ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Vũ Minh Nhật

NGHIÊN CỨU CÁC KỸ THUẬT KHUYẾN NGHỊ VÀ ÁP DỤNG CHO BÀI TOÁN GỢI Ý KHÓA HỌC

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Khoa học máy tính

HÀ NỘI - 2024

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

Vũ Minh Nhật

NGHIÊN CỨU CÁC KỸ THUẬT KHUYẾN NGHỊ VÀ ÁP DỤNG CHO BÀI TOÁN GỢI Ý KHÓA HỌC

KHÓA LUẬN TỐT NGHIỆP ĐẠI HỌC HỆ CHÍNH QUY
Ngành: Khoa học máy tính

Cán bộ hướng dẫn: PGS.TS Nguyễn Việt Anh

HÀ NỘI - 2024

VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Vu Minh Nhat

RESEARCHING RECOMMENDATION TECHNIQUES AND APPLYING FOR COURSES RECOMMENDATION PROBLEMS

GRADUATE THESIS
Major: Computer Science

Thesis Supervisor: Assoc. Prof. Nguyen Viet Anh

HANOI- 2024

Tóm tắt

Hiện nay, các hệ thống học online quy mô mở (Massive Open Online Course - MOOC) là một dạng hình thức giáo dục trực tuyến được thiết kế để cung cấp khóa học trực tuyến có sự tham gia lớn từ một lượng lớn người học, có thể là hàng ngàn hoặc thậm chí hàng triệu người. Điều này tạo ra một cơ hội cho việc học qua mạng với quy mô mở rộng, đưa kiến thức và giáo dục trực tuyến đến một đối tượng rộng lớn trên toàn cầu. Các nền tảng nổi tiếng cung cấp MOOCs bao gồm edX, Coursera, Udacity, Khan Academy và nhiều nền tảng khác. MOOCs đã đóng vai trò quan trọng trong việc cung cấp cơ hội học tập toàn cầu và mở rộng phạm vi truy cập đến giáo dục.

Để cho việc người dùng thuận lợi và cảm thấy thoải mái khi truy cập vào hệ thống học online hơn, các nền tảng học online đang áp dụng hệ thống gợi ý khóa học vào website của mình. Nắm bắt được những vấn đề này cùng với sự phát triển của các phương pháp tiên tiến, khóa luận của tôi hướng đến việc xây dựng hệ thống gợi ý bằng các phương pháp đang có hiện nay. Thử nghiệm trên bộ dữ liệu Kaggle và XuetangX đã đưa ra được kết quả khá tốt.

***Từ khóa**: Hệ thống gợi ý, Content-based, Collaborative Filtering, MOOCs.*

Abstract

Currently, Massive Open Online Course (MOOC) systems are a form of online education designed to provide large-scale online courses with extensive participation from a large number of learners, potentially numbering in the thousands or even millions. This creates an opportunity for online learning on an expanded scale, bringing knowledge and education to a wide audience globally. Well-known platforms offering MOOCs include edX, Coursera, Udacity, Khan Academy, and many others. MOOCs have played a significant role in providing global learning opportunities and expanding access to education.

To make user access to online learning systems more convenient and comfortable, online learning platforms are implementing course recommendation systems on their websites. Addressing these issues alongside the advancement of sophisticated methods, my thesis focuses on building a recommendation system using current methods. Testing on Kaggle and XuetangX datasets has yielded promising results.

***Key words**: Recommendation System, Content-based, Collaborative Filtering, MOOCs*

Lời cảm ơn

Lời đầu tiên em xin chân thành cảm ơn tất cả các thầy cô khoa Công nghệ thông tin đã truyền đạt những kiến thức nền tảng, hữu ích cho chúng em, tận tình hướng dẫn và giúp đỡ cho chúng em trong suốt quá trình 4 năm trên giảng đường Đại học và hoàn thành khóa luận tốt nghiệp. Do kiến thức còn hạn hẹp nên trong quá trình thực hiện khóa luận em không thể tránh khỏi những sai sót kính mong quý thầy cô trong hội đồng chỉ dẫn và giúp đỡ.

Em xin gửi lời cảm ơn đến thầy PGS.TS.Nguyễn Việt Anh đã trực tiếp hướng dẫn, góp ý, chia sẻ nhiều kinh nghiệm quý báu, tận tình giúp đỡ và tạo điều kiện để em hoàn thành tốt khóa luận.

Cảm ơn sự giúp đỡ của quý thầy cô trong khoa Công nghệ thông tin đã tạo điều kiện cho chúng em về thời gian, môi trường cũng như kiến thức để hoàn thành khóa luận. Em chắc rằng quyển báo cáo Khóa luận tốt nghiệp này không tránh khỏi những thiếu sót cũng như những những sai sót trong suốt thời gian hoàn thành, em rất cảm ơn nếu nhận được những ý kiến đóng góp của khoa, giảng viên và cũng tất cả các bạn đọc để em có thể hoàn thiện hơn thực hiện tốt hơn trong công việc sau này.

Cảm ơn toàn thể quý thầy cô, cán bộ công nhân viên chức trường Đại học Công Nghệ - Đại học Quốc gia Hà Nội đã tạo điều kiện tốt nhất để chúng em tiếp thu những kiến thức quý báu và thực hiện Khóa luận này. Mình cũng xin gửi lời cảm ơn đến các bạn lớp QH-I/CQ-2020-CA-CLC-3 khóa K65 đã trao đổi chia sẻ kiến thức cũng như những kinh nghiệm quý báu trong thời gian học tập tại trường để mình có nền tảng kiến thức cho bài Khóa luận này.

Em xin chân thành cảm ơn!

Lời cam đoan

Tôi là Vũ Minh Nhật, sinh viên lớp K65-CACLC-3, ngành Khoa học máy tính. Tôi xin cam đoan Khóa luận tốt nghiệp "Nghiên cứu các kỹ thuật khuyến nghị và áp dụng cho gợi ý khóa học" là kết quả của quá trình tự nghiên cứu của bản thân dưới sự hướng dẫn của thầy PGS.TS Nguyễn Việt Anh, không sao chép kết quả của khóa luận, luận văn tốt nghiệp nào trước đó. Khóa luận tốt nghiệp có tham khảo các tài liệu, thông tin theo danh mục tài liệu tham khảo trong báo cáo Khóa luận tốt nghiệp.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về Khóa luận tốt nghiệp của mình. Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong suốt thời gian làm khóa luận.

Sinh viên thực hiện

Vũ Minh Nhật

Mục lục

Danh mục từ viết tắt

CB	Content-based Filtering - Lọc dựa trên nội dung
CF	Collaborative Filtering - Lọc dựa trên cộng tác giữa người dùng và khóa học
DL	Deep Learning - Học sâu
HB	Hybrid - Lọc dựa trên sự kết hợp của nhiều phương pháp
MF	Matrix Factorization - Phương pháp Matrix Factorization
ML	Machine Learning - Học máy
MOOCs	Massive Open Online Course - Hệ thống học online quy mô mở

Danh sách hình vẽ

Danh sách bảng

Giới thiệu

Lý do chọn đề tài

Hệ thống học trực tuyến quy mô mở (MOOCs) đang thu hút sự quan tâm lớn từ cộng đồng người học, chủ yếu bởi vì chúng cung cấp khả năng học một cách linh hoạt và phong cách cởi mở, với sự tham gia và phân phối rộng rãi. Với sự gia tăng đáng kể sự chú ý từ phía người học, số lượng các khóa học online có sẵn ngày càng tăng, mang lại nhiều lựa chọn học tập hơn cho cộng đồng này. Điều này đã tạo ra nhu cầu cần thiết cho các hệ thống gợi ý, giúp người học lựa chọn những khóa học phù hợp với mục tiêu và mong muốn cá nhân của họ.

Từ khi Hệ thống học mở online đầu tiên xuất hiện vào năm 2008, nó đã mở ra một hướng mới cho hệ thống giáo dục, với điểm nổi bật là khả năng truy cập vào các khóa học giáo dục miễn phí. Số lượng các hệ thống học online và học viên đăng ký ngày càng tăng lên mỗi năm, sử dụng nền tảng như edX, Coursera, Udacity, NetEase và iCourse. Đến cuối năm 2019, hơn 900 trường đại học cung cấp MOOC với 13.500 khóa học, thu hút 110 triệu học viên đăng ký. Sau đại dịch Covid-19 năm 2020, xu hướng giáo dục trực tuyến tăng lên, với sự gia tăng đáng kể về số lượng người đăng ký mới và sự ra mắt các khóa học mới từ các trường đại học khác nhau. Số người đăng ký tham gia khóa học online tăng khoảng 25-30% sau đại dịch. Đến cuối năm 2020, có hơn 950 trường đại học cung cấp MOOC với 16.300 khóa học trực tuyến và khoảng 180 triệu người học mới đăng ký. [11]

Với sự gia tăng sử dụng MOOC, dữ liệu được tạo ra bởi MOOC cũng tăng lên, bao gồm thông tin về sở thích và hành vi của người học. Hệ thống gợi ý, được rộng rãi sử dụng trong thương mại và mạng xã hội, cũng có thể được áp dụng trong dữ liệu giáo dục để cung cấp gợi ý cho người học. Mục tiêu của hệ thống này là hỗ trợ người học thông qua các gợi ý liên quan đến đối tượng và thành phần học. Đồng thời, nó cũng góp phần quan trọng vào việc làm phổ biến MOOCs.

Hệ thống gợi ý, hoạt động như bộ lọc thông minh, giúp người dùng lọc thông tin hoặc mục từ số lượng lớn các dịch vụ và sản phẩm dựa trên sở thích và yêu cầu của họ. Gợi ý sẽ làm cho việc lựa chọn dễ dàng hơn, ngay cả khi đối mặt với nhiều tài nguyên học tập, giải quyết vấn đề quá tải thông tin. Nó có thể sử dụng trong MOOC để giúp học viên tìm kiếm khóa học phù hợp từ tập các tài nguyên có sẵn.

Với số lượng lớn khóa học, người học đối mặt với vấn đề lựa chọn khóa học mà không bị choáng ngợp. Hệ thống gợi ý có thể giúp vượt qua vấn đề này bằng cách giúp người học tìm ra các khóa học phù hợp từ một lượng lớn tài nguyên.

Mô tả bài toán

Trong Khóa luận tốt nghiệp này, từ những vấn đề nêu trên, tôi sẽ áp dụng một số phương pháp gợi ý đang hiện có vào bài toán "Gợi ý Khóa học" với nhiều loại khác nhau như gợi ý dựa trên nội dung và dựa trên cộng tác để tạo nên hệ thống gợi ý khóa học đến cho người dùng trên hệ thống học online.

Khó khăn và thách thức

Xây dựng nhãn: Các nhãn có thể được thể hiện một cách tường minh như việc mua sản phẩm hay không, việc đánh giá số sao của người dùng cho sản phẩm, hay việc chấp nhận kết bạn hay không. Những nhãn này còn được gọi là phản hồi tường minh (explicit feedback). Tuy nhiên, không phải hệ thống gợi ý nào cũng phục vụ cho việc mua bán sản phẩm hay không phải người dùng nào cũng sẵn sàng bỏ thời gian ra đánh giá sản phẩm. Rất nhiều trường hợp, nhãn được xây dựng dựa trên những phản hồi ẩn (implicit feedback) từ người dùng. Ví dụ, người dùng có thể không mua hàng nhưng họ đã click vào sản phẩm hoặc dành thời gian đọc về thông tin sản phẩm. Đôi khi, người dùng không click nhưng đã dừng lại ở phần quảng cáo sản phẩm đó trong một thời gian đủ lớn và bật âm thanh lớn để nghe về sản phẩm cũng là một tín hiệu hữu ích.

Dữ liệu lệch: Một khó khăn trong việc xây dựng các mô hình gợi ý là việc nhãn thường bị lệch một cách nghiêm trọng. Số lượng mẫu có nhãn dương (có đánh giá tốt, có click, có mua hàng, v.v.) thường rất nhỏ so với lượng mẫu không có phản hồi. Và việc không có phản hồi chưa chắc đã có nghĩa rằng người dùng không quan tâm tới sản phẩm. Sự chênh lệch nhãn này khiến việc xây dựng mô hình trở lên phức tạp hơn. Việc chọn phương pháp đánh giá cũng hết sức quan trọng.

Hiện tượng đuôi dài: Không những bị lệch về lượng mẫu có và không có phản hồi

mà lượng phản hồi cho các sản phẩm cũng chênh nhau đáng kể. Sẽ có những sản phẩm phổ biến có rất nhiều dữ liệu những cũng có nhiều lần số sản phẩm ít phổ biến có rất ít phản hồi.

Vòng phản hồi (feedback loop): Đôi khi, việc gợi ý cho người dùng dựa hoàn toàn vào phản hồi của họ lại không thực sự thú vị. Nếu một người xem một video về chó mèo và rồi hệ thống gợi ý đúng về các video chó mèo khác và người đó tiếp tục xem thì dần dần người đó sẽ hoàn toàn nhận được các gợi ý về chó mèo mà không có thể loại nào khác. Với các hệ thống gợi ý, nhãn thu được bị ảnh hưởng một phần từ những gì mà hệ thống đã gợi ý trong quá khứ. Nếu tiếp tục phụ thuộc hoàn toàn vào nhãn thì kết quả gợi ý sẽ dần hội tụ về một lượng nhỏ các video. Vòng phản hồi này có tác động tiêu cực tới trải nghiệm người dùng và cần được hạn chế.

Khởi đầu lạnh (cold start): Khởi đầu lạnh xảy ra khi hệ thống không thể đưa ra một gợi ý đáng tin cậy khi lượng dữ liệu có là quá ít. Khi bắt đầu xây dựng hệ thống, khi có người dùng mới, hoặc khi có sản phẩm mới là những trường hợp mà xuất phát lạnh xảy ra. Với những sản phẩm mới, chưa có người dùng nào tương tác với nó, lúc này hệ thống cần có càng nhiều thông tin mô tả về sản phẩm càng tốt để gán nó vào gần với những nhóm đã có tương tác với người dùng. Với những người dùng mới, hệ thống gần như không có thông tin gì về sở thích hay thói quen của họ. Lúc này, hệ thống cần đưa ra những quyết định dựa trên lượng thông tin ít ỏi mà nó có thể suy đoán được như vị trí địa lý, ngôn ngữ, giới tính, tuổi, v.v. Những quyết định ban đầu này có thể ảnh hưởng trực tiếp tới những gợi ý tiếp theo và trải dụng của người dùng. Nếu một hệ thống hoàn toàn mới chưa có cả người dùng và sản phẩm, tốt nhất chưa nên sử dụng ML/DL mà dựa vào các phương pháp đơn giản khác.

Tốc độ xử lý: Cá nhân hóa đồng nghĩa với việc hệ thống phải đưa ra những quyết định khác nhau với mỗi người dùng tại mỗi thời điểm khác nhau. Tại một thời điểm, có thể có hàng triệu người dùng và hàng triệu sản phẩm mà hệ thống cần gợi ý. Tốc độ tính toán là một điểm tối quan trọng trong một hệ thống gợi ý. Nếu bạn có thể xây dựng được một hệ thống có thể dự đoán với độ chính xác cao ở một bộ dữ liệu kiểm thử nhưng không thể triển khai trong thực tế thì cũng vô nghĩa. Nếu trải nghiệm người dùng bị ảnh hưởng bởi tốc độ hiển thị, họ sẽ dần rời khỏi nền tảng [17].

Mục tiêu của đề tài

Nhu cầu cá nhân hóa: Các hệ thống gợi ý khóa học sử dụng phương pháp khuyến nghị có thể tạo ra trải nghiệm học tập cá nhân hóa. Điều này giúp người học tiếp cận nội dung

mà họ quan tâm, tương thích với kiến thức hiện có và mục tiêu học tập cá nhân.

Tăng hiệu suất học tập: Các hệ thống gợi ý khóa học có thể giúp tối ưu hóa quá trình học tập bằng cách đề xuất các khóa học phù hợp và hữu ích dựa trên lịch sử học tập và sở thích của người học. Điều này giúp họ tiết kiệm thời gian và năng lượng khi tìm kiếm thông tin.

Giúp người học khám phá nội dung mới: Hệ thống khuyến nghị có thể giới thiệu người học đến các lĩnh vực mới và có thể quan tâm mà họ chưa biết đến. Điều này có thể thúc đẩy sự đa dạng trong việc tiếp cận kiến thức và kích thích sự tò mò.

Tối ưu hóa giáo dục trực tuyến: Trong ngữ cảnh giáo dục trực tuyến, việc sử dụng hệ thống gợi ý khóa học có thể giúp tối ưu hóa trải nghiệm học tập trực tuyến bằng cách giúp người học dễ dàng chọn lựa giữa các tùy chọn khóa học phù hợp.

Phát triển kinh nghiệm người dùng: Việc sử dụng các phương pháp khuyến nghị có thể tạo ra một trải nghiệm người dùng tích cực và thân thiện, giúp người học cảm thấy hài lòng và có khả năng quay lại sử dụng nền tảng học tập.

Tối ưu hóa doanh nghiệp giáo dục: Các tổ chức giáo dục và nền tảng học tập có thể tận dụng các hệ thống gợi ý để tối ưu hóa quá trình quản lý nguồn lực và nội dung, tăng cường sự hiệu quả và hiệu suất của họ.

Phạm vi đề tài

Trong dự án này, ban đầu ta sẽ đi thu thập và tiền xử lý dữ liệu về các khóa học, bao gồm thông tin về nội dung khóa học, thời lượng, độ khó của khóa học, dữ liệu về đánh giá từ người dùng và dữ liệu của người dùng. Đồng thời ta sẽ đi tìm hiểu và nghiên cứu một vài phương pháp tiên tiến dựa trên học máy, học sâu và đồ thị để áp dụng chúng cho hệ thống gợi ý khóa học. Sau đó, triển khai các thuật toán khuyến nghị để gợi ý các khóa học phù hợp với từng người dùng cụ thể. Các kỹ thuật khuyến nghị có thể bao gồm Collaborative Filtering, Content-Based Filtering.

Phương pháp nghiên cứu

Phương pháp nghiên cứu trong đề tài này là nghiên cứu về các kỹ thuật khuyến nghị cho gợi ý khóa học sau đó sẽ áp dụng một số kỹ thuật vào bài toán. Ban đầu, ta sẽ tập trung tìm hiểu về các thuật toán, mô hình học máy, học sâu liên quan đến hệ thống gợi ý để nắm rõ được nguyên lý và cơ sở lý thuyết của chúng.

Tiếp theo, công việc sẽ đi vào thu thập và phân tích dữ liệu nhằm đảm bảo dữ liệu đưa vào mô hình đủ tin cậy để đưa ra được những gợi ý tốt nhất cho người dùng. Sau đó quá trình tiếp theo sẽ là thử nghiệm mô hình để đánh giá hiệu suất và khả năng của mỗi mô hình bằng một vài phương pháp như là độ chính xác (accuracy), MAP, nDCG, MAE, . . .

Đóng góp và cấu trúc của khóa luận

Dự án tìm hiểu và áp dụng các kỹ thuật khuyến nghị vào gợi ý khóa học trong hệ thống học online sẽ đem lại cho ngành giáo dục đặc biệt là nền giáo dục từ xa đang trở thành xu hướng hiện nay một lợi thế lớn. Quan trọng nhất của một dịch vụ là trải nghiệm người dùng, vì vậy việc gợi ý cho người dùng những gì họ cần và những gì họ nên cần trong tương lai sẽ là một điểm nhấn trong hệ thống giáo dục từ xa (giáo dục online). Chúng sẽ giúp cho người dùng thuận tiện hơn và nhanh hơn trong việc đưa ra quyết định lựa chọn khóa học phù hợp với bản thân.

Ngoài ra, việc sử dụng các kỹ thuật khuyến nghị sử dụng học máy, học sâu sẽ giúp cho hệ thống gợi ý tối đa hóa được khả năng phân tích hành vi người học sâu sắc hơn, nhờ vậy những gợi ý được đưa ra sẽ cá nhân hóa hơn với người dùng. Hơn nữa sẽ tạo cho người dùng thoải mái và cảm giác thuận tiện khi sử dụng dịch vụ học từ xa.

Từ việc xây dựng hệ thống gợi ý khóa học để thỏa mãn được nhu cầu người dùng, ngành giáo dục từ xa sẽ ngày càng được quan tâm hơn trên thế giới và đặc biệt tại Việt Nam và đem lại nguồn thu đáng kể cho ngành giáo dục từ xa. Không những thế, nhờ việc tự học từ xa, người dùng ngày càng có nhiều kiến thức hơn, từ đó thế giới ngày càng phát triển và văn minh hơn.

Tóm lại, hệ thống gợi ý đóng vai trò quan trọng lớn trong việc tối ưu hóa trải nghiệm học tập trực tuyến bằng cách cá nhân hóa, tăng cường tiện ích và động lực cho người dùng.

Khóa luận tốt nghiệp của tôi gồm Phần Giới thiệu, ba Chương chính và Phần Kết luận, như sau:

Giới thiệu. Phần này là giới thiệu động lực thúc đẩy để tôi xây dựng lên khóa luận này, trình bày vấn đề của bài toán và cuối cùng là khó khăn và thách thức tôi gặp phải khi bắt tay vào làm khóa luận này.

Chương 1: Cơ sở lý thuyết và các nghiên cứu liên quan. Tôi sẽ nói qua về cơ sở lý thuyết, giới thiệu một vài những nghiên cứu liên quan trước đó có cùng định hướng

với chúng tôi. Cụ thể như làm cùng bài toán, làm cùng phương pháp, làm cùng dữ liệu hay làm cùng ngôn ngữ.

Chương 2: Phương pháp. Chương này sẽ đề cập, phân tích các phương pháp mà tôi sẽ sử dụng trong Khóa luận tốt nghiệp này

Chương 3: Thực nghiệm, kết quả và đánh giá mô hình. Chương này tôi sẽ thực nghiệm các phương pháp đã nêu ở Chương trước và đánh giá kết quả đầu ra của bài toán

Kết luận. Phần này kết thúc khóa luận nghiên cứu bằng cách tóm tắt các đề xuất và kết luận quan trọng. Ngoài ra, chúng tôi nêu ra những hạn chế của các mô hình của chúng tôi và chỉ ra một số phần mở rộng hơn nữa trong tương lai.

Chương 1

Cơ sở lý thuyết và các nghiên cứu liên quan

1.1 Một số định nghĩa tổng quan

1.1.1 Hệ thống khuyến nghị là gì?

1.1.1.1 Định nghĩa

Hệ thống khuyến nghị (Recommendation System - RS) là một loại phần mềm hoặc công nghệ được thiết kế để đề xuất và gợi ý sản phẩm, dịch vụ hoặc thông tin một cách tự động đến người dùng dựa trên các dữ liệu về sở thích, hành vi hoặc thuộc tính của họ. Mục tiêu chính của hệ thống khuyến nghị là cung cấp trải nghiệm cá nhân hóa cho người dùng và giúp họ khám phá những nội dung mới và phù hợp với mong muốn của mình.

1.1.1.2 Tại sao thế giới lại cần hệ thống gợi ý

Giả sử công ty của bạn hoạt động ở Việt Nam. Điều đó không có nghĩa rằng 95 triệu người Việt sẽ mua hàng của công ty bạn mà chỉ một phần trong số đó có nhu cầu. Trong số các khách hàng có nhu cầu thì không phải ai cũng sẽ mua tất cả các sản phẩm của công ty mà họ có khi chỉ mua một vài sản phẩm mà họ quan tâm.

Một công ty nếu muốn tối đa hóa lợi nhuận thì điều quan trọng nhất họ phải hiểu được khách hàng họ cần gì. Recommendation là một thuật toán kì diệu có thể giúp bạn thực hiện điều đó. Hãy tưởng tượng tình hình kinh doanh sẽ ra sao nếu không có thuật

toán này. Một loạt các hệ quả mà ta có thể hình dung ra:

- Công ty không thể tìm được đúng khách hàng tiềm năng khi người có nhu cầu đối với một sản phẩm lại không được chào bán. Người không có nhu cầu lại bị tiếp cận mời chào. Điều này gây lãng phí thời gian và dẫn tới mất thiện cảm của khách hàng về dịch vụ của công ty.
- Hiệu quả marketing gần như là không đáng kể nếu không tìm đúng tập khách hàng. Chi phí quảng cáo, chi phí cho nhân viên sale tăng lên nhưng doanh thu vẫn thế. Theo một nghiên cứu, một số doanh nghiệp sẵn sàng bỏ ra từ 30-40% lợi nhuận cho việc marketing. Đây là một chi phí không hề nhỏ nhưng để tồn tại họ không thể ngừng đốt tiền. Cuối cùng người hưởng lợi nhiều nhất lại là Google, Facebook.

Chính vì thể ngày nay thuật toán recommendation được phát triển và ứng dụng rộng rãi trong nhiều doanh nghiệp thuộc đa dạng các lĩnh vực khác nhau như thương mại điện tử, tài chính, ngân hàng, kinh doanh, bán lẻ, phim ảnh,....

Hệ thống gợi ý (Recommendation system) ngày càng trở thành một phần không thể thiếu trong các sản phẩm online với số lượng người dùng ngày càng tăng. Sự phổ biến của các sản phẩm cá nhân hóa trực tuyến có mục đích chính là mang đến cho người dùng những sản phẩm phù hợp nhất hoặc cải thiện trải nghiệm của họ. Việc quảng cáo sản phẩm đến đúng đối tượng sẽ tăng khả năng mua hàng.

Gợi ý một video mà người dùng có khả năng thích hoặc gợi ý kết bạn đến đúng đối tượng cũng giúp họ duy trì sự ở lại trên nền tảng của bạn lâu hơn. Khi họ ở lại trên nền tảng của bạn lâu hơn, họ sẽ thấy nhiều quảng cáo hơn và lợi nhuận từ quảng cáo cũng tăng lên.

1.1.2 MOOCs

1.1.2.1 MOOCs

Hệ thống học online quy mô mở (Massive Open Online Course - MOOC) là một dạng hình thức giáo dục trực tuyến được thiết kế để cung cấp khóa học trực tuyến có sự tham gia lớn từ một lượng lớn người học, có thể là hàng ngàn hoặc thậm chí hàng triệu người. Điều này tạo ra một cơ hội cho việc học qua mạng với quy mô mở rộng, đưa kiến thức và giáo dục trực tuyến đến một đối tượng rộng lớn trên toàn cầu.

Các nền tảng nổi tiếng cung cấp MOOCs bao gồm edX [18], Coursera [19], Udemy

[20], Khan Academy [21] và nhiều nền tảng khác. MOOCs đã đóng vai trò quan trọng trong việc cung cấp cơ hội học tập toàn cầu và mở rộng phạm vi truy cập đến giáo dục.

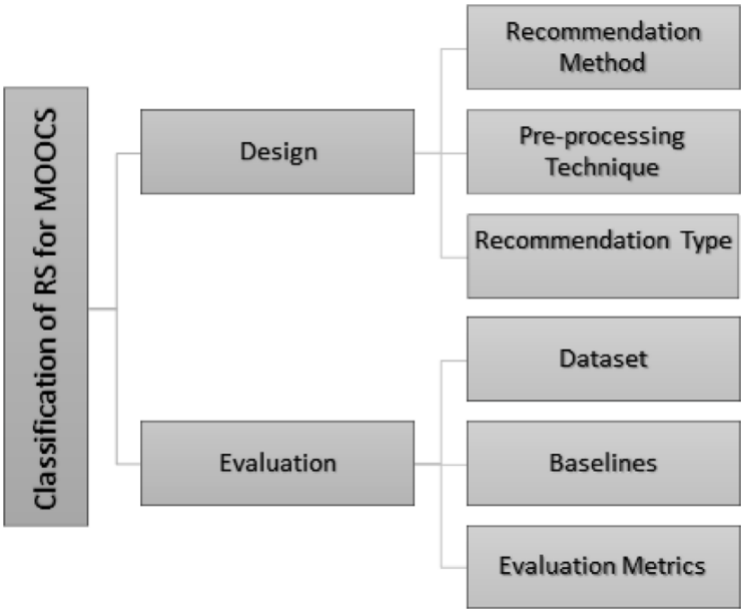
1.1.2.2 Sự cần thiết của hệ thống học online

Các khóa học online mang lại nhiều lợi ích đến người dùng, bao gồm:

- Tiện lợi và linh hoạt: Khóa học online cho phép người dùng học tập ở bất kỳ đâu và bất kỳ khi nào họ muốn, chỉ cần có kết nối internet. Điều này tạo điều kiện thuận lợi cho họ tự quản lý thời gian và tạo lịch học linh hoạt phù hợp với lịch trình cá nhân và chuyên môn của họ.
- Truy cập đến nhiều tài nguyên học tập: Các khóa học online thường cung cấp một loạt các tài liệu học tập, video, bài giảng, bài kiểm tra và các tài liệu bổ sung khác. Điều này giúp người học có thể tiếp cận nhiều nguồn tài liệu và phương tiện học tập để nâng cao kiến thức và kỹ năng của mình.
- Tiết kiệm chi phí: So với việc tham dự các khóa học truyền thống tại các trường học hay tổ chức đào tạo, các khóa học online thường có chi phí thấp hơn hoặc thậm chí miễn phí. Điều này giúp tiết kiệm chi phí cho việc đi lại, ăn ở và học phí.
- Học theo tốc độ cá nhân: Trong môi trường học online, người học có thể tiến triển qua các nội dung học tập theo tốc độ của họ. Họ có thể tóm tắt hoặc lặp lại các bài học khi cần thiết, tạo điều kiện cho việc hiểu sâu hơn về các chủ đề cụ thể.
- Học cùng lúc với công việc hoặc gia đình: Khóa học online cho phép người học điều chỉnh việc học theo lịch trình cá nhân của họ, điều này giúp họ cân bằng giữa học tập, công việc và cuộc sống gia đình.
- Học từ các chuyên gia hàng đầu: Nhiều khóa học online được thiết kế và giảng dạy bởi các chuyên gia hàng đầu trong lĩnh vực của họ. Điều này mang lại cho người học cơ hội tiếp cận kiến thức và kinh nghiệm từ những người có uy tín và kinh nghiệm sâu rộng.
- Mở rộng mạng lưới xã hội và chuyên môn: Tham gia vào các khóa học online cũng mở ra cơ hội để giao lưu và kết nối với những người học khác từ khắp nơi trên thế giới. Điều này giúp mở rộng mạng lưới xã hội và chuyên môn của người học.

1.2 Công việc chính khi áp dụng hệ thống khuyến nghị để gợi ý khóa học

Trong hệ thống khuyến nghị, ta sẽ có hai phần chính đó là: Thiết kế hệ thống và Đánh giá hệ thống khuyến nghị 1.1.



Hình 1.1: Các công việc khi triển khai hệ thống gợi ý. [11]

1.2.1 Thiết kế hệ thống gợi ý

Trong Khóa luận này, chúng ta sẽ triển khai về 3 phần sau:

- **Recommendation Type (Loại khuyến nghị):** Trong phân loại này, chúng ta đã quan sát được tổng thể của MOOC trong đó RS được triển khai. Những hệ thống gợi ý có thể được sử dụng trong nhiều loại khác nhau của MOOCs như là Course, Peer, Learning element, Teacher support và Thread Recommender. Trong khóa luận này, ta tập trung chủ yếu vào Course Recommendation (gợi ý khóa học).
- **Pre-processing Technique (Các kỹ thuật tiền xử lý dữ liệu):** Được dùng trên bộ dữ liệu để chuẩn bị dữ liệu cho hệ thống gợi ý. Phần này ta sẽ tập trung phân tích dữ liệu từ bộ dữ liệu Kaggle, XuetangX để có cái nhìn tổng quan về dữ liệu hơn.
- **Recommendation Method (Phương pháp khuyến nghị):** Phân loại này đề cập đến loại phương pháp đề xuất (tức là CF, CBF, HB) được sử dụng để thiết kế hệ thống gợi ý.

1.2.2 Đánh giá hệ thống gợi ý

Đánh giá hệ thống gợi ý, ta cũng sẽ nghiên cứu về 3 phần sau:

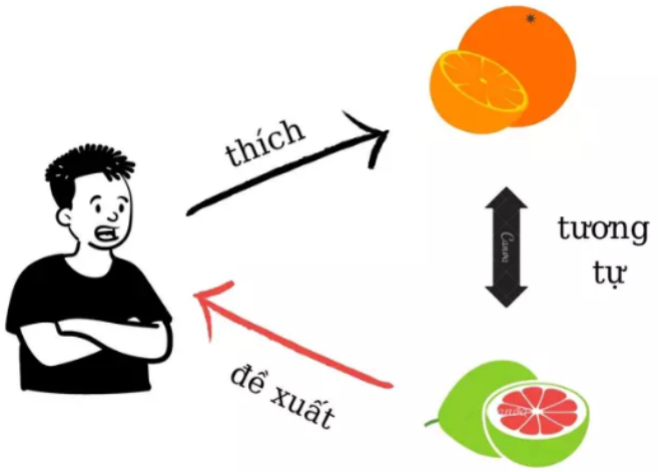
- **Dataset (Bộ dữ liệu):** Tìm hiểu về các bộ dữ liệu được sử dụng trong bài Khóa luận này.
- **Evaluation Metrics (Các metrics dùng để đánh giá):** Nghiên cứu các metrics thường dùng để đánh giá hệ thống khuyến nghị.
- **Baseline (Cơ sở đánh giá):** Khám phá các cơ sở thường được sử dụng để đánh giá hệ thống gợi ý.

1.3 Các phương pháp gợi ý hiện nay

CF (Collaborative Filtering), CBF (Content-based Filtering) và HB (Hybrid) là những phương pháp phổ biến nhất hiện nay. Trong đó HB bao gồm các công việc nghiên cứu bằng việc kết hợp hai thuật toán khuyến nghị đang tồn tại với nhau.

1.3.1 Content-based Filtering

Hệ thống gợi ý dựa trên nội dung là hệ thống đơn giản nhất. Trong hệ thống này, mô hình dự đoán liệu một người dùng có thích một sản phẩm không dựa trên lịch sử dữ liệu của người dùng đó đối với các sản phẩm tương tự. Độ quan tâm của những người dùng khác không được sử dụng 1.2.



Hình 1.2: Ví dụ về Content-based Filtering [22].

Nhìn dưới góc độ xây dựng mô hình dự đoán, hệ thống xây dựng một mô hình ML

cho mỗi người dùng. Mỗi sản phẩm sẽ được mô tả bởi một vector đặc trưng. Để dự đoán mức độ yêu thích của mỗi người dùng đối với một sản phẩm, ta chỉ cần đưa vector đặc trưng của sản phẩm vào mô hình đã được xây dựng cho người dùng đó.

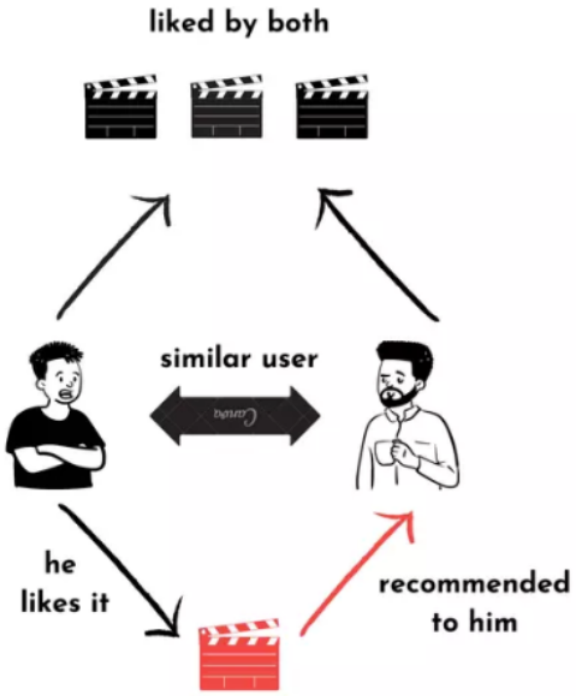
Ứng dụng của phương pháp gợi ý dựa trên nội dung cho gợi ý khóa học: Content-Based Filtering (CBF) là một kỹ thuật lọc trong hệ thống đề xuất (Recommendation Systems) mà sự gợi ý dựa trên nội dung hay đặc tính của các mục đã được đánh giá. Trong MOOCs (Massive Open Online Courses), Content-Based Filtering sử dụng thông tin về nội dung của các khóa học và sở thích của người học để tạo ra các khuyến nghị cá nhân hóa.

- **Xây dựng Hồ sơ Người Học (User Profile):** Thu thập thông tin về sở thích, kỹ năng, và lịch sử học tập của người học. Điều này có thể bao gồm việc theo dõi các khóa học họ đã tham gia, các chủ đề họ quan tâm, hoặc bất kỳ dữ liệu hành vi nào khác.
- **Xây dựng Hồ sơ Khóa Học (Item Profile):** Mô tả nội dung của các khóa học bằng cách sử dụng thông tin như chủ đề, mức độ khó, loại tài nguyên (bài giảng, bài đọc, bài kiểm tra), và các đặc điểm khác.
- **Vector Hóa (Vectorization):** Chuyển đổi thông tin trong hồ sơ người học và hồ sơ khóa học thành các vector số để có thể tính toán sự tương đồng.
- **Tính Điểm Tương Đồng (Similarity Score):** Sử dụng các phương pháp như cosine similarity [23] để đo lường sự tương đồng giữa hồ sơ người học và hồ sơ khóa học. Tạo Khuyến Nghị Dựa trên sự tương đồng, đưa ra các khóa học mà người học có thể quan tâm. Càng cao điểm tương đồng, càng được ưu tiên. Ví dụ, nếu người học thường xuyên đăng ký và hoàn thành các khóa học về Machine Learning, hệ thống có thể gợi ý những khóa học mới về Machine Learning hoặc các chủ đề liên quan.

1.3.2 Collaborative Filtering

Collaborative filtering hay còn gọi là lọc tương tác, sử dụng sự tương tác qua lại trong hành vi mua sắm giữa các khách hàng để tìm ra sở thích của một khách hàng đối với một sản phẩm. Hầu hết các hành vi hoặc sở thích của mọi người đều có những đặc điểm chung và có thể nhóm lại thành các nhóm tương đồng. Một phụ nữ A nếu đến siêu thị mua dầu ăn thường mua thêm nước tương và nước mắm. Hành vi này lặp lại đối với 100 lượt mua sắm là 90 lần thì khả năng cao một phụ nữ B nếu mua dầu ăn cũng sẽ mua

thêm nước tương và nước mắm. Từ đó sẽ khuyến nghị sản phẩm cho khách hàng dựa trên hành vi của các khách hàng khác liên quan nhất. 1.3



Hình 1.3: Ví dụ về Collaborative Filtering [24].

Có hai loại chính của Collaborative Filtering:

- **User-Based Collaborative Filtering:** Dựa trên việc tìm kiếm các người dùng có hành vi giống nhau với người dùng hiện tại. Nếu người dùng A và B có sở thích giống nhau, và người dùng A thích một mục mà người dùng B chưa thấy, thì hệ thống có thể gợi ý mục đó cho người dùng B.
- **Item-Based Collaborative Filtering:** Tìm kiếm các mục tương tự dựa trên lịch sử tương tác của người dùng. Nếu người dùng A thích một mục và có người dùng B giống A trong quá khứ, và người dùng B chưa thấy mục đó, hệ thống có thể gợi ý mục đó cho người dùng B.

Quy trình hoạt động của Collaborative Filtering bao gồm:

- Xây dựng Ma trận Xếp hạng (Rating Matrix): Mỗi hàng của ma trận biểu diễn một user và mỗi cột biểu diễn một course. Ô trong ma trận chứa điểm xếp hạng hoặc tương tác giữa user và course.

- Tính Điểm Tương Đồng: Sử dụng các phương pháp như cosine similarity [5] để đo lường sự tương đồng giữa user hoặc item.
- Dự đoán Điểm Xếp hạng: Dựa trên thông tin tương đồng, dự đoán điểm xếp hạng cho các mục chưa được xếp hạng.
- Tạo Khuyến Nghị: Gợi ý các mục với điểm dự đoán cao cho người dùng.

Ở mức độ lớn, Collaborative Filtering trong MOOCs giúp tạo ra các khuyến nghị dựa trên kinh nghiệm học tập của cộng đồng người học. Điều này có thể bao gồm việc gợi ý khóa học mới dựa trên sở thích và kết quả học tập của những người học có hành vi tương tự.

Tuy nhiên, Collaborative Filtering trong MOOCs cũng đối mặt với những thách thức như vấn đề lạc hậu khi có người học mới tham gia, và có thể không hiệu quả nếu người học có sự tương tác rất ít hoặc không có sự tương tác gì.

1.4 Công nghệ sử dụng

Trong hệ thống gợi ý này, tôi lựa chọn sử dụng Python [25] vì nó là công cụ hỗ trợ rất tốt cho các mô hình học máy học sâu mà tôi sẽ sử dụng trong Khóa luận này. Và các Framework nổi tiếng của Python được sử dụng tới như là Pandas [26], Numpy [27] để xử lý các ma trận dữ liệu đầu vào và tính toán dựa trên ma trận. Ngoài ra, tôi còn sử dụng thêm hai Framework nữa đó là Matplotlib [28] và Seaborn [29] để trực quan hóa dữ liệu giúp ta hiểu hơn về tập dữ liệu đầu vào. Sau đó để xây dựng mô hình ML/DL để phục vụ cho bài toán, Pytorch [30] và Tensorflow [31] là sự lựa chọn hữu hiệu nhờ tính linh hoạt và dễ sử dụng của nó.

1.5 Các nghiên cứu liên quan

Hiện nay, có nhiều nghiên cứu đang được thực hiện về hệ thống gợi ý, bao gồm các chủ đề như: Cải tiến các thuật toán gợi ý: Nghiên cứu về cách cải tiến các thuật toán gợi ý, bao gồm CF, CBF, và HB. Mục tiêu là tăng cường độ chính xác và hiệu suất của hệ thống gợi ý.

- Nghiên cứu về cách sử dụng các kỹ thuật học máy và học sâu để xây dựng các mô hình gợi ý phức tạp và hiệu quả hơn.
- Nghiên cứu về cách tích hợp yếu tố cảm xúc vào hệ thống gợi ý, bao gồm nhận

dạng cảm xúc từ phản hồi của người dùng và sử dụng thông tin này để cá nhân hóa gợi ý.

- Nghiên cứu về cách áp dụng hệ thống gợi ý trong các lĩnh vực đặc biệt như y tế, giáo dục, du lịch, và thương mại điện tử để cải thiện trải nghiệm người dùng và tăng cường hiệu suất kinh doanh.
- Nghiên cứu về cách tối ưu hóa việc đa dạng hóa gợi ý để người dùng nhận được các gợi ý phong phú và đa dạng hóa nội dung tiếp cận.
- Nghiên cứu về cách sử dụng thông tin ngữ cảnh như vị trí địa lý, thời gian, và trạng thái người dùng để cải thiện độ chính xác của hệ thống gợi ý.
- Nghiên cứu về cách bảo vệ thông tin cá nhân của người dùng trong quá trình gợi ý và giữ cho hệ thống an toàn và tin cậy.
- Những nghiên cứu này đóng vai trò quan trọng trong việc phát triển và cải thiện các hệ thống gợi ý để đáp ứng nhu cầu ngày càng đa dạng của người dùng và doanh nghiệp.

Việc gợi ý khóa học theo [5] đã duy trì đồ thị user-course bằng việc tạo một node điều hướng từ người dùng tới một khóa học mà người dùng đã học trong khi [2] sử dụng Markov Process. [9] sử dụng cây Monte-Carlo để duy trì các cụm khóa học của những khóa học tương đồng và cũng để đưa ra các khóa học được cá nhân hóa cho sinh viên dựa trên cơ sở về lịch sử của những sinh viên tương đồng.

[14] đã sử dụng TF-IDF để xác định những kỹ năng của người học từ profile LinkIn của họ và các thông tin khóa học, và sử dụng những thông tin này để đưa ra gợi ý. [15] đã mở rộng chức năng mục tiêu MF hiện có bằng cách thêm kỹ năng người dùng và ma trận kỹ năng khóa học cùng với ma trận khóa học và người dùng trong quá trình hình thành vấn đề đề xuất. Họ duy trì hai ma trận, ma trận người dùng-khóa học và ma trận kỹ năng người dùng khóa học và sử dụng cả hai ma trận này cho quy trình đề xuất khóa học.

Kỹ thuật lọc cộng tác là kỹ thuật hoàn thiện nhất và được triển khai phổ biến nhất. Tính năng lọc cộng tác đề xuất các mục bằng cách xác định những người dùng khác có cùng sở thích; nó sử dụng ý kiến của họ để giới thiệu các mặt hàng cho người dùng đang hoạt động. Hệ thống tư vấn hợp tác đã được triển khai trong các lĩnh vực ứng dụng khác nhau. GroupLens là một kiến trúc dựa trên tin tức sử dụng các phương pháp cộng tác để hỗ trợ người dùng định vị các bài viết từ cơ sở dữ liệu tin tức khổng lồ [1].

Một mô hình phổ biến của mô hình CF là tham số hóa người dùng và các mục dưới dạng các phần nhúng và tìm hiểu các tham số nhúng bằng cách xây dựng lại các tương tác lịch sử của người dùng. Ví dụ, các mô hình CF trước đây như matrix factorization (MF) [13] ánh xạ ID của một người dùng (hoặc một mục) thành một vector nhúng.

Vượt ra ngoài việc chỉ sử dụng thông tin ID, một loại phương pháp CF khác xem xét các mục lịch sử như là các đặc điểm tồn tại trước đó của một người dùng, nhằm mục tiêu cải thiện biểu diễn người dùng. Ví dụ, FISM [10] sử dụng trung bình có trọng số của các nhúng ID của các mục lịch sử như là nhúng của người dùng mục tiêu. Gần đây, các nhà nghiên cứu nhận ra rằng các mục lịch sử đóng góp khác nhau vào việc hình thành sở thích cá nhân. Với mục tiêu này, các cơ chế chú ý được giới thiệu để bắt kịp các đóng góp biến thiên, như ACF [4] và NAIS [7], để tự động học sự quan trọng của mỗi mục lịch sử. Khi xem xét lại các tương tác lịch sử như là một đồ thị hai phần người dùng-mục, các cải tiến hiệu suất có thể được quy về việc mã hóa khu vực láng giềng cục bộ - các hàng xóm một bước đi - cải thiện việc học nhúng.

Mặt khác, các kỹ thuật dựa trên nội dung sẽ kết nối tài nguyên nội dung với đặc điểm của người dùng. Các kỹ thuật lọc dựa trên nội dung thường dựa trên dự đoán của chúng về thông tin của người dùng và chúng bỏ qua những đóng góp từ những người dùng khác như trường hợp của các kỹ thuật cộng tác[3].

Tóm lại, trong phần cơ sở lý thuyết này và các nghiên cứu liên quan ta đã nắm rõ được những kỹ thuật cơ bản, các hệ thống gợi ý hiện nay, các công nghệ thường dùng để triển khai hệ thống gợi ý. Ngoài ra ta còn tìm hiểu được những công việc liên quan tới việc nghiên cứu về hệ thống gợi ý khóa học qua các bài báo. Đặc biệt là cơ sở để chúng ta áp dụng những kỹ thuật hiện có vào bài toán “Gợi ý khóa học”.

Chương 2

Phương pháp

Trong chương này, ban đầu ta sẽ tìm hiểu qua về tập dữ liệu sử dụng trong bài khóa luận này, sau đó, lần lượt ta sẽ đi nghiên cứu các phương pháp hiện có bao gồm: Gợi ý dựa theo nội dung mô tả khóa học, gợi ý dựa trên đánh giá của tất cả người dùng, gợi ý dựa trên sự tương tác giữa người dùng và khóa học. Trong phương pháp gợi ý dựa trên sự tương tác giữa người dùng và khóa học, ta sẽ đi nghiên cứu hai mô hình: Matrix Factorization, LightGCN.

2.1 Tổng quan về Dữ liệu

2.1.1 Dữ liệu dành cho gợi ý dựa trên nhân khẩu học và gợi ý dựa trên nội dung khóa học

Tập dữ liệu được lấy từ hệ thống Kaggle [32], và bộ dữ liệu bao gồm thông tin về các khóa học trên Coursera như là tiêu đề, mô tả về khóa học, thời lượng, độ khó, rating ...

Tập dữ liệu này bao gồm 1000 khóa học khác nhau, với 12 trường thông tin.

- course_title : Tiêu đề khóa học
- course_organization : Tổ chức cung cấp khóa học
- course_certificate_type: Loại chứng chỉ
- course_time: Thời lượng khóa học
- course_rating: Điểm số người dùng đánh giá cho khóa học
- course_reviews_num: Số lượng người dùng đánh giá khóa học

	course_title	course_organization	course_certificate_type	course_time	course_rating	course_reviews_num	course_difficulty	course_url	course_students_enrolled	course_skills	co
0	(ISC)² Systems Security Certified Practitioner...	ISC2	Specialization	3 - 6 Months	4.7	492	Beginner	https://www.coursera.org/specializations/sscp-...	6,958	['Risk Management', 'Access Control', 'Asset',...	
1	.NET FullStack Developer	Board Infinity	Specialization	1 - 3 Months	4.3	51	Intermediate	https://www.coursera.org/specializations/dot-n...	2,531	['Web API', 'Web Development', 'Cascading Styl...	
2	21st Century Energy Transition: how do we make...	University of Alberta	Course	1 - 3 Months	4.8	62	Beginner	https://www.coursera.org/learn/21st-century-en...	4,377	[]	
3	A Crash Course in Causality: Inferring Causal...	University of Pennsylvania	Course	1 - 3 Months	4.7	517	Intermediate	https://www.coursera.org/learn/crash-course-in...	39,004	['Instrumental Variable', 'Propensity Score Ma...	
5	AI & Law	Lund University	Course	1 - 4 Weeks	4.8	370	Beginner	https://www.coursera.org/learn/ai-law	26,023	[]	
...	...	...	...	...	...	...	...	...	...	...	
995	Étudier en France: French Intermediate course ...	École Polytechnique	Course	1 - 3 Months	4.8	4.6k	Intermediate	https://www.coursera.org/learn/etudier-en-france	519,333	[]	

Hình 2.1: Tổng quan về tập dữ liệu Kaggle

- course_difficulty: Độ khó của khóa học
- course_url: Liên kết dẫn tới khóa học
- course_students_enrolled: Số lượng người dùng tham gia khóa học
- course_skills: Kỹ năng mà khóa học đào tạo
- course_summary: Tổng kết chung về khóa học
- course_description: Mô tả chi tiết về khóa học

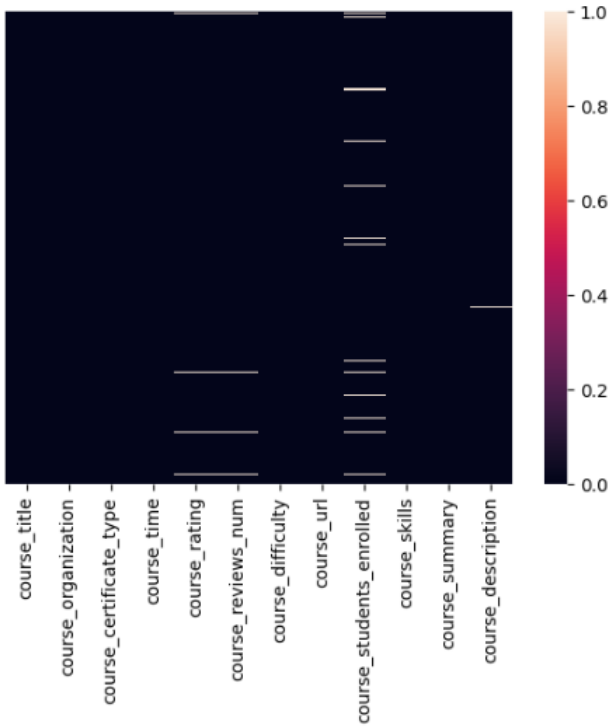
Đầu tiên trước khi đi vào áp dụng các kỹ thuật gợi ý, ta sẽ cùng đi xem xem tổng quan về bộ dữ liệu này có gì.

Ban đầu ta thấy được bộ dữ liệu này có số lượng bị thiếu ở phần course_student_enrolled, course_rating, course_review_num, course_description nên ta sẽ xóa những khóa học bị thiếu vì những trường thông tin trên rất quan trọng trong việc gợi ý cho nên những trường thông tin này không thể thiếu được.

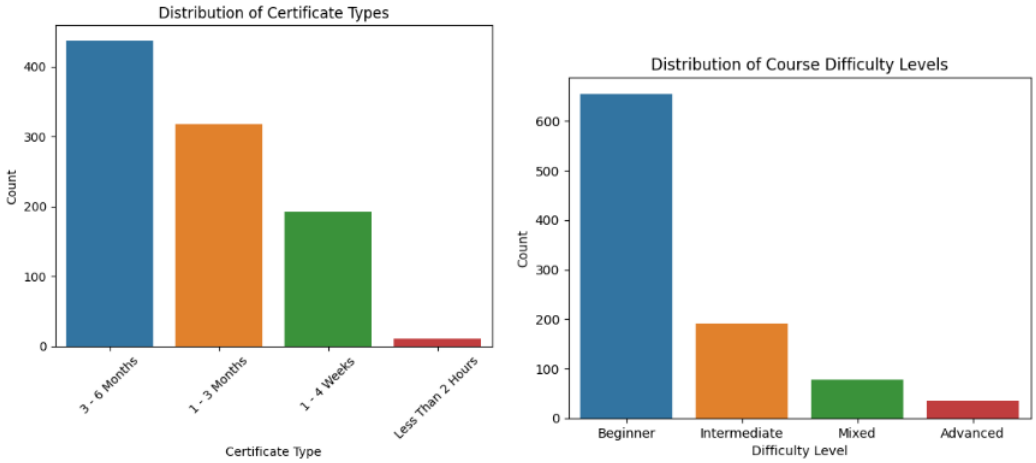
Ở đây, các khóa học đa phần có mức độ là dễ dàng và chủ yếu là các khóa học dài hạn từ 1-4 tuần cho tới 6 tháng (Hình 2.3).

Hình 2.4 cho ta thấy được top 10 khóa học có số lượng người dùng tham gia nhiều nhất.

Dưới đây (Hình 2.5) là top 10 tổ chức cung cấp khóa học nhiều nhất. Thật vậy, khi truy cập vào Coursera ta dễ dàng tìm được những khóa học trực tuyến của Google và IBM.



Hình 2.2: Các dữ liệu null trong tập dữ liệu

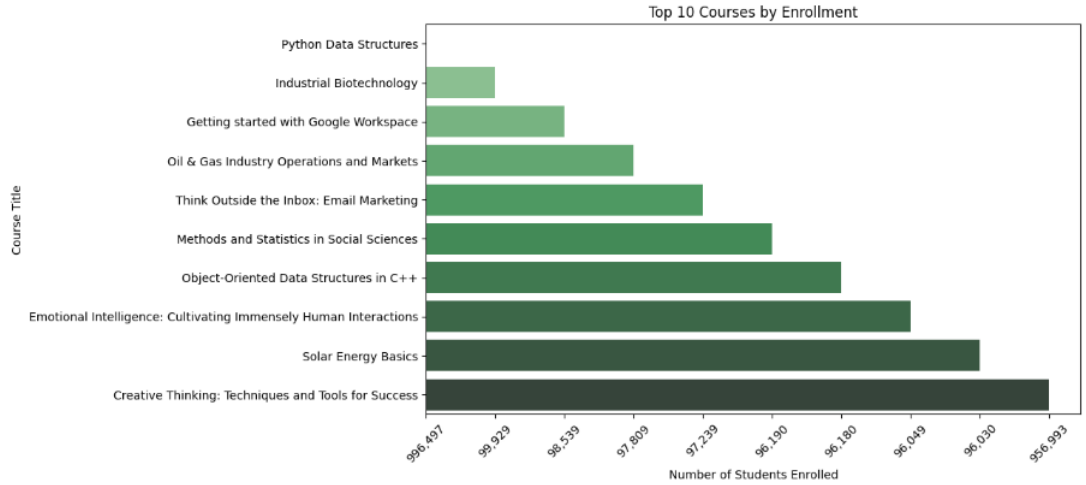


Hình 2.3: a. Các loại chứng chỉ theo thời gian, b. Độ khó

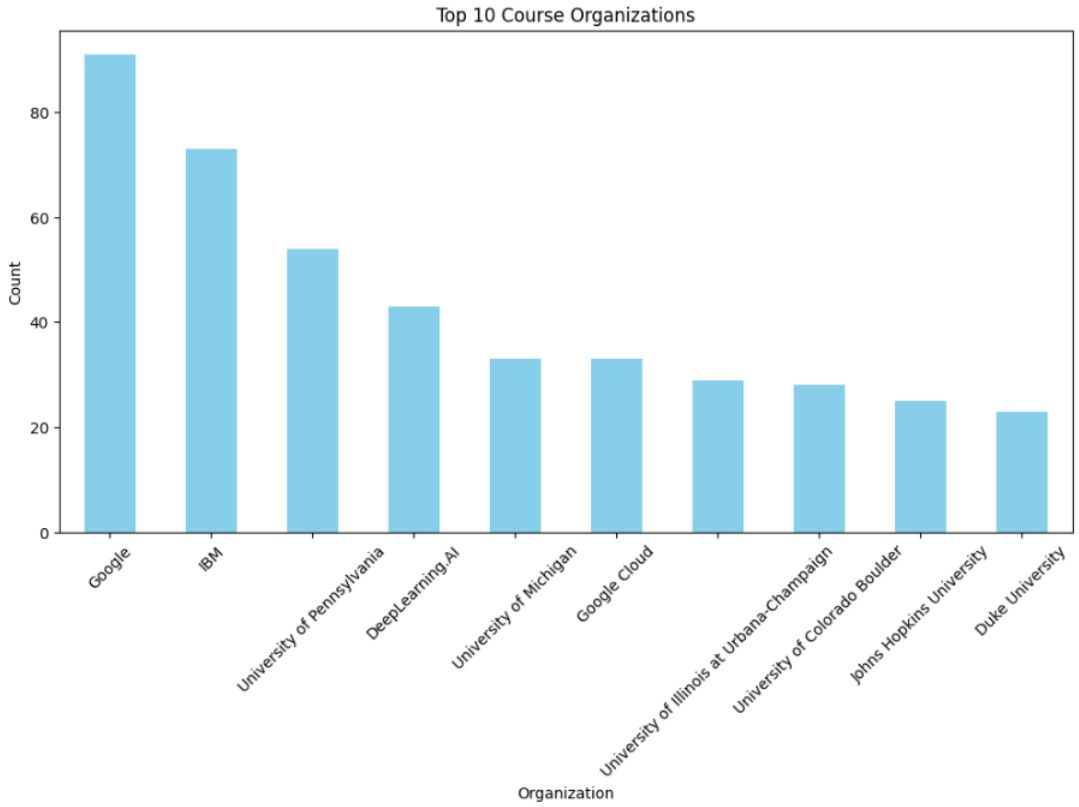
Như hình 2.6, ta thấy được đa phần các khóa học ở Coursera được đánh giá cao từ 4.3-5 sao. Vì vậy Coursera được coi là nền tảng hữu hiệu nhất cho việc tìm và học thêm kiến thức và kỹ năng từ xa.

2.1.2 Dữ liệu dành cho Gợi ý khóa học dựa trên cộng tác

Các tệp này là bộ dữ liệu XuetangX [33] được sử dụng trong bài viết "Understanding Dropouts in MOOCs" trong AAAI 2019 [6]. Các tệp log tracking (2015/08 - 2016/08,

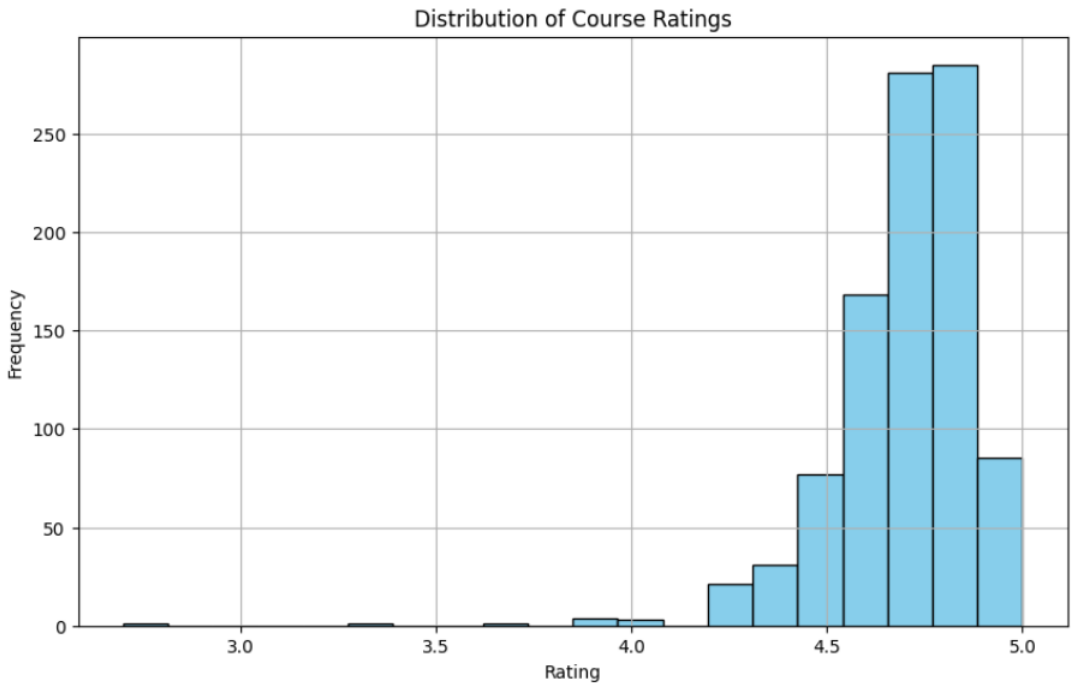


Hình 2.4: Top các khóa học có nhiều người tham gia.



Hình 2.5: Top các tổ chức cung cấp khóa học.

2016/08 - 2017/08) bao gồm tất cả các hoạt động học tập của người dùng trong nền tảng XuetangX từ 2015/08 đến 2017/08. Các nhật ký này là dữ liệu hỗ trợ cho các phân tích trong bài báo. User Profile: là thông tin của người dùng XuetangX, bao gồm giới tính, năm sinh và trình độ học vấn. Course Information: bao gồm ngày bắt đầu khóa học, ngày kết thúc khóa học, danh mục khóa học và loại khóa học.



Hình 2.6: Phân bố về rating trong tập dữ liệu Kaggle.

2.1.2.1 Tập dữ liệu Tracking log

Định dạng của các tệp json này như sau:

```
[
  [course_id,
    {user_id:
      { session_id:
        [
          [activity_event_1, time_1],
          [activity_event_2, time_2],
        ... ],
      ... },
    ...}
  ],
...]
```

Hình 2.7: Định dạng tệp json

Trong đó:

- cousrse_id: id khóa học
- user_id: id người dùng
- session_id: id phiên truy cập vào khóa học
- activity_event, time: hoạt động trong khóa học và thời gian hoạt động.

2.1.2.2 Tập dữ liệu User Profile

- user_id: id của người dùng
- gender: giới tính của người dùng
- education: bậc giáo dục người dùng
- birth: năm sinh người dùng

2.1.2.3 Tập dữ liệu Course Information

- id: id dạng số của khóa học
- course_id: id dạng chuỗi của khóa học, sử dụng cho lịch sử giám sát
- start: thời gian bắt đầu khóa học
- end: thời gian kết thúc khóa học
- course_type: dạng khóa học (0: khóa học có người hướng dẫn, 1: khóa tự học)
- category: loại khóa học

2.1.2.4 Tiền xử lý dữ liệu

Ở đây ta sẽ sử dụng chủ yếu bộ dữ liệu Tracking log, tệp dữ liệu này chứa lịch sử giám sát hành động của người học theo định dạng json.

Tuy nhiên, ở đây tracking log chỉ ghi lại các lần truy cập của người dùng vào một khóa học nào đó, chứ không phải là sự đánh giá của người học đối với khóa học. Do đó, dựa trên số lần truy cập vào khóa học, ta sẽ đưa ra được chỉ số đánh giá của người học đối với khóa học đó như thế nào, là tích cực tương tác hay tiêu cực, và chỉ số đánh giá được dựa trên thang điểm 5.

Theo phân tích trên bộ dữ liệu, 99% người dùng đều truy cập tối đa 10 lần trên một khóa học. Vì vậy, ta sẽ chọn 10 làm mốc, sau đó những người tương tác 10 lần trở lên sẽ được đánh giá là tương tác tốt nhất với khóa học là 5 điểm, còn lại ta sẽ tính theo tỉ lệ để cho ra được điểm số mong muốn.

Sau đó bộ dữ liệu được đưa lại dạng danh sách cạnh của ma trận có trọng số, và trọng số ở đây chính là điểm tương tác giữa người dùng và khóa học theo thang điểm 5. Ví dụ: {user_id[i], course_id[i], rating[i]} với rating[i] > 0 thì được coi là 1 cạnh của đồ thị với hai đỉnh là user_id và course_id.

Sau khi tiền xử lý ta sẽ có bộ dữ liệu sau:

	userID	courseId	rating
0	1660673	0	0
1	2021250	0	0
2	293249	0	0
3	3048969	0	0
4	81291	0	2

Hình 2.8: Tổng quan dữ liệu XuetangX sau khi tiền xử lý

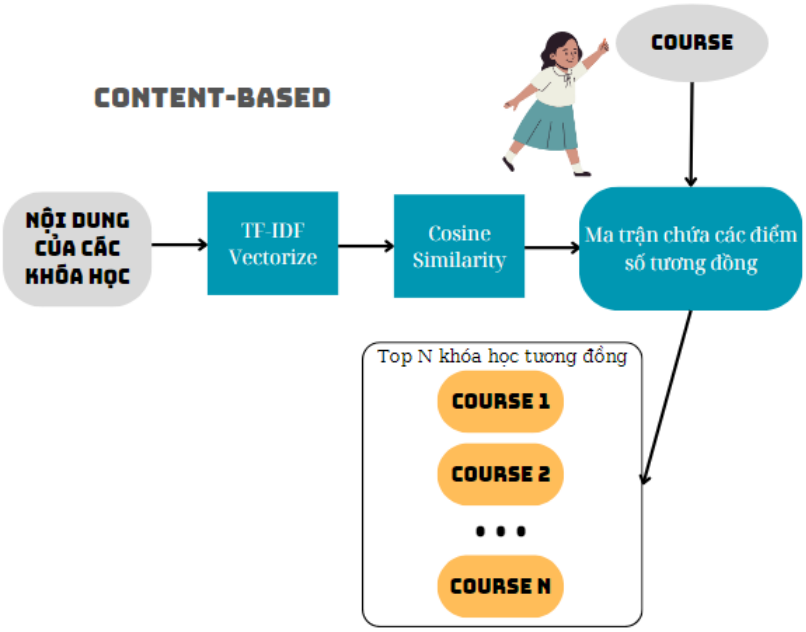
2.2 Phương pháp

2.2.1 Gợi ý dựa trên nội dung mô tả của khóa học

Như ta đã tìm hiểu ở phần cơ sở lý thuyết, gợi ý dựa trên nội dung được gọi là Content-based Filtering. Ta sẽ áp dụng chúng cho bài toán gợi ý khóa học ở đây. Ta có thể hình dung như sau, khi người dùng truy cập vào một khóa học nào đó mà người dùng muốn, ta sẽ gợi ý cho họ những khóa học có nội dung giống hoặc gần giống khóa học đó để cho người dùng có thêm những lựa chọn khác sao cho phù hợp với sở thích của họ.

Và sau đây ta sẽ đi áp dụng phương pháp Content-based Filtering vào việc gợi ý khóa học dựa trên nội dung khóa học người dùng lựa chọn. Dưới đây, ta thiết kế một mô hình gợi ý như hình 2.9:

Ta thấy rằng, ở đây, phần mô tả về khóa học thuộc dạng text, như vậy ta sẽ làm sao cho máy có thể hiểu được nội dung của từng mô tả trong khóa học, từ đó máy sẽ so sánh



Hình 2.9: Hệ thống gợi ý dựa trên nội dung

về sự tương đồng về mô tả khóa học. Điều đầu tiên ta nghĩ tới một kỹ thuật thường dùng trong NLP đó là ta sẽ đưa phần mô tả này thành một vector trong không gian vector N chiều, có nghĩa là ta sẽ chuyển từ dạng chuỗi các từ thành một vector chứa các số để máy có thể hiểu và so sánh được. Thuật toán nhúng được sử dụng đến ở đây là TF-IDF [34]. Lý do sử dụng TF-IDF là bởi vì thuật toán này ứng dụng mạnh mẽ cho việc tìm kiếm thông tin về chuỗi được nhúng và trong bài toán này, ta cũng sẽ đi tìm top 10 khóa học có phần mô tả gần giống nhất với khóa học mà người dùng lựa chọn.

Từ đây việc tìm kiếm 1 khóa học sau đó tìm những khóa học tương đồng của chúng đã trở nên dễ dàng. Trong toán học, để tính sự tương đồng của hai vector với nhau ta sẽ sử dụng cosine similarity [23]. Giả sử ta sẽ lựa chọn 1 khóa học, sau đó thuật toán sẽ tính toán cho chúng ta về sự tương đồng của các vector khác đối với nó.

Tuy nhiên ở đây, các khóa học được gợi ý chỉ dựa trên nội dung về khóa học mà người dùng lựa chọn tìm hiểu. Để có thể len lỏi sâu vào trong sở thích, hành vi của người dùng hơn thì Collaborative Filtering sẽ là lựa chọn hữu hiệu hơn. Nhưng để đề xuất những khóa học tương đồng thì Content-based đang làm tốt phần việc của mình.

2.2.2 Gợi ý dựa trên đánh giá của người dùng

Ở đây, ta sẽ chọn ra top 10 khóa học được tất cả người dùng đánh giá cao nhất. Bởi vì đa số chúng được đánh giá cao thì thường khóa học sẽ hữu ích cho phần đa người dùng trên hệ thống Coursera.

Như đã đề cập ở trên, ta sẽ sử dụng điểm trung bình người dùng vote trên từng khóa học (course_rating) và số lượng người dùng tham gia đánh giá điểm trên mỗi khóa học (course_reviews_num), sau đó sẽ tính ra được điểm số đánh giá cụ thể từng khóa học và đề xuất ra top 10 khóa học được đánh giá cao nhất.

Tuy nhiên trong nhiệm vụ này, ta có một vướng mắc đó là khi mà điểm đánh giá của khóa học khá cao nhưng số lượng người đánh giá lại ít hơn rất nhiều so với các khóa học khác. Vậy chúng ta phải làm như thế nào để công bằng hơn đối với điểm số của từng khóa học. Một thuật toán được áp dụng cho nhiệm vụ này được IMDb triển khai là Weighted Rating (IMDb Formula [35]) để tính được điểm số chính xác dựa trên số lượng đánh giá và điểm số đánh giá của toàn bộ tập dữ liệu của chúng ta để đưa ra được điểm số đánh giá chính xác nhất.

Ta có công thức sau đây:

$$WeightedRating(WR) = (\frac{v}{v+m}.R) + (\frac{m}{v+m}.C); \tag{2.1}$$

Trong đó:

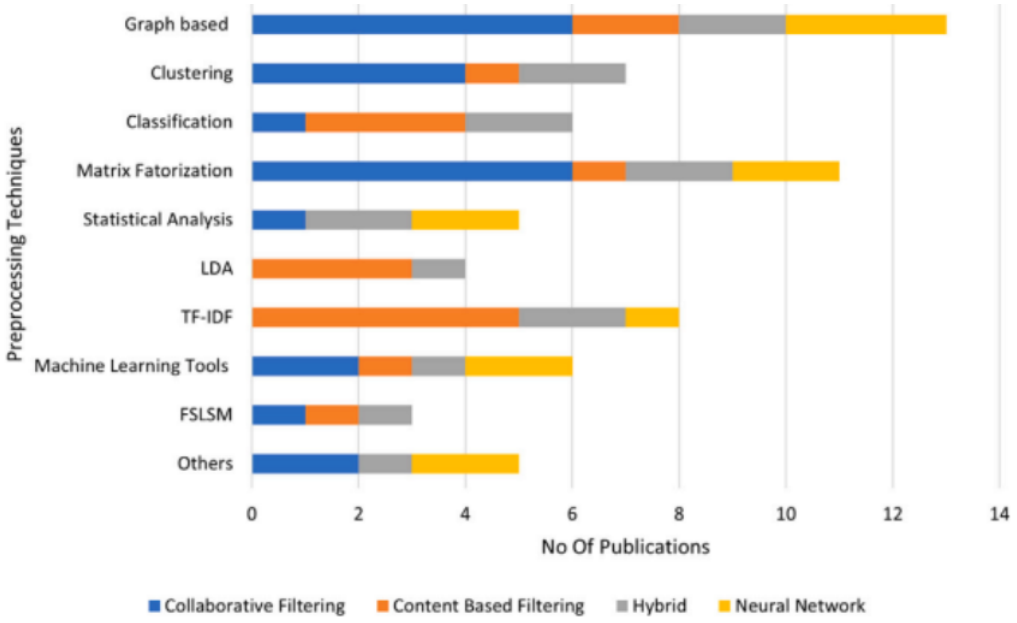
- v: số lượng bình chọn cho khóa học
- m: số lượng bình chọn tối thiểu yêu cầu được liệt kê trong bảng
- R: rating trung bình của khóa học
- C: rating trung bình trên toàn bộ tập dữ liệu

Nhờ vào việc tính toán điểm số này, các khóa học sẽ được đánh giá một cách khách quan hơn dựa trên điểm số trung bình mà chúng được đánh giá và số lượng lượt đánh giá trên khóa học đó.

2.2.3 LightGCN

Sau khi tìm hiểu về các phương pháp áp dụng cho Collaborative Filtering như Matrix Factorization, Clustering, Classification, Graph-based,... tính từ những năm đầu của hệ

thống gợi ý, càng về gần đây thì hệ thống gợi ý dựa trên mạng neural ngày càng được ứng dụng nhiều hơn do sự phát triển mạnh mẽ của Deep Learning. Theo bài báo [11] ta có biểu đồ sau:



Hình 2.10: Các nghiên cứu về hệ thống đề xuất và các kỹ thuật

Ta có thể thấy, như hình 2.10, đối với Collaborative Filtering, hầu hết ngày nay đều được triển khai dựa trên cấu trúc đồ thị.

Đồ thị là một cấu trúc có thể biểu diễn trực quan nhất cho những mối quan hệ, sự tương tác qua lại của các nút vì vậy việc áp dụng đồ thị vào hệ thống gợi ý đặc biệt là với Collaborative Filtering là điều hoàn toàn hợp lý. Do đó, việc học trên đồ thị đóng vai trò quan trọng trong hệ thống gợi ý vì nó cho phép hệ thống hiểu được mối quan hệ giữa các yếu tố khác nhau trong dữ liệu. Đồ thị là một cách biểu diễn dữ liệu dưới dạng các đối tượng (nút) kết nối với nhau thông qua các mối quan hệ (cạnh). Trong hệ thống gợi ý, đồ thị có thể được sử dụng để biểu diễn các thông tin về người dùng, sản phẩm, hoặc bất kỳ thực thể nào khác mà hệ thống quan tâm. Việc học trên đồ thị giúp hệ thống gợi ý hiểu được các mối quan hệ phức tạp giữa các yếu tố này, từ đó có thể đưa ra các gợi ý chính xác và cá nhân hóa hơn cho người dùng. Và trong bài KLTN này, ta sẽ nghiên cứu và áp dụng một mô hình dựa trên đồ thị và được coi là State of The Art của Hệ thống gợi ý vào năm 2020 đó là LightGCN [8]. Với hiệu suất được đưa ra trong bài báo:

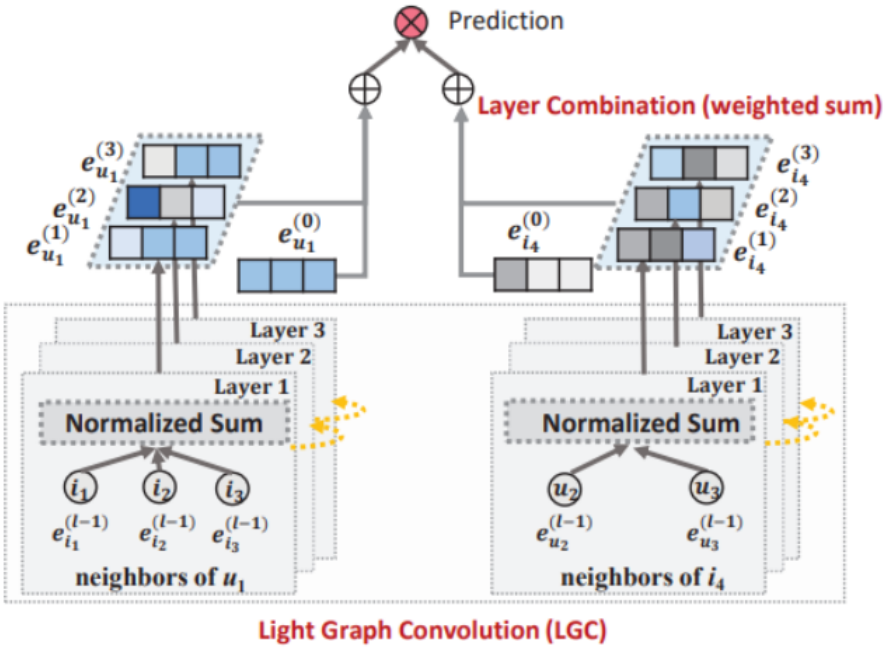
Tuy nhiên hiện tại đến năm nay, theo tìm hiểu trên Connected Paper [37], thấy rằng chỉ có một vài bài báo tối ưu cho mô hình LightGCN nhưng độ cải thiện không đáng kể, cho nên trong bài toán này, tôi vẫn sẽ nghiên cứu và áp dụng mô hình đơn giản

Dataset	Gowalla		Yelp2018		Amazon-Book	
Method	recall	ndcg	recall	ndcg	recall	ndcg
NGCF	0.1570	0.1327	0.0579	0.0477	0.0344	0.0263
Mult-VAE	0.1641	0.1335	0.0584	0.0450	0.0407	0.0315
GRMF	0.1477	0.1205	0.0571	0.0462	0.0354	0.0270
GRMF-norm	0.1557	0.1261	0.0561	0.0454	0.0352	0.0269
LightGCN	**0.1830**	**0.1554**	**0.0649**	**0.0530**	**0.0411**	**0.0315**

Hình 2.11: Kết quả thực nghiệm so sánh LightGCN và các phương pháp khác

LightGCN để giải quyết được yêu cầu của bài toán.

Tổng quan mô hình LightGCN như hình 2.12:



Hình 2.12: Mô hình LightGCN [8]

2.2.3.1 Nhúng - Embedding

Ban đầu, ta sẽ tạo một ma trận kề đối xứng có dạng:

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{R} \\ \mathbf{R}^{\mathbf{T}} & 0 \end{bmatrix} \tag{2.2}$$

Trong ma trận $\mathbf{R}$ thì mỗi $\mathbf{R_{ui}}$ là 1 nếu u và i tương tác với nhau và bằng 0 nếu không tương tác. Nhưng chỉ tạo ma trận A như vậy là chưa đủ, tín hiệu sẽ được tăng cường sau một số lớp, vì vậy chúng ta cần chuẩn hóa nó, có một số cách để thực hiện việc chuẩn

hóa, chúng tôi đang sử dụng phương pháp trong bài báo:

$$\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \tag{2.3}$$

Trong đó, $\tilde{A}$ là ma trận đối xứng được chuẩn hóa.

Sau 2 lần tạo, người dùng sẽ được kết nối với người dùng khác và item sẽ kết nối với một vài item khác. Và ma trận embedding tại lớp 0 được tính như sau:

$$\mathbf{E}^{(\mathbf{k+1})} = (\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}})\mathbf{E}^{(\mathbf{k})} \tag{2.4}$$

Trong đó $\mathbf{D}$ là một ma trận đường chéo kích thước $(M+N)(M+N)$ trong đó mỗi phần tử D_{ii} biểu thị số lượng phần tử khác không trong vector hàng thứ i của ma trận kề $\mathbf{A}$ (cũng được gọi là ma trận bậc).

2.2.3.2 Tổng hợp đặc trưng các nút trong mạng đồ thị - Aggregation

Tiền thân của mô hình LightGCN là mô hình GCN (Graph Convolution Network), trước đó, GCN được coi là SOTA cho Lọc cộng tác (Collaborative Filtering). Tuy nhiên tính hiệu quả của hệ thống gợi ý chưa được tốt. Trong thử nghiệm, thiết kế phổ biến trong GCN chứa các chuyển đổi đặc trưng và các hàm kích hoạt phi tuyến tính (non-linear activation) đóng góp rất ít vào hiệu suất của lọc cộng tác. Vì thế việc đưa chúng vào sẽ làm giảm hiệu suất của hệ thống đề xuất của chúng ta đi. GCN có rất nhiều biến thể, và trước khi LightGCN ra đời, có biến thể là NGCF được coi là hiệu suất tốt nhất lúc đó. Tuy nhiên điều mà NGCF cũng mắc phải đó là thiết kế của nó khá nặng và rắc rối điều này khá là phức tạp so với nhiệm vụ Gợi ý khóa học. Sau đó, bài báo [8] đã làm đơn giản hóa thiết kế của GCN, ngắn hơn và đơn giản hơn cho đề xuất. Trong đó LightGCN chỉ tổng hợp vùng lân cận, cụ thể, chúng tìm hiểu các phần user-item được embedding bằng lan truyền tuyến tính trên đồ thị biểu hiện sự tương tác giữa người dùng và khóa học (user-course). Nhờ đó hiệu suất của GCN được cải thiện hơn rất nhiều (Hình 2.11).

Ban đầu NGCF có dạng:

$$\begin{aligned}
\mathbf{e}_u^{(k+1)} &= \sigma\left(\mathbf{W}_1\mathbf{e}_u^{(k)} + \sum_{i\in\mathcal{N}_u}\frac{1}{\sqrt{|\mathcal{N}_u|\,|\mathcal{N}_i|}}\left(\mathbf{W}_1\mathbf{e}_i^{(k)} + \mathbf{W}_2\left(\mathbf{e}_i^{(k)}\odot\mathbf{e}_u^{(k)}\right)\right)\right), \\
\mathbf{e}_i^{(k+1)} &= \sigma\left(\mathbf{W}_1\mathbf{e}_i^{(k)} + \sum_{u\in\mathcal{N}_i}\frac{1}{\sqrt{|\mathcal{N}_u|\,|\mathcal{N}_i|}}\left(\mathbf{W}_1\mathbf{e}_u^{(k)} + \mathbf{W}_2\left(\mathbf{e}_u^{(k)}\odot\mathbf{e}_i^{(k)}\right)\right)\right),
\end{aligned} \tag{2.5}$$

NGCF chủ yếu tuân theo GCN tiêu chuẩn [12], bao gồm việc sử dụng hàm kích hoạt phi tuyến tính $\sigma(\cdot)$ và ma trận biến đổi đặc trưng $\mathbf{W1}$ và $\mathbf{W2}$. Tuy nhiên, hai hoạt

động này không quá hữu ích cho lọc cộng tác. Trong phân loại nút bán giám sát, mỗi nút có các đặc trưng ngữ nghĩa phong phú như đầu mục và từ ngữ trừu tượng của một bài báo. Do đó, việc thực hiện nhiều lớp biến đổi phi tuyến tính là có ích cho việc học đặc trưng. Tuy nhiên, trong lọc cộng tác, mỗi nút của đồ thị tương tác người dùng-mặt hàng chỉ có một ID làm đầu vào mà không có ý nghĩa cụ thể. Trong trường hợp này, việc thực hiện nhiều biến đổi phi tuyến tính sẽ không góp phần vào việc học các đặc trưng tốt hơn; thậm chí, nó có thể làm khó khăn hơn trong việc huấn luyện. Tác giả đã tiến hành một số thí nghiệm loại bỏ trên NGCF. Và chúng ta đã thấy:

- Mô hình hoạt động tốt hơn sau khi loại bỏ chỉ biến đổi đặc trưng.
- Mô hình hoạt động kém hơn sau khi loại bỏ chỉ hàm kích hoạt phi tuyến tính.
- Mô hình hoạt động tốt hơn rất nhiều sau khi loại bỏ cả hai!

Vì vậy, hãy loại bỏ hết chúng! Và đó chính là LightGCN.

Thay vào đó LightGCN có sự thay đổi hơn đó là bỏ hàm kích hoạt phi tuyến tính, bỏ các ma trận biến đổi đặc trưng, thay đổi từ cách thu được vector nhúng cuối cùng từ việc nối $\mathbf{e}_u^* = \mathbf{e}_u^{(0)} \| \cdots \| \mathbf{e}_u^{(L)}$ thành phép cộng tức là $\mathbf{e}_u^* = \mathbf{e}_u^{(0)} + \cdots + \mathbf{e}_u^{(L}$

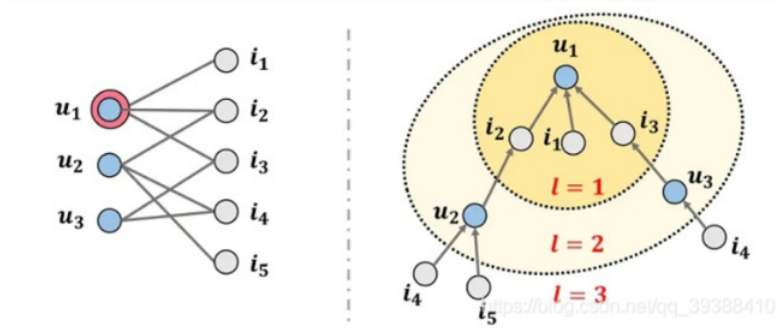
Trong LightGCN, áp dụng tính tổng trọng số đơn giản và bỏ qua việc sử dụng biến đổi đặc trưng và kích hoạt phi tuyến tính. Phép tích chập đồ thị (còn được gọi là quy tắc lan truyền [16]) trong LightGCN được định nghĩa như sau:

$$\begin{aligned} \mathbf{e}_u^{(k+1)} &= \sum_{i \in \mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} \mathbf{e}_i^{(k)}, \\ \mathbf{e}_i^{(k+1)} &= \sum_{u \in \mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} \mathbf{e}_u^{(k)}. \end{aligned} \tag{2.6}$$

Trong LightGCN, tham số mô hình chỉ được huấn luyện lại lớp thứ 0. Ví dụ $\mathbf{e_u^0}$ cho tất cả người dùng, $\mathbf{e_i^0}$ cho tất cả khóa học.

Và thành phần nhúng tại lớp cao hơn được tính thông qua LGC (light graph convolution) công thức trên. Sau đó K lớp ta sẽ kết hợp embedding thu được tại mỗi layer để tạo thành 1 đại diện cuối cùng cho người dùng hoặc khóa học:

$$\mathbf{e}_u = \sum_{k=0}^{K} \alpha_k \mathbf{e}_u^{(k)}; \quad \mathbf{e}_i = \sum_{k=0}^{K} \alpha_k \mathbf{e}_i^{(k)}, \tag{2.7}$$



Hình 2.13: Tổng hợp các nút theo từng layer

$\alpha_k > 0$ thể hiện tầm quan trọng trong lần embedding cuối cùng, nó được coi là 1 siêu tham số có thể điều chỉnh thủ công. Trong thử nghiệm [8], thống nhất $\alpha_k = \frac{1}{k+1}$ sẽ cho hiệu suất tốt hơn.

2.2.3.3 Dự đoán - Prediction

Ta có thể đã nhận thấy rằng LightGCN cuối cùng cho ra một vector embedding và sử dụng tích vô hướng của các embedding như là điểm số, vì vậy cơ bản, chúng ta có thể chỉ cần trích xuất kết quả nhúng và lưu nó dưới dạng một mô hình phân tích ma trận.

Không chỉ chúng ta có thể sử dụng MF[36] để dự đoán trực tiếp, THẬM CHÍ không gian tham số, và các mục điều chỉnh cũng hoàn toàn giống nhau.

$$E_{\text{out}} = \sum_{k=0}^{N} \alpha_k \hat{A}^k E = \left(\sum_{k=0}^{N} \alpha_k \hat{A}^k \right) E \tag{2.8}$$

Dự đoán của mô hình được định nghĩa là tích vô hướng của biểu diễn cuối cùng của người dùng và mặt hàng:

$$\hat{y}_{ui} = \mathbf{e}_u^T \mathbf{e}_i, \tag{2.9}$$

Điều này được sử dụng làm điểm xếp hạng cho việc tạo ra các đề xuất.

Chương 3

Thực nghiệm, kết quả và đánh giá mô hình

3.1 Quy trình thử nghiệm và kết quả

Quy trình thử nghiệm ở đây, ta sẽ tuân theo quy trình ở Hình 3.1:

Hình 3.1: Quy trình thử nghiệm

Đầu tiên ta sẽ đưa dữ liệu đã được tiền xử lý vào trong mô hình, sau đó ta sẽ bắt đầu quá trình huấn luyện để đưa ra kết quả thực nghiệm. Cuối cùng ta sẽ sử dụng các phương pháp đánh giá để đánh giá hiệu suất của mô hình.

3.1.1 Gợi ý dựa trên nội dung mô tả khóa học

Như vậy, đối với bộ dữ liệu Kaggle, ta sẽ gợi ý khóa học theo course_description của khóa học:

Sau khi sử dụng thuật toán TF-IDF ta sẽ có một ma trận gồm các vector như sau (Hình 3.3)

```
0      Pursue better IT security job opportunities an...
1      Develop the proficiency required to design and...
2      Affordable, abundant and reliable energy is fu...
3      We have all heard the phrase “correlation does...
5      About this Course\nThis four-week course title...
                           ...
995    Ce cours de français de niveau intermédiaire B...
996    За менш ніж шість місяців, без досвіду або спе...
997    استعد لمهنة جديدة في مجال تحليلات البيانات عال...
998    本系列課程從零開始，教授一般認為最適合初學者的程式語言「Python」，目標是讓大家在完成本...
999    接續用 Python 做商管程式設計（一）的內容，我們將在這個課程中繼續探討Python語言...
```

Hình 3.2: Nội dung mô tả của các khóa học trong bộ dữ liệu Kaggle

```
array([[0.06158368, 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       ...,
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.18957   , 0.        ,
        0.        ],
       [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
        0.        ]])
```

Hình 3.3: Ma trận sau khi sử dụng TF-IDF

Sau đó ta sẽ sử dụng Cosine Similarity như đã nêu ở phần phương pháp ta được kết quả sau (Hình 3.4)

```
array([4.45227640e-02, 3.15050307e-02, 1.34243757e-02, 1.14359231e-02,
       2.40472870e-02, 4.27229889e-02, 2.86497021e-02, 6.60311199e-02,
       4.45674169e-02, 0.00000000e+00, 4.21697174e-02, 1.80455136e-01,
       9.28238552e-02, 4.42760911e-02, 2.48287441e-02, 1.54579185e-02,
       6.85438984e-02, 1.23507651e-01, 5.56045774e-02, 4.50025594e-02,
       2.27040027e-02, 3.44726036e-02, 1.37387261e-02, 1.12782702e-02,
       1.96923329e-02, 3.13338463e-02, 2.16929730e-02, 1.73661631e-02,
       5.09442750e-02, 1.51003690e-01, 1.55425952e-02, 3.41729054e-02,
       1.00442596e-02, 4.89125055e-02, 8.83200050e-02, 4.76179015e-02,
       7.62261845e-02, 3.35386014e-02, 1.95264171e-02, 5.88179200e-02,
       4.67017722e-02, 4.49521592e-02, 4.17447042e-02, 3.63080238e-02,
       1.75538959e-02, 3.66848997e-02, 3.14260917e-01, 4.16387776e-02,
       6.09600804e-03, 1.55471491e-03, 6.75146998e-03, 3.02267189e-02,
       3.27633308e-02, 6.86627419e-03, 2.79582661e-02, 7.13882041e-02,
       9.48458360e-03, 1.68222575e-02, 8.77542204e-02, 9.04987948e-03,
       2.82793398e-02, 5.42301453e-03, 3.06841854e-02, 7.86942845e-02,
       4.20773040e-02, 5.82000799e-02, 5.21488383e-02, 9.50535467e-02,
       4.45027591e-02, 3.38945249e-02, 4.41829001e-02, 4.74510871e-02,
       5.94698972e-02, 2.57013347e-02, 2.03471920e-02, 4.16821493e-02,
       2.11324431e-03, 5.87497284e-04, 2.62183208e-03, 2.73553490e-02,
       1.78195183e-02, 2.64588654e-02, 3.58248134e-02, 4.88257224e-02,
       1.13789604e-02, 3.89631791e-02, 3.64965819e-02, 3.39374097e-02,
       3.69378745e-02, 5.07552706e-02, 1.11277633e-02, 3.55626630e-02,
       3.17664108e-02, 1.94591004e-02, 2.96259187e-02, 2.15516557e-02,
       6.61669187e-02, 1.14435938e-01, 1.17180025e-01, 3.39501479e-02,
...
       9.89590471e-03, 8.75561105e-03, 1.34326786e-02, 5.33395541e-02,
       2.20439352e-02, 9.46702818e-03, 2.51698835e-02, 4.70828367e-02,
       2.37641444e-02, 2.69517064e-02, 1.45636995e-02, 3.27157831e-02,
       1.93668142e-02, 1.26277719e-03, 6.08034941e-03, 4.55874458e-03,
       0.00000000e+00, 0.00000000e+00])
```

Hình 3.4: Kết quả sau khi sử dụng Cosine Similarity

Từ một danh sách các điểm số tương đồng, để đưa ra top 10 khóa học có nội dung gần giống với khóa học đã chọn, ta sẽ sắp xếp chúng từ cao đến thấp rồi chọn ra top 10. Và trong thực nghiệm, ta đã thử tìm xem top 10 khóa học có nội dung liên quan nhất đến khóa học "AI & Law" (Hình 3.5).

```
get_recommendations_course('AI & Law')
✓ 0.0s

8                                        AI For Business
479                                       IBM Applied AI
9                                        AI For Everyone
544                  Introduction to Back-End Development
966    Unsupervised Learning, Recommenders, Reinforce...
11                             AI Foundations for Everyone
75                  Artificial Intelligence: an Overview
659                                       Machine Learning
907    Supervised Machine Learning: Regression and Cl...
38                             Advanced Learning Algorithms
```

Hình 3.5: Kết quả khi gợi ý khóa học liên quan đến "AI & Law"

Ở đây, khi người dùng lựa chọn tìm hiểu về khóa học AI& Law, máy tính sẽ gợi ý thêm 10 khóa học khác có nội dung gần giống nhất với AI & Law, từ đây người dùng không cần phải tự mình tìm kiếm thêm những khóa học có cùng nội dung nữa.

3.1.2 Gợi ý dựa trên đánh giá người dùng

Sau khi tính toán từng điểm số chính xác và công bằng, ta có kết quả top các khóa học được người dùng đánh giá cao nhất (Hình 3.6).

	course_title	course_reviews_num	course_rating	score
742	Neuroscience and Neuroimaging	119900	4.9	4.843082
139	Calculus for Machine Learning and Data Science	94900	4.9	4.832765
822	Python Data Structures	93900	4.9	4.832274
363	Foundations of Public Health Practice	73000	4.9	4.820073
508	Improving Deep Neural Networks: Hyperparameter...	62500	4.9	4.812120
314	Excel for Beginners: Introduction to Spreadsheets	58900	4.9	4.809016
347	Fondamentaux de la cybersécurité	48500	4.9	4.798677
315	Excel for Beginners: Pivot Tables	47500	4.9	4.797557
830	Python for Everybody	268200	4.8	4.784027
800	Programming for Everybody (Getting Started wit...	224700	4.8	4.781432

Hình 3.6: Kết quả gợi ý dựa trên số lượng vote và điểm đánh giá trung bình

3.1.3 LightGCN

Dữ liệu ban đầu thu được từ tracking log trên nền tảng XuetangX như hình 3.7:

Tổng thể ở đây ta có 1577 khóa học khác nhau và mỗi khóa học đó sẽ chứa những id người dùng truy cập vào, trong mỗi id người dùng truy cập vào hệ thống ghi lại toàn bộ phiên truy cập và các thao tác trong khóa học đó.

0	course-v1:TsinghuaX+10610204_tv+2015_T1	{'1660673': {'22ef2c411f8cf49bfeeca748ce280679...
1	course-v1:TsinghuaX+AP000008X+sp	{'2143320': {'b3ef95b3f76ae6022e485505ebb9588c...
2	course-v1:TsinghuaX+10430484X+2016_T1	{'1095809': {'256afb6dcda237ee1a8deb0f08531983...
3	course-v1:RiceX+RELI157x+2017_T2	{'480009': {'d72dee409b2194d97018fc608295b651'...
4	UC_BerkeleyX/ColWri2_1x_2015_T1/2015_T1	{'6692611': {'9004616531a165c1069f6bbfbbad3972...
...	...	...
1572	course-v1:UQx+Think101x+sp	{'4497413': {'7129114f2d71f03c5e791d32d71a00e9...
1573	TsinghuaX/40050444X/2015_T2	{'186593': {'31abab07f63b048782ec8f88c361bd58'...
1574	course-v1:TsinghuaX+40040152X_tv+2015_T1	{'7051024': {'21ce3f3724423d647b4135bce8ba29bb...
1575	course-v1:NCTU+wym+2017_T2	{'328771': {'e78d6beec65c74dbf72434f841356e9c'...
1576	course-v1:TsinghuaX+00670122X+sp	{'7014912': {'50b920140bf2ad9a3a0790943b626d90...

1577 rows × 2 columns

Hình 3.7: Tổng quan về dữ liệu trong tracking log XuetangX

Sau tiền xử lý dữ liệu và đưa dữ liệu vào huấn luyện mô hình LightGCN ta sẽ có kết quả theo hình 3.8:

```
top 10 recommend courses for user  6280729 are:
 304              course-v1:NTHU+MOOC_00_005+sp
643    course-v1:TsinghuaX+80000901X_1+2017_T1
272      course-v1:TsinghuaX+00690863X+2017_T1
771      course-v1:TsinghuaX+00670122X+2017_T1
516           course-v1:TsinghuaX+00310222X+sp
712             course-v1:SEU+00690803+2017_T1
249          course-v1:TsinghuaX+00740123_X+sp
358           course-v1:JNUX+07009156X+2017_T1
474               course-v1:SEU+00034237_p2+sp
298      course-v1:TsinghuaX+10800163X+2017_T1
```

Hình 3.8: Kết quả gợi ý 10 khóa học dành cho người dùng có id 6280729

Trên đây, mô hình đã đưa ra được top 10 khóa học tốt nhất được gợi ý đến người dùng 6280729 (lấy ví dụ đối với người học có id là 6280729). Và 10 khóa học này hoàn toàn dựa trên việc sự tương đồng về sở thích của người dùng mà mô hình nhận thấy dựa trên sự tương tác của người dùng với những khóa học mà họ truy cập.

3.2 Các phương pháp đánh giá

Recall & Precision [40]

Với bài toán phân loại mà tập dữ liệu của các lớp là chênh lệch nhau rất nhiều, có một phép đó hiệu quả thường được sử dụng là Precision-Recall. Trước hết xét bài toán phân loại nhị phân. Ta cũng coi một trong hai lớp là positive, lớp còn lại là negative.

Với một cách xác định một lớp là positive, Precision được định nghĩa là tỉ lệ số

điểm true positive trong số những điểm được phân loại là positive (TP + FP).

Recall được định nghĩa là tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).

Một cách toán học, Precison và Recall là hai phân số có tử số bằng nhau nhưng mẫu số khác nhau:

$$\text{Precision } = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ Recall } = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.1}$$

Recall đánh giá khả năng của hệ thống trong việc không bỏ lỡ những mục quan trọng mà người dùng có thể quan tâm. Một giá trị Recall cao cho thấy hệ thống đề xuất một tỷ lệ lớn các mục liên quan, nhưng không cung cấp thông tin về số lượng mục không liên quan mà hệ thống có thể trả về. Thường thì, đánh giá Recall cùng với Precision giúp hiểu rõ hơn về hiệu suất tổng thể của hệ thống khuyến nghị.

Precision tập trung vào chất lượng của gợi ý bằng cách đánh giá xem bao nhiêu trong số những mục được đề xuất thực sự liên quan đến người dùng. Một giá trị Precision cao đồng nghĩa với việc hệ thống đề xuất một tỷ lệ lớn các mục mà người dùng thích và quan trọng.

nDCG [41]

nDCG (normalized Discounted Cumulative Gain) là một độ đo về hiệu suất của một hệ thống xếp hạng, xem xét đến vị trí của các mục liên quan trong danh sách xếp hạng. Nó dựa trên ý tưởng rằng các mục ở vị trí cao hơn trong danh sách xếp hạng nên được đánh giá cao hơn so với các mục ở vị trí thấp hơn. NDCG được tính bằng cách chia tổng giảm giá tích lũy (DCG) của danh sách xếp hạng cho DCG của danh sách xếp hạng lý tưởng, tức là danh sách với các mục liên quan được xếp hạng theo thứ tự tối ưu nhất.

Công thức nDCG:

$$\text{nDCG}_\text{p} = \frac{DCG_p}{IDCG_p} \tag{3.2}$$

nDCG có giá trị trong khoảng từ 0 đến 1, với các giá trị cao hơn cho thấy hiệu suất tốt hơn. nDCG@k càng tốt, chỉ ra rằng danh sách đề xuất có sự ưu tiên cao và các mục quan trọng được đặt ở những vị trí cao hơn trong danh sách.

MAP [42]

MAP (Mean Average Precision) là một chỉ số đánh giá thường được sử dụng trong

lĩnh vực Information Retrieval và Evaluation, đặc biệt là trong việc đánh giá hiệu suất của các hệ thống gợi ý và tìm kiếm.

Cụ thể, MAP đo lường chất lượng của một danh sách các mục được sắp xếp, như danh sách kết quả từ một hệ thống tìm kiếm hoặc gợi ý. Nó thường được sử dụng trong các bài toán mà có nhiều mục tiêu mong muốn cho mỗi truy vấn và việc xác định chính xác vị trí của các mục mong muốn là quan trọng.

Đối với hệ thống gợi ý, MAP (Mean Average Precision) được sử dụng để đánh giá chất lượng của các gợi ý được đưa ra. MAP tính toán độ chính xác của việc đưa ra danh sách các mục gợi ý dựa trên thứ bậc ưu tiên của chúng.

MAP đo lường khả năng của hệ thống đưa ra các mục có liên quan ở vị trí hàng đầu của danh sách gợi ý. Nó cung cấp một cái nhìn toàn diện về hiệu suất của hệ thống gợi ý bằng cách tính trung bình của độ chính xác của các danh sách gợi ý cho từng người dùng.

Việc tối ưu hóa MAP là mục tiêu để cải thiện độ chính xác và chất lượng của hệ thống gợi ý, đảm bảo rằng những mục có liên quan xuất hiện ở các vị trí đầu tiên trong danh sách gợi ý.

Công thức MAP:

$$\mathbf{MAP} = \frac{\sum_{q=1}^{Q} \frac{1}{\text{Số lượng mục liến quan đến rank k}} \sum_{k=1}^{K} \text{Precision tại k}}{Q} \tag{3.3}$$

Trong đó:

- Q: Số lượng truy vấn.
- K: Số lượng mục trong danh sách được xem xét (rank k)
- Precision tại K: Tính toán tỉ lệ số lượng mục mong muốn trong các mục đến rank k.

MAP thường nằm trong khoảng từ 0 đến 1, với giá trị càng cao, hệ thống gợi ý càng được đánh giá cao vì có nhiều mục mong muốn xuất hiện ở vị trí hàng đầu.

BPR Loss [43]

BPR Loss là sự mất mát xếp hạng được cá nhân hóa theo cặp bắt nguồn từ công cụ ước tính hậu nghiệm tối đa. Nó đã được sử dụng rộng rãi trong nhiều mô hình khuyến

nghị hiện có. Dữ liệu huấn luyện của BPR bao gồm cả cặp dương và âm (thiếu giá trị). Nó giả định rằng người dùng thích mục tích cực hơn tất cả các mục không được quan sát khác.

Trong công thức, dữ liệu huấn luyện được xây dựng bằng tuple dưới dạng (u,i,j), tức biểu diễn rằng người dùng u ưa thích sản phẩm i hơn sản phẩm j. Công thức Bayes trong BPR được cho dưới đây nhắm tới việc cực đại hóa xác suất hậu nghiệm:

$$p\left(\Theta \mid >_u\right) \propto p\left(>_u \mid \Theta\right) p(\Theta) \tag{3.4}$$

Các tham số có thể huấn luyện của LightGCN chỉ là các phần nhúng của lớp thứ 0, tức là $\Theta = \{E(0)\}$; nói cách khác, độ phức tạp của mô hình giống như hệ số ma trận tiêu chuẩn (MF). Tôi sử dụng BPR Loss theo bài báo [8], đây là tổn thất theo cặp nhằm khuyến khích dự đoán về một mục nhập được quan sát cao hơn so với các mục nhập không được quan sát của nó:

$$L_{BPR} = -\sum_{u=1}^{M} \sum_{i \in N_u} \sum_{j \notin N_u} \ln \sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \lambda \left\| \mathbf{E}^{(0)} \right\|^2 \tag{3.5}$$

3.3 Đánh giá

3.3.1 Gợi ý dựa trên nội dung mô tả khóa học

Phương pháp này dựa trên hoàn toàn về nội dung của khóa học, và được áp dụng cho toàn bộ người dùng. Do đó, khi người dùng tìm hiểu về một khóa học nào đó, đơn thuần hệ thống chỉ đưa ra những khóa học có độ tương đồng về nội dung mô tả của khóa học đó. Vì vậy, việc gợi ý dựa trên nội dung mô tả của khóa học sẽ đưa ra cho người học những khóa học có nội dung tương đồng mà học không cần phải mất công tìm kiếm. Nhưng nó cũng làm cho người học không tìm ra được những khóa học mới mẻ hơn.

Tổng kết lại Gợi ý dựa trên nội dung mô tả khóa học đã giúp cho ta giải quyết được vấn đề Khởi đầu lạnh (Cold start) trong phần Khó khăn và thách thức như đã nêu ở phần Giới thiệu khi hệ thống có ít dữ liệu về người dùng và người dùng là người dùng mới, chỉ tương tác ít.

3.3.2 Gợi ý dựa trên đánh giá người dùng

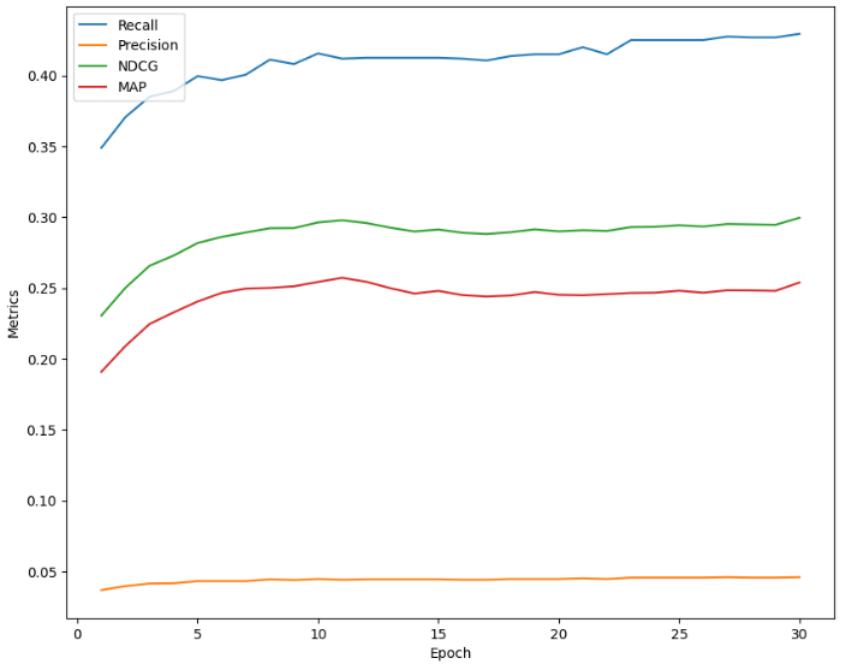
Việc lọc ra được 10 khóa học được người dùng đánh giá cao nhất sẽ giúp cho người dùng có thêm lựa chọn nữa là họ có thể biết được những khóa học nào đang thực sự cần thiết

và phù hợp với nhiều người dùng hơn. Tuy nhiên, việc lọc theo đánh giá cũng có nhược điểm và cũng chưa thực sự tối đa hóa được trải nghiệm của người dùng khi chưa đi sâu vào sự tương tác của người dùng đối với khóa học và của những người dùng khác.

Gợi ý dựa trên điểm số đánh giá của người dùng giúp cho người dùng có thể chọn được top các khóa học được đánh giá cao nhất hiện nay, và việc gợi ý này cũng giải quyết được vấn đề Khởi đầu lạnh khi người dùng hoàn toàn mới và chưa có một dữ liệu tương tác nào.

3.3.3 LightGCN

Ở đây, ta đã triển khai huấn luyện và đánh giá giống với bài báo gốc [11]. Kết quả Recall, Precision, nDCG, MAP trong quá trình huấn luyện: Trong đó điểm số Recall khá cao

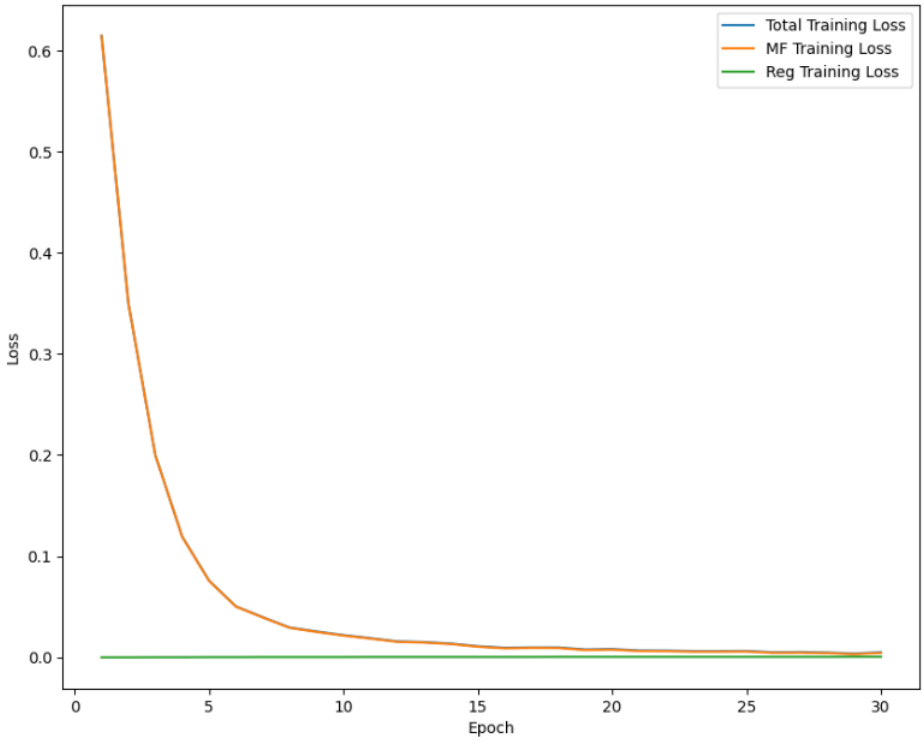


Hình 3.9: Recall, Precision, nDCG, MAP trong quá trình training

(xấp xỉ 0.45), điều này cho ta thấy được hệ thống gợi ý không bỏ lỡ những mục quan trọng mà người dùng có thể quan tâm.

Sau khi huấn luyện, ta thấy được loss function dường như hội tụ sau khoảng 10 epochs. Và BPR loss sau khi hội tụ xấp xỉ 0.03, điều này ta thấy được mô hình của chúng ta hoạt động khá ổn định (Hình 3.10).

Kết quả của mô hình sau khi thử nghiệm trên tập Test (Hình 3.11) ta thấy được kết quả khá là khả quan với BPR Loss rất nhỏ 0.005, và Recall khá cao với 0.4294.



Hình 3.10: Training Loss

```
Test Data Recall ->  0.4294
Test Data Precision ->  0.0459
Test Data NDCG ->  0.2996
Test Data MAP ->  0.2539
Train Data Loss ->  0.005
```

Hình 3.11: Kết quả các phép đo trên tập Test

Tổng kết lại, mô hình LightGCN là một mô hình dựa trên đồ thị thực sự tốt cho bài toán gợi ý khóa học theo phương pháp Collaborative Filtering, và nó vẫn giữ được hiệu suất cao, thời gian huấn luyện ít hơn, đối với bộ dữ liệu vừa thì tốc độ của nó khá tốt.

3.4 Thảo luận

Ở trong bài Khóa luận này, ta áp dụng những kĩ thuật khuyến nghị của các nghiên cứu đã có từ trước đến nay vào bài toán "Gợi ý khóa học", bước đầu tiên nhận thấy các mô hình đã có kết quả đầu ra khá khả quan. Đảm bảo cơ bản gợi ý được khóa học cho người dùng.

Đối với gợi ý dựa trên nội dung mô tả khóa học đã đưa ra được top các khóa học liên quan đến khóa học mà người dùng đang quan tâm tới bằng cách đo sự tương đồng về nội dung của các khóa học trong tập dữ liệu Kaggle. Gợi ý dựa trên sự đánh giá người

dùng đã đưa ra được một cái nhìn tổng quan về các khóa học được người dùng đánh giá cao dựa trên công thức mà IMDb đã đưa ra, từ đây, người dùng có một cái nhìn công bằng hơn so với trước đây là chỉ dựa trên điểm đánh giá trung bình. LightGCN được áp dụng trong Khóa luận này đã thể hiện được tính gọn nhẹ, hiệu suất cao, huấn luyện nhanh chóng của nó trên dữ liệu ma trận người dùng - khóa học rất thưa để đưa ra gợi ý các khóa học cho người dùng. LightGCN đã giữ được sức mạnh của nó như bài báo [11] đã nêu ra.

Kết luận

Khóa luận này đã thành công trong việc nghiên cứu và áp dụng các phương pháp khuyến nghị như là Gợi ý dựa trên nội dung mô tả khóa học, dựa trên đánh giá người dùng, LightGCN cho bài toán Gợi ý khóa học, đáp ứng chính xác và hiệu quả các yêu cầu đặt ra ban đầu là gợi ý khóa học cho người dùng. Nhờ đó, việc áp dụng các phương pháp vào bài toán "Gợi ý khóa học" sẽ giúp cho người dùng có thể lựa chọn nhiều hơn các khóa học liên quan đến nội dung mình đang tìm kiếm, ngoài ra còn gợi ý thêm những khóa học có thể liên quan đến bản thân.

Các phương pháp được sử dụng trong bài Khóa luận này đã giải quyết được một số vấn đề đã nêu trong mục Khó khăn và thách thức ở phần Giới thiệu:

Phương pháp	Chưa có dữ liệu tương tác người dùng	Dữ liệu người dùng rất ít	Đã có dữ liệu tương tác
Gợi ý dựa trên nội dung mô tả	0	v	0
Gợi ý dựa trên điểm đánh giá của người dùng	v	v	0
LightGCN	0	0	v

Bảng 3.1: Giải quyết vấn đề của các phương pháp

Như bảng trên, ta thấy được các Gợi ý dựa trên nội dung mô tả khóa học và dựa trên điểm đánh giá của người dùng đã giải quyết được vấn đề mà LightGCN gặp phải khi không có và có rất ít dữ liệu. Ngược lại, LightGCN lại giải quyết được vấn đề mà hai phương pháp còn lại gặp phải đó là có thể dự đoán được những khóa học mà người dùng có thể tương tác trong tương lai để gợi ý.

Vậy, hệ thống gợi ý xây dựng gồm 3 phương pháp trên về cơ bản đã Gợi ý thành công các khóa học cho người dùng. Và chúng đã giải quyết được những nhược điểm của nhau, từ đây, người dùng sẽ có một trải nghiệm tốt hơn khi học.

Tuy nhiên việc áp dụng các phương pháp cũng có một số hạn chế nhất định:

- Đối với dữ liệu XuetangX ghi lại hoạt động của người dùng truy cập vào khóa học

nhưng không biết thực sự họ có đang học hay không, điều này cũng có thể có hại cho kết quả của mô hình LightGCN.

- Đối với dữ liệu Kaggle thiếu đi dữ liệu về lịch sử người dùng, do đó với việc áp dụng LightGCN ta phải sử dụng dữ liệu XuetangX để làm ví dụ.

Hướng phát triển trong tương lai Từ những hạn chế đã nêu, con đường tiếp theo trong tương lai là:

- Thu thập dữ liệu nhiều hơn để làm rộng hơn tập dữ liệu.
- Tối ưu lại mô hình LightGCN hơn nữa để tăng được hiệu suất mô hình.
- Ứng dụng thêm phương pháp Hybrid tích hợp cả Content-based và Collaborative để giúp gợi ý hoàn thiện hơn.
- Phát triển thêm các bài toán khác trên nền tảng hệ thống học online như gợi ý các video bài giảng liên quan tới nội dung của người học thảo luận trên các diễn đàn của hệ thống học Online.

Tài liệu tham khảo

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, pp. 734–749, 07 2005.

[2] M. Agrebi, M. Sendi, and M. Abed, "Deep reinforcement learning for personalized recommendation of distance learning," 03 2019.

[3] O. Celma and X. Serra, "Foafing the music: Bridging the semantic gap in music recommendation," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 6, pp. 250–256, 11 2008.

[4] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," 08 2017, pp. 335–344.

[5] S. Fazeli, E. Rajabi, L. Lezcano, H. Drachsler, and P. Sloep, "Supporting users of open online courses with recommendations: An algorithmic study," 07 2016, pp. 423–427.

[6] W. Feng, J. Tang, T. X. Liu, S. Zhang, and J. Guan, "Understanding dropouts in moocs," in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, 2019.

[7] X. He, Z. He, J. Song, Z. Liu, Y.-G. Jiang, and T.-S. Chua, "Nais: Neural attentive item similarity model for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2354–2366, 2018.

[8] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," 2020.

[9] Y. Hou, P. Zhou, J. Xu, and D. O. Wu, "Course recommendation of mooc with big data support: A contextual online learning approach," in *IEEE INFOCOM 2018*

- IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2018, pp. 106–111.

[10] S. Kabbur, X. Ning, and G. Karypis, "Fism: factored item similarity models for top-n recommender systems," 08 2013, pp. 659–667.

[11] A. Khalid, K. Lundqvist, and A. Yates, "A literature review of implemented recommendation techniques used in massive open online courses," *Expert Systems with Applications*, vol. 187, p. 115926, 10 2021.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems. ieee, computer journal, 42(8), 30-37," *Computer*, vol. 42, pp. 30 – 37, 09 2009.

[14] G. Piao and J. G. Breslin, "Analyzing mooc entries of professionals on linkedin for user modeling and personalized mooc recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, ser. UMAP '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 291–292. [Online]. Available: https://doi.org/10.1145/2930238.2930264

[15] P. Symeonidis and D. Malakoudis, "Multi-modal matrix factorization with side information for recommending massive open online courses," *Expert Systems with Applications*, vol. 118, 10 2018.

[16] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '19. ACM, Jul. 2019. [Online]. Available: http://dx.doi.org/10.1145/3331184.3331267

[17] Hệ thống gợi ý: https://bit.ly/49zCxMY

[18] EDX: https://www.edx.org/

[19] Coursera: https://www.coursera.org/

[20] Udemy: https://www.udemy.com/

[21] Khan Academy: https://www.khanacademy.org/

[22] Content-based Filtering: https://bit.ly/4b9tgwK

[23] Cosine Similarity: https://bit.ly/4b9Iema

[24] Collaborative Filtering: https://bit.ly/3UBB9nk

[25] Tài liệu Python: https://docs.python.org/3/

[26] Pandas: https://pandas.pydata.org/docs/

[27] Numpy: https://numpy.org/doc/stable/

[28] Matplotlib: https://matplotlib.org/

[29] Seaborn: https://seaborn.pydata.org

[30] Pytorch: https://pytorch.org/docs/stable/index.html

[31] Tensorflow: https://www.tensorflow.org/

[32] Coursera Course Dataset: https://bit.ly/4b6RAzh

[33] XuetangX: https://www.xuetangx.com/

[34] TF-IDF: https://vi.wikipedia.org/wiki/Tf-idf

[35] Bayes Estimator: https://bit.ly/3wrOt5O

[36] Matrix Factorization: https://bit.ly/matrixfactorization

[37] Connected Papers: https://www.connectedpapers.com/

[38] MAE: https://bit.ly/3JWplHD

[39] MSE: https://bit.ly/3W8R7YN

[40] Recall và Precision: https://bit.ly/4dzVdzw

[41] nDCG: https://bit.ly/3UT0ScC

[42] MAP: https://bit.ly/3QEQRNv

[43] BPR Loss: https://bit.ly/4dvUP4U