HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN NGỌC HÙNG ANH

NGHIÊN CỨU GIẢI PHÁP PHÂN TÍCH HÀNH VI NGƯỜI DÙNG QUA MẠNG HỌC SÂU NHẰM THIẾT KẾ GIẢI THUẬT TƯ VẤN KÊNH CHO NGƯỜI XEM TRUYỀN HÌNH

Chuyên ngành: HỆ THỐNG THÔNG TIN
Mã số: 8.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ
(Theo định hướng ứng dụng)

TP. HỒ CHÍ MINH – NĂM 2022

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS TRẦN THU HÀ**

Phản biện 1: ……

Phản biện 2: ……...

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ..... ngày..... tháng ..... năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, Ngành Công nghệ thông tin đã và đang được phát triển rất mạnh về phần cứng và cũng như phần mềm. Với sự phát triển đó, có một lĩnh vực cũng đang phát triển rất mạnh, cũng là xu thế trong tương lai và là một sự kết hợp giữa sự phát triển của phần cứng lẫn phần mềm đó là lĩnh vực dịch vụ phát sóng Truyền hình trên Internet.

Để duy trì dịch vụ Truyền hình trên Internet, mô hình OTT (Over The Top) là giải pháp cung cấp nội dung cho người sử dụng dựa trên nền tảng Internet cung cấp bởi bên thứ ba. Công nghệ OTT cho phép cung cấp các nguồn Truyền hình có nội dung phong phú đa dạng theo yêu cầu của người sử dụng vào bất kì thời điểm nào, tại bấy kì nơi đâu chỉ với một thiết bị phù hợp với ứng dụng và có kết nối Internet.

Trên thế giới, công nghệ OTT đã làm thay đổi bộ mặt của dịch vụ truyền hình số cổ điển. Cùng với sự phát triển của các thiết bị công nghệ hiện đại như điện thoại, máy tính, Smart Tivi và các phương tiện kỹ thuật số.

Nhằm giúp cho người sử dụng có thể nhanh chóng tìm ra nội dung muốn xem, mô hình OTT đã có những tiện ích như sau:

- Tạo ứng dụng chương trình xem lại kênh vừa mới xem ngay trước đó. Tâm lý là người xem thường chọn cho mình thêm một chương trình dự bị khi kênh đang xem không còn cuốn hút (do quảng cáo, do trục trặc kỹ thuật), chính vì thế việc luân chuyển giữa hai kênh thường xem, chỉ sử dụng một nút nhấn là cách rất hiệu quả giúp người xem nhanh chóng xem được chọn lựa của mình.
- Tạo danh sách các kênh yêu thích, giảm số lượng hàng trăm kênh xuống thành một vài kênh mà người xem quan tâm nhất.
- Tạo các chủ đề để phân loại các chương trình xem lại như kênh tổng hợp, ca nhạc, phim, v.v… Nhờ đó mà người xem sẽ nhanh chóng hơn khi chọn được chủ đề và chương trình để xem.

Tiện ích thứ 3 chỉ áp dụng được cho các nội dung xem lại, VoD (Video on Demand). Đối với các kênh truyền hình trực tiếp, chưa thể xem các chương trình phát sóng theo chủ đề riêng. Việc sử dụng lịch phát sóng truyền thống vẫn là giải pháp được áp dụng rộng rãi ở các kênh truyền hình: các chương trình phát sóng được liệt kê theo lần lượt theo thứ tự thời gian và cho từng đài / kênh phát sóng. Người sử dụng phải chọn kênh phát sóng để xem

chương trình đang phát có đúng chủ đề mình cần xem hay không. Thông tin về nội dung chương trình phát sóng có thể được mô tả trong lịch phát sóng. Tuy nhiên người xem phải đọc một cách "thủ công" tất cả thông tin này cho từng chương trình phát sóng để tìm ra đúng nội dung yêu thích.

Với hạn chế nêu trên khi tìm kiếm chương trình truyền hình muốn xem, chúng ta có thể ứng dụng những tiến bộ của công nghệ để cung cấp dịch vụ cho người dùng một cách tối ưu hơn nên em chọn đề tài **"Nghiên cứu giải pháp phân tích hành vi người dùng qua mạng học sâu nhằm thiết kế giải thuật tư vấn kênh cho người xem truyền hình"** cho luận văn Thạc sĩ này. Mục đích là cải thiện chất lượng thời gian tìm kiếm thông tin của chủ đề và gợi ý những nội dung tiếp theo giúp cho người xem dễ dàng xem những chủ đề yêu thích một cách nhanh nhất.

2. Mục đích nghiên cứu

Nghiên cứu phân tích hành vi người dùng qua mạng học sâu và thiết kế giải thuật tư vấn kênh cho người xem truyền hình:

- Nghiên cứu, phân loại đoạn văn tiếp nhận đầu vào và dùng các mô hình phân tích biết trước để xử lý đoạn văn của chương trình truyền hình và phân loại nhóm theo tựa đề của chương trình phát sóng trong lịch phát

sóng truyền thống và gán thành các nhãn là tên của chủ đề trong giao diện dịch vụ tìm kiếm. Đây là một giải pháp nâng cao chất lượng dịch vụ trong Truyền hình sẽ tiết kiệm thời gian tra cứu kênh và nội dung theo chủ đề cho người xem.

- Nghiên cứu ứng dụng thuật toán Kmeans trên cơ sở các quy luật xác định, đề xuất các tiêu chí để đánh giá, phân loại nội dung, tần suất xuất hiện của các cụm từ, các cấu trúc văn phạm, cách dùng từ, các diễn giải để làm cơ sở xác định chủ đề của nội dung Truyền hình. [4]
- Nghiên cứu và thiết kế giải thuật phân biệt câu từ, ngữ pháp, động từ , danh từ thuộc cấu trúc câu và tiến hành "đào tạo" các thuộc tính. Các nội dung sẽ được huấn luyện và gán vào một chủ đề tương ứng [2].
- Tiến hành thử nghiệm sản phẩm giúp người dùng có thể tìm kiếm được kênh truyền hình và biết thông tin kênh sẽ có nội dung mong muốn xem tiết kiệm thời gian tạo cảm giác thoải mái cho người dung đầu cuối khi giải trí.

3. Đối tượng và phạm vi nghiên cứu

❖ ***Đối tượng nghiên cứu:***

Biến đổi dữ liệu thô thu được từ các trang web có lịch phát sóng Truyền hình để phục vụ mục đích nghiên cứu.

Sử dụng thuật toán K-means clustering để phân loại và bổ sung theo luật xác định để tìm ra chủ đề của chương trình Truyền hình.

Sử dụng phương pháp tự động phân loại và bổ sung theo từng chủ đề của chương trình Truyền hình dựa vào mô hình máy học PhoBERT.

So sánh các phương pháp phân loại đoạn văn như: Bert, PhoBert

❖ **Phạm vi nghiên cứu:**

Dựa vào các quy luật xác định để phân tích được số lần xuất hiện của các cụm từ, cấu trúc văn phạm của người dùng yêu cầu để làm cơ sở xác định cho việc quyết định nhóm gợi ý cho người xem.

Dựa vào hỗ trợ của mô hình máy học PhoBERT để phân tích tự động nội dung chủ đề và bổ sung theo từng chủ đề yêu thích của người xem.

Mô hình OTT được chia thành ba thành phần chính, thực hiện những chức năng một cách tuần tự như sau:

➢ Thu thập thông tin từ trạng thái của hệ thống.

- Nhận yêu cầu từ bộ phận người dùng, xây dựng mô hình và ra quyết định.
- Nhận lệnh và thực thi.

4. Phương pháp nghiên cứu

Luận văn này sử dụng các phương pháp nghiên cứu lý thuyết và kết hợp với xây dựng ứng dụng thử nghiệm:

- Thu thập các tài liệu, thông tin có liên quan tới đề tài để phục vụ nghiên cứu.
- Ứng dụng các công nghệ lập trình python và các công nghệ trong lĩnh vực máy học như: BERT, PhoBERT, v.v… để so sánh, phát triển hệ thống thử nghiệm
- Tiến hành đánh giá kết quả thử nghiệm, đưa ra hướng phát triển mở rộng của để tài để đáp ứng những nhu cầu triển khai thực tế.

CHƯƠNG 1: CƠ SỞ LÝ LUẬN

Chương này luận văn giới thiệu khái quát về vai trò của OTT trong dịch vụ truyền hình Internet. Hiệu quả của tính năng trong quá trình điều chỉnh nội dung để thích ứng với nguồn phát. Phân loại nội dung của chương trình phát theo từng nhóm của chủ đề. Hiệu quả của việc phân loại chương trình theo nội dung truyền tải. Giúp cho chúng ta thấy được tầm quan trọng của việc phân loại nội dụng của kênh Truyền hình. Gợi ý cho người xem thông qua sở thích và thói quen của họ.

1.1 Tổng quan về mô hình OTT

Mô hình OTT (Over The Top) là giải pháp cung cấp các nội dung cho người dùng như âm thanh, hình ảnh trên nền tảng Internet độc lập, với mô hình công nghệ OTT, những nội dung truyền hình được phân phối qua nhiều hạ tầng Internet, không nhất thiết sở hữu bởi nhà cung cấp dịch vụ. Đây là điểm khác biệt so với các dịch vụ truyền thống như truyền hình cáp, truyền hình vệ tinh.

Với sự phát triển của các thiết bị công nghệ như smartphone, Smart TV đã làm thay đổi các nhà mạng cũng như dịch vụ truyền hình, đặt biệt là trong khoảng 10 năm qua, và chắc chắn sẽ còn rất nhiều thay đổi trong những năm tiếp theo. Từ đó

mô hình OTT đang ngày càng sử dụng phổ biến trong lĩnh vực Internet và đã mở ra nhiều cơ hội mới cho các nhà cung cấp dịch vụ truyền hình như Netflix, VTVGo, SCTV Online, v.v…

Tại Việt Nam dịch vụ truyền hình Internet phát qua Smart TV và ứng dụng truyền hình phát trên các thiết bị di động ngày càng phổ biến và tăng mạnh, các nhà cung cấp truyền hình OTT luôn đầu tư và phát triển với nội dung chất lượng cao và đa dạng hơn, giúp cho người dùng dễ dàng xem và chọn lựa nội dung mình yêu thích dễ dàng nhất.

1.2 Mô hình IPTV truyền thống

IPTV có thể xem là thế hệ tiền thân của truyền hình trên nền tảng OTT. Trên hệ thống IPTV, dịch vụ truyền hình số được được cung cấp qua thiết bị đầu cuối Set-top-box (STB). Qua thiết bị này, thuê bao có thể xem các kênh, thực hiện dịch vụ thuê bao cũng như các dịch vụ tương tác đa phương tiện khác thông qua nền tảng kết nối trực tiếp – quản lý bởi chính nhà cung cấp dịch vụ (managed IP). Bản chất kết nối giữa STB và nhà cung cấp dịch vụ là dựa trên nền tảng IP, nên dịch vụ IPTV có thể dễ dàng được cung cấp cùng với dịch vụ Internet khác như truy cập trang Web, điện thoại qua Internet, v.v…

- Hỗ trợ truyền hình có tính tương tác 2 chiều: tạo điền kiện cho việc cung cấp đa dạng các ứng dụng truyền hình có tính tương tác cao như truyền hình trực tiếp với nhiều góc quay, truyền hình có độ nét cao theo yêu cầu, các trò chơi truyền hình tương tác, v.v…
- Xem lại chương trình của kênh truyền hình: kết hợp với chức năng ghi hình cho phép người dùng xem lại chương trình đã phát sóng ở một thời điểm khác trước đây.
- Cải thiện trải nghiệm riêng biệt khi xem truyền hình: nhờ tương tác 2 chiều với nhà cung cấp dịch vụ thông qua STB, người dùng có thể chọn lựa kênh muốn xem và thời gian xem cho phù hợp với thị hiếu của mình.
- Sử dụng băng thông một cách hiệu quả: công nghệ IPTV bảo đảm chỉ phát kênh lên hạ tầng truyền dẫn khi có người yêu cầu. Chính thế dù có khả năng cung cấp rất nhiều chương trình cùng một thời điểm, băng thông của hạ tầng cũng được sử dụng một cách hợp lý.
- Giải trí thư giãn xem truyền hình qua nhiều thiết bị đầu cuối, hệ thống IPTV cung cấp nội dung không chỉ trên TV mà còn có thể trên PC hay trên điện thoại thông minh kết nối trực tiếp với mạng nội bộ của STB.

1.3 Các khó khăn thách thức trong dịch vụ truyền hình Internet

Với những khó khăn và các bước kỹ thuật phát triển cũng như dịch vụ kinh doanh chính của một mô hình OTT tiêu biểu. Với bất cứ mô hình nào, các đặc điểm chính của việc triển khai OTT luôn đòi hỏi giải pháp cho các vấn đề sau:

- Số lượng truy cập lớn: không quá bất thường là hiện tượng các gói OTT tạo ra hơn 2,5 triệu người xem trong những tuần đầu triển khai.
- Mô hình mua bản quyền xem truyền hình: có thể mua bản quyền xem phim trên truyền hình tại 1 thiết bị và xem phim đấy qua các thiết bị khác trong nhà.
- Mô hinh OTT theo cơ chế bảo mật, chỉ cho phép người dùng đã có bản quyền xem có thể tận hưởng các phim có trong chương trình TV.

Trong quá trình khảo sát chi tiết các môi trường phát triển OTT khác nhau, và ta có những thách thức như sau:

- Khả năng cung cấp nội dung từ nhiều nguồn khác nhau và cho nhiều định dạng cũng như độ phân giải khác nhau.

- Sự đa dạng về số lượng, chất lượng và sự hỗ trợ tính năng khác nhau của thiết bị đầu cuối.
- Tính năng bảo mật nội dung, sự linh động trong việc mua quyền sử dụng.
- Khả năng tích hợp với các hệ thống hỗ trợ vệ tinh đang hoạt động với dịch vụ IPTV như CDN, CMS.
- Khả năng tìm kiếm, phát hiện và nhận tư vấn để có thể tìm ra các nội dung phù hợp.

1.4 Các phương pháp phân loại văn bản

Bài toán mô hình phân loại văn bản thường có hai cách phân loại khác nhau là: phân loại dựa trên luật và phân loại dựa trên máy học.

Phân loại dựa trên luật là cách phân loại được cho là đơn giản nhất để phân loại các dạng văn bản. Việc phân loại nội dung câu văn dựa vào các luật ngữ pháp tiếng việt. Các luật này có được là do nghiên cứu và đề xuất từ các chuyên gia. Đối với cách phân loại này, một loạt các biểu thức được tạo ra để so sánh với các nhãn từ đó đưa ra quyết định phân loại nội dung văn bản và nhãn của văn bản.

Phân loại dựa trên máy học là cách tiếp cận được sử dụng phổ biến rộng rãi để giải quyết bài toán phân loại nội dung văn

bản. Cách tiếp cận này sẽ thay thế các kiến thức chuyên môn bằng một tập dữ liệu lớn các nội dung tiêu đề đã được gán nhãn (tập dữ liệu mẫu).

Cách tiếp cận dựa trên học máy được chia làm hai nhóm là nhóm các phương pháp học máy truyền thống và nhóm các phương pháp sử dụng mạng nơ-ron (Neural NetWork). Nhóm các phương pháp học máy truyền thống thường được sử dụng như là tính xác suất Naïve Bayes, Maximum Entropy, Máy Vector hỗ trợ (Support Vector machine - SVM),... Cách tiếp cận bằng học máy đã giải quyết được các hạn chế trong cách tiếp cận dựa trên luật.

CHƯƠNG 2: PHÂN TÍCH THIẾT KẾ ỨNG DỤNG

Chương 2 tập trung vào thiết kế các phương pháp phân loại văn bản theo dạng chủ đề, dùng mô hình phân tích để xử lý các chủ đề và đưa ra kết quả phân loại theo từng nhóm của chủ đề.

2.1 Sơ lược về phân loại nội dung tiêu đề trong mô hình OTT

Phân loại nội dung chương trình phát sóng truyền hình có thể được quy đổi về bài toán lớn hơn là phân loại văn bản, phân loại câu văn hay từ vựng. Đây là các bài toán cơ bản về Xử lý Ngôn ngữ Tự nhiên (NLP Natural Language Processing). Bài toán phân loại tên chương trình được mô hình hóa qua mạng học sâu (deeplearning) với mô hình chuyển đổi giữa các câu văn (sequence-to-sequence Model). Dữ liệu đầu vào được gán nhãn và mô hình sẽ học từ dữ liệu được gán nhãn cho trước, sau đó sẽ được dùng để dự đoán các nhãn tương ứng cho các dữ liệu mới trong mô hình.

2.2 Quy trình phân loại nội dung tiêu đề trong mô hình OTT

Phân loại nội dung tiêu đề được xây dựng mô hình các thuật toán học máy sẽ huấn luyện một bộ phân loại sử dụng các

vector thuộc tính của dữ liệu ở trên. Có nhiều mô hình học máy có thể được sử dụng để huấn luyện tạo ra mô hình cuối cùng. Trong nghiên cứu này đã sử dụng mô hình PhoBert để huấn luyện bao gồm lớp đầu vào, các lớp ẩn và lớp đầu ra.

Mô hình phân loại dữ liệu gồm hai giai đoạn*:*

- Giai đoạn huấn luyện (training) là giai đoạn học tập của mô hình phân loại văn bản. Trong bước này, mô hình sẽ học từ dữ liệu có nhãn. Dữ liệu văn bản sẽ được số hóa thông qua bộ trích xuất đặc trưng (feature extractor) để mỗi mẫu dữ liệu trong tập huấn luyện trở thành một vector nhiều chiều (đặc trưng). Thuật toán máy học sẽ học và tối ưu các tham số để đạt được kết quả tốt trên tập dữ liệu. Nhãn của dữ liệu được dùng để đánh giá việc mô hình học tốt không và dựa vào đó để tối ưu.
- Giai dự đoán (prediction), là giai đoạn sử dụng mô hình học máy sau khi nó đã học xong. Ở giai đoạn này, dữ liệu cần dự đoán cũng vẫn thực hiện các bước trích xuất đặc trưng. Mô hình đã học sau đó nhận đầu vào là đặc trưng đó và đưa ra kết quả dự đoán.

2.3 Thuật toán K-Means

K-means là thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Ý tưởng của thuật toán phân cụm K-Means là phân chia một bộ dữ liệu thành các cụm khác nhau. Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý khác nhau. Các điểm dữ liệu trong cùng một cụm thì phải có cùng một số tính chất nhất định. Tức là giữa các điểm trong cùng một cụm phải có sự liên kết lẫn nhau. Đối với máy tính thì các điểm trong một cụm đó sẽ là các điểm dữ liệu gần nhau.

2.4 Giới thiệu mô hình BERT

BERT (Bidirectional Encoder Representations from Transformers) (tạm dịch: Mô hình mã hóa hai chiều dữ liệu từ các khối Transformer), là một phương pháp kỹ thuật được xây dựng dựa trên mô hình mạng mô phỏng theo hệ thống nơ-ron thần kinh của con người (neural network) dùng để đào tạo trước (pre-train) quá trình xử lý ngôn ngữ tự nhiên. Nói một cách đơn giản, thì nó có thể được sử dụng để giúp Google phân biệt rõ hơn ngữ cảnh của các từ xuất hiện trong truy vấn tìm kiếm.

CHƯƠNG 3: TRIỂN KHAI ỨNG DỤNG

Trong chương này trình bày các phương án cài đặt phương pháp tự động phân loại và bổ sung theo từng chủ đề của chương trình truyền hình đã được xây dựng ở Chương 2.

3.1 Xây dựng bộ dữ liệu

Với mô hình OTT thì khi người dùng muốn xem danh sách các chương trình có thể xem vào các thời điểm nào đó. Người dùng có thể nhập tên một thể loại bất kỳ thì đầu ra của danh sách sẽ khởi tạo một danh sách chương trình phát sóng theo thể loại của người dùng yêu cầu, không lệ thuộc vào khái niệm "Kênh" cổ điển, chỉ tập trung vào nội dung / chủ đề của nội dung mà người thuê bao quan tâm. Hình 3.1 thể hiện sơ đồ chức năng hiển thị danh sách các kênh cho người dùng theo từng chủ đề.

3.2 Thiết lập thực nghiệm

Với bộ dữ liệu chuẩn bị cho thiết lập thực nghiệm, luân văn lấy được 1000 tiêu đề của lịch phát sóng truyền hình theo pháp quy tiếng việt. Luận văn được chia thành 8 nhãn.

Để đánh giá kết quả của việc xác định thực thể và thuộc tính ta đánh giá thông qua độ chính xác (precision), độ bao phủ (recall), tính cân bằng của độ chính xác và độ bao phủ (F1) được xác định như sau:

$$\text{precision} = \frac{\text{số nhãn gán đúng}}{\text{tổng số nhãn được gán}}$$

$$\text{recall} = \frac{\text{số nhãn gán đúng}}{\text{tổng số nhãn thực tế}}$$

$$\text{F1} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

3.3 Công cụ thực nghiệm

Luận văn được lấy dữ liệu phát sóng từ trang web VTV: https://vtv.vn/lich-phat-song.htm

Dữ liệu khoảng 1000 tiêu đề phát sóng của dịch vụ truyền hình VTV mỗi ngày. Nội dung bao gồm các chủ đề như: phim truyện, thời sự, thể thao, giải trí, ca nhạc, kỹ năng sống, trẻ em, du lịch.

3.4 Các mô hình thực nghiệm

Luận văn sử dụng hai công cụ thực nghiệm là sklearn svm Linear SVC sử dụng cho mô hình SVM và simpletransformers sử dụng cho hai mô hình còn lại là BERT multilingual và PHOBERT

3.5 Kết quả thực nghiệm

Kết quả phân loại nhị phân theo từng nhãn của các mô hình khá trên lệch. Các nhãn được phân loại theo mô hình PhoBert điều đạt kết quả khá tốt, đều trên 85%.

CHƯƠNG 4: ĐÁNH GIÁ KẾT QUẢ THỬ NGHIỆM

Chương này mô tả chi tiết về việc thử nghiệm cải tiến mô hình OTT trong lĩnh vực phân loại nội dung chương trình phát sóng truyền hình.

4.1 Mô tả kết quả phân loại chương trình

Mô tả kết quả giao diện hiển thị danh sách các kênh truyền hình được trình chiếu trong ngày. Người dùng nhập vào ô tìm kiếm để tìm danh sách các kênh đang được trình chiếu trong khung giờ của lịch phát sóng VTV.

Ví dụ: Người dùng nhập "Thể thao 24 giờ" thì danh sách theo chủ đề thể thao sẽ hiện thị lên theo các kênh của VTV. Người dùng chọn kênh theo nôi dụng mà họ yêu thích hoặc tìm các khung giờ mà họ có thể xem.

4.2 Kết luận

Phân loại chương trình truyền hình theo nội dung / chủ đề có thể coi là 1 phương pháp giúp người xem tiếp cận với nội dung cần xem 1 cách nhanh chóng hơn, tiện lợi hơn.

Các tựa đề của chương trình truyền hình thường là 1 mệnh đề ngắn, không phải là 1 câu dài trọn vẹn. Chính vì thế việc phân tích ngữ nghĩa theo phương pháp xử lý ngôn ngữ tự nhiên (NLP) có thể hơi quá phức tạp và không hiệu quả. Đề tài sẽ tập trung

nhiều tuần dữ liệu các chương trình truyền hình, tập hợp các từ khóa thường gặp cho mỗi chủ đề cho trước. Việc phân loại một mệnh đề sẽ là cách tính "khoảng cách nhỏ nhất" của một mệnh đề và sự xuất hiện trong mệnh đề đó các từ khóa thuộc về một chủ đề / tiêu chí nhất định.

4.3 Kiến nghị hướng nghiên cứu tiếp theo

Tiến hành cài đặt đánh giá phương pháp tự động để phân bổ nội dung theo kênh phát trong môi trường Internet thực tế.

Đề tài cần nghiên cứu phát triển thêm các giải pháp, thuật toán AI (Deep learning, Machine learning) để phân tích điều khiển dạng Text và giọng nói để phân loại đáp ứng nhu cầu thực tế, giúp cho người dùng có thể tìm kiếm nhanh nội dung chương trình truyền hình muốn xem, gợi ý cho người dùng các chủ đề theo sở thích, thói quen trong khoảng thời gian ngắn nhất và chính xác nhất.

Cải thiện giao diện để thân thiện hơn cho người dùng bằng tính năng tương thích theo chương trình phát sóng. Đề tài đã đề xuất một phương thức khác hẳn với truyền hình truyền thống để xem Tivi. Người xem sẽ xác định chủ đề mình muốn xem khi bật Tivi.

KẾT LUẬN

Đề tài phân loại chương trình truyền hình Internet theo nội dung là một trong những phương pháp giúp cho người xem tiếp cận với nội dung Truyền hình trong khoảng thời gian ngắn nhất.

Với tính năng tương thích chương trình, lịch xem truyền hình quen thuộc được phân tích lại. Kênh truyền hình không còn giữ vai trò quan trọng trong việc tìm kiếm nội dung cần xem. Từ khóa tìm kiếm ở đây là sự phân loại thực sự của chương trình phát sóng. Việc tìm kiếm từ khóa có trong tên chương trình, việc sắp xếp các chương trình theo chủ đề của nội dung sẽ cải tạo hoàn toàn giao diện của lịch phát sóng cung cấp trong đề tài. Nhằm giảm khó khăn cho người xem trong việc phân biệt kênh không có tín hiệu và kênh có độ trễ để có thể nghe / nhìn nội dung lần đầu tiên khi vào xem truyền hình.

Đề tài đã nghiên cứu khái quát về các tựa đề của chương trình Truyền hình và tự động phân tích ngữ nghĩa theo phương pháp xử lý ngôn ngữ tự nhiên để phân lớp thành nhiều tuần dữ liệu của các chương trình truyền hình, tập hợp các từ khóa thường gặp cho mỗi chủ đề để gợi ý trước cho người xem. Tên tựa đề thường là 1 mệnh đề không trọn vẹn (không đủ thành 1

câu trọn nghĩa), việc phân tích bằng học sâu sẽ không thể cho kết quả tốt. Tựa đề là tiếng Việt Nam, các kiểu chơi chữ trong dùng từ để tựa đề thêm súc tích, v.v… cũng gây ảnh hưởng không nhỏ đến kết quả dự đoán của hệ thống.

Việc mất cân bằng trong chủ đề phát sóng của truyền hình (trẻ em và du lịch chỉ chiếm phần nhỏ trong toàn chương trình truyền hình) cũng tác động đến sự sai lệch trong việc dự đoán chủ đề. Việc sử dụng mô hình PhoBert được huấn luyện trước với nhiều từ việt ngữ hơn, việc tự động tìm kiếm qua các nguồn tìm kiếm trên Internet để hiểu thêm môi trường cho những tựa đề của chương trình, làm phong phú thêm về số lượng và chất lượng của tập tin học thêm cho mô hình huấn luyện là những hướng đi khả thi và hứa hẹn sẽ cải thiện tốt hơn việc nhận biết / phân loạt tên chương trình một cách nhanh chóng và chính xác.