

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN ĐỖ HẢI

NGHIÊN CỨU GIẢI PHÁP TƯ VẤN LẠI SỬ DỤNG ĐỒNG HUẤN LUYỆN

Chuyên ngành: Hệ thống thông tin

Mã số: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. Từ Minh Phương

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

MỤC LỤC

MỞ ĐẦU	1
CHƯƠNG 1 - GIỚI THIỆU	4
1.1 Hệ thống tư vấn.....	4
1.1.1 Mục đích.....	4
1.1.2 Các thành phần chính	4
1.1.3 Một số tác vụ chính của hệ thống tư vấn.....	4
1.2 Một số phương pháp tư vấn đã phát triển	4
1.2.1 Phương pháp tư vấn dựa trên lọc cộng tác	4
1.2.2 Phương pháp tư vấn dựa trên nội dung	5
1.2.3 Phương pháp tư vấn dựa trên thông tin cá nhân	5
1.2.4 Phương pháp tư vấn lai.....	5
1.3 Kết luận chương.....	5
CHƯƠNG 2 GIẢI PHÁP TƯ VẤN LAI SỬ DỤNG ĐỒNG HUẤN LUYỆN ...	6
2.1 Phương pháp đồng huấn luyện	6
2.1.1 Tổng quan về phương pháp đồng huấn luyện	6
2.1.2 Một số ứng dụng của phương pháp đồng huấn luyện	7
2.2 Đề xuất giải pháp tư vấn lai sử dụng đồng huấn luyện.....	7
2.2.1 Mô tả bài toán	7
2.2.2 Giải pháp tổng thể.....	7
2.2.3 Một số phương pháp tính độ tin cậy.....	7
2.3 Xây dựng giải pháp tư vấn.....	7
2.3.1 Xây dựng các bộ hồi quy	8
2.3.2 Đồng huấn luyện	9
2.3.3 Tổng hợp kết quả.....	9
2.4 Kết luận chương.....	9
CHƯƠNG 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ.....	11
3.1. Mô tả dữ liệu	11
3.1.1. Mô tả bộ dữ liệu MovieLens.....	11
3.1.2. Thu thập thông tin bổ sung về các bộ phim.....	11
3.2. Phương pháp thử nghiệm	12

3.2.1.	<i>Phân chia dữ liệu</i>	12
3.2.2.	<i>Xây dựng mô hình hồi quy lọc cộng tác</i>	12
3.2.3.	<i>Xây dựng mô hình hồi quy dựa trên nội dung</i>	12
3.2.4.	<i>Đồng huấn luyện</i>	12
3.2.5.	<i>Tổng hợp kết quả</i>	12
3.3.	Kết quả và đánh giá	12
3.3.1.	<i>Các chỉ số đánh giá</i>	12
3.3.2.	<i>Kết quả thực nghiệm</i>	13
3.3.3.	<i>Nhận xét và phân tích kết quả</i>	13
3.4.	Kết luận chương	14
KẾT LUẬN	16
TÀI LIỆU THAM KHẢO	18

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
API	Application Programming Interface	Thư viện lập trình ứng dụng
CONFINE	CONFidence estimation based on the Neighbors' Errors	Ước lượng độ tin cậy dựa trên sai số của các hàng xóm
CONFIVE	CONFidence estimation based on the Variance in the Environment	Ước lượng độ tin cậy dựa trên biến thiên của môi trường
MAE	Mean Absolute Error	Sai số tuyệt đối trung bình
RMSE	Root Mean Square Error	Sai số bình phương trung bình
SVM	Support Vector Machine	Máy vector tựa

DANH SÁCH BẢNG

Bảng 3.1 - Kết quả đánh giá hiệu quả của ba phương pháp qua từng vòng lặp.....13

Bảng 3.2 - Bảng so sánh hiệu quả của các phương pháp tư vấn.....13

DANH SÁCH HÌNH VẼ

Hình 2.1 - Các bước xây dựng mô hình hồi quy lọc cộng tác	8
Hình 2.2 - Các bước xây dựng mô hình hồi quy dựa trên nội dung	8

MỞ ĐẦU

Ngày nay, con người đang sống trong thời đại số, nơi mà mạng Internet được phổ biến khắp toàn cầu. Mỗi một người dùng Internet được tiếp cận với rất nhiều nguồn thông tin khác nhau. Do đó họ có thể tìm thấy bất cứ thứ gì mình muốn trên Internet. Tuy nhiên có một vấn đề được đặt ra đó là những thông tin nào thực sự cần thiết cho người dùng Internet trong hàng nghìn nguồn thông tin khác nhau trên mạng Internet? Ví dụ như khi vào một trang Web để xem phim như Netflix, sẽ có hàng trăm nghìn bộ phim trong cơ sở dữ liệu của Netflix được đưa ra cho người dùng lựa chọn. Và lúc này, người dùng sẽ rất khó khăn để tìm ra bộ phim mà họ muốn xem trong một số lượng lớn các bộ phim như vậy.

Để giải quyết vấn đề này, các hệ thống tư vấn [6][8][15][19] đã ra đời với nhiệm vụ đưa ra những gợi ý giúp cho người dùng dễ dàng tìm được thông tin mà mình cần tìm một cách nhanh chóng và chính xác. Đã có rất nhiều nghiên cứu được thực hiện để tìm ra phương pháp tư vấn hiệu quả. Một số phương pháp đã cho kết quả tốt như: phương pháp tư vấn dựa trên lọc cộng tác [8][19], phương pháp tư vấn theo nội dung [8], phương pháp tư vấn dựa trên thông tin cá nhân [15] và phương pháp lai giữa các phương pháp trên.

Từ trước đến nay, các phương pháp tư vấn như tư vấn dựa trên nội dung, tư vấn dựa trên lọc cộng tác có những hướng khai thác các khía cạnh của dữ liệu khác nhau để đưa ra tư vấn một cách tốt nhất. Mỗi một phương pháp tư vấn này lại có một số nhược điểm riêng như vấn đề cold-start của phương pháp lọc cộng tác, hay vấn đề thiếu thông tin của phương pháp lọc theo nội dung. Để giải quyết vấn đề này, các phương pháp tư vấn lai ra đời để kết hợp các phương pháp tư vấn này lại với nhau để khắc phục các nhược điểm của nhau.

Mục đích của luận văn này là đi xây dựng một hệ thống tư vấn lai như vậy, phương pháp lai được sử dụng trong luận văn này là phương pháp đồng huấn luyện. Đây là một phương pháp dùng để kết hợp hai phương pháp tư vấn là phương pháp

tư vấn dựa trên lọc cộng tác và phương pháp tư vấn dựa trên nội dung lại với nhau. Phương pháp đồng huấn luyện thường bổ sung các dự đoán của hai bộ hồi quy lẫn nhau để huấn luyện lại. Tuy nhiên việc bổ sung toàn bộ các dự đoán này chưa tính toán đến việc các dự đoán đó có thể sai dẫn đến việc sử dụng các dự đoán sai đó sẽ ảnh hưởng đến độ chính xác của mô hình hồi quy còn lại. Để giải quyết vấn đề này, phương pháp trong luận văn sử dụng một cải tiến: trước khi bổ sung giá trị dự đoán của bộ hồi quy này vào tập huấn luyện bộ hồi quy còn lại, ta sẽ thêm bước ước lượng độ tin cậy của dự đoán; chỉ những dự đoán có độ tin cậy cao mới được bổ sung vào dữ liệu huấn luyện cho bộ hồi quy còn lại. Cải tiến này cho phép giảm ảnh hưởng của dự đoán sai tới các vòng lặp đồng huấn luyện sau đó. Có rất nhiều cách để tính toán độ chính xác của một dự đoán, luận văn này lựa chọn hai phương pháp CONFINE và CONFIVE [4] để tính độ tin cậy của các dự đoán của các mô hình hồi quy.

Luận văn này có nội dung tiếp nối các nghiên cứu của học viên về hệ tư vấn, một số kết quả liên quan đã được đăng trong tạp chí Information sciences [3] và trình bày tại hội nghị SoCPAR 2013 [8].

Với mục tiêu như vậy, bố cục của luận văn sẽ bao gồm bốn chương theo cấu trúc như sau:

Chương 1: Giới thiệu

Trình bày một cách tổng quan về mục tiêu, ý nghĩa cũng như các thành phần chính của một hệ tư vấn. Giới thiệu qua về một số phương pháp tư vấn đã được phát triển cũng như những ưu nhược điểm của nó.

Chương 2: Giải pháp tư vấn lai sử dụng đồng huấn luyện

Nội dung của chương này sẽ làm sáng tỏ về mặt lý thuyết cho hệ thống tư vấn mà luận văn này định xây dựng. Phần đầu của chương này sẽ đi sâu tìm hiểu về phương pháp đồng huấn luyện. Sau đó, luận văn sẽ đề xuất ra một giải pháp tư vấn lai sử dụng đồng huấn luyện có cải tiến. Cuối cùng luận văn sẽ đi sâu vào từng bước để xây dựng nên hệ thống tư vấn sử dụng giải pháp đề xuất.

Chương 3: Thực nghiệm và đánh giá

Chương 3 sẽ mô tả các bước để triển khai giải pháp đề xuất ở chương 2 vào thực tiễn. Bên cạnh đó nội dung chương 3 sẽ tiến hành đánh giá những kết quả đạt được thông qua một số độ đo thường được sử dụng cho bài toán tư vấn.

Kết luận

Tổng kết bài toán, tóm tắt những kết quả đã đạt được và còn chưa đạt được. Từ đó đề xuất mục tiêu hướng tới cũng như hướng nghiên cứu, phát triển tiếp theo.

CHƯƠNG 1 - GIỚI THIỆU

Chương này sẽ giới thiệu một cách tổng quan về hệ thống tư vấn bao gồm mục tiêu mà bài toán hướng tới, các thành phần của hệ thống tư vấn. Đồng thời cũng trình bày sơ lược về những giải pháp đã được phát triển cho bài toán tư vấn.

1.1 Hệ thống tư vấn

1.1.1 Mục đích

Các hệ thống tư vấn là một tập hợp các kỹ thuật, công cụ phần mềm có nhiệm vụ cung cấp các gợi ý cho người dùng về các sản phẩm mà họ có thể quan tâm và muốn sử dụng. Những gợi ý này có thể rất hữu ích cho người dùng trong quá trình đưa ra quyết định, ví dụ như người dùng nên xem bộ phim nào, nghe bài hát nào, mua sản phẩm nào, hay đọc tin tức nào.

Các hệ thống tư vấn thường chỉ đưa ra gợi ý cho người dùng về một loại sản phẩm cụ thể nào đó (ví dụ: phim, sách, hay tin tức). Các kỹ thuật tư vấn sẽ dựa trên những thông tin về từng loại sản phẩm để đưa ra các tư vấn hiệu quả, hữu ích nhất cho từng loại sản phẩm.

Ngoài ra, các hệ thống tư vấn cũng thường tập trung vào tư vấn cho người dùng cá nhân, vì những người dùng này thường không đủ khả năng hay công sức để tìm kiếm trong một số lượng lớn các sản phẩm hiện có trên một trang Web.

1.1.2 Các thành phần chính

1.1.2.1 Tập người dùng:

1.1.2.2 Tập đối tượng tư vấn:

1.1.2.3 Tập các phản hồi từ người dùng:

1.1.3 Một số tác vụ chính của hệ thống tư vấn

1.1.3.1 Dự đoán đánh giá

1.1.3.2 Tư vấn Top-N sản phẩm

1.2 Một số phương pháp tư vấn đã phát triển

1.2.1 Phương pháp tư vấn dựa trên lọc cộng tác

1.2.1.1 Giới thiệu chung

1.2.1.2 Phân loại phương pháp tư vấn dựa trên lọc cộng tác

1.2.2 Phương pháp tư vấn dựa trên nội dung

1.2.2.1 Giới thiệu chung

1.2.2.2 Phương pháp biểu diễn đối tượng tư vấn

1.2.2.3 Một số phương pháp xây dựng mô hình sở thích người dùng

1.2.3 Phương pháp tư vấn dựa trên thông tin cá nhân

1.2.4 Phương pháp tư vấn lai

1.3 Kết luận chương

Tóm lại, một hệ thống tư vấn được xây dựng nhằm mục đích đưa ra gợi ý cho người dùng về một số sản phẩm mà họ có thể quan tâm dựa trên nhiều nguồn thông tin khác nhau. Các hệ thống tư vấn có ý nghĩa rất quan trọng không chỉ với những người sử dụng các hệ thống tư vấn, mà còn có ý nghĩa với những nhà cung cấp dịch vụ tư vấn.

Có bốn phương pháp tư vấn hay được sử dụng là phương pháp tư vấn dựa trên lọc cộng tác, phương pháp tư vấn dựa trên nội dung, phương pháp tư vấn dựa trên thông tin cá nhân và phương pháp lai. Trong số đó, phương pháp tư vấn lai nổi lên là phương pháp tốt nhất do phương pháp này là sự kết hợp của nhiều phương pháp tư vấn khác lại với nhau để khắc phục nhược điểm của nhau. Do đó, nội dung luận văn này sẽ đi xây dựng một phương pháp tư vấn lai kết hợp cả ba phương pháp tư vấn còn lại với nhau bằng một phương pháp mới có tên là phương pháp đồng huấn luyện. Chương tiếp theo của luận văn sẽ đi xây dựng những cơ sở lý thuyết cho phương pháp tư vấn lai sử dụng đồng huấn luyện này.

CHƯƠNG 2 GIẢI PHÁP TƯ VẤN LAI SỬ DỤNG ĐỒNG HUẤN LUYỆN

Nội dung của chương hai sẽ đi tìm hiểu một phương pháp thường được sử dụng để xây dựng các hệ thống tư vấn lai có khả năng tận dụng những dữ liệu chưa gán nhãn có tên là phương pháp đồng huấn luyện. Từ những tìm hiểu đó, luận văn sẽ đề xuất ra một giải pháp tư vấn lai sử dụng phương pháp đồng huấn luyện để giải quyết bài toán dự đoán đánh giá. Phương pháp đồng huấn luyện trước đây [2][5] thường thêm tất cả các giá trị dự đoán được vào bộ huấn luyện cho bộ phân loại sau, tuy nhiên trong các dự đoán này sẽ tồn tại một số dự đoán lỗi mà ảnh hưởng đến sự chính xác của các bước tiếp theo.

Để giải quyết vấn đề này, luận văn sẽ cải tiến thuật toán đồng huấn luyện thông thường bằng cách thêm vào quá trình đồng huấn luyện một bước tính toán độ tin cậy của các dự đoán để loại bỏ đi những dự đoán sai, từ đó chỉ có những dự đoán đúng mới được sử dụng để huấn luyện cho các bước đồng huấn luyện tiếp theo. Có rất nhiều phương pháp tính độ tin cậy khác nhau, nhưng trong luận văn này sẽ sử dụng hai phương pháp tính độ tin cậy thích hợp cho bài toán hồi quy là phương pháp CONFINE [4] và phương pháp CONFIVE [4]. Sau khi loại bỏ những ô dự đoán sai thì kết quả sẽ chính xác hơn.

2.1 Phương pháp đồng huấn luyện

2.1.1 Tổng quan về phương pháp đồng huấn luyện

Học bán giám sát là một kỹ thuật học máy thu hút được nhiều sự chú ý của các nhà nghiên cứu bởi một số lượng lớn các dữ liệu chưa được gán nhãn có thể được sử dụng để cải thiện hiệu năng của các thuật toán học máy trong khi nếu chỉ sử dụng các dữ liệu có gán nhãn là không đủ để làm điều này. Blum và Mitchell [5] lần đầu tiên xem xét đến vấn đề chia những thông tin về một mẫu dữ liệu thành hai phần dưới hai góc nhìn độc lập. Ví dụ, một trang Web có thể được biểu diễn bởi các từ xuất hiện trong trang Web đó, hoặc cũng có thể được biểu diễn bằng các từ xuất hiện trong một siêu liên kết trỏ tới trang Web đó. Do đó chúng ta có thể chỉ cần sử dụng

một trong hai cách biểu diễn trên để phân loại một trang Web. Phương pháp phân chia đặc điểm của một đối tượng thành hai phần trên đây được gọi là đồng huấn luyện.

2.1.2 Một số ứng dụng của phương pháp đồng huấn luyện

2.1.2.1 Bài toán phân loại

2.1.2.2 Bài toán dự đoán đánh giá

2.2 Đề xuất giải pháp tư vấn lai sử dụng đồng huấn luyện

2.2.1 Mô tả bài toán

Đầu vào:

- Tập người dùng U ,
- Tập bộ phim I ,
- Ma trận đánh giá của người dùng cho một số bộ phim $R_{U \times I}$.

Đầu ra:

- Ma trận dự đoán đánh giá của người dùng cho tất cả các bộ phim.

Giải pháp: Sử dụng phương pháp tư vấn để xây dựng nên mô hình dự đoán đánh giá cho các ô còn thiếu trong ma trận $R_{U \times I}$ ban đầu. Cụ thể trong luận văn này sẽ sử dụng phương pháp đồng huấn luyện để kết hợp phương pháp tư vấn lọc cộng tác và phương pháp tư vấn dựa trên nội dung.

2.2.2 Giải pháp tổng thể

2.2.3 Một số phương pháp tính độ tin cậy

2.2.3.1 CONFINE

2.2.3.2 CONFIVE

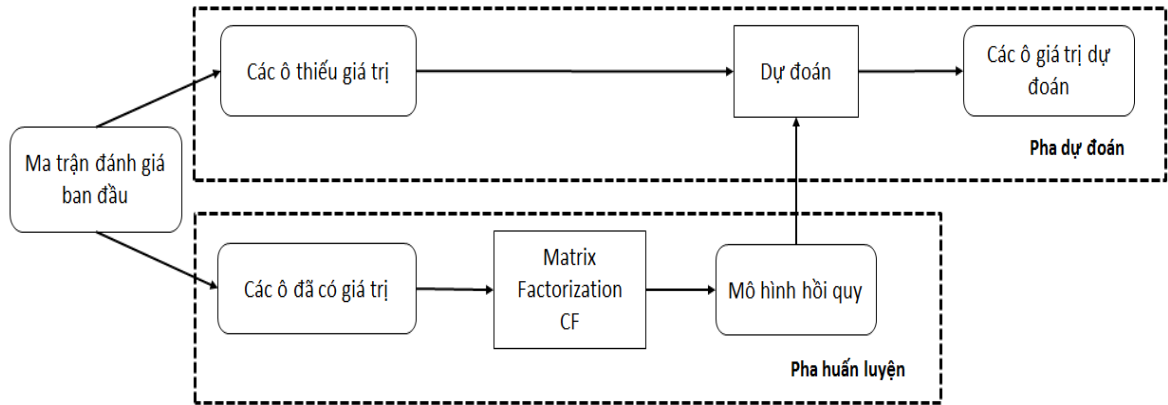
2.3 Xây dựng giải pháp tư vấn

Nội dung phần này sẽ đi trình bày quá trình thực nghiệm để triển khai các bước trong giải pháp tư vấn đề xuất. Nội dung đầu tiên mô tả cách huấn luyện hai bộ hồi quy dựa trên những dữ liệu đã có. Tiếp theo, mô tả cách triển khai phương

pháp đồng huấn luyện. Cuối cùng, mô tả cách kết hợp kết quả các bộ hồi quy để đưa ra kết quả dự đoán cuối cùng.

2.3.1 Xây dựng các bộ hồi quy

2.3.1.1 Bộ hồi quy lọc cộng tác



Hình 2.1 - Các bước xây dựng mô hình hồi quy lọc cộng tác

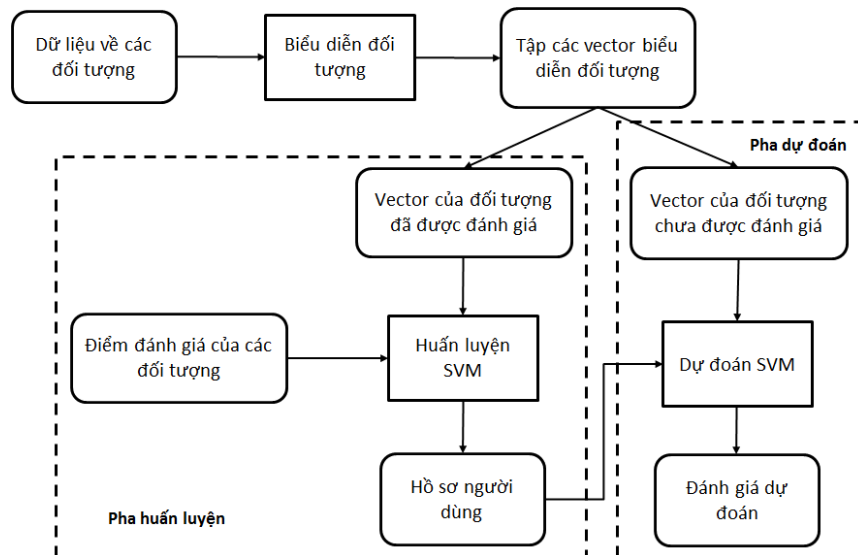
a. Đầu vào:

b. Đầu ra:

c. Pha huấn luyện:

d. Pha dự đoán:

2.3.1.2 Bộ hồi quy dựa trên nội dung



Hình 2.2 - Các bước xây dựng mô hình hồi quy dựa trên nội dung

a. Đầu vào

b. Đầu ra

c. Biểu diễn đối tượng tư vấn

d. Pha huấn luyện

e. Pha dự đoán

2.3.2 Đồng huấn luyện

Như đã trình bày ở phần đề xuất giải pháp, ở bước đồng huấn luyện này chúng ta sẽ lặp đi lặp lại k lần một số bước. Việc đồng huấn luyện này có mục đích tận dụng những dự đoán có độ tin cậy cao của bộ hồi quy này làm dữ liệu đầu vào cho bộ dữ liệu kia để cải thiện độ chính xác của thuật toán. Cụ thể, có ba bước cần lặp lại qua mỗi lần đồng huấn luyện bao gồm (1) tính toán với bộ hồi quy lọc cộng tác h_1 , (2) tính toán với bộ hồi quy dựa trên nội dung h_2 , và (3) cập nhật lại hai bộ hồi quy. Nội dung phần này của luận văn sẽ đi trình bày cụ thể những việc cần làm ở ba bước này.

2.3.2.1 Tính toán với bộ hồi quy lọc cộng tác

2.3.2.2 Tính toán với bộ hồi quy dựa trên nội dung

2.3.2.3 Cập nhật hai bộ hồi quy

2.3.3 Tổng hợp kết quả

Kết thúc quá trình đồng huấn luyện, ta thu được hai bộ hồi quy h_1 và h_2 . Hai bộ hồi quy này có khả năng đưa ra dự đoán cho các ô còn thiếu bất kỳ trong ma trận đầu vào R . Như vậy, lúc này ta có hai giá trị dự đoán cho cùng một ô r_{ui} là $h_1(u, i)$ và $h_2(u, i)$ tương ứng với giá trị dự đoán của hai bộ hồi quy h_1 và h_2 . Bên cạnh đó ta cũng tính được độ tin cậy của hai dự đoán này là $c_1(u, i)$ và $c_2(u, i)$. Nhiệm vụ của bước tổng hợp kết quả này là kết hợp hai giá trị dự đoán lại để được một giá trị dự đoán cuối cùng $h(u, i)$.

2.4 Kết luận chương

Như vậy, nội dung chương 2 đã đi tìm hiểu về phương pháp đồng huấn luyện, từ đó đề xuất ra một giải pháp tư vấn lai sử dụng phương pháp đồng huấn luyện.

Giải pháp này được đề xuất để giải quyết bài toán dự đoán đánh giá bằng cách kết hợp phương pháp tư vấn lọc cộng tác và phương pháp tư vấn dựa trên nội dung lại với nhau. Phần cuối cùng của chương trình bày nội dung cụ thể các bước để xây dựng nên giải pháp đề xuất.

Cụ thể, phương pháp đồng huấn luyện là một phương pháp học bán giám sát, được sử dụng để tận dụng một số lượng lớn các dữ liệu chưa được gán nhãn để cải thiện hiệu năng của các thuật toán học máy trong khi nếu chỉ sử dụng các dữ liệu có gán nhãn là không đủ để làm điều này. Phương pháp đồng huấn luyện có thể được áp dụng để giải quyết bài toán phân loại văn bản, hay bài toán tư vấn. Cụ thể trong luận văn này sẽ đề xuất một giải pháp tư vấn sử dụng phương pháp đồng huấn luyện kết hợp hai phương pháp tư vấn dựa trên lọc cộng tác và tư vấn theo nội dung.

Giải pháp tư vấn lai sử dụng phương pháp đồng huấn luyện được đề xuất trong luận văn sẽ đi giải quyết bài toán dự đoán đánh giá của người dùng cho các bộ phim. Đầu vào của hệ thống bao gồm tập các người dùng, tập các bộ phim, và ma trận đánh giá của người cho một số bộ phim. Giải pháp được đề xuất bao gồm ba bước: 1 - Xây dựng hai bộ hồi quy: bộ hồi quy lọc cộng tác và bộ hồi quy dựa trên nội dung, 2 – Đồng huấn luyện, 3 - Tổng hợp kết quả.

Như vậy toàn bộ chương 2 đã xây dựng được một hệ thống tư vấn lai sử dụng phương pháp đồng huấn luyện. Nhiệm vụ tiếp theo ở chương 3 đó là đi xây dựng một hệ thống thực nghiệm để kiểm tra tính đúng đắn của giải pháp đã đề ra.

CHƯƠNG 3 - THỰC NGHIỆM VÀ ĐÁNH GIÁ

Để kiểm chứng độ chính xác của giải pháp đề xuất, nội dung chương 3 sẽ đi trình bày việc triển khai một hệ thống thực nghiệm triển khai giải pháp này. Sau đó sẽ sử dụng một số chỉ số để đánh giá hiệu quả của hệ thống. Cuối cùng là một số nhận xét về kết quả thu được từ hệ thống.

3.1. Mô tả dữ liệu

3.1.1. Mô tả bộ dữ liệu *MovieLens*

Trong luận văn này, hệ thống thực nghiệm sẽ sử dụng dữ liệu đầu vào là bộ dữ liệu *MovieLens*. Bộ dữ liệu *MovieLens* là một bộ dữ liệu được thu thập bởi dự án nghiên cứu *GroupLens*.

Mỗi một người dùng trong bộ dữ liệu này sẽ được gán một ID định danh duy nhất, ngoài ra họ còn có kèm theo một số thông tin cá nhân cơ bản như tuổi, giới tính, nghề nghiệp và zipcode. Trong đó mức tuổi của người dùng là từ 7 đến 73 tuổi, giới tính của người dùng có hai lựa chọn là nam hoặc nữ. Nghề nghiệp của người dùng có 21 loại. Mỗi một người dùng này có ít nhất 20 đánh giá cho các bộ phim.

Tương tự như người dùng, mỗi một bộ phim được gán một ID định danh duy nhất, thông tin kèm theo mỗi bộ phim còn có tiêu đề phim, ngày phát sóng, đường dẫn tới trang *IMDb*¹, thể loại của phim (bao gồm 19 thể loại). Mỗi một bộ phim có thể được xếp vào một hay nhiều thể loại khác nhau. Tuy nhiên những thông tin này vẫn chưa đủ để mô tả nội dung một bộ phim, do đó ở phần tiếp theo của luận văn sẽ trình bày cách thu thập một số nội dung của các bộ phim này.

3.1.2. Thu thập thông tin bổ sung về các bộ phim

Như đã trình bày ở trên, những thông tin về một bộ phim trong bộ dữ liệu *MovieLens* là không đủ để chúng ta có thể xây dựng một phương pháp tư vấn dựa

¹ <http://www.imdb.com/>

trên nội dung. Do đó, luận văn này sẽ đi thu thập thêm thông tin về các bộ phim thông qua một Web API có tên là OMDb².

3.2. Phương pháp thử nghiệm

3.2.1. Phân chia dữ liệu

3.2.2. Xây dựng mô hình hồi quy lọc cộng tác

3.2.2.1. Chuẩn bị đầu vào

3.2.2.2. Xây dựng mô hình hồi quy

3.2.2.3. Phương pháp dự đoán

3.2.3. Xây dựng mô hình hồi quy dựa trên nội dung

3.2.3.1. Chuẩn bị đầu vào

3.2.3.2. Xây dựng mô hình hồi quy

3.2.3.3. Phương pháp dự đoán

3.2.4. Đồng huấn luyện

3.2.4.1. Chuẩn bị đầu vào

3.2.4.2. Đồng huấn luyện

3.2.5. Tổng hợp kết quả

3.3. Kết quả và đánh giá

3.3.1. Các chỉ số đánh giá

Để đánh giá độ chính xác của giải pháp đề xuất, chúng ta sẽ sử dụng hai chỉ số đánh giá là Root Mean Square Error (RMSE) và Mean Absolute Error (MAE). RMSE và MAE là hai chỉ số thường được sử dụng để đánh giá độ chính xác của các bộ hồi quy.

² <http://www.omdbapi.com/>

3.3.2. Kết quả thực nghiệm

3.3.2.1. Kết quả với bộ dữ liệu tổng thể

Bảng 3.1 - Kết quả đánh giá hiệu quả của ba phương pháp qua từng vòng lặp

Loop	confine + no round		confine + round		confive + round	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
1	0.91631	0.71726	0.91495	0.71639	0.91443	0.71548
2	0.91524	0.71671	0.91445	0.71568	0.91451	0.71620
3	0.91461	0.71593	0.92060	0.71838	0.91452	0.71587
4	0.91533	0.71632	0.91623	0.71714	0.91486	0.71637
5	0.91643	0.71784	0.91484	0.71633	0.91557	0.71640
10	0.91619	0.71715	0.91538	0.71644	0.91495	0.71639

Bảng 3.2 - Bảng so sánh hiệu quả của các phương pháp tư vấn

		RMSE	MAE
Baseline	CF	0.91635	0.71754
	CBF	1.02161	0.78950
Cotraining	confine + no round	0.91461	0.71593
	confine + round	0.91445	0.71568
	confive + round	0.91443	0.71548

3.3.2.2. Kết quả với trường hợp Cold-start

3.3.3. Nhận xét và phân tích kết quả

3.3.3.1. So sánh với các phương pháp baseline

Dựa vào Bảng 3.5 ta có thể thấy phương pháp tư vấn lai dựa trên đồng huấn luyện được đề xuất đã có tỷ lệ lỗi giảm đi so với các phương pháp tư vấn cơ bản là phương pháp tư vấn dựa trên lọc cộng tác và phương pháp tư vấn dựa trên nội dung. Với cả ba cách lựa chọn độ tin cậy thì cả hai chỉ số RMSE và MAE đều thấp hơn so với hai phương pháp cơ bản.

3.3.3.2. So sánh giữa các phương pháp lựa chọn độ tin cậy

Theo Bảng 3.4, chúng ta có thể thấy, việc lựa chọn phương pháp CONFIVE có các chỉ số RMSE và MAE thấp hơn so với phương pháp CONFINE trong tất cả các vòng lặp từ 1 đến 10 của bước đồng huấn luyện, các chỉ số RMSE của phương pháp CONFIVE đều lớn hơn, đặc biệt ở vòng đầu tiên, chỉ số này lớn hơn so với phương pháp CONFINE khoảng 0.0019.

Từ Bảng 3.4 ta cũng thấy được việc làm tròn đánh giá trước khi cho vào huấn luyện bộ hồi quy khác. Để kiểm chứng điều này, chúng ta nhìn vào so sánh hai phương pháp CONFINE+No round và CONFINE+Round, ở đây hai phương pháp này đều sử dụng một phương pháp tính độ tin cậy. Chúng ta thấy rõ là các chỉ số RMSE và MAE của phương pháp CONFINE+Round đều thấp hơn.

3.3.3.3. Vấn đề Cold-start

Từ Bảng 3.6 ta có thể thấy giải pháp tư vấn lại sử dụng đồng huấn luyện của chúng ta đã cải thiện được độ chính xác cho các dự đoán với các đối tượng ít được đánh giá. Cụ thể, trong 10 bin dữ liệu, bin1 có số lượng đánh giá trung bình cho mỗi bộ phim là ít nhất, nhưng lại là bin có chỉ số RMSE của phương pháp tư vấn lại là 1.0148, chỉ số này là thấp nhất trong ba phương pháp.

3.4. Kết luận chương

Tóm lại, nội dung chương 3 mô tả lại cách xây dựng nên một hệ thống tư vấn bằng phương pháp lại sử dụng đồng huấn luyện để kết hợp phương pháp tư vấn lọc cộng tác và phương pháp tư vấn dựa trên nội dung lại với nhau.

Bộ dữ liệu MovieLens được sử dụng làm bộ dữ liệu đầu vào. MovieLens là bộ dữ liệu cung cấp thông tin về các đánh giá của các người dùng cho các bộ phim. Bộ dữ liệu này còn chứa một số thông tin về người dùng và các bộ phim. Ngoài thông tin từ bộ dữ liệu MovieLens, hệ thống còn thu thập một số thông tin về các bộ phim trên IMDb.

Hệ thống tư vấn được xây dựng qua ba bước. Bước đầu tiên là đi xây dựng các bộ hồi quy lọc cộng tác và bộ hồi quy dựa trên nội dung. Bộ hồi quy dựa trên lọc cộng tác được xây dựng bằng phương pháp phân tích ma trận sử dụng bộ công cụ

MyMedialite, bộ hồi quy dựa trên nội dung được xây dựng bằng mô hình học máy SVM sử dụng bộ công cụ SVM-Light. Bước thứ hai của quá trình này là đồng huấn luyện. Ở bước này, có ba thủ tục được lặp đi lặp lại là (1) tính toán trên bộ hồi quy lọc cộng tác, (2) tính toán trên bộ hồi quy dựa trên nội dung và (3) cập nhật lại hai bộ hồi quy. Bước cuối cùng là tổng hợp kết quả, là bước sử dụng hai bộ hồi quy đã được tăng cường dữ liệu bằng đồng huấn luyện để dự đoán ra các giá trị đánh giá chưa biết. Bước tổng hợp kết quả sẽ sử dụng phương pháp voting để kết hợp hai giá trị dự đoán của hai bộ hồi quy rồi đưa ra giá trị dự đoán cuối cùng.

Kết quả thu được từ quá trình thực nghiệm là rất khả quan. Cụ thể, các chỉ số RMSE và MAE của giải pháp được đề xuất giảm đi so với hai phương pháp tư vấn lọc cộng tác và phương pháp tư vấn theo nội dung. Ngoài ra việc sử dụng độ tin cậy CONFIVE và làm tròn đánh giá trước khi bổ sung vào bộ hồi quy cũng làm tăng độ chính xác của hệ thống. Cuối cùng, giải pháp đề xuất cũng phần nào giải quyết được vấn đề cold-start khi mà chỉ số RMSE của phương pháp đề xuất nhỏ hơn so với hai phương pháp còn lại.

KẾT LUẬN

Trên cơ sở tìm hiểu về việc xây dựng một hệ thống tư vấn, và cụ thể là tác vụ dự đoán đánh giá của người dùng cho các đối tượng chưa được đánh giá, luận văn đã đạt được một số kết quả sau:

- Tìm hiểu một cách cụ thể về một hệ thống tư vấn, các khía cạnh tìm hiểu bao gồm mục đích, các thành phần chính, các tác vụ chính và một số phương pháp tư vấn đã được phát triển.
- Nghiên cứu phương pháp đồng huấn luyện là một phương pháp học bán giám sát có khả năng tận dụng các dữ liệu chưa gán nhãn để tăng độ chính xác của hệ thống. Đưa ra ý tưởng về việc sử dụng phương pháp đồng huấn luyện để xây dựng một hệ thống tư vấn.
- Đề xuất ra một giải pháp tư vấn lai sử dụng phương pháp đồng huấn luyện kết hợp phương pháp tư vấn lọc cộng tác và phương pháp tư vấn dựa trên nội dung để giải quyết bài toán dự đoán đánh giá.
- Ứng dụng hai phương pháp tính độ tin cậy CONFINE và CONFIVE để tính toán độ tin cậy của các dự đoán của mô hình hồi quy sử dụng trong phương pháp đồng huấn luyện.
- Tiến hành thực nghiệm cho giải pháp đề xuất dựa trên bộ dữ liệu MovieLens, từ đó đưa ra những nhận xét về kết quả thu được.

Bên cạnh những kết quả thu được thì luận văn vẫn còn một số hạn chế đó là:

- Độ chính xác của các dự đoán của hệ thống vẫn chưa được cải thiện quá nhiều do việc kết hợp các kết quả dự đoán chưa hoàn toàn phù hợp.
- Luận văn mới giải quyết được vấn đề cold-start cho trường hợp những bộ phim mới chứ chưa giải quyết được vấn đề cold-start cho những người dùng mới.

Từ những kết quả thu được và những hạn chế nêu trên, có thể thấy việc xây dựng nên một hệ thống tư vấn nói chung và việc giải quyết bài toán dự đoán đánh

giá nói riêng vẫn cần một quá trình nghiên cứu dài nữa để cải tiến được chúng. Nội dung luận văn mới chỉ trình bày việc kết hợp hai phương pháp tư vấn lại với nhau, tuy nhiên một hệ thống tư vấn lai có thể kết hợp nhiều phương pháp hơn nữa.

Hướng phát triển tiếp theo của luận văn đó là:

- Tìm kiếm phương pháp kết hợp kết quả dự đoán của các mô hình một cách tốt hơn để tăng độ chính xác của các dự đoán.
- Áp dụng thêm phương pháp tư vấn dựa trên thông tin cá nhân vào mô hình đồng huấn luyện để giải quyết trường hợp cold-start cho người dùng mới.

TÀI LIỆU THAM KHẢO

Tài liệu tiếng Việt:

- [1] Từ Minh Phương, Học viện Công nghệ Bưu chính Viễn thông (2014), *Giáo trình Trí tuệ nhân tạo*.
- [2] Nguyễn Việt Tân, Hoàng Vũ, Đặng Vũ Tùng, Từ Minh Phương (2014), “Phân loại dữ liệu có liên kết sử dụng phương pháp đồng huấn luyện”, *Tạp chí Khoa học DHQGHN: Khoa học Tự nhiên và Công nghệ*, Tập 30, Số 4, trang 48-57.

Tài liệu tiếng Anh:

- [3] Ngo Xuan Bach, Nguyen Do Hai, Tu Minh Phuong (2016), “Personalized recommendation of stories for commenting in forum-based social media”, *Information Sciences*, pp. 48-60.
- [4] S. Briesemeister, J. Rahrienführer, O. Kohlbacher (2012), “No longer confidential: estimating the confidence of individual regression predictions”, *PLoS ONE*.
- [5] A. Blum, T. Mitchell (1998), “Combining Labeled and Unlabeled Data with Co-Training”, In *Proceedings of COLT*, pp. 92-100.
- [6] A. Felfernig, M. Jeran, G. Ninaus, F. Reinfrank, S. Reiterer, M. Stettinger (2014), “Basic approaches in recommendation systems”, In *Recommendation Systems in Software Engineering*, pp. 15-37.
- [7] Ed Greengrass(2000), *Information Retrieval: A Survey*.
- [8] Nguyen Do Hai, Tran Quang An, Ngo Xuan Bach, Tu Minh Phuong (2013), “What Should I Comment: Recommending Posts for Commenting”, 5th International Conference of Soft Computing and Pattern Recognition, Hà Nội.
- [9] J. Han, M. Kamber (2000), *Data Mining: Concepts and Techniques*.
- [10] W. Hill, L. Stead, M. Rosenstein, and G. Furnas (1995), “Recommending and evaluating choices in a virtual community of use”, In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 194–201.

- [11] Joachims, Freitag, Mitchell (1997), “WebWatcher: A Tour Guide for the World Wide Web”, In Proceedings of the 15th International Joint Conference on Artificial Intelligence, Nagoya, Japan, pp. 770 -775.
- [12] Pazzani, Billsus (1997), “Learning and revising user profiles: The identification of interesting web sites”, Machine learning, 27(3), pp. 313-331.
- [13] Prasad, Kumari (2012), “A categorical review of recommender systems”, International Journal of Distributed and Parallel Systems (IJDPS) Vol.3, No.5.
- [14] F. Ricci, Lior Rokach, and Bracha Shapira (2011), Introduction to recommender systems handbook, Springer US.
- [15] Laila Safoury, Akram Salah (2013), “Exploiting User Demographic Attributes for Solving Cold-Start Problem in Recommender System”, Lecture Notes on Software Engineering Vol. 1, No. 3.
- [16] Shardanand, Maes, (1995, May), “Social information filtering: algorithms for automating “word of mouth” ”. *In Proceedings of the SIGCHI conference on Human factors in computing systems* , pp. 210-217.
- [17] Z. Tao, M. Cheung, J. She, R. Lam (2014), “Item Recommendation Using Collaborative Filtering in Mobile Social Games: A Case Study”, In *Big Data and Cloud Computing (BdCloud), 2014 IEEE International Conference*, pp. 293-297.
- [18] Loren Terveen, Will Hill (2001), “Beyond Recommender Systems: Helping People Help Each Other”, *HCI in the New Millennium 1*, pp. 487-509.
- [19] L. Ungar, D. Foster (1998), “Clustering methods for collaborative filtering”, In Proceedings of *the Workshop on Recommendation Systems*, AAAI Press, Menlo Park California.