023. QIVISE nhằm mục đích hỗ trợ người dùng xử lý các tác vụ Tìm kiếm mục đã biết và Tìm kiếm video đặc biệt với hiệu quả và độ chính xác cao. QIVISE dựa trên bộ mã hóa hình ảnh văn bản để đạt được khả năng nhúng đa phương thức và giới thiệu nhiều khả năng tương tác, bao gồm tương tác lấy cảm hứng từ lượng tử mới trên mô hình, tìm kiếm nhãn và tìm kiếm đa phương thức để tinh chỉnh kết quả truy xuất thông qua tương tác và phản hồi của người dùng.

KeyWord: Video browser showdown · Interactive video retrieval · Quantum theory

1/ Intro:

Video Browser Showdown [6] là một cuộc thi tìm kiếm video quốc tế hàng năm được tổ chức bởi Hội nghị Quốc tế về Mô hình hóa Đa phương tiện (MMM), nhằm mục đích đánh giá hệ thống truy xuất video hiện đại và thúc đẩy hiệu quả của tìm kiếm video quy mô lớn. Trong VBS2023, những người tham gia được yêu cầu giải quyết các nhiệm vụ Tìm kiếm mục đã biết (KIS) và Tìm kiếm video đặc biệt (AVS), trong đó các truy vấn được trình bày dưới dạng clip trực quan hoặc mô tả văn bản. Truy vấn cũng có thể được mở rộng và xây dựng bởi người dùng thông qua tương tác sau tìm kiếm ban đầu.

Trong bài báo này, chúng tôi trình bày Công cụ tìm kiếm video tương tác lấy cảm hứng từ lượng tử (QIVISE). QIVISE là một hệ thống kế thừa các phương pháp tiên tiến và mới nhất trong nghiên cứu truy xuất video và các hệ thống nổi bật trước đây trong VBS cũng như giới thiệu một mô hình tương tác lấy cảm hứng từ lượng tử mới nổi. Tóm lại, những đóng góp của QIVISE là: (1) Được thúc đẩy bởi các khái niệm và tìm kiếm từ khóa của hệ thống VBS hiện có [7], QIVISE giới thiệu tìm kiếm nhãn tương tác, cho phép người dùng chọn nhãn được trích xuất từ ​​đầu vào truy vấn văn bản và thêm nhãn phụ để nhấn mạnh thông tin chính. (2) QIVISE áp dụng mô hình tương tác lấy cảm hứng từ lượng tử để lập mô hình phản hồi của người dùng và tinh chỉnh kết quả truy xuất. (3) QIVISE hỗ trợ tìm kiếm video theo nhiều phương thức mà người dùng có thể sử dụng đầy đủ các thông tin phương thức khác nhau để nhấn mạnh thông tin chính.

**2/ Interface and Navigation:**

**A screenshot of a computer

Description automatically generated**

Tổng quan về giao diện hiện tại được hiển thị trong Hình 1. Về tương tác giữa người và máy tính và tham chiếu đến hệ thống hiện có như Vibro [3], giao diện minh họa có thể xử lý hiệu quả cả tác vụ AVS và KIS. Người dùng có thể tìm kiếm các video clip được nhắm mục tiêu bằng cách nhập mô tả bằng văn bản hoặc tải lên một video clip. Ngoài ra, hệ thống hỗ trợ hai chế độ tương tác rõ ràng. Một là người dùng có thể thêm nhãn chủ quan cho đầu vào truy vấn, được giới thiệu chi tiết trong Phần. 3.1. Và cách khác là người dùng có thể cung cấp phản hồi dựa trên kết quả truy vấn để cải thiện tỷ lệ thu hồi cuối cùng. Phần 1, đề cập đến các nhãn trong Hình 1, được dành riêng cho việc tùy chỉnh truy vấn. Một mô tả chi tiết về tất cả các phương pháp có sẵn có thể được tìm thấy trong Phần. 3. Ở Phần 2, theo yêu cầu truy vấn của người dùng, hệ thống lấy điểm tương đồng của các khung hình chính tương ứng với video trong cơ sở dữ liệu thông qua công cụ truy xuất vector [11] và lọc ra 4000 kết quả theo thứ tự giảm dần. Quá trình này mất vài micro giây.

Để nâng cao hiệu quả tìm kiếm, chúng tôi cung cấp cho người dùng bản xem trước ở dạng 49 ảnh mỗi nhóm trong một khu vực giới hạn và bằng cách sử dụng Masonry Layouts, người dùng có thể nhanh chóng duyệt qua tất cả các kết quả có thể và xem thông tin chi tiết bằng cách nhấp vào ảnh. Trong Phần 3, hệ thống giới thiệu thuật toán phân cụm [2] để chọn ra 10 khung hình chính có khả năng phù hợp nhất nhằm tránh việc người dùng phải lựa chọn và so sánh lâu dài. Người dùng đánh giá mức độ liên quan của các clip được trình chiếu và cung cấp phản hồi bằng cách nhấp vào nút được ưa thích hoặc không được ưa thích bên cạnh video clip đã chọn. Nếu người dùng không hài lòng với tất cả các kết quả được trình bày, họ có thể nhận được kết quả vòng mới thông qua nút làm mới. Sau khi người dùng hoàn thành phản hồi và nhấp vào nút xếp hạng lại, hệ thống sẽ cung cấp kết quả sắp xếp thứ cấp được cải thiện dựa trên thuật toán xếp hạng lại lấy cảm hứng từ lượng tử, được giới thiệu chi tiết trong Phần. 3.2.

**3/ Multiple Ways of User Interactions:**

Hệ thống sử dụng BLIP [5], một mô hình Đào tạo trước ngôn ngữ-hình ảnh Bootstrapping, để nhúng văn bản và hình ảnh [4]. Với kết quả truy xuất ban đầu thu được bằng cách so sánh độ tương tự cosine giữa vectơ ding truy vấn và hình ảnh, hệ thống của chúng tôi áp dụng một số phương pháp để tinh chỉnh kết quả truy xuất vòng đầu tiên thông qua tương tác người dùng, về cơ bản bao gồm tìm kiếm nhãn, xếp hạng lại lấy cảm hứng từ lượng tử và tìm kiếm đa phương thức.

**3.1/ Label Search:**

So với một truy vấn văn bản đơn lẻ, các nhãn được trích xuất từ ​​văn bản truy vấn và được người dùng chọn có thể chính xác hơn trong việc mô tả nhu cầu tìm kiếm của người dùng. Hệ thống của chúng tôi áp dụng phương pháp mới nhất [10], một khung từ đầu đến cuối hoạt động để trích xuất nhãn, hay còn gọi là cụm từ khóa, từ các câu truy vấn. Mô hình trích xuất các cụm từ chính hiện tại từ văn bản và cả những cụm từ ẩn trong văn bản. Nó giúp người dùng có thể nắm bắt thông tin còn thiếu nhưng hữu ích. Với các cụm từ khóa hiện tại và tiềm ẩn được trích xuất từ ​​​​truy vấn, người dùng chọn các nhãn phù hợp với nhu cầu tìm kiếm của họ và xếp hạng chúng theo mức độ liên quan của chúng. Bên cạnh đó, người dùng cũng có thể thêm các nhãn chứa thông tin mà người dùng muốn nhấn mạnh. Bằng cách nhúng nhãn vào không gian chung và tính toán độ tương tự cosin giữa nhãn và video clip, hệ thống sẽ nhận được điểm phù hợp bổ sung để sửa đổi và cải thiện điểm ban đầu để xếp hạng chính xác.

**3.2/ Quantum Inspired Interaction for Clips Re-Ranking:**

Sau bước nhảy vọt trong truy xuất thông tin lấy cảm hứng từ lượng tử [9], hệ thống của chúng tôi giới thiệu một mô hình tương tác lấy cảm hứng từ lượng tử mới để lập mô hình tương tác của người dùng.

Trong lý thuyết lượng tử, trạng thái lượng tử của sự phân cực được biểu diễn bằng sự kết hợp tuyến tính của phân cực dọc và ngang, được biểu thị bằng công thức sau:



trong đó |ϕ là trạng thái lượng tử của sự phân cực, |1 và |0 là sự phân cực dọc và ngang, α và β là các hệ số phụ thuộc vào |α|­2+|β|2 = 1. Trong các cách tiếp cận truy xuất thông tin lấy cảm hứng từ lượng tử, một tài liệu được đại diện lại theo mức độ liên quan và không liên quan của một truy vấn [8]:

Ta có |α| 2 = p(d|q) và |β|2 = 1 − p(d|q), trong đó p(d|q) là xác suất liên quan trong truy xuất vòng đầu tiên, nó có thể được xác định bằng điểm xếp hạng chuẩn hóa.

Sau đó, chúng ta có thể ước tính về |r với phương trình r|d = α:



Với các công thức trên, bây giờ chúng tôi giới thiệu chiến lược của chúng tôi để xếp hạng lại. Sau khi truy xuất vòng đầu tiên, người dùng có thể chọn các video clip được hiển thị phù hợp cao hoặc không phù hợp với nhu cầu của người dùng. Với những clip đã chọn đó, chúng ta có thể ước tính được nhu cầu thực sự của người dùng bằng cách áp dụng (3). Nhu cầu thực sự của người dùng có thể được coi là không gian kéo dài bởi vectơ clip đã chọn và không gian con phần bù của nó, như trong Hình 2. Để đơn giản hóa việc tính toán và tăng tốc quy trình tương tác, không gian con phần bù được xác nhận bằng tổng chuẩn hóa của giá trị cơ bản bộ giải pháp của phương trình. (4):



Tổng công thức như sau:

A math equations and formulas

Description automatically generated

trong đó |r đại diện cho nhu cầu của người dùng, n là kích thước nhúng, |d đại diện cho video clip đã chọn, |xi là vectơ cơ sở của phương trình. (4), |¬d đại diện cho không gian con phần bù và α, β là các xác suất liên quan trong phép truy xuất vòng trước. Với |r , điểm cuối cùng để xếp hạng lại được xác định như sau:



trong đó veva là video clip sẽ được đánh giá ở vòng tiếp theo, cosine(·) là độ tương tự cosine, q là vectơ nhúng truy vấn, r if là nhu cầu của người dùng được tạo bởi video nhất quán thứ i và r inf là nhu cầu của người dùng ước tính nhu cầu được tạo bởi video không nhất quán thứ i. Trong trường hợp xác suất liên quan của tất cả các video clip đã chọn bằng một, phương trình. (6) là thuật toán phản hồi mức độ phù hợp của Rocchio [1]. So với Thuật toán Rocchio, Eq. (6) sử dụng |rf , có liên quan đến xác suất liên quan trong truy xuất vòng trước, thay vì chỉ sử dụng |d để tạo điểm xếp hạng lại và có thể cải thiện độ bền của nó đối với các mẫu nhiễu ở một mức độ nào đó. Để thuận tiện cho người dùng có thể đánh giá và lựa chọn clip từ toàn bộ kết quả tìm kiếm, chúng tôi áp dụng Hierarchical Clustering để chia kết quả tìm kiếm thành các cấp độ khác nhau để người dùng tương tác nhanh.

A diagram of a diagram of a graph

Description automatically generated

**3.3/ Multi modal Search:**

Phần tìm kiếm đa phương thức có thể coi là phần mở rộng của phần tìm kiếm trên. Ví dụ: khi xử lý các tác vụ KIS-V, người dùng có thể thêm nhãn hoặc mô tả văn bản cho hình ảnh truy vấn. Tương tự, khi thực hiện các tác vụ KIS-T hoặc AVS, người dùng có thể nhập hình ảnh dưới dạng mô tả bổ sung cho truy vấn ban đầu. Sau đó, điểm số cuối cùng để xếp hạng sẽ được xác định là tổng trọng số của các điểm số khác nhau được tính toán độc lập. Bằng cách này, các phần tìm kiếm cho các phương thức khác nhau không còn bị cô lập mà được kết hợp chặt chẽ đến mức có thể sử dụng đầy đủ thông tin về các phương thức khác nhau.

**4/ Conclusion:**

Bài báo này trình bày một công cụ tìm kiếm video tương tác lấy cảm hứng từ lượng tử để tham gia Video Browser Showdown 2023. Công cụ này áp dụng phương pháp nhúng hình ảnh văn bản mới nhất BLIP [5] và giới thiệu một số cách cho hành động tương tác của người dùng bao gồm tìm kiếm nhãn, tương tác lấy cảm hứng từ lượng tử cho xếp hạng lại và tìm kiếm đa phương thức