Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00096

ỨNG DỤNG KỸ THUẬT ĐỊNH DANH TỪ DỮ LIỆU VIDEO VÀO VIỆC NHẬN DẠNG CON NGƯỜI, HÀNH ĐỘNG VÀ ĐỊA ĐIỂM XUẤT HIỆN

Phạm Thế Phi [1], Đỗ Thanh Nghị [1]

[1] Khoa Công nghệ thông tin và Truyền thông, Trường Đại học Cần Thơ

ptphi@cit.ctu.edu.vn, dtnghi@cit.ctu.edu.vn

TÓM TẮT— *Bài viết này đề xuất một phương pháp mới để truy hồi video với các chú thích nội dung bằng văn bản không hoàn chỉnh. Ý tưởng chính là việc sử dụng cơ chế suy diễn Bayes để dự đoán định danh của con người, hành động và địa điểm họ xuất hiện trong các khung hình video. Một vài mô hình truy hồi video với khả năng tích hợp các chứng cứ xuất hiện của ảnh và văn bản không hoàn chỉnh được đề xuất và so sánh. Trong các thí nghiệm, bài viết sử dụng các tập của bộ phim truyền hình Buffy the Vampire Slayer làm các tập dữ liệu huấn luyện và kiểm thử. Mô hình mạng Bayes được đề xuất có khả năng cho phép kết hợp nhiều thuộc tính của video như hình ảnh và văn bản, xử lý các câu truy vấn trong đó có nhiều thực thể có quan hệ ngữ nghĩa với nhau và quan trọng nhất là có khả năng suy luận ra các thực thể nếu chúng không được nhắc tới trong văn bản nhưng lại xuất hiện trong các khung hình.*

Từ khóa— *Khai thác dữ liệu đa phương tiện, lập chỉ mục và truy hồi video*.

I. GIỚI THIỆU

Web đang ngày càng trở thành một nguồn thông tin sống còn với khả năng thoả mãn hầu như mọi nhu cầu tra cứu thông tin của con người. Các bộ máy tìm kiếm Web, đến lượt mình, trở thành công cụ trích lọc phải có nhằm giúp cho người dùng thu hẹp phạm vi tìm kiếm trong một biển thông tin, để cuối cùng có được vài thông tin mà họ thực sự cần. Các bộ máy tìm kiếm có lịch sử phát triển lâu đời dựa trên nhiều kết quả nghiên cứu thành công trong các lĩnh vực lập chỉ mục và truy hồi thông tin (information indexing and retrieval). Lập chỉ mục và truy hồi thông tin dạng văn bản là một ví dụ điển hình. Kỹ thuật này so khớp câu truy vấn của người dùng không chỉ với siêu dữ liệu (metadata) dùng để mô tả tài liệu (trên Web thường là thiếu hoặc không hoàn chỉnh) mà còn với chính nội dung bên trong tài liệu đó. Các kỹ thuật lập chỉ mục và truy hồi thông tin hiện đại thường đạt đến mức rút trích và trình bày thông tin theo ngữ nghĩa (semantics extraction and representation). Kỹ thuật này tập trung vào việc so khớp câu truy vấn và tài liệu theo ngữ nghĩa. Đây rõ ràng là một thành tựu quan trọng vì hệ thống có thể hiểu được ý nghĩa thực sự đằng sau câu truy vấn của người dùng và sau đó trả về cho người dùng các tài liệu có ý nghĩa tương đồng mà rất có thể họ sẽ cảm thấy thoả mãn.

Với sự phổ biến nhanh chóng của truyền thông đại chúng, nhu cầu thông tin của người dùng không còn giới hạn ở các văn bản thuần tuý mà mở rộng ra đến hình ảnh, âm thanh, video, thông tin y sinh,... Rất nhiều nhà nghiên cứu nắm bắt xu hướng này đã và đang nỗ lực không ngừng để xây dựng các hệ thống truy hồi thông tin đa phương tiện nhằm thỏa mãn nhu cầu tìm hiểu thông tin đa phương tiện của người dùng. Xử lý thông tin đa phương tiện hướng nội dung thực sự khó hơn nhiều so với thuần tuý xử lý văn bản. Thứ nhất, cần nhiều nỗ lực để rút trích các đặc trưng quan trọng của dữ liệu đa phương tiện. Thứ hai, những đặc trưng này thường không tương ứng với trực giác của con người vì chúng thường là các đặc trưng thô thể hiện dữ liệu ở mức thấp mà chứa đựng rất ít ngữ nghĩa. Thứ ba, số lượng cũng như các mối liên hệ giữa các khái niệm mang tính ngữ nghĩa sẽ khác nhau tuỳ vào các nguồn dữ liệu, vì vậy làm sao để chọn ra các khái niệm mang tính ngữ nghĩa vừa đủ và vừa có ý nghĩa để trình bày cho người dùng là một thử thách lớn. Những thách thức nói trên thường được đề cập đến như là khoảng trống về ngữ nghĩa (semantic gap) giữa nội dung ở mức thấp và các khái niệm mức cao.

Câu hỏi đặt ra là làm sao để lấp đầy khoảng trống ngữ nghĩa này? Các nỗ lực ban đầu là nhằm vòng tránh không giải quyết trực tiếp vấn đề mà chỉ đơn thuần chuyển đổi thể hiện câu truy vấn ở mức thấp hơn. Ví dụ, một số hệ thống truy hồi video sử dụng cơ chế "truy vấn qua ví dụ" (query by example). Tuy nhiên, với câu truy vấn tương đối phức tạp như "Buffy is fighting in the graveyard", rất khó để tìm ra tấm ảnh ví dụ tương ứng trong cơ sở dữ liệu hiện có để có thể rút trích ra những đặc trưng cấp thấp nhằm phản ánh toàn bộ tập các khái niệm "Buffy", "fighting" và "graveyard". Một cơ chế xử lý thông dụng khác là sử dụng các câu truy vấn dạng văn bản. Với cách này, việc nắm bắt ngữ nghĩa là dễ dàng hơn đối với câu truy vấn. Tuy nhiên, chúng ta cũng phải rút trích ra các ý nghĩa từ các phương tiện khác với văn bản, ánh xạ chúng tới các ngữ nghĩa từ câu truy vấn và trả về cho người dùng. Phương tiện âm thanh trong video có thể được chuyển đổi thành văn bản để làm giàu thêm thông tin cho nguồn văn bản, nhưng đối với dữ liệu hình ảnh thì vấn đề vẫn đang được nghiên cứu.

Vấn đề còn lại là làm sao để ánh xạ từ các đặc trưng cấp thấp sang các ngữ nghĩa cấp cao? Giải pháp được áp dụng phổ biến trong thực tiễn là sử dụng việc chú thích bằng tay (bởi nhân viên lưu trữ hoặc sử dụng các dịch vụ gán nhãn trực tuyến bởi con người, chẳng hạn dịch vụ Amazon Mechanical Turk[1]). Tuy nhiên, với quy mô của vấn đề, chú thích bằng tay không phải là lựa chọn thích hợp. Vì thế, sẽ hợp lý hơn nếu các kỹ thuật máy học được áp dụng để giải quyết bài toán

[1] http://www.mturk.com

mà ở đó các tiến trình học có thể là có giám sát một phần hoặc không có giám sát. Các kỹ thuật có giám sát một phần thực hiện việc rút trích đặc trưng của dữ liệu ảnh một cách tự động, định nhãn bằng tay một phần dữ liệu và gán nhãn tự động phần còn lại của dữ liệu dựa trên sự tương đồng của các đặc trưng cấp thấp. Các kỹ thuật không giám sát thực hiện tự động tất cả các thủ tục rút trích đặc trưng cấp thấp và các ngữ nghĩa cấp cao, liên kết chúng lại dựa trên các kiểu mẫu được phát hiện trong toàn bộ dữ liệu. Hoạt động nghiên cứu các kỹ thuật máy học này đang tiến triển và thực sự cần nhiều hơn nữa các đóng góp, cải tiến để có thể được áp dụng rộng rãi. Bài viết này sẽ thảo luận một số đóng góp vào hướng nghiên cứu này. Cụ thể là chúng tôi muốn khám phá sự đồng xuất hiện của dữ liệu văn bản và hình ảnh để xây dựng các mô hình có thể lấp đầy khoảng trống ngữ nghĩa giữa các đặc trưng cấp thấp và các ngữ nghĩa mức cao. Ở đây chúng tôi sẽ tập trung vào nghiên cứu các hệ thống truy hồi video.

Các hệ thống truy hồi video hiện tại thường dựa vào các chú thích bằng tay của các video tương ứng. Việc có được các chú thích này thường được tổ chức thông qua các hoạt động gán nhán dựa trên số đông người đóng góp. Tuy nhiên, trong một số trường hợp, các chú thích được thêm vào vì lý do yêu thích cá nhân. Với dữ liệu video mà chúng tôi sử dụng trong bài viết này (các tập phim truyền hình "*Buffy the Vampire Slayer*" [1]), người hâm mộ đã thêm vào các mô tả theo dạng văn bản ngôn ngữ tự nhiên nhằm kể lại những gì đang diễn ra trong video. Những mô tả này rất không hoàn thiện vì rất nhiều khung hình không có mô tả hoặc các mô tả có được chỉ thể hiện một phần nội dung của một khung hình video (ví dụ như con người và hành động của họ được mô tả, nhưng địa điểm xuất hiện lại không có). Thêm nữa, các mốc thời gian chỉ đạt được sự trùng khớp tương đối. Trong bối cảnh này, các phương pháp truy hồi video dựa theo nội dung có xem xét cả nội dung ảnh và các mô tả bằng văn bản tương ứng là có giá trị vì chúng có khả năng cải thiện độ chính xác khi truy vấn cũng như cung cấp cái nhìn rõ ràng hơn bên trong các tập dữ liệu video [2]. Lý do là vì các phương pháp truy hồi thông tin hướng nội dung có nhiều khả năng có thể tóm tắt nội dung của phương tiện chứa thông tin thành các mệnh đề mô tả súc tích và phân lớp các mô tả này.

Nghiên cứu của chúng tôi tập trung vào việc truy vấn các video về *Buffy* với các khái niệm ngữ nghĩa như con người, hành động của họ và địa điểm mà họ xuất hiện (ví dụ: *Buffy is fighting in the graveyard).* Để có thể truy hồi được các khung hình tương ứng, lý tưởng nhất là mỗi khung hình được chú thích một cách chi tiết và đầy đủ như ở ví dụ trên. Ở đây chúng tôi sẽ trình bày một mô hình dùng để chú thích nội dung các khung hình một phần dựa trên một số nhận dạng thông tin không chắc chắn từ các nguồn văn bản, các khung hình và một số thông tin so khớp không chắc chắn của chúng, từ đó sử dụng mạng Bayes để suy diễn ra các mối liên hệ còn thiếu, lấp đầy khoảng trống ngữ nghĩa. Các phương pháp tích hợp chứng cứ này có khả năng tích hợp tốt vào các mô hình truy hồi thông tin dựa theo xác suất. Chúng tôi đề xuất ba mô hình truy hồi thông tin. Mô hình truy hồi cơ sở đầu tiên (gọi là Unigram Language Model) truy hồi các khung hình video chỉ dựa trên các chú thích bằng văn bản hiện có nhưng không hoàn chỉnh. Mô hình thứ hai (gọi là Unimodal Entity-Relation Model) sẽ rút trích tên con người, tên hành động và tên địa điểm từ văn bản, liên kết các tên thực thể thành các bộ dữ liệu quan hệ (relational tuple) và sử dụng các bộ dữ liệu này trong các mô hình truy hồi thông tin hướng nội dung. Mô hình này biểu diễn nội dung của các chú thích của người hâm mộ thành túi các bộ dữ liệu theo một mô thức dữ liệu duy nhất là văn bản. Cuối cùng, ngoài thông tin được rút trích từ văn bản, chúng tôi cũng tích hợp thông tin rút trích từ nội dung các khung hình video để xây dựng mô hình tích hợp nội dung phức tạp hơn. Mô hình thứ ba (Multimodal Entity-Relation Model) tương tự như mô hình dùng túi các bộ nhưng tích hợp các chứng cứ từ nhiều mô thức dữ liệu khác nhau. Tất cả các mô hình suy luận với tri thức chưa chắc chắn được rút ra từ các mô thức dữ liệu với độ phức tạp trải rộng từ các mô hình xác suất hướng nội dung đơn giản đến các mô hình suy luận sử dụng mạng Bayes đầy đủ. Hơn nữa, các mạng Bayes cho phép suy diễn ra các ngữ nghĩa còn thiếu ở các khung hình mà ở đó các chú thích bằng văn bản không đầy đủ.

Đóng góp chính của bài viết là mô hình truy hồi khung hình video mới, có thể hoạt động với cả trường hợp các mô tả bằng văn bản là không có hoặc không đầy đủ. Thêm vào đó, chúng tôi so sánh một số mô hình truy hồi thông tin với câu truy vấn theo dạng một bộ các quan hệ (relational, nghĩa là một người thực hiện một hành động nào đó tại một địa điểm nào đó) và việc thể hiện tài liệu để truy vấn là không chắc chắn mở đường cho việc tích hợp chứng cứ có được từ nhiều mô thức dữ liệu khác nhau.

Phần còn lại của bài viết được tổ chức như sau. Phần II thảo luận các nghiên cứu có liên quan trong lĩnh vực lập chỉ mục và truy hồi thông tin hướng ngữ nghĩa. Phần III giới thiệu các khái niệm và thuật ngữ được sử dụng xuyên suốt bài báo, sau đó giới thiệu các công việc mà chúng tôi sẽ giải quyết. Phần IV trình bày các cách tiếp cận giải quyết các công việc nêu ở phần III. Phần V mô tả các thiết kế thực nghiệm, các kết quả và các khám phá của chúng tôi. Chúng tôi kết thúc bài viết ở phần VI.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Việc truy hồi video bằng các từ khoá tìm kiếm có ngữ nghĩa là một trong những thử thách lớn nhất trong lĩnh vực xử lý và quản lý video. Nhiệm vụ quan trọng nhất trong hướng nghiên cứu này là lấp đầy khoảng trống giữa các đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao. Về nguyên tắc, một hệ thống truy hồi video cần phải làm được các công việc sau: 1) tìm kiếm một mục thông tin cụ thể và 2) duyệt qua và tóm tắt một tập các dữ liệu thông tin [2]. Để có thể tương tác được với người dùng, một hệ thống truy hồi thông tin đa phương tiện cần có một sơ đồ ánh xạ từ các đặc trưng cấp thấp hàm chứa nội dung của các mục thông tin đến các khái niệm hay điều khoản ở mức cao dễ hiểu hơn đối với người dùng. Người ta đề cập đến khái niệm "khoảng trống về ngữ nghĩa" như là sự thiếu tính trùng hợp (coincidence) giữa những thông tin mà người ta rút ra từ dữ liệu hình ảnh và những diễn giải cho chính dữ liệu đó để cung cấp cho người dùng trong một hoàn cảnh cho trước [3]. Hơn nữa, số lượng các khái niệm ngữ nghĩa là rất lớn và đa dạng. Ví dụ

như các khuôn mặt con người, núi đồi, cảnh bãi biển, bầu trời, đường phố, nhà cửa và nhiều khái niệm nữa. Việc xây dựng một hệ thống truy hồi thông tin mà có thể thỏa mãn mọi truy vấn của người dùng với tất cả các loại khái niệm thường là vượt quá khả năng của các công trình nghiên cứu đương đại. Vì vậy, nghiên cứu của chúng tôi tập trung vào 03 loại khái niệm cơ bản nhưng hữu ích: con người, hoạt động, địa điểm và mối quan hệ giữa chúng. Thực tế, có nhiều nhà nghiên cứu đã và đang tập trung nghiên cứu các phương pháp dùng để học nhằm nhận dạng ba loại khái niệm này từ dữ liệu video. Chẳng hạn, [4, 5, 6, 1, 7, 8, 9] đã biểu diễn những kết quả thú vị trong việc định nhãn cho con người. Việc phát hiện và phân loại hành động của con người đã được nghiên cứu bởi [10, 11, 12], trong khi [13, 14] giải quyết vấn đề phát hiện các địa điểm trong video.

Tuy nhiên, rất ít công trình xem xét kết hợp cả ba loại khái niệm này. Luo và cộng sự [15] đề xuất mô hình kết hợp Expectation – Maximization để định nhãn khuôn mặt và dáng điệu của con người một cách đồng thời. Nitta và cộng sự [16] gán nhãn cho con người và hành động của họ trong các video thể thao bằng cách đầu tiên là dùng văn bản (phụ đề đóng – closed caption) để trích ra các phân cảnh (scenes) cùng với con người, hành động và sự kiện họ xuất hiện, sau đó phân đoạn lại video tương ứng bằng cách sử dụng các đầu mối từ hình ảnh, cuối cùng liên kết các phân đoạn video với các phân đoạn văn bản. Marszalek và cộng sự [17], theo cách khác, trình bày kết quả nghiên cứu về phát hiện hành động và địa điểm trong video chủ yếu dựa trên giả thiết rằng hành động của con người có liên quan cao đến địa điểm mà họ xuất hiện. Trọng tâm trong hướng tiếp cận của họ là khuôn khổ túi các đặc trưng (bag-of-features) dùng cho các mô hình xử lý ảnh nhằm phát hiện các khung cảnh và hành động. Các khuôn khổ và mô hình này được kết hợp với nhau trong một bộ phân loại hỗn hợp hành động-khung cảnh dựa trên kỹ thuật SVM. Bằng cách đề xuất các phương pháp nhằm phát hiện một tập hợp các khái niệm, [15, 16, 17] đã tận dụng mối liên hệ giữa các khái niệm này – điều rất có giá trị trong việc chú thích tự động các khái niệm trong video. Ví dụ như, một khung cảnh "*dưới nước"* thường xuất hiện với một con "*cá mập*" thay vì là một "*con chim*"; hoặc một tập các khuôn mặt giống nhau đồng xuất hiện có hệ thống cùng với cái tên *Bush* trong mô tả văn bản tương ứng nên được gán tên là *Bush*. Nghiên cứu của chúng tôi cũng cố gắng học các mối tương quan giữa các khái niệm và mở rộng ra mối quan hệ giữa 03 khái niệm (con người, hành động, địa điểm) thay vì các mối quan hệ tự thân hoặc hai chiều.

Với ý định tổ chức các tài liệu thành một cấu trúc của các khái niệm ngữ nghĩa, chúng tôi tìm hiểu các mô hình truy hồi có hỗ trợ việc lập chỉ mục và hoạt động được trên một cấu trúc tài liệu như vậy. Mô hình có liên quan nhiều nhất (mô hình đồ thị xác xuất dùng để lập chỉ mục và truy vấn các tài liệu hướng nội dung) được giới thiệu bởi Turtle và Croft [18]. Họ sử dụng các mạng Bayes để mô tả các sự phụ thuộc về xác suất giữa các khái niệm ngữ nghĩa. Các mạng Bayes này được biểu diễn như là các đồ thị có hướng và không có chu trình. Mô hình này bao gồm 02 phần: một mạng của tập các tài liệu (DN) và một mạng truy vấn (QN). Trong mạng các tài liệu, mỗi tài liệu (d) được trình bày như là một cấu trúc phân cấp của các nút thể hiện các tài liệu, các từ của tài liệu và các khái niệm ngữ nghĩa của chúng. Các nút thể hiện các từ và các khái niệm ngữ nghĩa có thể được chia sẻ bởi nhiều tài liệu với xác suất khác nhau. Mạng truy vấn sẽ được xây dựng mỗi khi người dùng đệ trình câu truy vấn của họ (q). Đây cũng là cấu trúc phân cấp của các từ, các khái niệm ngữ nghĩa thể hiện yêu cầu thông tin của người dùng. Sau đó mạng truy vấn sẽ được gắn vào mạng các tài liệu bằng cách so khớp các khái niệm ngữ nghĩa của câu truy vấn và của tài liệu. Các nút trong mô hình này có giá trị nhị phân, nghĩa là nhận giá trị từ tập {true, false}. Việc ước lượng điểm số xếp hạng được thực hiện tách biệt cho từng nút tài liệu. Nghĩa là một nút tài liệu được bật còn các nút tài liệu khác được tắt và điểm số xếp hạng được tính bằng $P(q \mid d)$. Các mô hình truy hồi ngôn ngữ (language retrieval models), theo phân tích của Croft và Laferty [19], có thể được xem như các dạng mô hình đồ thị đơn giản của mô hình mà Turtle và Croft đề xuất. Ở đó, các tài liệu cũng như câu truy vấn được trình bày như là các đồ thị của các nút thể hiện các từ (bag-of-words) mà không có lớp các nút chứa các quan niệm ngữ nghĩa. Và cơ chế so khớp câu truy vấn – tài liệu chỉ đơn thuần sử dụng kỹ thuật so khớp các từ với nhau.

Khởi đầu, mô hình được đề xuất bởi Turtle và Croft được áp dụng cho các tài liệu thuần văn bản. Graves và Lalmas [20] đã mở rộng nó cho các tài liệu video ở khuôn dạng MPEG-7. Ở đó, họ tận dụng các chú thích được kết hợp sẵn trong video (màn – scene, cảnh – shot, đối tượng – object, con người, hành động, địa điểm,…), khai thác các đặc tính của chuẩn MPEG-7 và xây dựng một hệ thống truy hồi video hiệu quả. Coelho và đồng sự [21] trình bày nghiên cứu của họ trong lĩnh vực truy hồi ảnh mà cũng chia sẻ sự quan tâm đến việc sử dụng một mạng Bayes để mô hình hoá các tập tài liệu ảnh, ảnh truy vấn và việc nối kết chúng. Cụ thể là các tài liệu ảnh được trình bày như là các túi từ, nhưng dựa trên các nguồn chứng cứ khác nhau (các thẻ mô tả, thẻ meta, văn bản đầy đủ hoặc đoạn văn bản xung quanh các bức ảnh).

Nghiên cứu của chúng tôi mở rộng các công trình đi trước bằng cách kết hợp các chứng cứ từ các khung hình video và các văn bản đi kèm, bằng cách gán thuộc tính đa trị cho các nút trong mạng Bayes thay vì chỉ là nhị phân và bằng cách suy diễn ra các mô tả cho các khung hình nơi mà các mô tả văn bản không có hoặc thiếu.

Sau cùng, chúng tôi sử dụng kỹ thuật truy hồi video dựa trên các khung hình chính (nghĩa là chúng tôi truy hồi các khung hình chính - dữ liệu được cho là tiêu biểu cho mỗi cảnh – shot). Kỹ thuật này thường được sử dụng trong các hệ thống truy hồi video [22, 23].

III. CÁC ĐỊNH NGHĨA CƠ BẢN VỀ CHÚ THÍCH NGỮ NGHĨA CHO VIDEO VÀ CÁC BƯỚC THỰC HIỆN

Nhiệm vụ của chúng tôi là xây dựng và đánh giá một hệ thống lập chỉ mục và truy hồi video mà nó có thể tự động rút trích ra các khái niệm ngữ nghĩa trong video (con người, hành động và địa điểm), học các mối tương quan giữa chúng, lập chỉ mục cho chúng cùng với các đơn vị video tương ứng (khung hình chính) và xử lý các câu truy vấn của người dùng.

Trong các mô hình truy hồi phía sau, chúng tôi sử dụng các thuật ngữ sau đây:

- Khái niệm ngữ nghĩa (semantic concept): Chúng tôi định nghĩa khái niệm ngữ nghĩa trong video theo ba loại: con người, hành động, địa điểm.
- Bộ khái niệm ngữ nghĩa (semantic concept tuple): Một tổ hợp của một con người đang thực hiện một hành động tại một địa điểm được trình bày như là một bộ <person, action, location>. Các bộ có thể là hoàn chỉnh hoặc không hoàn chỉnh.
- Tài liệu hướng ngữ nghĩa: Trong một đoạn video (ở đây thể hiện bằng một khung hình chính) chúng ta có thể thấy vài bộ khái niệm ngữ nghĩa. Chúng tôi gọi một khung hình video cùng với các bộ tương ứng của nó là một tài liệu hướng ngữ nghĩa.
- Câu truy vấn hướng ngữ nghĩa: Một yêu cầu thông tin của người dùng được trình bày như một bộ quan hệ (ví dụ: một người đang thực hiện một hành động tại một địa điểm, một người tại một địa điểm hoặc chỉ là một con người…).

Cho một video và văn bản mô tả (không đầy đủ) tương ứng, các công việc của chúng tôi là:

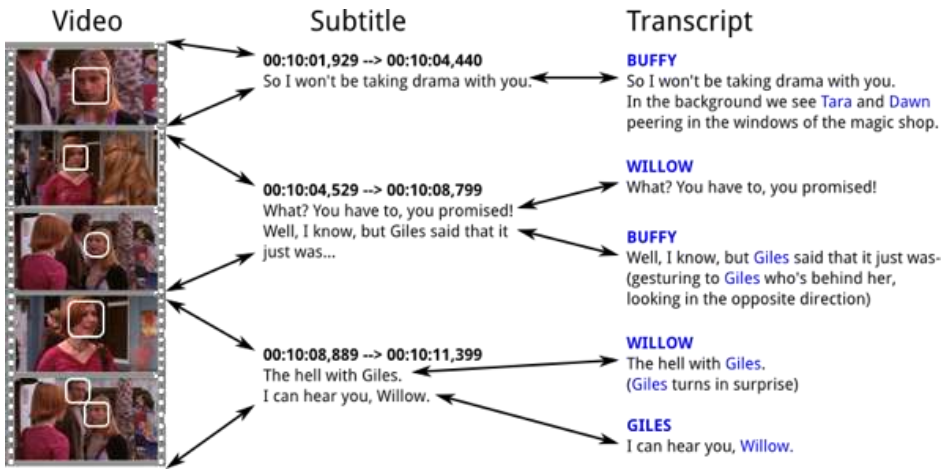
- Tiền xử lý dữ liệu để tạo ra một tập các tài liệu $S=\{d_1,d_2,...,d_{|S|}\}$, trong đó d_i có thể là một túi từ hay một túi các bộ (bag-of-tuples) tuỳ thuộc vào mô hình xử lý.
- Trong quá trình lập chỉ mục cho nội dung của *S*, tất cả các mô hình suy luận với tri thức không chắc chắn rút ra từ dữ liệu với nhiều mô hình xác suất hướng nội dung khác nhau.
- Với câu truy vấn *q*, các mô hình trên sử dụng hàm xếp hạng $r(q,d)$ để sắp xếp tài liệu được truy hồi d_i tương ứng theo độ tương đồng với *q*. Hàm xếp hạng $r(q,d)$ được tính như sau:

$$r(q,d)=logP(q|d) \quad (1)$$

IV. PHƯƠNG PHÁP THỰC HIỆN

Trong các phần tiếp theo, chúng tôi thảo luận quá trình tiền xử lý video nhằm tạo ra các khối dữ liệu cần thiết cho các mô hình lập chỉ mục và truy hồi tiếp theo. Sau đó chúng tôi thảo luận ba mô hình truy hồi theo cấp độ từ đơn giản (Unigram Language Model và Unimodal Entity-Relation Model) đến phức tạp (Multimodal Entity-Relation Model).

A. Tiền xử lý video



Hình 1. Sự cân chỉnh về thời gian xuất hiện của các khuôn mặt và tên

Do chúng tôi mong muốn có thể truy hồi các khung hình video với sự xuất hiện của con người, hoạt động của họ và địa điểm họ xuất hiện, nhiệm vụ đầu tiên của chúng tôi là phát hiện ra các khuôn mặt con người. Xuất phát từ các khuôn mặt người, chúng tôi khám phá các chú thích (phụ đề - subtitles hoặc kịch bản - transcripts) được canh thời gian với video để tìm ra các tên người có thể có. Từ đây, chúng tôi tiếp tục tìm kiếm trong các chú thích này tất cả các hành động hoặc hoạt động có thể có của con người. Cuối cùng chúng tôi xác định các nơi chốn hoặc địa điểm mà những người đã được nhận dạng có thể xuất hiện trong đó.

Trong các tập phim *Buffy the Vampire Slayer*, tên người được rút trích từ hai nguồn thông tin dạng văn bản: các phụ đề có sẵn trong DVD và các kịch bản có được từ website của người hâm mộ [1]. Các kịch bản chứa thông tin về người đang nói, câu nói là gì và mô tả một phần hành động của các nhân vật. Nhưng không có thông tin về thời gian tương ứng với những gì diễn ra trong phim. Còn phụ đề lại chứa thời điểm các câu nói phát ra. Một giải thuật cân chỉnh thời gian [1] được áp dụng để cuối cùng có được thông tin hoàn chỉnh về ai, đang làm gì/nói gì và khi nào. Hình 1 mô tả lại giải thuật cân chỉnh thời gian.

Với các kịch bản phim cùng với thông tin về thời gian có được từ quá trình áp dụng giải thuật cân chỉnh thời gian, chúng tôi so khớp về mặt thời gian xuất hiện của các khuôn mặt trong các khung hình và các tên người trong kịch bản như sau: Với một phân đoạn thời gian của kịch bản, nếu một chuỗi các khuôn mặt có sự giao thoa về thời gian xuất hiện với phân đoạn kịch bản này, tất cả các khuôn mặt trong chuỗi sẽ nhận được tất cả các tên người trong phân đoạn kịch bản làm ứng viên để đặt tên. Hình 1 thể hiện các ví dụ về việc giới thiệu các tên ứng viên cho các khuôn mặt trong video. Lược đồ gán tên chi tiết cho từng khuôn mặt được thực hiện thông qua giải thuật mà chúng tôi đề xuất trong [8]. Tuy nhiên ở đây chúng tôi khởi động giải thuật Expectation Maximization (EM) bằng cách sử dụng bộ phân loại Naïve Bayes, sử dụng dữ liệu huấn luyện bị nhiễu (như đề cập ở trên).

Chúng tôi dựa trên nguồn văn bản để rút trích ra các động từ và địa điểm bằng cách sử dụng giải thuật gán nhãn vai trò ngữ nghĩa có giám sát một phần (semi-supervised semantic role labeling) [26, 27, 14]. Áp dụng kỹ thuật gán nhãn vai trò ngữ nghĩa trong các kịch bản phim chính là việc liên kết tên của nhân vật với hành động và địa điểm. Và sau quá trình gán nhãn tên người cho các khuôn mặt trong các khuôn hình video, các hành động và địa điểm cũng được gắn với từng người cụ thể trong khung hình. Kết quả là việc gán nhãn khởi đầu cho các khung hình với các tên của các khuôn mặt, hành động và địa điểm xuất hiện – nếu những thông tin này là hiện hữu trong kịch bản. Tất cả các chú thích ban đầu được thể hiện dưới dạng các phân phối xác suất. Để ý rằng sẽ có nhiều khung hình không có phần kịch bản đi kèm, hoặc kịch bản không hoàn chỉnh. Vì thế chúng tôi hy vọng rằng khả năng suy diễn của một mạng Bayes có thể giúp luôn tạo ra các chú thích đầy đủ ngay cả khi các khung hình không có nguồn thông tin văn bản đi kèm.

B. Mô hình ngôn ngữ unigram

Khi chỉ sử dụng duy nhất nguồn thông tin văn bản cho việc truy hồi các khung hình video, mô hình ngôn ngữ unigram (Unigram Language Model) cung cấp một cách thức phù hợp và phổ biến để xếp hạng các khung hình (là các tài liệu ngữ nghĩa) ứng với sự tương thích với câu truy vấn. Chúng tôi sẽ sử dụng mô hình này làm mô hình cơ sở trong các thí nghiệm của mình.

Với tập gồm $|S|$ các khung hình video mà ở đây gọi là các tài liệu ngữ nghĩa, $S=\{d_1,d_2,...,d_{|S|}\}$, bài toán trở thành: Làm sao để xếp hạng những tài liệu này với câu truy vấn q bao gồm $|q|$ từ? Chúng tôi giả sử rằng câu truy vấn q hoặc tài liệu d được cấu thành từ một tập các thuộc tính độc lập nhau (ở đây là unigram hoặc các từ đơn): $q=(q^1,q^2,...,q^{|q|})$ và $d=(d^1,d^2,...,d^{|d|})$. Hàm truy hồi trở thành:

$$P(q|d)=P(q^1,q^2,...,q^{|q|}|d)=\prod_{i=1}^{|q|}(\lambda P(q^i|d)+(1-\lambda)P(q^i|S)) \quad (2)$$

với $P(q^i|S)$ là xác suất để rút ra được từ q^i một cách ngẫu nhiên từ tập S; λ là tham số làm mượt. Việc làm mượt này là cần thiết vì có rất nhiều khung hình trong phim Bufy không có kịch bản đi kèm hoặc kịch bản đó không cung cấp thông tin gì hữu ích cho câu truy vấn.

Giá trị của $P(q^i|d)$ và $P(q^i|S)$ có được bằng cách sử dụng phép ước lượng khả năng xuất hiện cao nhất (maximum likelihood) của một từ trong một tài liệu và trong tập $|S|$ tài liệu. λ thông thường được ước lượng từ một tập các phản hồi liên quan (relevant feedback) của người dùng. Ở đây chúng tôi đặt giá trị thực nghiệm $\lambda=0.8$ bởi vì chúng tôi không có phản hồi liên quan cho câu truy vấn q.

C. Mô hình thực thể - quan hệ đơn dạng thức

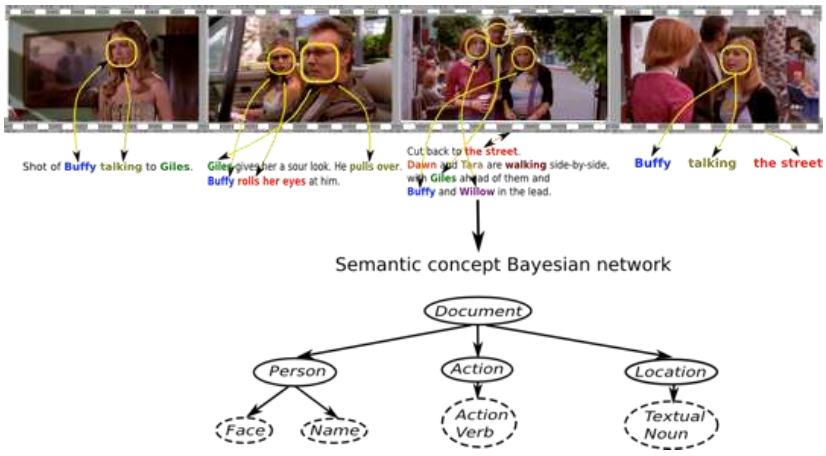
Trong bước tiền xử lý, chúng tôi đã thực hiện phân tích ban đầu cho các khung hình video và đã rút trích ra được tên người, hành động và địa điểm từ các kịch bản tương ứng. Với những thông tin này, chúng ta có thể tạo nên các bộ quan hệ. Bộ gán nhãn vai trò ngữ nghĩa đã cung cấp cho chúng ta mối quan hệ giữa hành động và người thực hiện. Chúng tôi giả sử rằng một khung hình chỉ phản ánh một địa điểm. Vì thế chúng ta có thể tạo ra nhiều bộ chứa các thực thể cũng như các mối quan hệ của chúng (ví dụ như các bộ "*con người, hành động, địa điểm" - <person, action, location>*, "*con người, địa điểm" - <person, location>, ...*).

Với các bộ được rút ra từ các tài liệu và cũng giả sử rằng câu truy vấn cũng được trình bày như một bộ các các khái niệm, *qc*, chúng tôi có thể xây dựng mô hình truy hồi thực thể - quan hệ đơn dạng thức như sau:

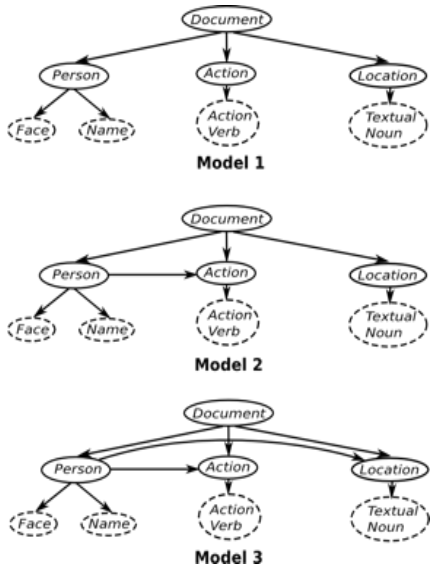
$$P(qc|d)=P(qc^1,qc^2,...,qc^{|qc|}|d)=\prod_{i=1}^{|qc|}(\lambda P(qc^i|d)+(1-\lambda)P(qc^i|S)) \quad (3)$$

Giá trị của $P(qc^i|d)$ và $P(qc^i|S)$ cũng được tính dựa trên phép ước lượng khả năng xuất hiện cao nhất, nhưng ở đây là một khái niệm trong một tài liệu ngữ nghĩa và trong tập các tài liệu ngữ nghĩa S. Ở đây ngưỡng λ cũng được đặt là 0.8.

D. Mô hình thực thể - quan hệ đa dạng thức



Hình 2. Mô hình mạng Bayes dùng để truy hồi tài liệu ngữ nghĩa



Hình 3. Các mô hình mạng Bayes

Trong mô hình thực thể quan hệ đa dạng thức, chúng tôi suy luận trên nguồn thông tin được rút ra từ bước tiền xử lý trong cả dữ liệu hình ảnh và văn bản, ở cả hai cấp độ: cục bộ trong tài liệu ngữ nghĩa và toàn cục bao gồm toàn bộ các tài liệu trong tập phim. Thêm vào đó, một số thông tin ẩn chứa có thể được học bởi mô hình này với hy vọng có thể truy hồi các khung hình còn thiếu thông tin mô tả từ văn bản. Các phương pháp kết hợp chứng cứ bây giờ được cài đặt trong một mạng Bayes đầy đủ. Một mạng Bayes [28] có thể được định nghĩa như là một cặp (G,Q) với: G là một đồ thị có hướng không có chu trình trên các biến Z, được gọi là kiến trúc mạng; Q là tập các bảng xác suất có điều kiện, mỗi bảng dùng cho một biến trong Z, được gọi là tham số hoá mạng.

Mạng suy diễn dùng cho việc truy hồi tài liệu ngữ nghĩa (chúng tôi gọi là *Model 1*), được mô tả trong hình 2, bao gồm một nút biểu diễn tài liệu ngữ nghĩa (Document), các nút biểu diễn các khái niệm ngữ nghĩa (Person, Action and Location) và các nút lá (Face, Name, Action Verb, Textual Noun). Các nút lá thể hiện các quan sát vật lý của ba khái niệm trừu tượng là Person, Action và Location. Những quan sát này có được từ hệ thống tiền xử lý được trình bày ở phần IV.A.

Mô hình mạng 1 được xem là cơ bản vì đã cài đặt được định nghĩa về một tài liệu ngữ nghĩa nhưng không cho phép biểu diễn các mối quan hệ ngữ nghĩa giữa các khái niệm – điều mà nội dung dữ liệu đã thể hiện rõ. Có hai mối quan hệ chính yếu mà chúng tôi mong muốn có thể khai thác từ dữ liệu: 1) con người và hành động của họ, 2) con người và địa điểm xuất hiện của họ. Bởi vì hệ thống truy hồi của chúng tôi tập trung vào việc trả lời cho câu hỏi *ai, làm gì, ở đâu*, nó nên coi con người là trung tâm. Như đã trình bày trong phần IV.A, bộ gán nhãn vai trò ngữ nghĩa rõ ràng đã xác định mối quan hệ giữa con người và hành động của họ. Việc khai thác mối quan hệ này sẽ đem lại lợi lích khi mà hệ thống cần phải dự đoán hành động của một người trong một khung hình mà thông tin mô tả bằng văn bản không có. Thay vì gán tất cả hành động cho một người với xác suất là bằng nhau, hành động mà người đó thường thực hiện như trong dữ liệu mô tả nên là ứng viên nổi trội nhất. Con người và địa điểm xuất hiện thường khó để ghép lại với nhau, tuy nhiên chúng tôi ước lượng mối quan hệ bằng cách sử dụng thông tin về thời gian đồng xuất hiện của hai khái niệm ngữ nghĩa này.

Hình 3 mô tả ba mô hình mạng Bayes mà, ngoài mô hình cơ bản 1, mô hình 2 cài đặt mối quan hệ *con người – hành động*, mô hình 3 cài đặt cả hai mối quan hệ *con người – hành động* và *con người – địa điểm* qua các quan hệ phụ thuộc về xác suất. Chúng tôi sẽ so sánh tác dụng của ba mô hình trong phần thí nghiệm.

E. Ước lượng các tham số của mạng Bayes

Trong mạng tài liệu được trình bày trong hình 2, nút gốc (*Document*) nhận các giá trị rời rạc thể hiện chỉ mục của tài liệu ngữ nghĩa (viết tắt là *d*). Các nút *Person*, *Action* and *Location* cũng nhận các giá trị rời rạc thể hiện các chỉ mục duy nhất của các lớp con người (*pm*), hành động (*am*) và địa điểm (*lm*). Giá trị của các nút lá thể hiện tính chất vật lý về hình ảnh/văn bản của các khái niệm ngữ nghĩa.

Chúng tôi gọi tập các nút X, U là một gia đình nếu các nút U là cha của nút X. Mỗi tập giá trị xu của gia đình này được gọi là một thể hiện của gia đình XU. Tham số mạng $\Theta_{x|u}$ được định nghĩa là $\Theta_{x|u}=P(x|u)$. Công việc của chúng tôi là ước lượng những tham số này biết trước cấu trúc của mạng. Các tham số của mạng có thể được ước lượng theo các cách tiếp cận sau đây.

1. Ước lượng ngẫu nhiên các tham số mạng

Để tìm ra các tham số mạng ngẫu nhiên, Θ, trong mỗi gia đình *XU* biết trước thể hiện *u* của các nút cha, các tham số $\Theta_{x|u}$ được lấy mẫu từ phân phối Dirichlet đối xứng $Dir(p_1,p_2,...,p_{|X|})$ với p_i=0.1 là một giá trị giả lập lại một xác xuất có điều kiện xác định, với một giá trị gần với 1 và các giá trị khác gần với 0.

2. Ước lượng các tham số mạng dựa trên dữ liệu đã được tiền xử lý

Xác suất của một khuôn mặt biết trước một người *P*(*fm*|*pm*) được ước lượng bằng cách sử dụng phương pháp đo mật độ khuôn mặt quanh một nhân (một lớp con người) có chuẩn hoá (normalized kernel density estimation). Việc ước lượng xác suất của một tên biết trước một người *P*(*nm*|*pm*) được thực hiện bằng cách tính khả năng xuất hiện cao nhất (maximum likelihood estimation) của một tên đối với một lớp con người. Khi *nm* không được gán cho khuôn mặt nào trong tài liệu d_i bởi tiến trình gán tên trước đó, *P*(*nm*|*pm*) được ước lượng một cách đồng nhất biết trước số lượng con người trong một tập phim. Với số lượng hành động và địa điểm là có giới hạn trong một tập phim chúng tôi đã xây dựng một từ điển cho các từ chỉ hành động av_i và lớp hành động của chúng, tương tự là từ điển cho các từ chỉ địa điểm lo_i và lớp của chúng dựa trên WordNet (ví dụ như xác định từ "cemetery" thuộc về lớp chỉ địa điểm "graveyard"). Các từ điển này được dùng để ước lượng xác suất của một động từ cho trước lớp hành động *P*(*av*|*am*) và xác suất của một từ chỉ địa điểm biết trước lớp địa điểm *P*(*lo*|*lm*). Các xác suất của một người biết trước một tài liệu *P*(*pm*|*d*), của một lớp hành động biết trước một tài liệu *P*(*am*|*d*) và của một lớp địa điểm biết trước một tài liệu *P*(*lm*|*d*), được ước lượng bằng cách tính khả năng xuất hiện cao nhất trong tài liệu này. Cách tiếp cận tương tự cũng được áp dụng để ước lượng xác suất của một lớp hành động biết trước một người và một tài liệu *P*(*am*|*pm*,*d*). Tất cả các ước lượng đều sử dụng cơ chế làm mượt để giải quyết các trường hợp xác suất bằng không.

3. Ước lượng các tham số mạng sử dụng kỹ thuật EM

Từ các bước tiền xử lý trước đây nhằm nhận dạng con người, hành động và địa điểm, chúng ta có được tập các tài liệu ngữ nghĩa $S=\{d_1,d_2,...,d_{|S|}\}$. Tuy nhiên tập dữ liệu này là không hoàn chỉnh vì trong nhiều trường hợp chúng tôi không thể có được bộ quan hệ đầy đủ ⟨person, action, location⟩.

Giải thuật EM dùng cho việc học các tham số của mạng Bayes được đề xuất bởi [29] là một giải thật rõ ràng và trực quan để ứng phó với việc dữ liệu bị mất hoặc thiếu. Giải thuật này có thể được khởi tạo bởi các phân phối xác suất có điều kiện được định nghĩa trước đó (được ước lượng ngẫu nhiên hoặc bởi các dữ liệu từ bước tiền xử lý) và sau đó có nhiều hy vọng là sẽ tối ưu hoá các phân phối xác suất này. Đầu tiên, bước E sẽ hoàn thiện tập dữ liệu bằng cách điền vào các phần dữ liệu bị thiếu sử dụng các phân phối hiện có. Bước M tiếp sau đó sử dụng tập dữ liệu đã hoàn thiện này để ước lượng lại các tham số mô hình mạng Bayes.

Giải thuật chi tiết được trình bày như sau.

Bước E: Với mỗi tài liệu d_i, tìm tập các biến C_i với giá trị bị thiếu trong d_i. Ước lượng xác suất cho mỗi biến cần hoàn thiện c_i trong d_i, $P_{\Theta^k}(c_i|d_i)$, dựa trên tập các tham số hiện tại Θ^k.

Bước M: Ước lượng lại các tham số Θ^{k+1} biết trước tập dữ liệu *S* và các tham số hiện hành Θ^k như sau [28]:

$$\Theta_{x|u}^{k+1} = \frac{\sum_{i=1}^{|S|} P_{\Theta^k}(xu|d_i)}{\sum_{i=1}^{|S|} P_{\Theta^k}(u|d_i)} \tag{4}$$

Giải thuật EM sẽ hội tụ trong một số hữu hạn các vòng lặp [29].

Các tham số mạng khởi tạo, Θ^0, có thể được ước lượng bởi một tiến trình lấy giá trị ngẫu nhiên (xem phần IV.E.1) hoặc sử dụng dữ liệu ở bước tiền xử lý (xem phần IV.E.2).

F. Truy vấn mạng Bayes

Với mạng Bayes được đề xuất, các tài liệu *d* được xếp hạng tương ứng với câu truy vấn *q* bằng cách ước lượng phân phối hậu biên (*posterior marginal* distribution) *P*(*d*|*q*) [28]. Để cho hàm xếp hạng trong mô hình Multimodal Entity-Relation Model tương thích với các mô hình truy hồi khác, chúng tôi sử dụng công thức Bayes để biến đổi *P*(*d*|*q*) thành *P*(*q*|*d*), mà ở đó, với mục đích là xếp hạng tài liệu, chúng tôi có thể gỡ bỏ một cách an toàn các hằng số (xác suất tiên nghiệm - prior probability của một tài liệu, được xem như là đồng nhất đối với tất cả các tài liệu).

V. THÍ NGHIỆM VÀ KẾT QUẢ

Chúng tôi thực hiện các thí nghiệm trên bộ phim truyền hình *Buffy the Vampire Slayer* season 5, episode 2 (cũng được sử dụng trong [1]). Từ kết quả của các quá trình gán tên cho khuôn mặt, rút trích hành động trong văn bản và nhận dạng địa điểm, chúng tôi lập ra một tập dữ liệu gồm 1474 tài liệu ngữ nghĩa. Trong đó có 14 con người (phân biệt), 51 lớp hành động và 14 lớp địa điểm. Trong tổng số 1474 tài liệu ngữ nghĩa, 128 tài liệu không có thông tin văn bản đi kèm hoặc văn bản đi kèm không chứa thông tin về con người, hành động và địa điểm.

Chúng tôi lập ra các tập dữ liệu sau để sử dụng trong các thí nghiệm: 1) **Dataset A** bao gồm 630 tài liệu ngữ nghĩa đa dạng thức, được lựa chọn ngẫu nhiên từ 1474 tài liệu gốc. Tập dữ liệu này sẽ được dùng để thử nghiệm mô hình ngôn ngữ unigram - Unigram Language Model (*UL-Text*), mô hình thực thể quan hệ đơn dạng thức - Unimodal Entity-Relation Model (*UER-Text*) và mô hình thực thể quan hệ đa dạng thức - Multimodal Entity-Relation Model; 2) **Dataset B** bao gồm 128 tài liệu ngữ nghĩa đa dạng thức không có thông tin văn bản đi kèm. Tập dữ liệu này sẽ được sử dụng để xem xét mô hình Multimodal Entity-Relation Model sẽ học các mối quan hệ giữa các thực thể ngữ nghĩa như thế nào và sau đó dự đoán các thực thể bị thiếu như thế nào. Việc truy hồi tài liệu với các mô hình *UL-Text* and *UER-Text* sẽ không cho kết quả khả quan vì xác suất tương quan giữa tài liệu và câu truy vấn chỉ dựa vào thao tác làm mượt các phân phối xác suất.

Riêng với mô hình Multimodal Entity-Relation Model, chúng tôi so sánh hiệu quả của mô hình với bốn phương pháp khởi tạo các tham số: khởi tạo ngẫu nhiên (*MER-Ran*), khởi tạo bằng các dữ liệu có được từ bước tiền xử lý (*MER-Pre*), khởi tạo ngẫu nhiên và khởi tạo sử dụng dữ liệu ở bước tiền xử lý sau đó cải thiện các phân phối xác suất bằng giải thuật EM (gọi tương ứng là *MER-EM-Ran* and *MER-EM-Pre*).

Trong các thí nghiệm dưới đây, chúng tôi xem xét sáu loại câu truy vấn: $q1$ truy vấn một tên người, $q2$ truy vấn một hành động, $q3$ truy vấn địa điểm, $q4$ truy vấn một người thực hiện một hành động, $q5$ truy vấn một người tại một địa điểm và $q6$ truy vấn một người thực hiện một hành động tại một địa điểm. Với mỗi loại câu truy vấn, tất cả các thể hiện (instantiations) của câu truy vấn được sử dụng và đường cong *precision-recall* trung bình được xác định.

Các kết quả thí nghiệm trên Dataset A với nhiều loại câu truy vấn được trình bày trong hình 5. Chúng thể hiện rằng việc kết hợp các chứng cứ có trong văn bản và hình ảnh (*MER-Ran*, *MER-Pre* and *MER-EM-Pre*) luôn luôn cho hiệu suất truy hồi tài liệu tốt hơn. Thông tin từ bước tiền xử lý giúp tăng thêm hiệu quả truy hồi tài liệu (*MER-Pre* and *MER-EM-Pre*) khi so sánh với các kết quả đạt được từ các mô hình *MER-Ran* và *MER-EM-Ran*. Về hiệu quả của bước tiền xử lý dữ liệu, bộ gán nhãn vai trò ngữ nghĩa nhận dạng các hành động và người thực hiện với độ chính xác hơn 90% [30], trong khi các địa điểm được nhận dạng với độ chính xác 69% [14] . Chúng tôi nhận thấy trong tập dữ liệu rằng các mối quan hệ giữa con người và hành động có thể được tận dụng trong kiến trúc mạng Bayes.

Hình 6 trình bày kết quả truy hồi tài liệu trên Dataset B với nhiều phương pháp ước lượng tham số mạng Bayes. Mô hình *MER-EM-Pre* cho phép lấp đầy các phần thông tin bị thiếu dựa vào nguồn thông tin hiện có. Hình 4 mô tả các khung hình được xếp hạng cao nhất (từ trái sang phải) bởi mô hình truy hồi *MER-EM-Pre* từ Dataset B với nguồn thông tin văn bản bị thiếu.

Chúng tôi sẽ nhận xét các kết quả trên một cách chi tiết như sau:

***Nhận xét* 1**: Việc kết hợp các chứng cứ từ dữ liệu văn bản và hình ảnh sẽ làm tăng hiệu suất truy hồi tài liệu.

Hình 5 thể hiện rõ nhận xét này trên dataset A. Ta thấy rằng mô hình Multimodal Entity-Relation Model vượt trội các mô hình khác theo các số đo precision và recall. Việc kết hợp cả hai mô thức dữ liệu là hình ảnh và văn bản cung cấp cho người dùng nhiều khung hình thích hợp hơn so với việc chỉ sử dụng mô thức văn bản. Đặc biệt trong trường hợp truy hồi con người và địa điểm, mô hình xử lý đa dạng thức luôn luôn vượt trội so với cac mô hình khác. Giá trị của mô hình xử lý đa dạng thức cũng được thể hiện khi chúng ta xét trường hợp truy vấn các hành động. Khi mà các hành động được rút trích ra chỉ từ mô thức văn bản thì các chứng cứ có được không đủ để truy hồi một cách đáng tin cậy các tài liệu liên quan đến hành động.

***Nhận xét* 2**: Thông tin có được từ bước tiền xử lý giúp tăng hiệu suất truy hồi tài liệu.

Hình 5 cũng thể hiện rõ rằng việc sử dụng dữ liệu hình ảnh và văn bản đã được tiền xử lý giúp cải thiện các điểm số recall và precision của quá trình truy hồi tài liệu. Việc sử dụng dữ liệu hình ảnh và văn bản đặc trưng để học các mối quan hệ giữa các khái niệm là quan trọng trong việc khởi tạo mạng Bayes. Trong vài trường hợp, việc ước lượng các tham số mạng sử dụng dữ liệu tiền xử lý còn đạt hiệu quả cao hơn việc dùng giải thuật EM để cải thiện chính những dữ liệu tiền xử lý này. Chúng tôi nghi ngờ rằng tiến trình EM đã làm mượt quá mức các phân phối xác suất.

***Nhận xét* 3**: Việc truy hồi các khung hình được cải thiện bằng cách xem xét thêm các mối quan hệ giữa con người và hành động.

Chúng tôi thực hiện sáu loại câu truy vấn trên mô hình Multimodal Entity-Relation Model với ba kiến trúc mạng (hình 3) trên tập dữ liệu Dataset A. Mặc dù chúng tôi không thấy sự khác biệt rõ nét về hiệu năng trên ba mô hình này, mô hình 2 và 3 dường như thể hiện tốt hơn trong việc phản ánh mối quan hệ giữa con người và hành động (trung bình cải thiện 5% precision). Đây bởi vì các hành động được phát hiện thông qua công cụ định nhãn vai trò ngữ nghĩa mà ở đó hành động được gắn với thực thể thực hiện, chủ yếu là con người. Việc kết hợp giữa con người và địa điểm cải thiện hiệu

năng không nhiều (cải thiện nhiều nhất là 0.7%, diễn ra giữa mô hình 1 và mô hình 3 với loại câu truy vấn *person - location*), vì thế chúng ta cũng không thấy nhiều điểm khác biệt trong Hình 5.

Điều cốt lõi mà chúng tôi rút ra được qua các thí nghiệm trên là việc sử dụng các chứng cứ đa dạng thức trong các thiết kế truy hồi dữ liệu trở nên đầy hứa hẹn, ngay cả khi ở đâu đó trong dữ liệu không có sự hiện diện của một số trường thông tin hoặc các trường thông tin này bị nhiễu. Quan trọng hơn nữa là việc tận dụng các mối quan hệ giữa các khái niệm ngữ nghĩa trong quá trình học của hệ thống là rất hữu ích cho các dự đoán cuối cùng của các trường thông tin bị thiếu này.

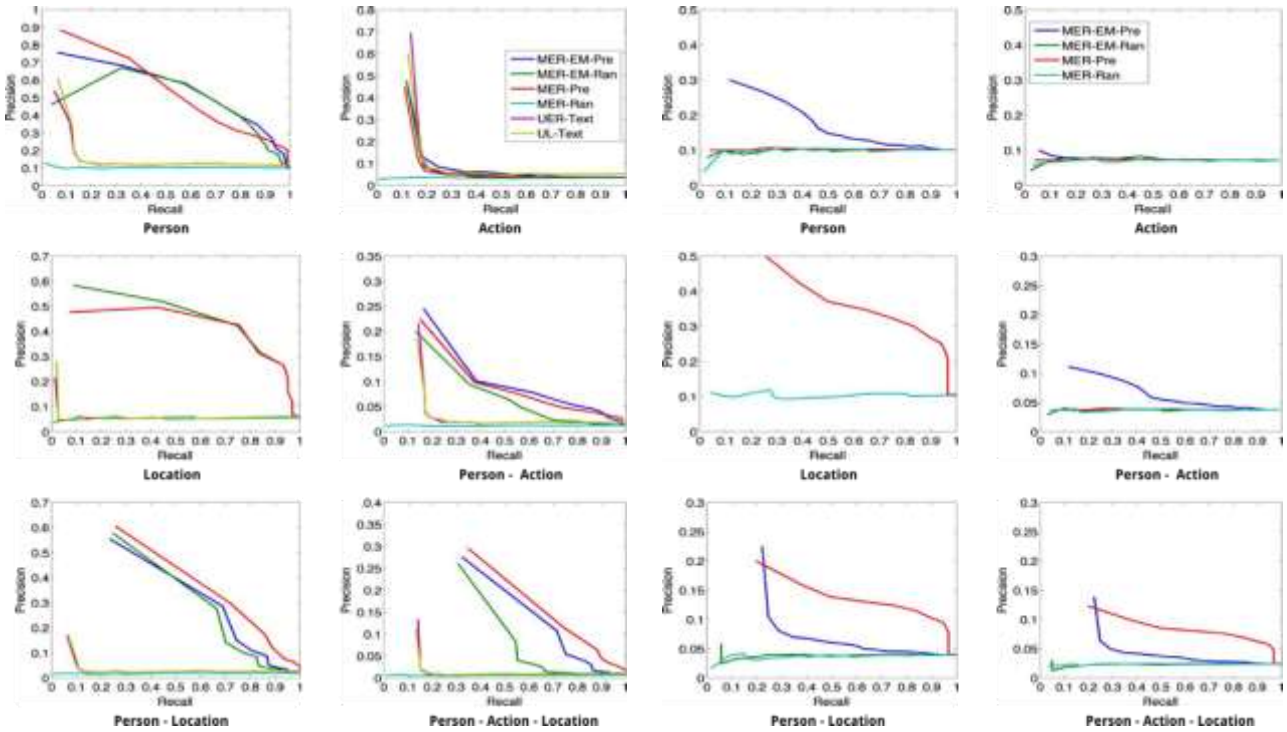
Hình 4. Các khung hình xếp hạng cao nhất được truy hồi từ Dataset B với các ví dụ của 6 loại câu truy vấn

VI. KẾT LUẬN

Chúng tôi đã đề xuất ba mô hình truy hồi video, trong đó mô hình Multimodal Entity-Relation, một mô hình mới dựa trên các chứng cứ có được từ các dạng thức dữ liệu ảnh và văn bản, đã thể hiện vượt trội so với các mô hình còn lại (Unigram Language Model và Unimodal Entity-Relation Model) bởi vì nó đã tận dụng được nguồn thông tin đa phương tiện trong quá trình lập chỉ mục video nhằm học tốt hơn các mối quan hệ giữa các khái niệm ngữ nghĩa. Việc sử dụng năng lực suy luận của mạng Bayes để học các mối quan hệ này được chứng minh là có lợi.

Khi phải dựa vào nguồn thông tin văn bản không hoàn chỉnh hoặc bị thiếu, việc truy hồi các khung hình video có liên quan trở nên khó khăn và đã cho kết quả không khả quan. Chúng tôi đã trình bày cách thức mà mô hình mạng Bayes cho phép việc kết hợp các nguồn thông tin đa dạng thức từ ảnh và văn bản, xử lý các câu truy vấn dạng quan hệ và suy luận các mô tả cho các khung hình video trong trường hợp các mô tả này bị thiếu. Để so sánh với các mô hình Unigram Language và Unimodal Entity-Relation, tất cả các mô hình được thực thi và đánh giá trên một tập của bộ phim truyền hình Buffy the Vampire Slayer, sự gia tăng về hiệu năng truy hồi các khung hình là rõ ràng, đặc biệt là các khung hình thiếu nguồn thông tin văn bản.

Chúng tôi tin tưởng rằng công nghệ mạng Bayes, với khả năng kết hợp chứng cứ để suy luận và khả năng ứng phó với việc thiếu hoặc mất thông tin, có tiềm năng lớn nếu được sử dụng trong các hệ thống truy hồi thông tin đa phương tiện trong tương lai.



Hình 5. Đánh giá việc truy hồi tài liệu trong tập dữ liệu Dataset A với 6 loại câu truy vấn và các kỹ thuật ước lượng tham số khác nhau. Chú ý sự khác nhau ở thang điểm trong các đồ thị

Hình 6. Đánh giá việc truy hồi tài liệu trong tập dữ liệu Dataset B với 6 loại câu truy vấn và các kỹ thuật ước lượng tham số khác nhau. Chú ý sự khác nhau ở thang điểm trong các đồ thị

TÀI LIỆU THAM KHẢO

[1] M. Everingham, J. Sivic, A. Zisserman, "Hello! My name is... Buffy"– automatic naming of characters in TV video, in: Proceedings of the 17th British Machine Vision Conference, 2006, pp. 889–908.

[2] M. S. Lew, N. Sebe, C. Djeraba, R. Jain, Content-based multimedia information retrieval: State of the art and challenges, ACM Transactions on Multimedia Computing, Communications and Applications. 2 (1) (2006) 1–19. doi:10.1145/1126004.1126005. URL http://doi.acm.org/10.1145/1126004.1126005.

[3] A. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content- based image retrieval at the end of the early years, Pattern Analysis and Machine Intelligence, IEEE Transactions on 22 (12) (2000) 1349 –1380. doi:10.1109/34.895972.

[4] S. Satoh, Y. Nakamura, T. Kanade, Name-it: Naming and detecting faces in news videos, IEEE Multimedia 1 (1999) 22–35.

[5] J. Vendrig, M. Worring, Multimodal person identification in movies, in: Proceedings of CIVR 2002: International Conference on Image and Video Retrieval, 2002, pp. 175–185.

[6] J. Yang, A. G. Hauptmann, Naming every individual in news video monologues, in: Proceedings of the ACM Multimedia 2004, 2004, pp. 580–587.

[7] M. Everingham, J. Sivic, A. Zisserman, Taking the bite out of automated naming of characters in TV video, Image and Vision Computing 27 (5) (2009) 545–559.

[8] P. T. Pham, M.-F. Moens, T. Tuytelaars, Cross media alignment of names and faces, IEEE Transactions on Multimedia 12 (1) (2010) 13– 27.

[9] T. Cour, B. Sapp, B. Taskar, Learning from partial labels, Journal of Machine Learning Research 12 (2011) 1501–1536. URL http://dl.acm.org/citation.cfm?id=1953048.2021049.

[10] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of CVPR 2008, 2008, pp. 1 –8. doi:10.1109/CVPR.2008.4587756.

[11] J. Yuan, Z. Liu, Y. Wu, Discriminative subvolume search for efficient action detection, in: Computer Vision and Pattern Recogni- tion, 2009. CVPR 2009. IEEE Conference on, 2009, pp. 2442 –2449. doi:10.1109/CVPR.2009.5206671.

[12] M. Hoai, Z.-Z. Lan, F. D. la Torre, Joint segmentation and classi- fication of human actions in video, Computer Vision and Pattern Recognition, IEEE Computer Society Conference on (2011) 3265– 3272doi:http://doi.ieeecomputersociety.org/10.1109/CVPR.2011.5995470.

[13] F. Schaffalitzky, A. Zisserman, Automated location matching in movies, Computer Vision and Image Understanding 92 (2-3) (2003) 236–264. doi:10.1016/j.cviu.2003.06.008. URL http://dx.doi.org/10.1016/j.cviu.2003.06.008.

[14] C. Engels, K. Deschacht, J. H. Becker, T. Tuytelaars, S. Moens, L. Van Gool, Automatic annotation of unique locations from video and text, in: Proceedings of the British Machine Vision Conference, BMVA Press, 2010, pp. 115.1–115.11.

[15] J. Luo, B. Caputo, V. Ferrari, Who's doing what: Joint modeling of names and verbs for simultaneous face and pose annotation, in: Pro- ceedings of the Twenty-Fourth Annual Conference on Neural Informa- tion Processing System, 2009, pp. 1168–1176.

[16] N. Nitta, N. Babaguchi, T. Kitahashi, Extracting actors, actions and events from sports video - a fundamental approach to story tracking, International Conference on Pattern Recognition 4 (2000) 4718–4721. doi:http://doi.ieeecomputersociety.org/10.1109/ICPR.2000.903018.

[17] M. Marszalek, I. Laptev, C. Schmid, Actions in context, in: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 2929–2936.

[18] H. R. Turtle, W. B. Croft, Inference networks for document retrieval, in: Proceedings of SIGIR 1990, 1990, pp. 1–24.

[19] W. B. Croft, J. Lafferty, Language Modeling for Information Retrieval, Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[20] A. Graves, M. Lalmas, Video retrieval using an MPEG-7 based inference network, in: Proceedings of SIGIR 2002, ACM, New York, NY, USA, 2002, pp. 339–346. doi:10.1145/564376.564436.URL http://doi.acm.org/10.1145/564376.564436.

[21] T. Coelho, P. Calado, L. Souza, B. Ribeiro-Neto, R. Muntz, Image retrieval using multiple evidence ranking, Knowledge and Data Engineering, IEEE Transactions on 16 (4) (2004) 408 – 417. doi:10.1109/TKDE.2004.1269666.

[22] M. J. Pickering, S. Ruger, Evaluation of key frame-based retrieval techniques for video, Computer Vision and Image Understanding 92 (2-3) (2003) 217 – 235. doi:10.1016/j.cviu.2003.06.002. URL http://www.sciencedirect.com/science/article/pii/S1077314203001206

[23] G. C. de Silva, T. Yamasaki, K. Aizawa, Evaluation of video summa- rization for a large number of cameras in ubiquitous home, in: Proceed- ings of Multimedia '05: The 13th Annual ACM International Conference on Multimedia, ACM, New York, NY, USA, 2005, pp. 820–828. doi:10.1145/1101149.1101329. URL http://doi.acm.org/10.1145/1101149.1101329.

[24] P. Viola, M. Jones, Robust realtime object detection vector quantization, International Journal of Computer Vision 57 (2) (2004) 137–154.

[25] J. Shi, C. Tomasi, Good features to track, in: Proceedings of the Conference on Computer Vision and Pattern Recognition, IEEE Computer Society Press, 1994, pp. 593–600.

[26] K. Deschacht, J. De Belder, M.-F. Moens, The latent words language model, Computer Speech and Language 26 (5) (2012) 384 – 409. doi:10.1016/j.csl.2012.04.001.URL http://www.sciencedirect.com/science/article/pii/S0885230812000277.

[27] L. Ma`rquez, X. Carreras, K. C. Litkowski, S. Stevenson, Semantic role labeling: An introduction to the special issue, Computational Linguistics 34 (2) (2008) 145–159. doi:10.1162/coli.2008.34.2.145. URL http://dx.doi.org/10.1162/coli.2008.34.2.145.

[28] A. Darwiche, Modeling and Reasoning with Bayes Networks, Cambridge University Press, 2009.

[29] S. L. Lauritzen, The EM algorithm for graphical association models with missing data, Computational Statistics & Data Analysis 19 (2) (1995) 191 – 201.

[30] K. Deschacht, M.-F. Moens, Semi-supervised semantic role labeling us ing the latent words language model, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1, EMNLP '09, ACL, Stroudsburg, PA, USA, 2009, pp. 21–29. URL http://dl.acm.org/citation.cfm?id=1699510.1699514.

VIDEO RETRIEVAL: WHO IS DOING WHAT AND WHERE?

Pham The Phi, Do Thanh Nghi

ABSTRACT*— This paper proposes a novel method for the retrieval of video frames when incomplete textual annotations are available. The idea is using Bayesian inference for guessing potential textual descriptors about the actors, actions and locations in the frames. Several probabilistic retrieval models that incorporate evidence from the visual and incomplete textual data are evaluated and compared. For our experiments we use the soap videos of Buffy the Vampire Slayer.*

Keywords*— Multimedia data mining, video indexing and retrieval.*