

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN



ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

**ĐỀ TÀI: Xây dựng hệ khuyến nghị sách trong bán hàng trực tuyến
sử dụng kỹ thuật lọc dựa trên nội dung
và kỹ thuật lọc cộng tác**

Giảng viên hướng dẫn: Nguyễn Hoàng Anh
Sinh viên: Nguyễn Thị Lan Anh
Mã sinh viên: B16DCCN010
Lớp: D16CNPM1
Khóa: 2016-2021
Hệ đại học: Đại học chính quy

Hà Nội, Tháng 12/2020

LỜI CẢM ƠN

Lời đầu tiên em xin gửi lời cảm ơn chân thành tới tất cả thầy cô đang giảng dạy tại trường Học viện Công nghệ Bưu chính Viễn thông đã tận tình truyền đạt những kinh nghiệm và kiến thức quý báu giúp em hoàn thành nhiệm vụ học tập trong suốt thời gian là sinh viên của trường.

Em xin gửi lời biết ơn sâu sắc đến thầy giáo **ThS. Nguyễn Hoàng Anh**, người đã tận tình hướng dẫn, chỉ bảo, nhắc nhở em trong suốt quá trình học tập và hoàn thành đồ án này.

Cho con được gửi lời cảm ơn chân thành đến bố mẹ, ông bà, anh chị em đã luôn động viên, ủng hộ, cổ vũ và tạo mọi điều kiện tốt nhất cho con trong suốt những năm tháng ngồi trên ghế nhà trường.

Cho tôi gửi lời cảm ơn đến những người bạn của tôi, những người luôn chia sẻ, động viên, giúp đỡ và ở bên tôi mỗi khi tôi gặp khó khăn nhất.

Hà Nội, ngày ... tháng ... năm 20...

Sinh viên thực hiện

Nguyễn Thị Lan Anh

[illegible]

Hà Nội, ngày ... tháng ... năm 2021
Giảng viên phản biện

NHẬN XÉT, ĐÁNH GIÁ, CHO ĐIỂM
(Của giảng viên hướng dẫn)

This image shows a full page of white paper with horizontal dotted lines. The lines are evenly spaced and run across the width of the page, providing a guide for handwriting or typing. There are no margins, text, or other markings on the page.

Điểm.....(bằng chữ:)

Hà Nội, ngày ... tháng ... năm 2021
Giảng viên hướng dẫn

MỤC LỤC

LỜI CẢM ƠN.....	i
NHẬN XÉT, ĐÁNH GIÁ, CHO ĐIỂM	ii
NHẬN XÉT, ĐÁNH GIÁ, CHO ĐIỂM	iii
MỤC LỤC	iv
DANH MỤC HÌNH VẼ.....	vi
DANH MỤC BẢNG BIỂU.....	viii
BẢNG THUẬT NGỮ TIẾNG ANH	ix
LỜI NÓI ĐẦU	1
CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN	2
1.1. Giới thiệu hệ khuyến nghị – Recommender System	2
1.2. Lí do chọn đề tài	3
1.3. Giới thiệu một số hệ khuyến nghị sách.....	3
1.4. Dữ liệu và các nguồn tri thức của hệ khuyến nghị	8
1.5. Kết chương	8
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT	9
2.1. Kỹ thuật khuyến nghị dựa vào phương pháp lọc theo nội dung.....	9
2.1.1. Khái niệm	9
2.1.2. Phát biểu bài toán khuyến nghị lọc theo nội dung.....	9
2.1.3. Xây dựng hồ sơ sản phẩm – Item Profiles	9
2.1.4. Các phương pháp lọc theo nội dung	10
2.2. Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác.....	13
2.2.1. Khái niệm	14
2.2.2. Phát biểu bài toán lọc cộng tác	14
2.2.3. Phương pháp khuyến nghị lọc cộng tác dựa trên bộ nhớ.....	15
2.2.4. Phương pháp khuyến nghị lọc cộng tác dựa trên mô hình Matrix Factorization	19
2.3. So sánh đánh giá phương pháp lọc nội dung và lọc cộng tác.....	22
2.4. Kết chương	23
CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH.....	24
3.1. Thu thập dữ liệu.....	24
3.2. Xử lý dữ liệu.....	28
3.2.1. Tiền xử lý dữ liệu	28

3.2.2. Chia dữ liệu thực nghiệm.....	29
3.3. Thực nghiệm mô hình khuyến nghị.....	30
3.3.1. Thực nghiệm mô hình khuyến nghị sách theo kỹ thuật lọc dựa trên nội dung ..	30
3.3.2. Thực nghiệm mô hình khuyến nghị sách theo kỹ thuật lọc cộng tác	31
3.4. Đánh giá thuật toán.....	32
3.4.1. Các thông số đánh giá giải thuật.....	32
3.4.2. Kết quả thực nghiệm.....	33
3.4.3. Nhận xét kết quả và phân tích lỗi trong quá trình thực nghiệm	40
3.5. Kết chương	40
CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN.....	41
4.1. Mô tả hệ thống.....	41
4.2. Phân tích thiết kế hệ thống	41
4.3. Thiết kế hệ thống	43
4.3.1. Các công nghệ sử dụng	43
4.3.2. Mô hình tổng quát của hệ thống	43
4.3.3. Chương trình demo	44
4.4. Kết chương	48
KẾT LUẬN	49
DANH MỤC TÀI LIỆU THAM KHẢO	50

DANH MỤC HÌNH VẼ

Hình 1.1: Các công ty tích hợp hệ gợi ý trong sản phẩm.....	3
Hình 1.2: Minh họa về hệ gợi ý của Amazon.....	4
Hình 1.3: Minh họa về hệ gợi ý của Goodread	4
Hình 1.4: Minh họa sản phẩm của Tiki.....	5
Hình 1.5: Minh họa sản phẩm gợi ý của Tiki - 1.....	5
Hình 1.6: Minh họa sản phẩm gợi ý của Tiki - 2.....	5
Hình 1.7: Minh họa sản phẩm của Fahasa	6
Hình 1.8: Minh họa sản phẩm gợi ý của Fahasa - 1	6
Hình 1.9: Minh họa sản phẩm gợi ý của Fahasa - 2	6
Hình 1.10: Minh họa sản phẩm của Nhã Nam	7
Hình 1.11: Minh họa sản phẩm gợi ý của Nhã Nam	7
Hình 2.1: Minh họa phương pháp lọc nội dung dựa vào bộ nhớ.....	10
Hình 2.2: Minh họa khoảng cách cosine giữa hai vector.....	11
Hình 2.3: Minh họa phương pháp lọc cộng tác dựa trên người dùng.....	16
Hình 2.4: Minh họa chuẩn hóa ma trận.....	17
Hình 2.5: Minh họa chuẩn hóa utility matrix	19
Hình 2.6: Kỹ thuật phân rã ma trận	19
Hình 2.7: Bài toán tối ưu ma trận	20
Hình 3.1: Các bước xây dựng mô hình khuyến nghị	24
Hình 3.2: Biểu đồ phân phối đánh giá (dữ liệu thô)	25
Hình 3.3: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu thô).....	25
Hình 3.4: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu thô).....	26
Hình 3.5: Biểu đồ phân phối đánh giá (dữ liệu sau khi xử lý)	27
Hình 3.6: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu sau khi xử lý).....	27
Hình 3.7: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu sau khi xử lý).....	28
Hình 3.8: Dữ liệu đánh giá của người dùng	28
Hình 3.9: Tiền xử lý dữ liệu văn bản	29
Hình 3.10: Minh họa cách chia dữ liệu	29
Hình 3.11: Ma trận tương đồng giữa các sản phẩm.....	30
Hình 3.12: Biểu diễn precision, recall theo k với ngưỡng bằng 4	37
Hình 3.13: Biểu diễn precision, recall theo k với ngưỡng bằng 4.5	37
Hình 3.14: Biểu diễn precision, recall theo k với ngưỡng bằng 4	39

Hình 3.15: Biểu diễn precision, recall theo k với ngưỡng bằng 4.5	40
Hình 4.1: Sơ đồ usecase hệ thống khuyến nghị sách	41
Hình 4.2: Biểu đồ lớp toàn hệ thống	42
Hình 4.3: Lược đồ cơ sở dữ liệu	42
Hình 4.4: Mô hình tổng quát của hệ khuyến nghị sách.....	43
Hình 4.5: Kịch bản demo	44
Hình 4.6: Giao diện trang chủ - gợi ý theo lọc nội dung dựa trên mô hình	45
Hình 4.7: Giao diện trang chủ - gợi ý theo SVD	45
Hình 4.8: Giao diện trang chủ - gợi ý theo lọc cộng tác theo sản phẩm	46
Hình 4.9: Giao diện danh mục sách	46
Hình 4.10: Giao diện thông tin sản phẩm.....	47
Hình 4.11: Giao diện sản phẩm - gợi ý sản phẩm tương tự tên.....	47
Hình 4.12: Giao diện sản phẩm - gợi ý sản phẩm tương tự nội dung	48
Hình 4.13: Giao diện sản phẩm - sách thường được mua cùng.....	48

DANH MỤC BẢNG BIỂU

Bảng 2.1: Ví dụ Utility matrix.....	12
Bảng 2.2: Utility matrix với vector đặc trưng của sản phẩm	12
Bảng 2.3: Ma trận đánh giá của người dùng	14
Bảng 2.4: Ma trận đánh giá R.....	15
Bảng 2.5: So sánh ưu nhược điểm của lọc nội dung và lọc cộng tác	22
Bảng 3.1: Một vài mẫu dữ liệu sách	26
Bảng 3.2: Tham số mô hình k láng giềng gần nhất	31
Bảng 3.3: Tham số mô hình SVD.....	32
Bảng 3.4: Kết quả độ đo RMSE của mô hình lọc nội dung	33
Bảng 3.5: So sánh đánh giá cho một số người dùng.....	33
Bảng 3.6: Kết quả lọc cộng tác dựa trên sản phẩm trên tập huấn luyện	34
Bảng 3.7: Kết quả lọc cộng tác dựa trên người dùng trên tập huấn luyện	34
Bảng 3.8: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc theo bộ nhớ	35
Bảng 3.9: Kết quả đo precision và recall theo k bằng phương pháp lọc theo bộ nhớ	35
Bảng 3.10: Kết quả RMSE và MAE của mô hình phân rã ma trận.....	37
Bảng 3.11: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc phân rã ma trận	37
Bảng 3.12: Kết quả đo precision và recall theo k bằng phương pháp lọc phân rã ma trận.....	38
Bảng 3.13: Tổng kết kết quả huấn luyện.....	40
Bảng 4.1: Mô tả các lớp trong hệ thống.....	42

BẢNG THUẬT NGỮ TIẾNG ANH

Tên viết tắt	Tên tiếng anh	Tên tiếng Việt
API	Application Programming Interface	Giao tiếp lập trình ứng dụng
CB	Content-based Filtering	Kỹ thuật lọc theo nội dung
CF	Collaborative filtering	Kỹ thuật lọc cộng tác
FN	False Negative	Đánh giá không thích bị phân loại sai
FP	False Positive	Đánh giá thích bị phân loại sai
KNN	K nearest neighbor	Kỹ thuật k láng giềng gần nhất
MAE	Mean absolute error	Sai số tuyệt đối trung bình
MSE	Mean squared error	Sai số bình phương trung bình
RMSE	Root mean square error	Căn bậc hai của sai số bình phương trung bình
RS	Recommender System	Hệ thống khuyến nghị
SVD	Singular value decomposition	Kỹ thuật phân rã ma trận
TP	True Positive	Đánh giá thích được phân loại đúng
TF-IDF	Term frequency – Inverse document frequency	Phép đo tần suất kết hợp với tần suất xuất hiện ngược
TN	True Negative	Đánh giá không thích được phân loại đúng

LỜI NÓI ĐẦU

Với sự gia tăng chưa từng thấy lượng thông tin trên Internet hiện nay làm cho vấn đề quá tải thông tin trở nên trầm trọng đối với người dùng các dịch vụ trực tuyến. Có hàng trăm đến hàng triệu kết quả trả về cho một từ khóa khi ta tìm kiếm. Điều này khiến người dùng gặp rất nhiều khó khăn. Chính vì vậy, hệ thống khuyến nghị ra đời giúp chúng ta dễ dàng tìm thấy thông tin thích hợp nhất trên mạng internet.

Hệ khuyến nghị ra đời hướng đến việc giảm tải thông tin cho mỗi người dùng bằng cách đưa ra những gợi ý thông tin phù hợp và gỡ bỏ những thông tin không phù hợp cho mỗi người dùng. Đối với người dùng, hệ khuyến nghị tự động trợ giúp lựa chọn thông tin phù hợp trong vô số thông tin không phù hợp. Đối với các nhà cung cấp thông tin, hệ khuyến nghị không chỉ trợ giúp việc xác định những loại thông tin nào cần cung cấp cho mỗi người dùng đơn lẻ mà còn nó còn là nhân tố nâng cao hiệu quả và chất lượng dịch vụ cung cấp thông tin.

Cùng với sự phổ biến của mạng internet và sự phát triển vượt bậc của khoa học kỹ thuật, trải nghiệm người dùng web được cải thiện rất nhiều. Hệ khuyến nghị được áp dụng rất nhiều trong các lĩnh vực: giải trí: nghe nhạc, xem phim, đọc báo... đến thương mại điện tử, mua sắm online hay mạng xã hội... Tùy vào đặc điểm của mỗi lĩnh vực thì hệ khuyến nghị sẽ sử dụng các phương pháp tư vấn khác nhau. Nhưng mục tiêu đều là đưa ra những gợi ý mà người dùng sẽ click chọn ngay khi nhìn thấy.

Đọc sách là một thói quen tốt cần được khuyến khích và phát triển. Tuy nhiên, hiện nay trên thị trường có hàng trăm nghìn quyển sách với các thể loại, mẫu mã khác nhau. Mục đích là ta chọn đọc được những cuốn sách ý nghĩa, phù hợp với nhu cầu và sở thích của mỗi cá nhân. Chính vì vậy, tôi lựa chọn đề tài **“Xây dựng hệ khuyến nghị sách trong bán hàng trực tuyến sử dụng kỹ thuật lọc dựa trên nội dung và kỹ thuật lọc cộng tác”** để thực hiện trong khuôn khổ đồ án tốt nghiệp ngành Công nghệ thông tin. Đồ án được cấu trúc thành 4 chương, trong đó Chương 1 giới thiệu chung về hệ thống khuyến nghị và hệ khuyến nghị sách trong bán hàng trực tuyến. Chương 2 trình bày cơ sở lý thuyết các kỹ thuật khuyến nghị. Chương 3 mô tả thực nghiệm quá trình xây dựng một mô hình khuyến nghị sử dụng các kỹ thuật ở chương 2 và đánh giá với dữ liệu sách. Phần cuối cùng là ứng dụng mô hình khuyến nghị sách vào hệ thống bán hàng trực tuyến.

CHƯƠNG 1: GIỚI THIỆU CHUNG VỀ HỆ KHUYẾN NGHỊ VÀ HỆ KHUYẾN NGHỊ SÁCH TRONG BÁN HÀNG TRỰC TUYẾN

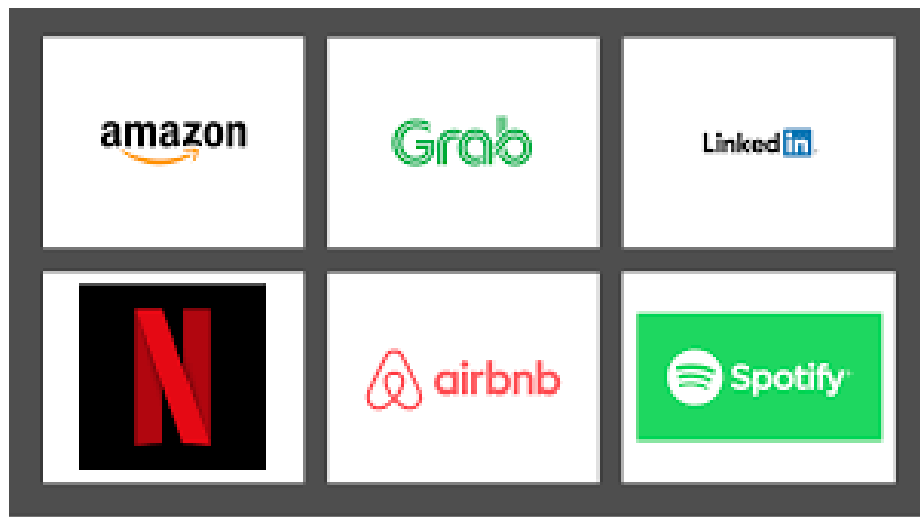
1.1. Giới thiệu hệ khuyến nghị – Recommender System

Wikipedia định nghĩa, hệ thống khuyến nghị, hay còn gọi là hệ thống tư vấn là một hệ thống lọc thông tin nhằm dự đoán, đánh giá sở thích, mối quan tâm, nhu cầu của người dùng để đưa ra một hoặc nhiều mục, sản phẩm, dịch vụ mà người dùng có thể sẽ quan tâm với xác suất lớn nhất. Những gợi ý được cung cấp nhằm hỗ trợ cho người sử dụng đưa ra quyết định lựa chọn những sản phẩm, dịch vụ phù hợp với nhu cầu và thị hiếu của mình, chẳng hạn như: mua sản phẩm nào, nghe thể loại nhạc nào, hay tin tức nào nên đọc... [1]

Chức năng của hệ khuyến nghị:

- Đối với nhà cung cấp:
 - + Tăng số lượng mặt hàng bán ra cho các hệ thống thương mại điện tử: thay vì người dùng chỉ mua 1 sản phẩm mà họ cần, họ được gợi ý mua những sản phẩm ‘có thể họ cũng quan tâm’ mà bản thân họ không nhận ra. Bằng cách tìm ra những mối quan tâm ẩn của người dùng, hệ thống gợi ý làm gia tăng số lượng mặt hàng bán ra. Tương tự đối với các hệ thống phi thương mại, hệ thống gợi ý sẽ giúp người dùng tiếp cận với nhiều đối tượng thông tin hơn.
 - + Tăng sự hài lòng của người dùng: vai trò chủ đạo của hệ khuyến nghị là hiểu nhu cầu của người dùng, gợi ý cho họ những thứ họ cần. Họ sẽ tìm thấy các gợi ý thú vị, có hiệu quả, chính xác, gợi ý kịp thời và một giao diện đẹp có thể tối ưu việc sử dụng và tăng sự hài lòng của người dùng trong hệ thống.
 - + Tăng độ tin cậy, độ trung thực của người dùng: khi người dùng càng tin cậy vào hệ thống, đưa ra những đánh giá trung thực cho các sản phẩm, hệ thống càng mang lại cho người dùng những gợi ý chính xác hơn
- Đối với người sử dụng:
 - + Tìm ra một số sản phẩm tốt nhất: Hệ thống gợi ý tới người dùng một số sản phẩm được xếp hạng, từ đó người dùng có thể tìm được sản phẩm tốt nhất cho bản thân.
 - + Gợi ý liên tục: Thay vì tập trung vào gợi ý đơn, các hệ thống khuyến nghị tạo các gợi ý liên tục tới người dùng đến khi họ tìm được sản phẩm ưng ý

Một vài ứng dụng nổi tiếng về hệ thống gợi ý của các công ty có tên trong hình 1.1 có thể kể đến như: gợi ý sản phẩm Amazon, hệ gợi ý phim Netflix, gợi ý nhạc Spotify, Facebook gợi ý bạn bè và quảng cáo, các trang thương mại điện tử phổ biến ở Việt Nam như Tiki, Lazada, Shopee... Hệ khuyến nghị đã chứng minh được ý nghĩa to lớn trong việc giúp người sử dụng giải quyết với tình trạng quá tải thông tin, trở thành một trong những công cụ mạnh mẽ và phổ biến trong thương mại điện tử và nhiều lĩnh vực khác.



Hình 1.1: Các công ty tích hợp hệ gợi ý trong sản phẩm

1.2. Lí do chọn đề tài

Đọc sách là một hành động mang lại lợi ích cho cá nhân và xã hội. Văn hóa đọc sách ở Việt Nam trong bối cảnh hội nhập và phát triển đã mở ra rất nhiều cơ hội mới và cả những khó khăn và thách thức.

Số lượng sách xuất bản mỗi năm lên tới 35000 cuốn và hơn 400 triệu bản sách, với thể loại, mẫu mã đa dạng. Hình thức mua sách qua các kênh online tăng đáng kể so với hình thức mua hàng trực tiếp. Một số đơn vị khác như Fahasa, Anphabook, Nhã Nam, Thái Hà Book đều ghi nhận sự tăng trưởng này với mức từ 20-30%. [2]

Bên cạnh những tích cực về mặt số lượng sách bán ra, chúng ta cũng cần quan tâm đến việc mở rộng số lượng người đọc sách, bằng cách giới thiệu và định hướng cho người dùng những sản phẩm phù hợp với sở thích, lứa tuổi đối với người dùng. Và hệ khuyến nghị sách có thể giúp ngăn chặn sự suy giảm đọc sách và giúp hướng người dùng đến với những cuốn sách phù hợp, từ đó tăng số lượng sản phẩm được bán ra.

Mục tiêu của luận văn là nghiên cứu, áp dụng một số các phương pháp trong kỹ thuật lọc nội dung và lọc cộng tác nhằm nâng cao kết quả dự đoán nhu cầu người dùng của hệ thống gợi ý. Luận văn trình bày các bước trong quá trình đưa ra dự đoán gợi ý sản phẩm đến người dùng, từ đó đánh giá và tối ưu mô hình để cải thiện chất lượng của mô hình.

1.3. Giới thiệu một số hệ khuyến nghị sách

Dưới đây là một số nền tảng ứng dụng bán sách đã tích hợp hệ khuyến nghị:

- Quốc tế:
 - + Amazon.com: Amazon được coi là 1 trong những hệ khuyến nghị tiên phong được bắt đầu từ năm 1995 với lĩnh vực kinh doanh ban đầu là sách. Amazon sử dụng các gợi ý như một công cụ tiếp thị qua email và hầu hết trên các trang web của họ. Các mục đề xuất bao gồm: gợi ý cá nhân, các sản phẩm thường được mua kèm, lịch sử xem, các sản phẩm liên quan, gợi ý theo sản phẩm đã mua, sản phẩm bán chạy... Một thống kê chỉ ra rằng 35% doanh thu của amazon được tạo ra dựa trên hệ khuyến nghị. Hình 1.2 mô tả 2 trong các mục đề xuất của Amazon.

Frequently bought together



Total price: **\$21.11**

[Add all three to Cart](#)

[Add all three to List](#)

- ✓ **This item:** The Deep End (Diary of a Wimpy Kid Book 15) by Jeff Kinney Hardcover **\$7.98**
- ✓ Dog Man: Grime and Punishment: From the Creator of Captain Underpants (Dog Man #9) (9) by Dav Pilkey Hardcover **\$6.99**
- ✓ Rowley Jefferson's Awesome Friendly Adventure by Jeff Kinney Hardcover **\$6.14**

Customers who viewed this item also viewed

Page 1 of 9






Dog Man: Grime and Punishment: From the Creator of Captain...
 > Dav Pilkey

Rowley Jefferson's Awesome Friendly Adventure
 > Jeff Kinney

Wrecking Ball (Diary of a Wimpy Kid Book 14)
 > Jeff Kinney
 ★★★★★ 10,984


The Meltdown (Diary of a Wimpy Kid Book 13)
 > Jeff Kinney
 ★★★★★ 8,274

The Getaway (Diary of a Wimpy Kid Book 12)
 > Jeff Kinney
 ★★★★★ 7,333

Hình 1.2: Minh họa về hệ gợi ý của Amazon

- + Goodreads.com: Goodread là một mạng xã hội cho phép người dùng chia sẻ thông tin về những cuốn sách mà họ đang đọc, và nhận đề xuất từ những người dùng khác. Mỗi người dùng có một bộ sưu tập những cuốn sách mà họ đã và đang đọc, đánh giá và gợi ý cho những người khác. Hình 1.3 minh họa các gợi ý cho 1 quyển sách tại Goodreads.com.

Books similar to Real Life



Real Life
by Brandon Taylor


★★★★★ 3.96 avg. rating • 7,056 Ratings

A FINALIST FOR THE 2020 BOOKER PRIZE


A NEW YORK TIMES EDITORS' CHOICE

A novel of startling intimacy, violence, and mercy among friends in a Midwestern university town, from an electric new voice.

Almost... [More](#)

[Want to Read](#)  Rate it: ★★★★★


Goodreads members who liked this book also liked:




Shuggie Bain
by Douglas Stuart

★★★★★ 4.35 avg. rating • 2,879 Ratings

Shuggie Bain is the unforgettable story of young Hugh "Shuggie" Bain, a sweet and lonely boy who spends his 1980s childhood in run-down public housing in Glasgow, Scotland. Thatcher's policies have pu... [More](#)

[Want to Read](#)  Rate it: ★★★★★



The New Wilderness
by Diane Cook

★★★★★ 3.82 avg. rating • 2,054 Ratings

A debut novel that explores a mother-daughter relationship in a world ravaged by climate change and overpopulation, a suspenseful second book from the author of the story collection, *Man V.*

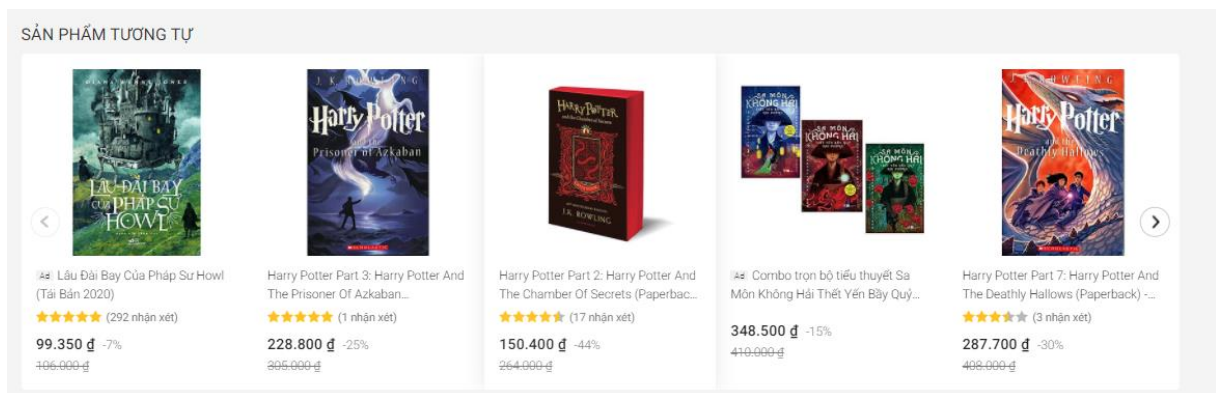
Hình 1.3: Minh họa về hệ gợi ý của Goodread

- Tại Việt Nam:

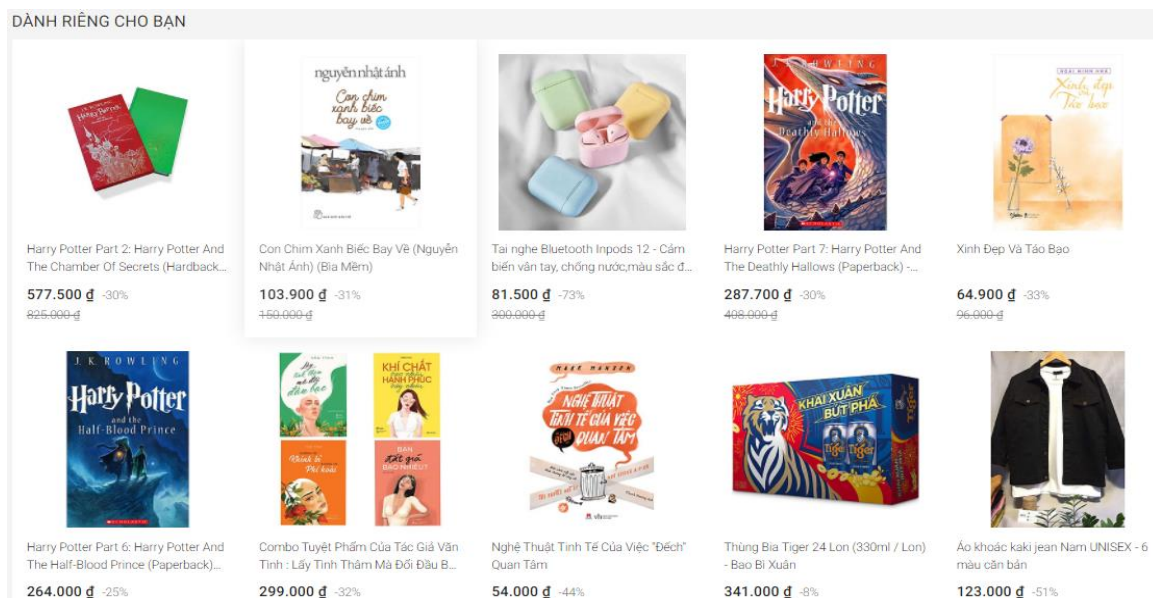
+ Tiki.vn: Gợi ý hiển thị tại trang thông tin sách (Hình 1.4, 1.5) và tại trang chủ đối với người dùng đã có lịch sử duyệt web (Hình 1.6)



Hình 1.4: Minh họa sản phẩm của Tiki



Hình 1.5: Minh họa sản phẩm gợi ý của Tiki - 1



Hình 1.6: Minh họa sản phẩm gợi ý của Tiki - 2

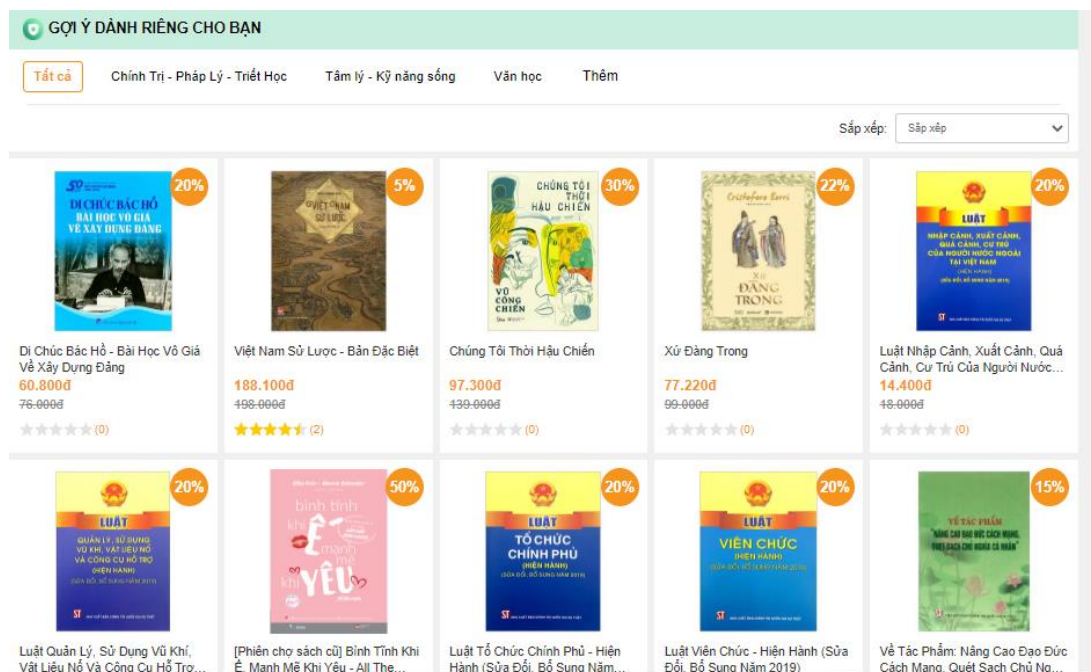
+ Fahasa.com: gợi ý hiển thị ở trang thông tin sách (Hình 1.7, 1.8) và tại trang chủ khi người dùng đăng nhập vào hệ thống (Hình 1.9)



Hình 1.7: Minh họa sản phẩm của Fahasa



Hình 1.8: Minh họa sản phẩm gợi ý của Fahasa - 1

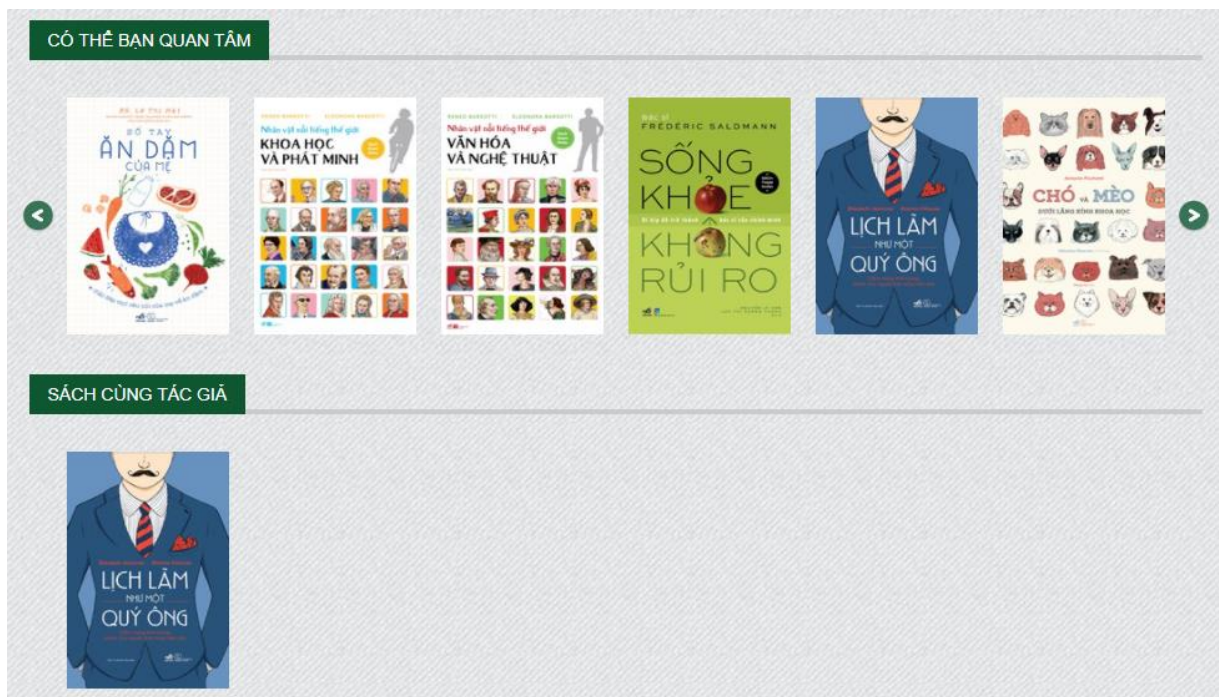


Hình 1.9: Minh họa sản phẩm gợi ý của Fahasa - 2

+ Nhanam.com.vn: hiển thị gợi ý tại trang thông tin sách (Hình 1.10, 1.11)



Hình 1.10: Minh họa sản phẩm của Nhã Nam



Hình 1.11: Minh họa sản phẩm gợi ý của Nhã Nam

➔ Các website đều có gợi ý tại trang chi tiết sản phẩm, một số trang web có sử dụng lịch sử xem của người dùng và dựa trên cả nội dung của sản phẩm về thể loại, nội dung từ đó gợi ý cho người dùng những quyển sách tương tự.

Để hiểu các hệ thống khuyến nghị hoạt động ra sao, tôi sẽ đi sâu vào trình bày bài toán trong các mục và chương tiếp theo.

1.4. Dữ liệu và các nguồn tri thức của hệ khuyến nghị

Dữ liệu được sử dụng bởi hệ khuyến nghị thuộc ba loại: sản phẩm (Item), người dùng (User), và các giao dịch (Transaction) [3]

- Sản phẩm là các đối tượng được gợi ý. Đó có thể là sản phẩm tiêu dùng, đồ ăn, sách báo hoặc các nội dung giải trí như tin tức, âm nhạc, phim ảnh...
- Người dùng của một hệ khuyến nghị có thể có đặc điểm và mục tiêu rất đa dạng, một hệ khuyến nghị tốt là hệ cá nhân hóa các gợi ý của người dùng.
- Giao dịch là sự tương tác hay phản hồi giữa người dùng và sản phẩm. Dữ liệu phản hồi đó được chia làm hai loại là phản hồi tường minh (explicit feedback) bằng cách yêu cầu người dùng phản hồi trực tiếp (đánh giá yêu thích, xếp hạng số sao, hành động thêm vào giỏ hàng,...) và phản hồi tiềm ẩn (implicit feedback) bằng cách tự động suy luận dựa trên những tương tác của người dùng với hệ thống (số lần nhấp chuột vào sản phẩm, thời gian quan sát...).

Sau khi thu thập và xử lý các nguồn tri thức trên, hệ khuyến nghị sẽ phân tích và từ đó đưa ra các quyết định, gợi ý đến người dùng.

1.5. Kết chương

Chương 1 đã giới thiệu về hệ khuyến nghị và đưa ra khái niệm, vai trò, tầm quan trọng của hệ khuyến nghị trong đời sống. Bên cạnh đó, chương 1 cũng đưa ra các ví dụ thực tế về hệ khuyến nghị. Có thể thấy, hệ khuyến nghị đang ngày được áp dụng rộng rãi trong đời sống hiện nay. Để hiểu sâu hơn cách hoạt động của hệ khuyến nghị, chương tiếp theo sẽ trình bày cơ sở lý thuyết về các kỹ thuật khuyến nghị được sử dụng trong luận văn là lọc nội dung và lọc cộng tác cùng với các mô hình liên quan.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

Dựa vào phương pháp lọc thông tin, các hệ khuyến nghị được chia làm 3 loại: khuyến nghị dựa vào phương pháp lọc theo nội dung (Content-based Filtering Recommendation), khuyến nghị dựa vào phương pháp lọc cộng tác (Collaborative Filtering Recommendation) và khuyến nghị dựa vào phương pháp lọc kết hợp (Hybrid Filtering Recommendation). Trong phạm vi đồ án, tôi sẽ trình bày chi tiết về 2 dạng khuyến nghị chính là lọc theo nội dung và lọc cộng tác, các kỹ thuật được sử dụng ở mỗi dạng, cũng như đưa ra những thế mạnh và nhược điểm của mỗi phương pháp.

2.1. Kỹ thuật khuyến nghị dựa vào phương pháp lọc theo nội dung

2.1.1. Khái niệm

Lọc theo nội dung (Content-based filtering) là phương pháp dựa trên những đặc điểm chính của sản phẩm. Gợi ý dựa trên nội dung khai thác những khía cạnh có liên quan đến nội dung thông tin sản phẩm người dùng đã từng sử dụng hay truy cập trong quá khứ để tìm ra những sản phẩm tương tự. Các phương pháp tiếp cận cho lọc nội dung có nguồn gốc từ lĩnh vực truy vấn thông tin, trong đó mỗi sản phẩm được biểu diễn bằng một hồ sơ sản phẩm, mỗi người dùng được biểu diễn bằng một hồ sơ người dùng.

2.1.2. Phát biểu bài toán khuyến nghị lọc theo nội dung

Cho $I = \{i_1, i_2, \dots, i_n\}$ là tập gồm n sản phẩm. Nội dung sản phẩm $i \in I$ được ký hiệu là $ItemProfile(i)$ được biểu diễn thông qua tập K đặc trưng nội dung của I . Tập các đặc trưng sản phẩm I được xây dựng bằng các kỹ thuật truy vấn thông tin để thực hiện mục đích dự đoán những sản phẩm khác tương tự với i

Cho $U = \{u_1, u_2, \dots, u_m\}$ là tập gồm m người dùng. Với mỗi người dùng $u \in U$, gọi $Transactions(u)$ là hồ sơ người dùng u . Hồ sơ của người dùng u thực chất là lịch sử truy cập hoặc đánh giá của người đó đối với các sản phẩm mà người dùng u đã từng truy nhập hoặc đánh giá.

Bài toán lọc theo nội dung khi đó là dự đoán những sản phẩm có nội dung thích hợp với người dùng dựa trên tập hồ sơ sản phẩm $ItemProfiles(i)$ và hồ sơ người dùng $Transactions(u)$

2.1.3. Xây dựng hồ sơ sản phẩm – Item Profiles

Dựa trên nội dung của mỗi sản phẩm, chúng ta cần xây dựng một bộ hồ sơ cho mỗi sản phẩm hay còn gọi là trích chọn đặc trưng sản phẩm. Hồ sơ này được biểu diễn dưới dạng toán học là một vector đặc trưng - feature vector để máy tính có thể tự động phân tích, tính toán trong số các đặc trưng. Đặc trưng này có thể trích xuất trực tiếp từ sản phẩm.

Trong phạm vi bài luận, đối tượng nghiên cứu có dạng văn bản, nên kỹ thuật thường được sử dụng là phép đo tần suất kết hợp với tần suất xuất hiện ngược TF-IDF (Term Frequency/ Inverse Document Frequency)

TF-IDF (Term Frequency – Inverse Document Frequency) [4] là một con số thu được qua thống kê, thể hiện mức độ quan trọng của một từ trong một văn bản. Độ quan trọng của một từ sẽ tỉ lệ thuận với số lần xuất hiện của nó trong văn bản và tỷ lệ nghịch với số lần xuất hiện của nó trong các văn bản khác của tập dữ liệu.

- TF (Term frequency): là tần suất xuất hiện của một từ trong văn bản

$$TF(t, d) = \frac{\text{Số lần từ } t \text{ xuất hiện trong văn bản } d}{\text{Tổng số từ trong văn bản } d} \quad (2.1)$$

- IDF (Inverse Document Frequency): là tần suất nghịch của một từ trong văn bản, dùng để ước lượng mức độ quan trọng của một từ. Khi tính tần suất xuất hiện TF, các từ đều được coi là quan trọng như nhau. Tuy nhiên, trong văn bản thường xuất hiện nhiều từ không quan trọng trong việc thể hiện ý nghĩa của văn bản như: từ nối (và, nhưng...), giới từ (trên, trong...), từ chỉ định (ấy, đó...). Do đó, ta cần giảm đi mức độ quan trọng của những từ đó bằng cách sử dụng IDF.

$$IDF(t, D) = \log \frac{\text{Tổng số văn bản trong tập } D}{\text{Số văn bản có chứa từ } t} \quad (2.2)$$

- Ta có công thức TF-IDF để xác định mức độ quan trọng của một từ trong tập dữ liệu:

$$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D) \quad (2.3)$$

Sau khi tính chỉ số TF-IDF cho mỗi từ, mỗi sản phẩm được biểu diễn thông qua các từ, và được gọi là TF-IDF vector. Ma trận các vector TF-IDF thu được chính là hồ sơ sản phẩm cần xây dựng.

2.1.4. Các phương pháp lọc theo nội dung

Sau khi xây dựng hồ sơ sản phẩm, ta tiến hành phân tích để đưa ra gợi ý. Lọc theo nội dung có thể tiếp cận theo 2 xu hướng: lọc dựa trên bộ nhớ và lọc dựa trên mô hình. Nội dung cụ thể các phương pháp được trình bày dưới đây:

2.1.4.1. Lọc nội dung dựa vào bộ nhớ

Ý tưởng: hệ thống sẽ tiến hành phân tích nội dung của sản phẩm và tính toán độ tương tự giữa các sản phẩm, từ đó đưa ra gợi ý cho người dùng. Phương pháp này sử dụng toàn bộ tập hồ sơ sản phẩm để thực hiện huấn luyện và dự đoán.

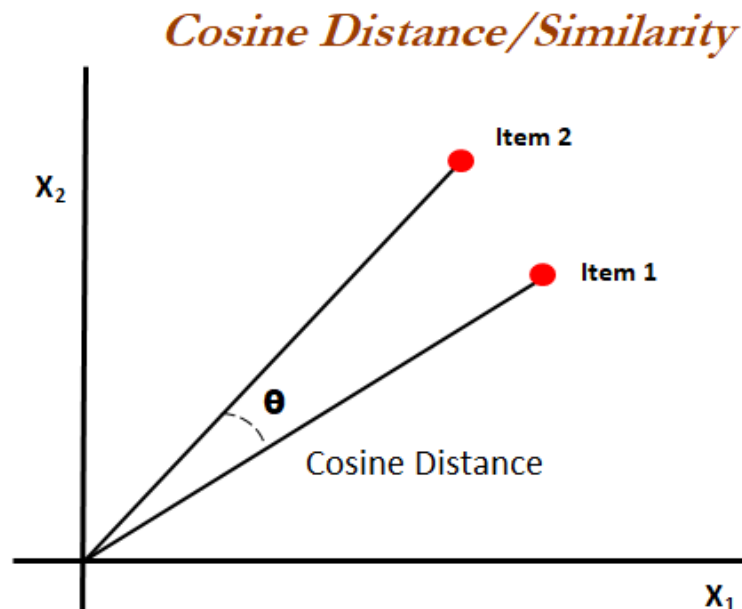


Hình 2.1: Minh họa phương pháp lọc nội dung dựa vào bộ nhớ

Trong pha xây dựng mô hình, hệ thống sẽ tính toán độ tương tự giữa các cặp sản phẩm, và sử dụng danh sách sản phẩm tương đồng nhất với sản phẩm người dùng đã đánh giá hoặc đã xem để gợi ý tới người dùng.

Các bước thực hiện gợi ý lọc nội dung dựa trên bộ nhớ:

- Bước 1 xây dựng hồ sơ sản phẩm đã được trình bày ở mục 2.1.3.
- Bước 2, sau khi đã có ma trận hồ sơ sản phẩm, ta tiến hành tính toán độ tương tự của một sản phẩm với các sản phẩm khác trong ma trận bằng cách áp dụng các hàm đo độ tương tự khác nhau như cosine, khoảng cách Euclidean. Thực nghiệm cho thấy, khoảng cách cosine thể hiện tốt trong các bài toán xử lý dữ liệu dạng văn bản, nên trong phạm vi bài luận, tôi sử dụng khoảng cách cosine làm phương pháp chính tính toán độ tương đồng giữa 2 đối tượng. Một cách toán học, nó tính toán cosine góc giữa 2 vector chiếu lên không gian đa chiều như hình 2.2.
- Công thức cosine tính góc giữa 2 vector [5]



Hình 2.2: Minh họa khoảng cách cosine giữa hai vector

$$similarity = \cos\theta = \frac{A \cdot B}{||A|| ||B||} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (2.4)$$

- + Giá trị cosine trong khoảng $[-1, 1]$, giá trị này càng lớn thể hiện độ tương đồng giữa 2 vector càng lớn.
- + Ta tiến hành tính toán toàn bộ độ tương tự giữa các cặp sản phẩm, với mỗi sản phẩm, chọn ra danh sách sản phẩm có độ tương đồng lớn nhất và đưa ra gợi ý cho sản phẩm.

2.1.4.2. Lọc nội dung dựa trên mô hình

Lọc nội dung dựa trên mô hình là phương pháp sử dụng tập hồ sơ sản phẩm và hồ sơ người dùng để xây dựng mô hình huấn luyện. Mô hình sau đó sẽ sử dụng kết quả của mô hình huấn luyện để sinh ra tư vấn cho người dùng. Trong cách tiếp cận này, lọc nội dung có thể sử dụng các kỹ thuật học máy để đưa ra dự đoán. Các bước xây dựng mô hình được thực hiện theo các bước dưới đây:

- **Bước 1:** Xây dựng ma trận đánh giá người dùng – sản phẩm - Utility Matrix
 - + Như đã đề cập, có 2 thực thể chính trong các hệ gợi ý là người dùng và sản phẩm. Mỗi người dùng sẽ có một mức độ quan tâm khác nhau đến từng sản

- phẩm. Mức độ quan tâm này, nếu đã biết, được gán cho một giá trị tương ứng với mỗi cặp user – item, giả sử mức độ quan tâm này được đo bằng đánh giá - rating - của người dùng với sản phẩm. Tập hợp tất cả ratings, bao gồm cả những giá trị chưa biết cần được dự đoán, tạo nên ma trận gọi là Utility matrix.
- + Xét ví dụ utility matrix trong bảng 2.1:

Bảng 2.1: Ví dụ Utility matrix

	A	B	C	D	E	F
Nhà giả kim	1	?	3	5	4	3
Mắt biếc	?	2	?	?	4	?
Kính vạn hoa	4	?	3	2	4	4
Nghĩ giàu, làm giàu	?	?	?	?	?	5
Thám tử lừng danh Conan tập 1	5	1	3	3	?	1

- + Trong ví dụ này, có 6 người dùng A-F và 5 cuốn sách. Các cuốn sách được đánh số theo mức độ từ 1-5 sao. Các dấu ‘?’ nền màu xám ứng với việc dữ liệu chưa tồn tại trong cơ sở dữ liệu. Công việc của hệ gợi ý là dự đoán các giá trị trong các ô màu xám.
- + Thông thường, có rất nhiều người dùng và sản phẩm trong hệ thống, và mỗi người dùng chỉ đánh giá một số lượng rất nhỏ các sản phẩm. Vì vậy, số lượng các ô xám càng ít, thì độ chính xác của hệ thống sẽ càng tốt.
- + Các sản phẩm được biểu diễn dưới dạng vector đặc trưng – TFIDF vector trong bảng 2.2.

Bảng 2.2: Utility matrix với vector đặc trưng của sản phẩm

	A	B	C	D	E	F	Vector đặc trưng
Nhà giả kim	1	?	3	5	4	3	$x_1 = [0.99, 0.52, \dots, 0.25, 0.05]$
Mắt biếc	?	2	?	?	4	?	$x_2 = [0.12, 0.43, \dots, 0.15, 0.95]$
Kính vạn hoa	4	?	3	2	4	4	$x_3 = [0.64, 0.59, \dots, 0.31, 0.09]$
Nghĩ giàu, làm giàu	?	?	?	?	?	5	$x_4 = [0.82, 0.12, \dots, 0.43, 0.75]$
Thám tử lừng danh Conan tập 1	5	1	3	3	?	1	$x_5 = [0.03, 0.02, \dots, 0.85, 0.55]$
Mô hình người dùng	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	

- + Bài toán đi tìm mô hình θ_i cho mỗi người dùng có thể coi là một bài toán hồi quy trong trường hợp ratings là một dải giá trị, hoặc là bài toán phân loại trong trường hợp ratings có dạng nhị phân (thích/không thích, 0/1). Dữ liệu huấn luyện để xây dựng mô hình θ_i là các cặp (vector đặc trưng sản phẩm, ratings) tương ứng với mỗi sản phẩm mà người dùng đó đã đánh giá. Việc điền các giá trị còn thiếu trong ma trận Utility matrix chính là việc dự đoán đầu ra cho các sản phẩm chưa được đánh giá khi áp dụng mô hình θ_i lên chúng.

- **Bước 2:** Xây dựng hàm mất mát
 - + Giả sử rằng số người dùng là N , số sản phẩm là M , utility matrix được mô tả bởi ma trận Y . Thành phần ở hàng thứ m , cột thứ n là đánh giá của người dùng thứ n lên sản phẩm thứ m . R là ma trận thể hiện việc một người dùng đã đánh giá một sản phẩm hay chưa. Cụ thể nếu $r_{ij} = 1$ nếu sản phẩm thứ i đã được đánh giá bởi người dùng thứ j và ngược lại $r_{ij} = 0$
 - + Mô hình tuyến tính [6]: giả sử rằng ta có thể tìm ra được một mô hình cho mỗi người dùng, minh họa bởi vector cột hệ số w_n và bias b_n sao cho rating của một người dùng cho một sản phẩm có thể tính được bằng hàm tuyến tính:

$$y_{mn} = x_m w_n + b_n \quad (2.5)$$

Trong đó x_m là một vector hàng, w_n là một vector cột

- + Xét một người dùng bất kỳ, nếu ta gọi tập huấn luyện là tập hợp các thành phần đã được điền của y_n , ta có thể xây dựng hàm mất mát theo Ridge Regression như sau:

$$\mathcal{L}_n = \frac{1}{2} \sum_{m: r_{mn}=1} (x_m w_n + b_n - y_{mn})^2 + \frac{\lambda}{2} \|w_n\|_2^2 \quad (2.6)$$

Trong đó thành phần thứ hai là regularization term và λ là một số dương. Trong thực hành, trung bình cộng của lỗi thường được dùng, và mất mát \mathcal{L}_n được viết lại thành:

$$\mathcal{L}_n = \frac{1}{2s_n} \sum_{m: r_{mn}=1} (x_m w_n + b_n - y_{mn})^2 + \frac{\lambda}{2s_n} \|w_n\|_2^2 \quad (2.7)$$

Trong đó s_n là số lượng các sản phẩm mà người dùng thứ n đã đánh giá. Nói cách khác s_n là tổng các phần tử trên cột thứ n của ma trận R .

$$s_n = \sum_{m=1}^M r_{mn} \quad (2.8)$$

- + Vì hàm mục tiêu chỉ phụ thuộc vào các sản phẩm đã được đánh giá, ta rút gọn nó bằng cách đặt \widehat{y}_n là sub vector của y được xây dựng bằng cách trích các thành phần khác dấu '?' ở cột thứ n , tức đã được đánh giá bởi người dùng thứ n trong Utility matrix Y . Đồng thời đặt \widehat{X}_n là sub matrix của ma trận đặc trưng X , được tạo bằng các trích các hàng tương ứng với các sản phẩm đã được đánh giá bởi người dùng thứ n . Khi đó, biểu thức hàm mất mát của mô hình cho người dùng thứ n được viết gọn thành:

$$\mathcal{L}_n = \frac{1}{2s_n} \|\widehat{X}_n w_n + b_n e_n - \widehat{y}_n\|_2^2 - \frac{\lambda}{2s_n} \|w_n\|_2^2 \quad (2.9)$$

Trong đó e_n là vector cột chứa s_n phần tử 1.

- + Mục tiêu của chúng ta là tìm cặp nghiệm w_n, b_n để tối ưu hàm mất mát. Từ đó thay vào hàm tuyến tính dự đoán đánh giá của người dùng cho sản phẩm.

2.2. Kỹ thuật khuyến nghị sách dựa trên lọc cộng tác

Qua tìm hiểu về hệ khuyến nghị dựa trên nội dung, có thể thấy, khi xây dựng mô hình cho cho một người dùng, các hệ thống lọc nội dung không tận dụng được thông tin từ

các khách hàng khác. Những thông tin này thường rất hữu ích trong việc hướng người dùng đến những sản phẩm mới. Bên cạnh đó, không phải lúc nào chúng ta cũng có bản mô tả sản phẩm. Tuy nhiên, những nhược điểm đó có thể giải quyết bằng lọc cộng tác. Trong mục 2.2, tôi sẽ trình bày chi tiết về phương pháp khuyến nghị lọc cộng tác và các kỹ thuật gợi ý của phương pháp này.

2.2.1. Khái niệm

Không giống với phương pháp lọc theo nội dung, phương pháp lọc cộng tác gợi ý dựa trên sự tương quan giữa các người dùng và/hoặc sản phẩm. Hệ thống sẽ so sánh, tính toán độ tương tự giữa những người dùng hay mặt hàng, từ đó người dùng sẽ được gợi ý những thông tin, mặt hàng được ưa chuộng nhất bởi những người dùng có cùng sở thích. Trong phương pháp này, hệ thống thường xây dựng các ma trận đánh giá bởi người dùng lên các mặt hàng, bản tin. Từ đó tính toán độ tương tự giữa họ.

Đầu vào của bài toán là ma trận thể hiện những hành vi quá khứ, gọi là ma trận người dùng – sản phẩm. Các hàng trong ma trận đại diện cho người dùng, các cột đại diện cho sản phẩm, giá trị mỗi ô là đánh giá của người dùng lên sản phẩm đó. Đầu ra của bài toán là: đánh giá của người dùng lên những sản phẩm mà họ chưa đánh giá. Hệ thống gợi ý dựa trên các đánh giá này mà xếp hạng các sản phẩm và gợi ý cho người dùng

Tùy theo hệ thống mà đánh giá của người dùng được quy ước những giá trị nào. Trong ví dụ trong bảng 2.3, các đánh giá có giá trị từ 1->5.

Bảng 2.3: Ma trận đánh giá của người dùng

	Sản phẩm 1	Sản phẩm 2	Sản phẩm 3
Người dùng 1	1	0	5
Người dùng 2	4	2	2
Người dùng 3	0	0	0

Ở ma trận này, đánh giá của người dùng 1 đối với sản phẩm 1 là 1, sản phẩm 3 là 5, sản phẩm 2 chưa được đánh giá. Hệ thống gợi ý phải đưa ra dự đoán: người dùng 1 đánh giá sản phẩm 2 là bao nhiêu, người dùng 3 đánh giá sản phẩm 1, 2, 3 là bao nhiêu.

2.2.2. Phát biểu bài toán lọc cộng tác

Ký hiệu $U = \{u_1, u_2, \dots, u_m\}$ là tập gồm m người dùng, $P = \{p_1, p_2, \dots, p_n\}$ là tập gồm n sản phẩm mà người dùng có thể lựa chọn.

Ký hiệu $R = \{r_{ij}\}$, $i = 1 \dots m$, $j = 1 \dots n$. Trong đó mỗi người dùng $u_i \in U$ đưa ra đánh giá của mình cho một số sản phẩm $p_j \in P$ bằng một số r_{ij} . Giá trị r_{ij} phản ánh mức độ ưa thích của người dùng u_i đối với sản phẩm p_j , giá trị r_{ij} có thể được thu thập trực tiếp bằng cách hỏi ý kiến người dùng hoặc thu thập gián tiếp thông qua cơ chế phản hồi của người dùng. Giá trị $r_{ij} = 0$ trong trường hợp người dùng u_i chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j

Với một người dùng cần được gợi ý u_a (được gọi là người dùng hiện thời, người dùng cần được tư vấn, hay người dùng tích cực), bài toán lọc cộng tác là bài toán dự đoán đánh giá của u_a đối với mặt hàng mà u_a chưa đánh giá ($r_{ij} = 0$), trên cơ sở đó gợi ý cho u_a những sản phẩm được đánh giá cao.

Bảng 2.4 thể hiện một ví dụ với ma trận đánh giá R trong hệ gồm 5 người dùng $U = \{u_1, u_2, u_3, u_4, u_5\}$ và 4 sản phẩm $P = \{p_1, p_2, p_3, p_4\}$. Mỗi người dùng đều đưa ra các đánh giá của mình về các sản phẩm theo thang bậc $\{0, 1, 2, 3, 4, 5\}$. Giá trị $r_{ij} = 0$ được hiểu là người dùng chưa đánh giá hoặc chưa bao giờ biết đến sản phẩm p_j . Giá trị $r_{5,2} = ?$ là sản phẩm hệ thống cần dự đoán cho người dùng u_5

Bảng 2.4: Ma trận đánh giá R

	p_1	p_2	p_3	p_4
u_1	2	1	3	5
u_2	4	2	1	0
u_3	3	0	2	4
u_4	4	4	0	0
u_5	4	?	5	5

Ma trận đánh giá R là thông tin đầu vào duy nhất của phương pháp lọc cộng tác. Dựa trên ma trận đánh giá, các phương pháp lọc cộng tác thực hiện hai tác vụ: Dự đoán quan điểm của người dùng hiện thời về các sản phẩm mà họ chưa đánh giá, đồng thời đưa ra một danh sách các sản phẩm có đánh giá cao nhất phân bổ cho người dùng hiện thời.

Phương pháp lọc cộng tác bao gồm các kỹ thuật như kỹ thuật láng giềng, kỹ thuật mạng Bayes, mạng neuron kết hợp SVD...

Lọc cộng tác tiếp cận theo hai xu hướng chính: lọc cộng tác dựa trên bộ nhớ và lọc cộng tác dựa trên mô hình. Mỗi phương pháp tiếp cận có những ưu điểm và hạn chế riêng, khai thác các mối liên hệ trên ma trận đánh giá người dùng.

2.2.3. Phương pháp khuyến nghị lọc cộng tác dựa trên bộ nhớ

Phương pháp này hay còn được gọi với cái tên là gợi ý dựa trên láng giềng gần nhất (Neighborhood-based Collaborative Filtering). Các phương pháp lọc dựa trên bộ nhớ sử dụng toàn bộ ma trận đánh giá để sinh ra dự đoán các sản phẩm cho người dùng hiện thời. Phương pháp thực hiện theo hai bước: Tính toán mức độ tương tự và bước tạo nên dự đoán

- Tính toán độ tương tự $\text{sim}(x, y)$: mô tả khoảng cách, sự liên quan, hay trọng số giữa hai người dùng x và y hoặc giữa hai sản phẩm x và y
- Dự đoán: đưa ra dự đoán cho người dùng cần được tư vấn bằng cách xác định tập láng giềng của người dùng này. Tập láng giềng của người dùng cần tư vấn được xác định dựa trên mức độ tương tự giữa các cặp người dùng hoặc sản phẩm.

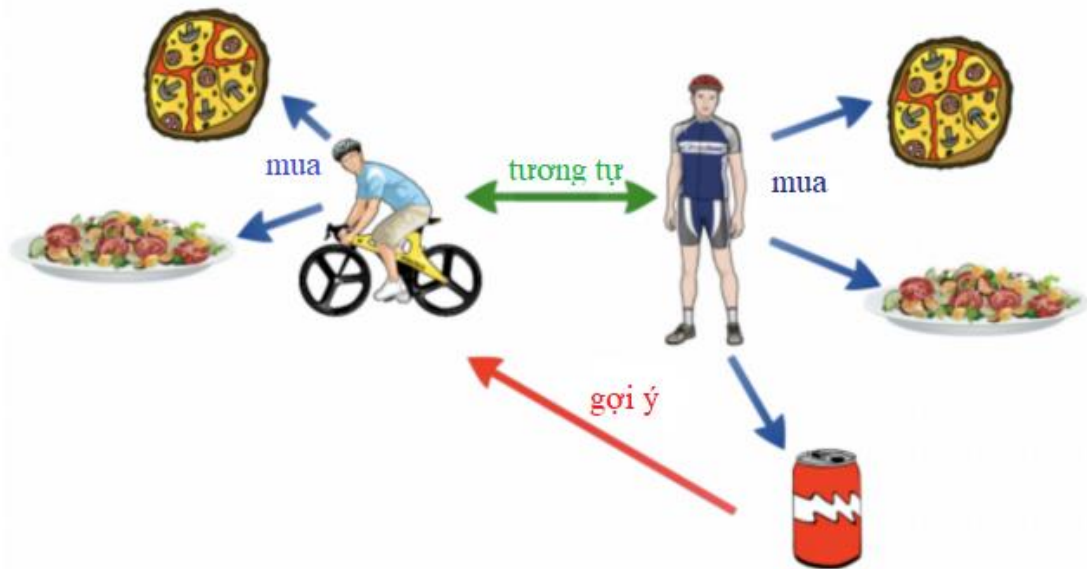
Các phương pháp tính toán mức độ tương tự: có nhiều phương pháp khác nhau tính toán mức độ tương tự $\text{sim}(x, y)$ giữa các cặp người dùng hoặc các cặp sản phẩm. Hai phương pháp phổ biến được sử dụng là độ tương quan Pearson và giá trị cosin giữa hai vector

Để đo mức độ tương tự giữa hai người dùng hoặc hai sản phẩm, cách thường làm là xây dựng vector đặc trưng cho mỗi người dùng/sản phẩm, việc xây dựng vector đặc

trung này được xây dựng trực tiếp dựa trên ma trận đánh giá Utility matrix chứ không dùng dữ liệu ngoài như xây dựng hồ sơ sản phẩm.

Lọc cộng tác dựa trên bộ nhớ được tiếp cận theo hai phương pháp chính: lọc dựa vào người dùng (User Based) và lọc dựa vào sản phẩm (Item Based)

2.2.3.1. Phương pháp lọc cộng tác dựa trên người dùng



Hình 2.3: Minh họa phương pháp lọc cộng tác dựa trên người dùng

Chuẩn hóa ma trận Utility matrix:

- Để có thể sử dụng ma trận này vào việc tính toán, chúng ta cần thay những dấu ‘?’ bởi một giá trị. Đơn giản nhất thì chúng ta có thể thay vào đó giá trị 0 hoặc giá trị trung bình ratings là 2.5. Tuy nhiên, những giá trị này sẽ hạn chế với cách đánh giá của mỗi người dùng. Với những người dùng khó tính, họ thậm chí chỉ đánh giá 3 sao cho một sản phẩm họ thích và dưới 3 sao khi không thích sản phẩm đó. Do đó, ta chọn giá trị trung bình cộng đánh giá của mỗi người dùng để chuẩn hóa ma trận. (Hình 2.4-a)
- Tuy nhiên, thay vì trực tiếp sử dụng các giá trị này thay cho dấu ‘?’ của mỗi người dùng, chúng ta sẽ trừ đi đánh giá của mỗi người dùng cho giá trị trung bình tương ứng của người dùng đó và thay dấu ‘?’ thành giá trị 0. Mục đích của cách xử lý này là phân loại ratings thành 2 loại: giá trị âm (người dùng không thích sản phẩm) và giá trị dương (người dùng thích sản phẩm). Các giá trị bằng 0 tương ứng với những đánh giá chưa được thực hiện. Một lí do nữa là do số chiều của Utility matrix rất lớn, trong khi lượng ratings biết trước thường rất nhỏ so với kích thước toàn bộ ma trận. Nếu thay những giá trị chưa biết bằng 0, ta có được sparse matrix (ma trận thưa thớt), tức ma trận chỉ lưu các giá trị khác 0 và vị trí của giá trị đó, việc lưu trữ sẽ tối ưu hơn. (Hình 2.4-b)
- Sau khi chuẩn hóa ma trận utility, tiến hành tính toán độ tương tự giữa 2 người dùng:
 - + Độ tương quan Pearson giữa 2 người dùng x và y (User-based similarity) được tính toán theo công thức sau [3]:

$$sim(x, y) = \frac{\sum_{p \in P_{xy}} (r_{x,p} - \bar{r}_x)(r_{y,p} - \bar{r}_y)}{\sqrt{\sum_{p \in P_{xy}} (r_{x,p} - \bar{r}_x)^2 \sum_{p \in P_{xy}} (r_{y,p} - \bar{r}_y)^2}} \quad (2.10)$$

Trong đó $P_{xy} = \{p \in P | r_{x,p} \neq 0 \cap r_{y,p} \neq 0\}$ là tập tất cả các sản phẩm người dùng x và y cùng đánh giá \bar{r}_x, \bar{r}_y là trung bình cộng các đánh giá khác 0 của người dùng x và người dùng y

- + Độ tương tự vector giữa hai người dùng x, y là cosine của hai vector x và y theo công thức dưới đây. Trong đó, hai người dùng x và y được xem xét như hai vector m chiều, $m = |P_{xy}|$ là số lượng các sản phẩm cả hai người dùng cùng đánh giá [3]. (Hình 2.4-c)

$$sim(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| ||\vec{y}||} = \frac{\sum_{p \in P_{xy}} r_{x,p} r_{y,p}}{\sqrt{\sum_{p \in P_{xy}} r_{x,p}^2} \sqrt{\sum_{p \in P_{xy}} r_{y,p}^2}} \quad (2.11)$$

- Để xác định mức độ quan tâm của một người dùng lên một sản phẩm dựa trên các người dùng gần nhất sử dụng thuật toán k láng giềng gần nhất
- + Công thức phổ biến thường được sử dụng để dự đoán đánh giá của người dùng u cho sản phẩm i là [7]:

$$\hat{y}_{i,u} = \frac{\sum_{u_j \in N(u,i)} \bar{y}_{i,u_j} sim(u, u_j)}{\sum_{u_j \in N(u,i)} |sim(u, u_j)|} \quad (2.12)$$

Trong đó $N(u, i)$ là tập k người dùng có độ tương đồng cao nhất với người dùng u và đã từng đánh giá sản phẩm i (Hình 2.4-d, 2.4-e)

- + Cuối cùng, cộng lại các giá trị đánh giá với giá trị trung bình theo từng cột, thu được ma trận hoàn thiện. (Hình 2.4-f)

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	4	?	?	0	?	2	?
i_2	?	4	1	?	?	1	1
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5
	↓	↓	↓	↓	↓	↓	↓
\bar{u}_j	3.25	2.75	2.5	1.33	2.5	1.5	3.33

a) Original utility matrix \mathbf{Y} and mean user ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0	0
i_1	0.75	0	0	-1.33	0	0.5	0
i_2	0	1.25	-1.5	0	0	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0	0.67
i_4	-1.25	-2.75	1.5	0	0	0	1.67

b) Normalized utility matrix $\bar{\mathbf{Y}}$.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
u_0	1	0.83	-0.58	-0.79	-0.82	0.2	-0.38
u_1	0.83	1	-0.87	-0.40	-0.55	-0.23	-0.71
u_2	-0.58	-0.87	1	0.27	0.32	0.47	0.96
u_3	-0.79	-0.40	0.27	1	0.87	-0.29	0.18
u_4	-0.82	-0.55	0.32	0.87	1	0	0.16
u_5	0.2	-0.23	0.47	-0.29	0	1	0.56
u_6	-0.38	-0.71	0.96	0.18	0.16	0.56	1

c) User similarity matrix \mathbf{S} .

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	1.75	2.25	-0.5	-1.33	-1.5	0.18	-0.63
i_1	0.75	0.48	-0.17	-1.33	-1.33	0.5	0.05
i_2	0.91	1.25	-1.5	-1.84	-1.78	-0.5	-2.33
i_3	-1.25	-0.75	0.5	2.67	1.5	0.59	0.67
i_4	-1.25	-2.75	1.5	1.57	1.56	1.59	1.67

d) $\hat{\mathbf{Y}}$

Predict normalized rating of u_1 on i_1 with $k = 2$

Users who rated i_1 : $\{u_0, u_3, u_5\}$

Corresponding similarities: $\{0.83, -0.40, -0.23\}$

\Rightarrow most similar users: $\mathcal{N}(u_1, i_1) = \{u_0, u_5\}$

with normalized ratings $\{0.75, 0.5\}$

$$\Rightarrow \hat{y}_{i_1, u_1} = \frac{0.83 \cdot 0.75 + (-0.23) \cdot 0.5}{0.83 + |-0.23|} \approx 0.48$$

e) Example

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	1.68	2.70
i_1	4	3.23	2.33	0	1.67	2	3.38
i_2	4.15	4	1	-0.5	0.71	1	1
i_3	2	2	3	4	4	2.10	4
i_4	2	0	4	2.9	4.06	3.10	5

f) Full \mathbf{Y}

Hình 2.4: Minh họa chuẩn hóa ma trận

2.2.3.2. Phương pháp lọc cộng tác dựa trên sản phẩm

Chuẩn hóa ma trận Utility:

- Thay vì tính trung bình cộng ratings của các người dùng, chúng ta sẽ đánh trung bình cộng ratings của các sản phẩm
- Thực hiện chuẩn hóa bằng cách trừ các ratings đã biết của sản phẩm cho giá trị trung bình vừa tính được, đồng thời thay các giá trị chưa biết bằng 0. Từ đó thu được ma trận utility chuẩn hóa

Một số công thức tính độ tương tự giữa 2 sản phẩm

- Độ tương quan Pearson giữa 2 sản phẩm x và y (Item-based similarity) được tính toán theo công thức sau [3]:

$$sim(x, y) = \frac{\sum_{u \in U_{xy}} (r_{u,x} - \bar{r}_x)(r_{u,y} - \bar{r}_y)}{\sqrt{\sum_{u \in U_{xy}} (r_{u,x} - \bar{r}_x)^2 \sum_{u \in U_{xy}} (r_{u,y} - \bar{r}_y)^2}} \quad (2.13)$$

Trong đó $U_{xy} = \{u \in U | r_{u,x} \neq 0 \cap r_{u,y} \neq 0\}$ là tập tất cả người dùng cùng đánh giá sản phẩm x và y, \bar{r}_x, \bar{r}_y là đánh giá trung bình cho sản phẩm x và sản phẩm y

- Độ tương tự vector giữa hai sản phẩm x, y là cosine của hai vector x và y theo công thức dưới đây. Trong đó, hai sản phẩm x và y được xem xét như hai vector cột n chiều, $n = |U_{xy}|$ là số lượng các người dùng đánh giá cả 2 sản phẩm x và y [3].

$$sim(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}|| ||\vec{y}||} = \frac{\sum_{u \in U_{xy}} r_{u,x} r_{u,y}}{\sqrt{\sum_{u \in U_{xy}} r_{u,x}^2} \sqrt{\sum_{u \in U_{xy}} r_{u,y}^2}} \quad (2.14)$$

Dự đoán đánh giá của người dùng u cho sản phẩm i, ta thực hiện:

- Tìm tập $N(i, u)$ các sản phẩm mà người dùng u đã đánh giá
- Tính độ tương tự của sản phẩm i với các sản phẩm trong tập $N(i, u)$. Chọn ra k sản phẩm có độ tương đồng cao nhất với i
- Tính ratings theo công thức [7]:

$$\widehat{y_{i,u}} = \frac{\sum_{i_j \in N(i,u)} \overline{y_{u,i_j}} sim(i, i_j)}{\sum_{i_j \in N(i,u)} |sim(i, i_j)|} \quad (2.15)$$

- Sau khi tính toán xong, cộng lại các giá trị đánh giá với giá trị trung bình theo từng hàng, thu được ma trận hoàn thiện. Ta có ví dụ minh họa các bước như hình 2.5.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6	
i_0	5	5	2	0	1	?	?	→ 2.6
i_1	4	?	?	0	?	2	?	→ 2
i_2	?	4	1	?	?	1	1	→ 1.75
i_3	2	2	3	4	4	?	4	→ 3.17
i_4	2	0	4	?	?	?	5	→ 2.75

a) Original utility matrix \bar{Y} and mean item ratings.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-6	-2.6	-1.6	0	0
i_1	2	0	0	-2	0	0	0
i_2	0	2.25	-0.75	0	0	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0	0.83
i_4	-0.75	-2.75	1.25	0	0	0	2.25

b) Normalized utility matrix \bar{Y} .

	i_0	i_1	i_2	i_3	i_4
i_0	1	0.77	0.49	-0.89	-0.52
i_1	0.77	1	0	-0.64	-0.14
i_2	0.49	0	1	-0.55	-0.88
i_3	-0.89	-0.64	-0.55	1	0.68
i_4	-0.52	-0.14	-0.88	0.68	1

c) Item similarity matrix S .

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	2.4	2.4	-6	-2.6	-1.6	-0.29	-1.52
i_1	2	2.4	-0.6	-2	-1.25	0	-2.25
i_2	2.4	2.25	-0.75	-2.6	-1.20	-0.75	-0.75
i_3	-1.17	-1.17	-0.17	0.83	0.83	0.34	0.83
i_4	-0.75	-2.75	1.25	1.03	1.16	0.65	2.25

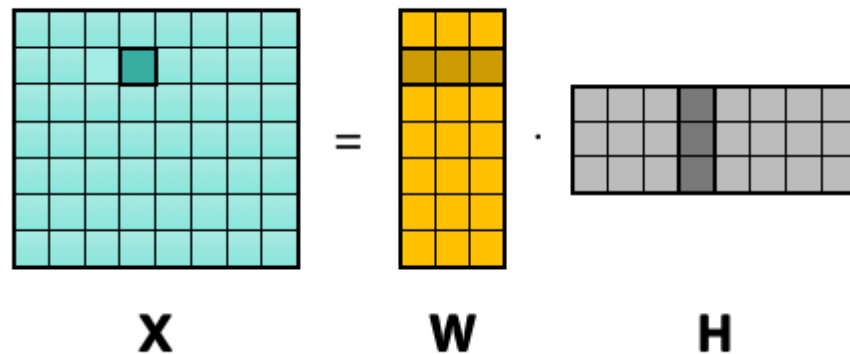
d) Normalized utility matrix \bar{Y} .

Hình 2.5: Minh họa chuẩn hóa utility matrix

2.2.4. Phương pháp khuyến nghị lọc cộng tác dựa trên mô hình Matrix Factorization

Matrix Factorization là một hướng tiếp cận khác của lọc cộng tác, còn gọi là Matrix Decomposition, nghĩa là gợi ý bằng kỹ thuật phân rã ma trận.

Kỹ thuật phân rã ma trận là phương pháp chia một ma trận lớn X thành hai ma trận có kích thước nhỏ hơn là W và H như minh họa hình 2.6, sao cho ta có thể xây dựng lại X từ hai ma trận nhỏ hơn này càng chính xác càng tốt, nghĩa là $X \sim WH^T$ [8]



Hình 2.6: Kỹ thuật phân rã ma trận

Có thể hiểu rằng, ý tưởng chính của Matrix Factorization là đặt người dùng và sản phẩm vào trong cùng một không gian thuộc tính ẩn. Trong đó, $W \in R^{|U| \times K}$ là một ma

trận mà mỗi dòng u là một vector bao gồm K nhân tố tiềm ẩn (latent factors) mô tả người dùng u và $H \in R^{|I| \times K}$ là một ma trận mà mỗi dòng i là một vector bao gồm K nhân tố tiềm ẩn mô tả cho sản phẩm i .

Áp dụng phương pháp này vào bài toán gợi ý, ta có x là một hồ sơ sản phẩm item profiles

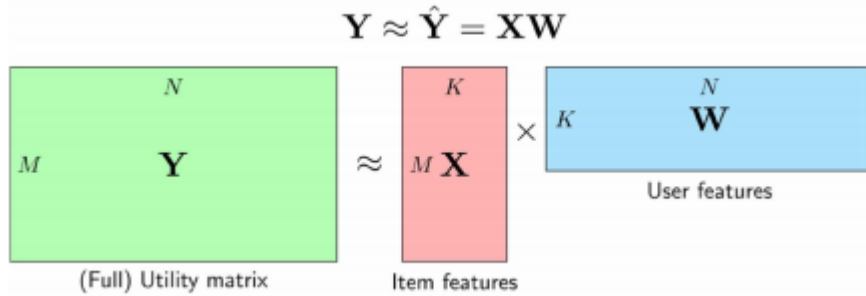
Mục tiêu của chúng ta là tìm một vector w tương ứng với mỗi người dùng sao cho ratings đã biết của người dùng đó cho sản phẩm (gọi là y) xấp xỉ với $y \sim xw$

Mở rộng với Y là Utility matrix, giả sử đã được điền hết giá trị, ta có:

$$Y \approx \begin{bmatrix} x_1 w_1 & x_1 w_2 & \dots & x_1 w_N \\ x_2 w_1 & x_2 w_2 & \dots & x_2 w_N \\ \vdots & \vdots & \ddots & \vdots \\ x_M w_1 & x_M w_2 & \dots & x_M w_N \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} [w_1 \ w_2 \ \dots \ w_N] = XW \quad (2.16)$$

Với M, N lần lượt là số người dùng và số sản phẩm [8]

Bài toán đưa về bài toán tối ưu các ma trận X và W , trong đó X là ma trận hồ sơ sản phẩm, còn W là ma trận các mô hình người dùng, mỗi cột tương ứng với một người dùng. Mục tiêu xấp xỉ Utility matrix $Y \in R^{M \times N}$ bằng tích của hai ma trận con là $X \in R^{M \times K}$ và $W \in R^{K \times N}$



Hình 2.7: Bài toán tối ưu ma trận

Xây dựng hàm mất mát [8]:

$$\mathcal{L}(X, W) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} (\|X\|_F^2 + \|W\|_F^2) \quad (2.17)$$

Trong đó $r_{mn} = 1$ nếu sản phẩm thứ m đã được đánh giá bởi người dùng thứ n , $\|\cdot\|_F^2$ là Frobenius norm, tức căn bậc hai của tổng bình phương tất cả các phần tử của ma trận, s là toàn bộ số rating đã có. Thành phần thứ nhất là trung bình sai số của mô hình, thành phần thứ hai là l2 regularization, giúp tránh overfitting

Việc tối ưu đồng thời X, W là tương đối phức tạp, thay vào đó, phương pháp được sử dụng lần lượt là tối ưu một ma trận trong khi cố định ma trận kia, tới khi hội tụ.

Tối ưu hàm mất mát [8]:

- Khi cố định X , việc tối ưu W chính là bài toán tối ưu trong lọc cộng tác dựa trên mô hình:

$$\mathcal{L}(W) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|W\|_F^2 \quad (2.18)$$

- Khi cố định W, việc tối ưu X được đưa về tối ưu hàm

$$\mathcal{L}(X) = \frac{1}{2s} \sum_{n=1}^N \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|X\|_F^2 \quad (2.19)$$

Hai bài toán trên được tối ưu bằng phương pháp Gradient Descent [8].

- Bài toán tối ưu W khi cố định X có thể được tách thành N bài toán nhỏ, mỗi bài toán ứng với việc tối ưu một cột của ma trận W.

$$\mathcal{L}(w_n) = \frac{1}{2s} \sum_{m:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|w_n\|_2^2 \quad (2.20)$$

Bằng cách đạo hàm hàm mất mát, ta có công thức cập nhật cho mỗi cột của W là

$$w_n = w_n - \eta \left(-\frac{1}{s} \widehat{X}_n^T (\widehat{y}_n - \widehat{X}_n w_n) + \lambda w_n \right) \quad (2.21)$$

- Tương tự cho bài toán tối ưu X khi cố định W, mỗi cột của X, tức vector cho mỗi sản phẩm, sẽ được tính bằng cách tối ưu:

$$\mathcal{L}(x_m) = \frac{1}{2s} \sum_{n:r_{mn}=1} (y_{mn} - x_m w_n)^2 + \frac{\lambda}{2} \|x_m\|_2^2 \quad (2.22)$$

Ta tính được công thức cập nhật cho mỗi hàng của X có dạng:

$$x_m = x_m - \eta \left(-\frac{1}{s} (\widehat{y}_m - x_m \widehat{W}_m) \widehat{W}_m^T + \lambda x_m \right) \quad (2.23)$$

- Trong đó \widehat{X}_n là ma trận được tạo bởi các hàng tương ứng với các sản phẩm đã được đánh giá đó, và \widehat{y}_n là các ratings tương ứng, \widehat{W}_m là ma trận được tạo bằng các cột của W ứng với các người dùng đã đánh giá sản phẩm đó và \widehat{y}_m là vector ratings tương ứng

Sau mỗi lần lặp, điểm hiện tại sẽ giảm một lượng tùy thuộc vào tốc độ η và lặp đến khi có được điểm cực tiểu chính là điểm cần tìm để hàm mất mát có giá trị nhỏ nhất.

2.3. So sánh đánh giá phương pháp lọc nội dung và lọc cộng tác

Qua tìm hiểu và phân tích các kỹ thuật trong lọc cộng tác và lọc nội dung, ta rút ra các ưu điểm và nhược điểm của 2 phương pháp trên và mô tả trong bảng 2.5 như sau:

Bảng 2.5: So sánh ưu nhược điểm của lọc nội dung và lọc cộng tác

	Lọc nội dung	Lọc cộng tác
Ưu điểm	<ul style="list-style-type: none"> - Mô hình không cần bất cứ dữ liệu gì về người dùng khác, gợi ý riêng theo từng đối tượng người dùng → hệ thống không yêu cầu số lượng người dùng lớn để đạt được độ chính xác đề nghị hợp lý - Mô hình có thể nắm bắt được sở thích cụ thể của người dùng và có thể đề xuất các mặt hàng thích hợp mà rất ít người dùng khác quan tâm 	<ul style="list-style-type: none"> - Không giới hạn về loại đối tượng dùng để gợi ý: phương pháp này dựa hoàn toàn vào đánh giá của những người dùng để đưa ra các nhận định về sở thích của người dùng → hệ thống không phụ thuộc vào sản phẩm - Gợi ý đa dạng: Khắc phục được giới hạn của phương pháp tiếp cận dựa trên nội dung, phương pháp lọc cộng tác có thể đưa ra các đối tượng sản phẩm khuyến nghị hoàn toàn khác so với các sản phẩm mà người dùng u đã thích trong quá khứ.
Nhược điểm	<ul style="list-style-type: none"> - Biểu diễn đặc trưng của sản phẩm đòi hỏi các kiến thức liên quan về lĩnh vực sản phẩm được gợi ý, đòi hỏi chuyên gia phân tích và mang tính thủ công → gặp khó khăn với các hệ thống có sản phẩm mang tính chất trừu tượng, phức tạp như chứng khoán, bất động sản... - Mô hình chỉ có thể đưa ra đề xuất dựa trên sở thích hiện có của người dùng. Nói cách khác mô hình có hạn chế trong việc hướng người dùng đến những sở thích mới. Thông thường những hệ thống gợi ý dựa trên nội dung đề xuất những đối tượng tương tự mà người dùng đã đánh giá trước đó. Trong một số trường hợp đặc biệt, đối tượng không nên được gợi ý vì chúng có độ tương tự gần như là tuyệt đối với sản phẩm người dùng vừa xem. Khi đó người dùng sẽ không quan tâm đến những sản phẩm đó nữa. - Vấn đề người dùng mới, khi đó họ chưa cung cấp đủ dữ liệu để 	<ul style="list-style-type: none"> - Vấn đề người dùng mới: để phân bổ chính xác sản phẩm người dùng quan tâm, lọc cộng tác phải ước lượng được sở thích của người dùng đối với các sản phẩm thông qua những đánh giá của họ trong quá khứ. Trong trường hợp một người dùng mới, số đánh giá của người dùng cho các sản phẩm là 0. Khi đó, phương pháp lọc cộng tác khó để đưa ra những khuyến nghị chính xác cho người dùng này. - Vấn đề sản phẩm mới: trong lọc thông tin, các sản phẩm thường xuyên được bổ sung, cập nhật vào hệ thống. Khi xuất hiện một sản phẩm mới, tất cả đánh giá người dùng cho sản phẩm này đều là 0. Do đó, lọc cộng tác không thể khuyến nghị sản phẩm cho bất kỳ người dùng nào trong hệ thống. - Vấn đề dữ liệu thưa: Kết quả dự đoán của lọc cộng tác phụ thuộc chủ yếu vào số các đánh giá có trước của người dùng đối với các sản phẩm. Tuy nhiên, đối với các hệ thống thực tế, số lượng người dùng và sản phẩm là rất lớn, số những đánh giá biết trước thường rất nhỏ so với số lượng các đánh giá cần được dự đoán.

	mô hình dự đoán, mô hình gặp vấn đề trong việc đưa ra gợi ý.	
--	--	--

Nhận xét thấy rằng, ưu điểm của lọc nội dung có thể khắc phục nhược điểm của lọc cộng tác, cũng như nhược điểm của lọc nội dung cũng được khắc phục bởi lọc cộng tác. Hai kỹ thuật này kết hợp tạo nên một hệ khuyến nghị đủ tốt cho bất kỳ hệ thống nào.

Kỹ thuật lọc nội dung phù hợp với các hệ thống không yêu cầu người dùng lớn, tuy nhiên sản phẩm gợi ý đòi hỏi rõ ràng về đặc trưng. Kỹ thuật lọc nội dung cũng được sử dụng để nhóm các sản phẩm tương đồng về đặc trưng, từ đó gợi ý cho người dùng mỗi khi họ chọn xem sản phẩm. Các hệ khuyến nghị lọc nội dung có gợi ý mang tính chất cá nhân.

Đối với lọc cộng tác, kỹ thuật này phù hợp với các hệ thống đã có nhiều người dùng trong hệ thống, càng nhiều người dùng thì hệ gợi ý càng tốt, và không phụ thuộc vào đối tượng sản phẩm gợi ý là gì. Do vậy, các hệ thống này mang tính chất đa dạng hơn hệ thống gợi ý lọc nội dung. Lọc cộng tác cũng có thể chia nhóm sản phẩm, nhóm người dùng dựa trên lịch sử giao dịch của người dùng, nên gợi ý mang tính chất bao quát hơn.

2.4. Kết chương

Chương 2 đã trình bày chi tiết về khái niệm, các kỹ thuật gợi ý, cơ sở lý thuyết và ưu nhược điểm của hai phương pháp lọc cộng tác và lọc nội dung. Có thể thấy mỗi phương pháp đều có những ưu thế riêng, phương pháp này khắc phục nhược điểm của phương pháp còn lại, từ đó phân tích các trường hợp sử dụng của hai phương pháp.

Chương 3 sẽ mô tả thực nghiệm với cả 2 phương pháp với dữ liệu sách, từ đó ta có cái nhìn trực quan hơn về từng phương pháp.

CHƯƠNG 3: THỰC NGHIỆM VÀ ĐÁNH GIÁ VỚI DỮ LIỆU SÁCH

Phương pháp thực nghiệm được mô tả theo các pha như hình 3.1.

- Bước 1: Thu thập dữ liệu
- Bước 2: Xử lý dữ liệu
- Bước 3: Áp dụng các thuật toán khuyến nghị
- Bước 4: Đánh giá thuật toán



Hình 3.1: Các bước xây dựng mô hình khuyến nghị

3.1. Thu thập dữ liệu

Dữ liệu là một thành phần vô cùng quan trọng trong các hệ gợi ý. Dữ liệu và nguồn tri thức có sẵn cho các hệ gợi ý có thể rất đa dạng. Tuy nhiên, trong bất kỳ trường hợp nào, dữ liệu được sử dụng bởi hệ gợi ý thuộc ba loại: sản phẩm (Item), người dùng (User) và tương tác giữa người dùng và sản phẩm (Transaction), trong phạm vi bài luận, dữ liệu tương tác chính là số sao đánh giá (rating) của người dùng với sản phẩm

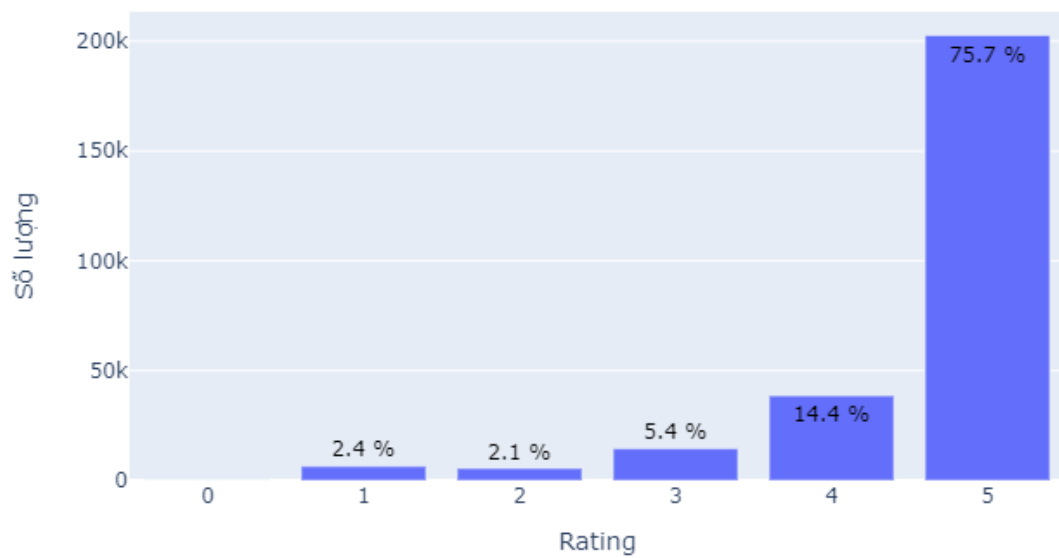
Qua thời gian tìm hiểu, tôi nhận thấy trang web Tiki có thể đáp ứng các điều kiện về dữ liệu cần có: số mẫu sản phẩm (sách) cũng như lượt đánh giá của người dùng với sản phẩm lớn. Tiki cung cấp API mở cho các nhà phát triển và nghiên cứu. Tài liệu chi tiết có thể tham khảo tại trang web <https://open.tiki.vn/#getting-started>

Sau khi tiến hành thu thập dữ liệu, dữ liệu thực nghiệm bao gồm 2 tập dữ liệu: 1 bộ dữ liệu thông tin sách và 1 bộ dữ liệu lưu trữ đánh giá bình luận của người dùng cho sách

Dữ liệu thô ban đầu bao gồm:

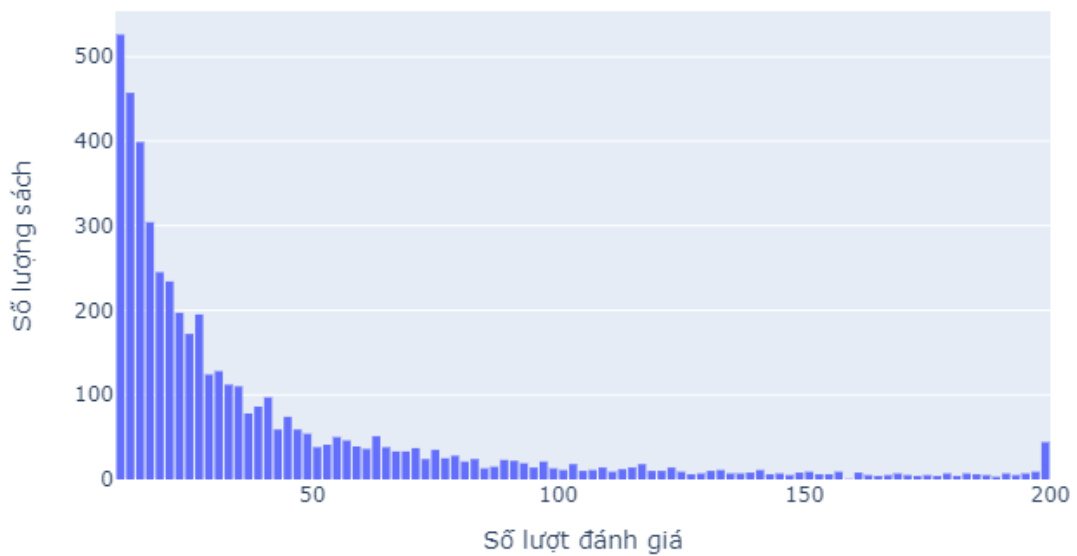
- Dữ liệu sách: 10738 bản ghi, có bao gồm các bản ghi bị trùng lặp, 21 trường dữ liệu.
- Dữ liệu đánh giá: gồm 267742 lượt đánh giá cho các sản phẩm trong bộ dữ liệu sách. Hình 3.2, 3.3, 3.4 thể hiện một vài phân phối của đánh giá theo sản phẩm cũng như người dùng từ dữ liệu thô ban đầu.

Biểu đồ phân phối của 267742 ratings

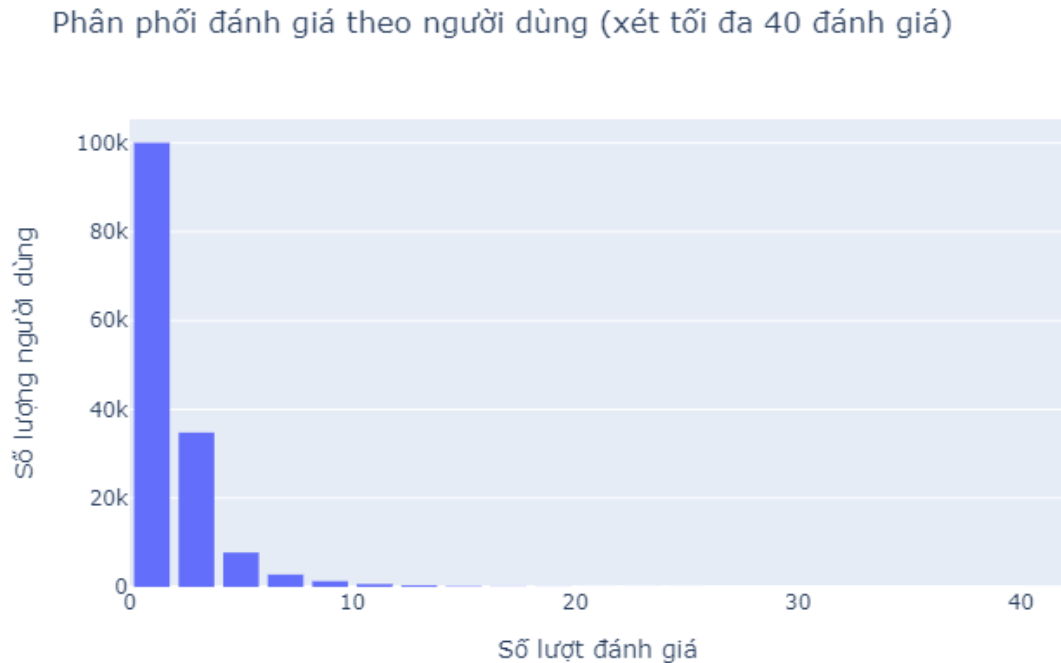


Hình 3.2: Biểu đồ phân phối đánh giá (dữ liệu thô)

Biểu đồ phân phối số sao đánh giá (Xét tối đa là 200 đánh giá)



Hình 3.3: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu thô)



Hình 3.4: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu thô)

Dữ liệu sách sau khi loại bỏ các bản ghi trùng lặp, thu được 10330 bản ghi, gồm các đặc trưng: id, tên, loại bìa, mô tả, giá bán, đánh giá trung bình, thể loại, thể loại con, tác giả, nhà xuất bản, số lượt bán. Bảng 3.1 minh họa một vài mẫu dữ liệu sách sau khi xử lý.

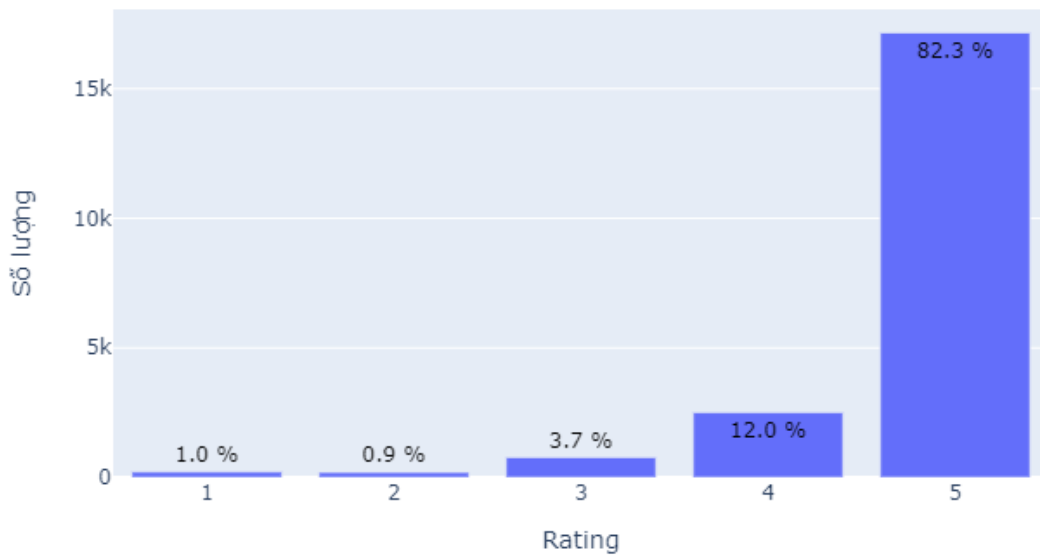
Bảng 3.1: Một vài mẫu dữ liệu sách

id	name	big_category	sub_category	price	authors	publisher	description
1	Python cơ bản	Sách công nghệ thông tin	Lập trình	125000	Bùi Việt Hà	NXB Đại học quốc gia HN	Hiện nay ngôn ngữ lập trình...
2	Đắc nhân tâm	Sách kỹ năng sống	Sách tư duy, kỹ năng sống	49800	Dale Carnegie	First New – Trí Việt	Đắc nhân tâm là cuốn sách...
3	Không gia đình	Sách văn học	Tác phẩm kinh điển	106900	Hector Malot	Huy Hoàng bookstore	Không gia đình là ...
4	Hackers Ielts: Writing	Sách học ngoại ngữ	Sách học tiếng anh	142900	Nhiều tác giả	Alphabooks	Bộ sách luyện thi ...

Nhận xét về dữ liệu đánh giá thô, đa phần người dùng chỉ đánh giá dưới 10 quyển sách. Để tránh vấn đề cold start của dữ liệu người dùng, cũng như đảm bảo tính gợi ý

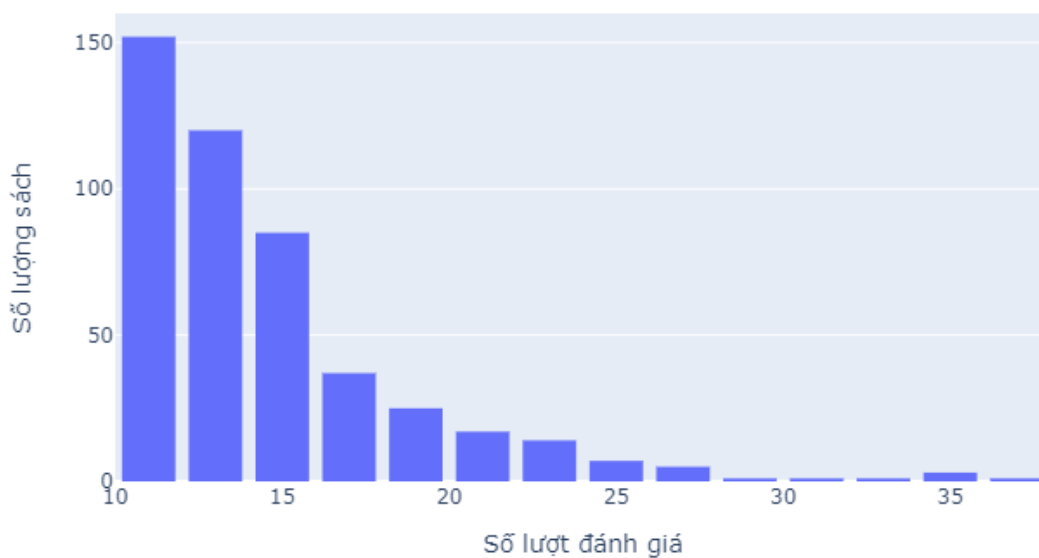
của mô hình, tôi chỉ giữ lại những người dùng có đánh giá tối thiểu là 10 quyển sách. Sau khi xử lý, hình 3.5, 3.6, 3.7 thể hiện phân phối của đánh giá theo người dùng và sản phẩm. Có thể thấy, dữ liệu đã giảm được sự thừa thớt.

Biểu đồ phân phối của 20849 ratings



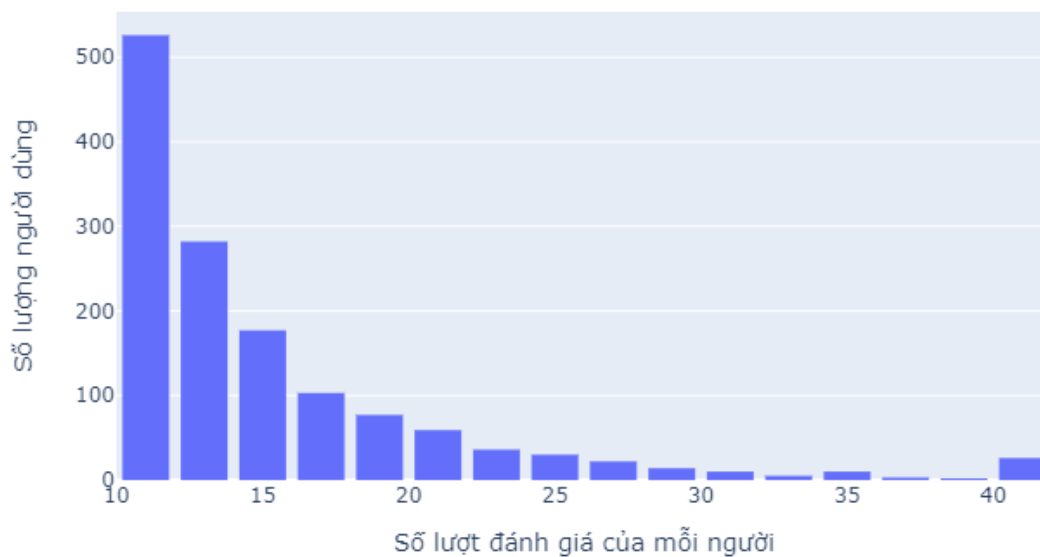
Hình 3.5: Biểu đồ phân phối đánh giá (dữ liệu sau khi xử lý)

Biểu đồ phân phối số sao đánh giá mỗi sản phẩm (Xét tối đa là 200 đánh giá)



Hình 3.6: Biểu đồ phân phối số sao đánh giá theo sản phẩm (dữ liệu sau khi xử lý)

Phân phối đánh giá theo người dùng (xét tối đa 40 đánh giá)



Hình 3.7: Biểu đồ phân phối số sao đánh giá theo người dùng (dữ liệu sau khi xử lý)

Sau khi xử lý, dữ liệu tương tác giữa người dùng và sách còn lại 20849 bản ghi như hình 3.8, gồm các cột mã sản phẩm, mã người dùng, số sao đánh giá, nội dung bình luận, tên người dùng, mã bình luận.

	product_id	content	customer_id	user_name	stars	comment_id
0	9672	NaN	53270	Nhung Phạm	5	4864338
1	9672	NaN	91213	Nguyen Nang Quang	4	4589409
2	3038	Sách có hình vẽ và nội dung rất đáng yêu, là m...	56289	Phan Thi Hoan	5	2831147
3	3038	Sách dễ thương. Bé nhà mình rất thích :)	88524	Yến Nhi	5	3477988
4	3038	Sách dễ thương	72126	Ngoc Dung	5	4633043
5	3038	NaN	131996	Đạt Đoàn	5	4913035
6	3038	NaN	9077	Lê Thu Hằng	5	3856816
7	3038	NaN	113186	Nguyễn Thị Luyện	5	3704500
8	9940	Sách 7+	34988	Mai Quốc Trung	5	4756951
9	9940	NaN	119325	Lâm Văn Xự	5	4750659

Hình 3.8: Dữ liệu đánh giá của người dùng

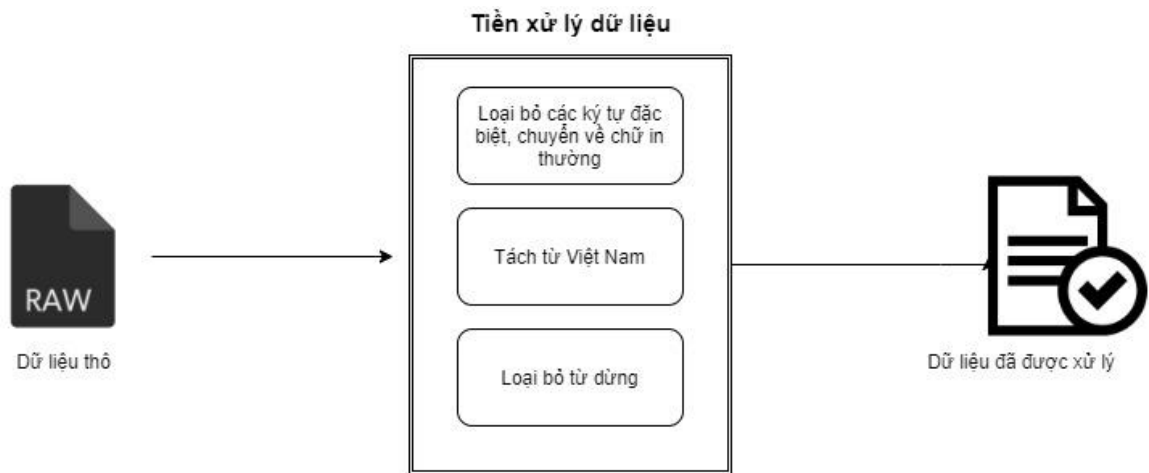
3.2. Xử lý dữ liệu

3.2.1. Tiền xử lý dữ liệu

Đối với dữ liệu dạng văn bản tiến hành xử lý, ví dụ một số trường trong dữ liệu sách như: phân mô tả sách, còn tồn tại các ký tự đặc biệt, thẻ html, từ dừng

Quá trình tiền xử lý dữ liệu bao gồm các bước như hình 3.9:

- Làm sạch dữ liệu bằng cách loại bỏ một số dấu câu, ký tự không cần thiết (trừ dấu hỏi chấm), loại bỏ thẻ html
- Chuẩn hoá từ: Chuẩn hoá dạng ký tự viết hoa về ký tự viết thường.
- Loại bỏ từ dừng



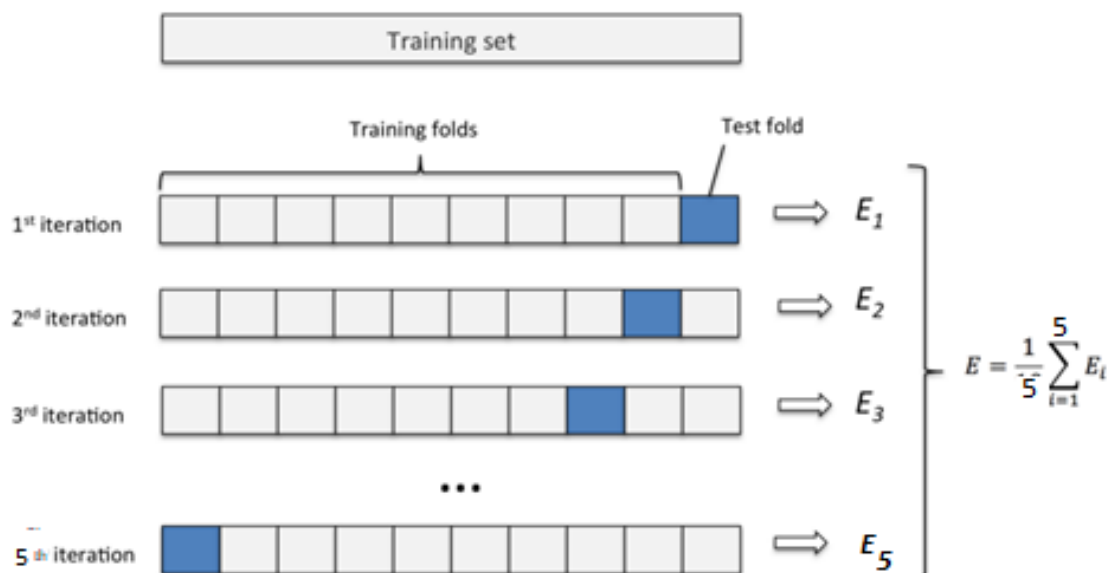
Hình 3.9: Tiền xử lý dữ liệu văn bản

Dữ liệu đánh giá trích chọn 3 trường làm đầu vào huấn luyện là: mã người dùng, mã sản phẩm, số sao đánh giá

Số lượng người dùng đánh giá	1382
Số lượng sản phẩm được đánh giá	5289

3.2.2. Chia dữ liệu thực nghiệm

Việc chia tập dữ liệu dựa vào phương pháp kiểm tra chéo (Cross validation test). Bộ dữ liệu đánh giá sau bước tiền xử lý được chia thành 5 bộ dữ liệu nhỏ hơn và có số lượng mẫu ngẫu nhiên tương đương nhau. Từ năm bộ dữ liệu nhỏ hơn, lần lượt lấy 4 bộ để thành tập huấn luyện và 1 bộ còn lại tạo thành tập kiểm tra. Đảm bảo, tất cả 5 bộ đều được đóng vai trò như một bộ kiểm tra 1 lần. Việc chia bộ dữ liệu thông qua phương pháp kiểm tra chéo được biểu diễn như hình 3.10, kết quả cuối cùng là trung bình cộng của 5 lần đo.



Hình 3.10: Minh họa cách chia dữ liệu

3.3. Thực nghiệm mô hình khuyến nghị

3.3.1. Thực nghiệm mô hình khuyến nghị sách theo kỹ thuật lọc dựa trên nội dung

Công cụ thực nghiệm: thư viện sklearn (hay scikit-learn) [9] – một thư viện mạnh mẽ nhất dành cho các thuật toán học máy được viết trên ngôn ngữ Python. Sklearn cung cấp một tập các công cụ xử lý các bài toán học máy và mô hình thống kê như phân lớp, hồi quy, phân cụm

Xây dựng hồ sơ sản phẩm: Chuyển đổi từ sang vector đặc trưng TF-IDF

- Sử dụng gói `sklearn.feature_extraction.text.TfidfVectorizer` [9]
- Tham số sử dụng mặc định.
- Sau khi vector hóa, thu được ma trận hồ sơ sản phẩm có kích thước 10320x7193. Trong đó 10320 tương ứng với 10320 sản phẩm, 7193 là kích thước từ điển trích từ bộ dữ liệu.

3.3.1.1. Lọc nội dung theo bộ nhớ

Sau khi đã có ma trận hồ sơ sản phẩm, tiến hành tính toán độ tương tự giữa các vector hàng. Để tiết kiệm chi phí tính toán và tính hợp lý, ta chỉ tính toán giữa các sản phẩm cùng thể loại chứ không tính toán hết trên toàn bộ ma trận.

Sử dụng gói hỗ trợ `sklearn.metrics.pairwise.cosine_similarity` [9], có đầu vào là ma trận hồ sơ sản phẩm.

Sau khi đã tính toán, với mỗi sản phẩm, sắp xếp các sản phẩm theo độ tương tự giảm dần. Những sản phẩm tương đồng nhất sẽ được sử dụng đề gợi ý.

Ma trận tương đồng có dạng như hình 3.11, có kích thước $n \times n$, trong đó n là số lượng sản phẩm, các đơn vị 0, 1, 2, 3, 4, 5 đại diện cho số thứ tự của sản phẩm. Giá trị mỗi ô là độ tương đồng giữa sản phẩm x và sản phẩm y . Do vậy, các hàng chéo có giá trị bằng 1.

	0	1	2	3	4	5
0	1.00	0.57	0.51	0.26	0.31	0.33
1	0.57	1.00	0.54	0.25	0.31	0.43
2	0.51	0.54	1.00	0.19	0.25	0.36
3	0.26	0.25	0.19	1.00	0.50	0.38
4	0.31	0.31	0.25	0.50	1.00	0.56
5	0.33	0.43	0.36	0.38	0.56	1.00

Hình 3.11: Ma trận tương đồng giữa các sản phẩm

3.3.1.2. Lọc nội dung theo mô hình

Mô hình sử dụng: hồi quy Ridge, gói hỗ trợ `sklearn.linear_model.Ridge`

Đầu vào mô hình là vector TF-IDF của sản phẩm và nhãn đánh giá từ (1-5).

Thực hiện huấn luyện dữ liệu bằng cách gọi phương thức `fit(X, y)`. Khi đó mô hình sẽ thực hiện học tập dữ liệu để tìm ra các hệ số cần tìm để tối ưu hàm mất mát.

Sau khi xây dựng xong hàm tuyến tính, tiến hành đánh giá mô hình trên tập kiểm thử.

3.3.2. Thực nghiệm mô hình khuyến nghị sách theo kỹ thuật lọc cộng tác

3.3.2.1. Công cụ thực nghiệm

Giới thiệu thư viện `sciki-surprise` [10]: Surprise là một thư viện Python mạnh mẽ trong việc xây dựng và phân tích các hệ khuyến nghị sử dụng dữ liệu phản hồi trực tiếp. *SurPRISE* đại diện cho Simple Python Recommendation System Engine

Surprise được thiết kế nhằm mục đích hỗ trợ nhà phát triển có thể tiến hành thử nghiệm các thuật toán và xây dựng một mô hình gợi ý một cách nhanh chóng. Ngoài ra thư viện cũng cung cấp các tập dữ liệu có sẵn, và cung cấp các phương thức đánh giá mô hình khuyến nghị.

Cài đặt thư viện: `pip install surprise`

Thực hiện huấn luyện dữ liệu bằng cách gọi phương thức `fit(X)`. Trong đó X là dữ liệu huấn luyện

Mỗi người dùng sau khi huấn luyện ta thu được w và b , từ đó suy ra hàm dự đoán đánh giá ứng với người dùng đó, có dạng $y_{mn} = x_m w_n + b_n$ (3.1)

3.3.2.2. Lọc cộng tác dựa trên bộ nhớ

Mô hình sử dụng: `KNNBasic`

Lần lượt thực nghiệm với các cặp tham số trong bảng 3.2:

Bảng 3.2: Tham số mô hình k láng giềng gần nhất

Tham số	Ý nghĩa
<code>name: cosine, user_based: True</code>	Lọc cộng tác dựa trên người dùng sử dụng độ đo cosine
<code>name: cosine, user_based: False</code>	Lọc cộng tác dựa trên sản phẩm sử dụng độ đo cosine
<code>name: pearson, user_based: True</code>	Lọc cộng tác dựa trên người dùng sử dụng độ đo pearson
<code>name: pearson, user_based: False</code>	Lọc cộng tác dựa trên sản phẩm sử dụng độ đo pearson

3.3.2.3. Lọc cộng tác bằng phương pháp matrix factorization

Qua quá trình thử nghiệm và đánh giá trên tập dữ liệu, mô hình SVD được đánh giá là có kết quả cao nhất trên bộ dữ liệu so với các thuật toán matrix factorization khác.

Các tham số tối ưu của mô hình SVD sau khi thực hiện chạy dữ liệu thử nghiệm với các tham số trong bảng 3.3

Bảng 3.3: Tham số mô hình SVD

Tham số	Ý nghĩa	Giá trị tối ưu
n_factors	Số lượng các nhân tử	25
n_epochs	Số lần lặp đào tạo	20
lr_all	Tốc độ học cho các tham số	0.008
reg_all	Regularization cho các tham số	0.08

3.4. Đánh giá thuật toán

3.4.1. Các thông số đánh giá giải thuật

Có rất nhiều thông số để đánh giá một mô hình khuyến nghị. Trong phạm vi đồ án, tôi sẽ trình bày các tiêu chí định lượng được sử dụng nhằm đánh giá số lượng các gợi ý liên quan.

3.4.1.1. Các tiêu chí đánh giá độ chính xác của các dự đoán

Việc đánh giá tính chính xác các dự đoán có thể sử dụng sai số bình phương trung bình (Mean Square Error – MSE), căn của sai số bình phương trung bình (Root Mean Square Error – RMSE), sai số tuyệt đối trung bình (Mean Absolute Error – MAE). Tính chính xác của dự đoán được đo trên n quan sát, trong đó p_i là giá trị dự đoán đánh giá của sản phẩm i và r_i là giá trị đánh giá thực tế của sản phẩm i . [11]

$$MSE = \frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2 \quad (3.2)$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2} \quad (3.3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n (p_i - r_i) \quad (3.4)$$

Các chỉ số này càng cao thì hiệu quả của hệ thống càng thấp và bằng 0 khi hệ thống đạt được hiệu quả tốt nhất.

3.4.1.2. Các tiêu chí đánh giá việc sử dụng các dự đoán

Ngoài việc đánh giá tính chính xác của các dự đoán, một số chỉ số khác như precision, recall và F-score, R-score được dùng để đánh giá việc sử dụng của các dự đoán trong trường hợp cơ sở dữ liệu nhị phân. Các chỉ số này đánh giá các gợi ý phù hợp cho mỗi người dùng thay vì đánh giá số điểm liên quan đến từng đề nghị. Đề nghị được coi là

phù hợp khi người dùng chọn mục dữ liệu từ danh sách những gợi ý đưa ra cho người dùng.

Precision là tỉ lệ giữa số lượng các gợi ý phù hợp và tổng số các gợi ý đã tạo ra [11].

$$Precision = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng gợi ý tạo ra}} \quad (3.5)$$

Recall là tỉ lệ giữa số lượng các gợi ý phù hợp và số lượng các sản phẩm mà người dùng đã chọn lựa. Recall được sử dụng để đo khả năng hệ thống tìm được những gợi ý phù hợp so với những gì người dùng cần [11].

$$Recall = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng sản phẩm chọn bởi người dùng}} \quad (3.6)$$

Precision và recall trong một số trường hợp có giá trị tỉ lệ nghịch với nhau. Giả sử số lượng gợi ý mà hệ thống tạo ra là 10, số lượng gợi ý phù hợp là 3, số lượng sản phẩm mua bởi người dùng là 3, thì độ chính xác thấp (30%), tuy nhiên giá trị recall lại cao (100%). Trong tình huống đó, chỉ số F-score được sử dụng để đánh giá hiệu quả tổng thể của hệ thống [11]

$$Fscore = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.7)$$

Giá trị của các độ đo này càng cao thì độ chính xác của phương pháp càng cao.

3.4.2. Kết quả thực nghiệm

3.4.2.1. Mô hình gợi ý theo lọc nội dung dựa trên mô hình

Bảng 3.4 thể hiện kết quả thử nghiệm độ đo RMSE trên bộ dữ liệu huấn luyện và bộ dữ liệu kiểm thử

Bảng 3.4: Kết quả độ đo RMSE của mô hình lọc nội dung

	RMSE
Tập huấn luyện	0.02533
Tập kiểm thử	0.53477

Sau khi xây dựng hàm tính toán đánh giá, tiến hành thực nghiệm dự đoán số sao trên kiểm thử và ghi lại kết quả tại bảng 3.5

Bảng 3.5: So sánh đánh giá cho một số người dùng

userid	Số sao thực tế	Số sao dự đoán
98	[3, 3, 2]	[3.65, 3.63, 3.76]
325	[5, 4, 5, 5]	[4.48, 4.72, 4.13, 4.38]
120	[5, 5]	[5. 5.]

→ Có thể thấy số sao thực tế và dự đoán số sao của mô hình không có sự chênh lệch quá lớn.

3.4.2.2. Mô hình gợi ý theo lọc cộng tác dựa trên bộ nhớ

Bảng 3.6 biểu diễn kết quả lọc cộng tác dựa trên sản phẩm trên tập huấn luyện (Item Based)

Bảng 3.6: Kết quả lọc cộng tác dựa trên sản phẩm trên tập huấn luyện

Độ tương tự		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Trung bình	Độ lệch chuẩn
Cosine	RMSE	0.6311	0.6325	0.6481	0.6179	0.6076	0.6274	0.0138
	MAE	0.3140	0.3156	0.3255	0.3184	0.3039	0.3155	0.0070
	Fit time	1.93	1.76	1.58	2.16	1.62	1.81	0.21
	Test time	0.15	0.15	0.23	0.17	0.14	0.17	0.03
Pearson	RMSE	0.6613	0.6699	0.6639	0.6789	0.6353	0.6619	0.0146
	MAE	0.4141	0.4160	0.4106	0.4149	0.3996	0.4110	0.0060
	Fit time	2.52	2.57	2.25	2.18	2.07	2.32	0.19
	Test time	0.19	0.14	0.40	0.46	0.13	0.26	0.14

Bảng 3.7 biểu diễn kết quả lọc cộng tác dựa trên người dùng trên tập huấn luyện (Item Based)

Bảng 3.7: Kết quả lọc cộng tác dựa trên người dùng trên tập huấn luyện

Độ tương tự		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Trung bình	Độ lệch chuẩn
Cosine	RMSE	0.7653	0.7810	0.7680	0.7469	0.7564	0.7635	0.0115
	MAE	0.4253	0.4306	0.4259	0.4105	0.4066	0.4198	0.0094
	Fit time	0.23	0.21	0.23	0.25	0.60	0.30	0.15

	Test time	0.14	0.14	0.16	0.18	0.14	0.15	0.02
Pearson	RMSE	0.6779	0.6425	0.6768	0.6897	0.6984	0.6771	0.0190
	MAE	0.4363	0.4254	0.4378	0.4407	0.4438	0.4368	0.0063
	Fit time	0.29	0.23	0.33	0.23	0.29	0.28	0.04
	Test time	0.08	0.08	0.11	0.08	0.12	0.09	0.02

Recall và Precision là các chỉ số đánh giá cho nhãn nhị phân, do đó, ta cần chuyển các đánh giá từ giá trị 1-5 về giá trị nhị phân 0/1, trong đó 0 có nghĩa là gợi ý không liên quan và 1 là gợi ý liên quan. Do đó ta có định nghĩa threshold (ngưỡng đánh giá), với những đánh giá thực lớn hơn ngưỡng mang nhãn 1 và ngược lại. Bảng 3.8 là kết quả đo được tương ứng với các giá trị ngưỡng.

Bảng 3.8: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc theo bộ nhớ

threshold	TP	FP	TN	FN	Precision	Recall	F1
0.0	4170	0	0	0	1.000000	1.000000	1.000000
0.5	4170	0	0	0	1.000000	1.000000	1.000000
1.0	4170	0	0	0	1.000000	1.000000	1.000000
1.5	4131	23	10	6	0.994463	0.998550	0.996502
2.0	4131	23	10	6	0.994463	0.998550	0.996502
2.5	4091	60	11	8	0.985546	0.998048	0.991758
3.0	4087	58	13	12	0.986007	0.997072	0.991509
3.5	3923	168	34	45	0.958934	0.988659	0.973570
4.0	3903	162	40	65	0.960148	0.983619	0.971742
4.5	3300	516	207	147	0.864780	0.957354	0.908715
5.0	1719	125	598	1728	0.932213	0.498695	0.649783

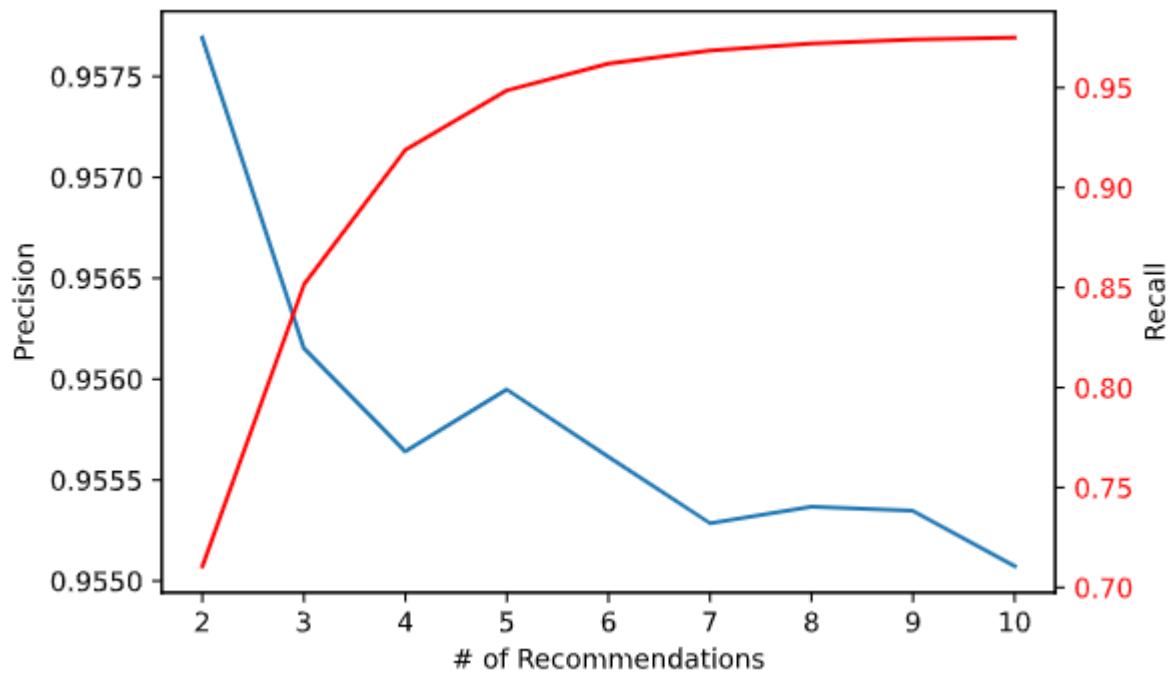
Có thể thấy, với ngưỡng bằng 4 hoặc 4.5, ta thu được kết quả tối ưu nhất.

Ngoài ngưỡng, việc chọn k – số lượng láng giềng gần nhất cũng có thể tác động đến kết quả của hệ gợi ý như trong bảng 3.9 dưới đây.

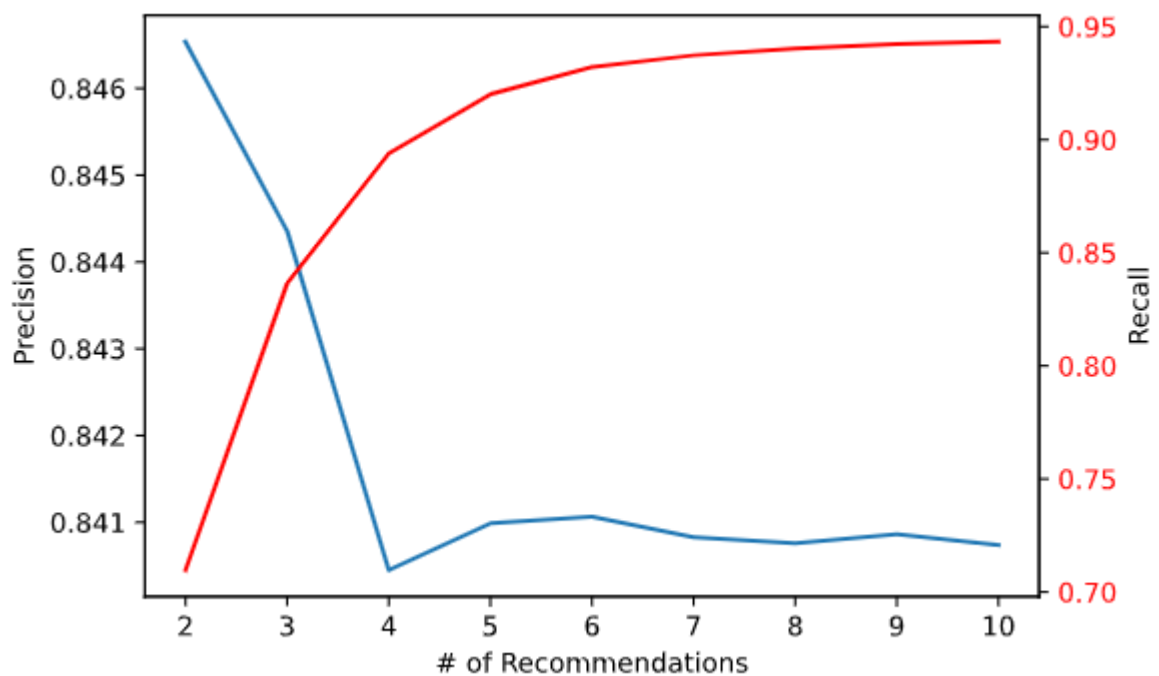
Bảng 3.9: Kết quả đo precision và recall theo k bằng phương pháp lọc theo bộ nhớ

Threshold	k	Precision	Recall	F1
4.0	2	0.95769	0.71069	0.81591
	3	0.95615	0.85140	0.90074
	4	0.95564	0.91889	0.93691
	5	0.95595	0.94867	0.95230
	6	0.95562	0.96196	0.95878
	7	0.95528	0.96847	0.96183
	8	0.95537	0.97199	0.96360
	9	0.95535	0.97406	0.96461
	10	0.95507	0.97502	0.96494
4.5	2	0.84654	0.70965	0.77207
	3	0.84436	0.83652	0.84042
	4	0.84045	0.89398	0.86639
	5	0.84099	0.92009	0.87876
	6	0.84106	0.93220	0.88429
	7	0.84083	0.93731	0.88645
	8	0.84076	0.94034	0.88777
	9	0.84086	0.94234	0.88871
	10	0.84074	0.94335	0.88909

Đồ thị biểu diễn precision và recall theo k với ngưỡng bằng 4 và 4.5 qua hình 3.12 và 3.13.



Hình 3.12: Biểu diễn precision, recall theo k với ngưỡng bằng 4



Hình 3.13: Biểu diễn precision, recall theo k với ngưỡng bằng 4.5

3.4.2.3. Mô hình gợi ý theo lọc cộng tác dựa trên phương pháp SVD

Bảng 3.10: Kết quả RMSE và MAE của mô hình phân rã ma trận

	RMSE	MAE	Fit time	Test time
SVD	0.5795	0.3383	1.528	0.112

Bảng 3.11: Kết quả đo precision và recall theo các ngưỡng bằng phương pháp lọc phân rã ma trận

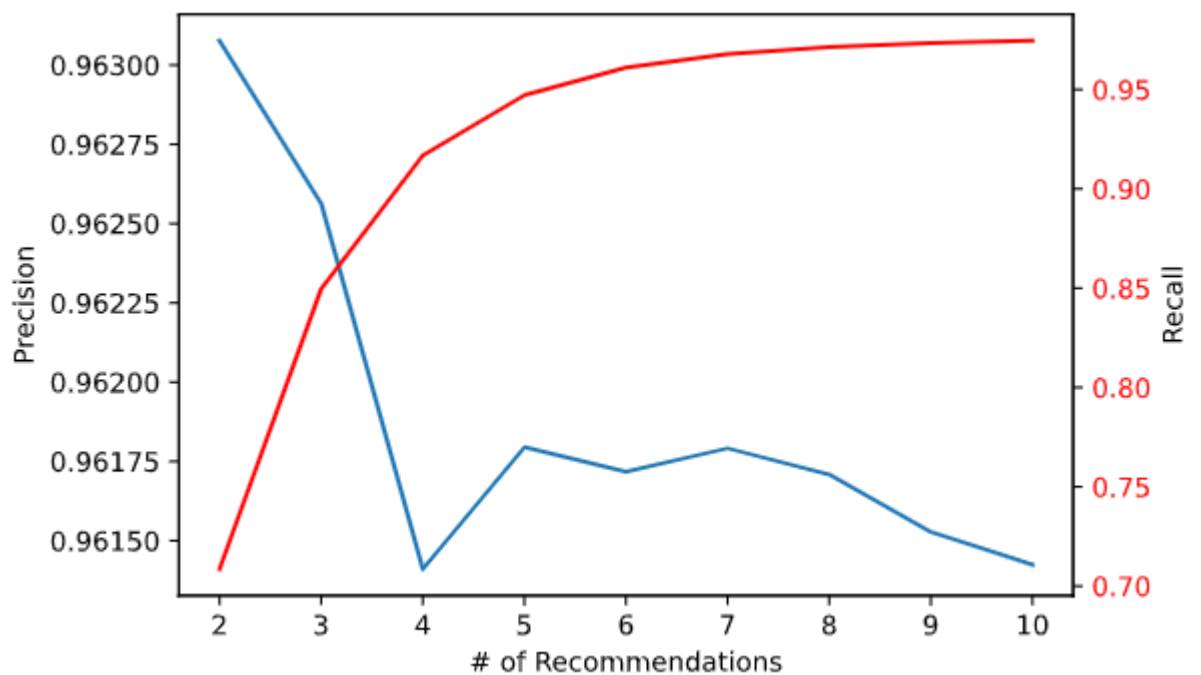
threshold	TP	FP	TN	FN	Precision	Recall	F1
0.0	4170	0	0	0	1.000000	1.000000	1.000000
0.5	4170	0	0	0	1.000000	1.000000	1.000000
1.0	4170	0	0	0	1.000000	1.000000	1.000000
1.5	4135	28	5	2	0.993274	0.999517	0.996386
2.0	4135	25	8	2	0.993990	0.999517	0.996746
2.5	4097	61	10	2	0.985329	0.999512	0.992370
3.0	4097	60	11	2	0.985567	0.999512	0.992490
3.5	3964	184	18	4	0.955641	0.998992	0.976836
4.0	3909	150	52	59	0.963045	0.985131	0.973693
4.5	3148	329	394	299	0.905378	0.913258	0.909301
5.0	552	21	702	2895	0.963351	0.160137	0.274627

Bảng 3.12: Kết quả đo precision và recall theo k bằng phương pháp lọc phân rã ma trận

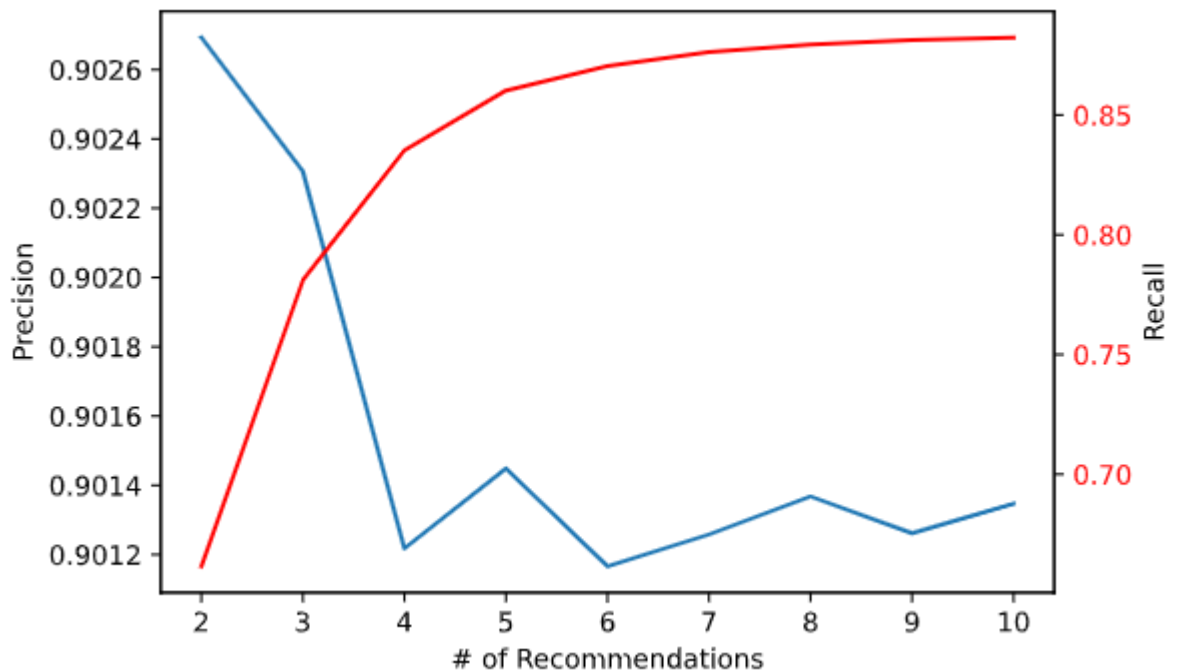
Threshold	k	Precision	Recall	F1
4.0	2	0.96308	0.70853	0.81642
	3	0.96256	0.84976	0.90265
	4	0.96141	0.91687	0.93861
	5	0.96179	0.94718	0.95443
	6	0.96172	0.96102	0.96137
	7	0.96179	0.96786	0.96482
	8	0.96171	0.97148	0.96657
	9	0.96153	0.97346	0.96746
	10	0.96142	0.97459	0.96796
4.5	2	0.90269	0.66165	0.76360
	3	0.90231	0.78106	0.83732
	4	0.90122	0.83515	0.86693

5	0.90145	0.86021	0.88035
6	0.90117	0.87050	0.88557
7	0.90126	0.87625	0.88858
8	0.90137	0.87947	0.89028
9	0.90126	0.88120	0.89112
10	0.90135	0.88229	0.89172

Đồ thị biểu diễn precision và recall theo k với ngưỡng bằng 4 và 4.5 qua hình 3.14, 3.15



Hình 3.14: Biểu diễn precision, recall theo k với ngưỡng bằng 4



Hình 3.15: Biểu diễn precision, recall theo k với ngưỡng bằng 4.5

3.4.3. Nhận xét kết quả và phân tích lỗi trong quá trình thực nghiệm

Qua thử nghiệm, ta thu được các kết quả tốt nhất như trong bảng 18.

Bảng 3.13: Tổng kết kết quả huấn luyện

	Content-based	KNN (CF)	SVD (CF)
RMSE	0.53477	0.6274	0.579
F1	-	~0.95	~0.96

Phương pháp gợi ý theo lọc nội dung cho kết quả thực nghiệm tốt nhất với bộ dữ liệu sách và dữ liệu đánh giá thu được. Các phương pháp lọc cộng tác thường cho kết quả với phương pháp phân rã ma trận tốt hơn.

Tuy nhiên, do bộ dữ liệu chủ yếu với đánh giá = 5, dữ liệu còn thưa thớt nên các mô hình gợi ý chưa thể cho kết quả tối ưu nhất.

3.5. Kết chương

Trong chương này, tôi đã trình bày được cách tiến hành các thực nghiệm, mô tả các mô hình thực nghiệm, với những công cụ thực nghiệm và đưa ra kết quả, cũng như những phân tích đánh giá về kết quả thực nghiệm đạt được. Các kết quả thu được đều cho sai số ở mức trung bình. Các mô hình sau đó được học trên toàn bộ dữ liệu và thực hiện đưa ra khuyến nghị, phần biểu diễn kết quả sẽ được trình bày trong chương tiếp theo.

Chương 4 tập trung xây dựng hệ khuyến nghị sách trong ứng dụng bán hàng trực tuyến sử dụng những mô hình gợi ý ở chương 3, nhằm mục đích mô phỏng những kết quả và ưu điểm của hệ khuyến nghị.

CHƯƠNG 4: ỨNG DỤNG MÔ HÌNH KHUYẾN NGHỊ SÁCH VÀO HỆ THỐNG BÁN HÀNG TRỰC TUYẾN

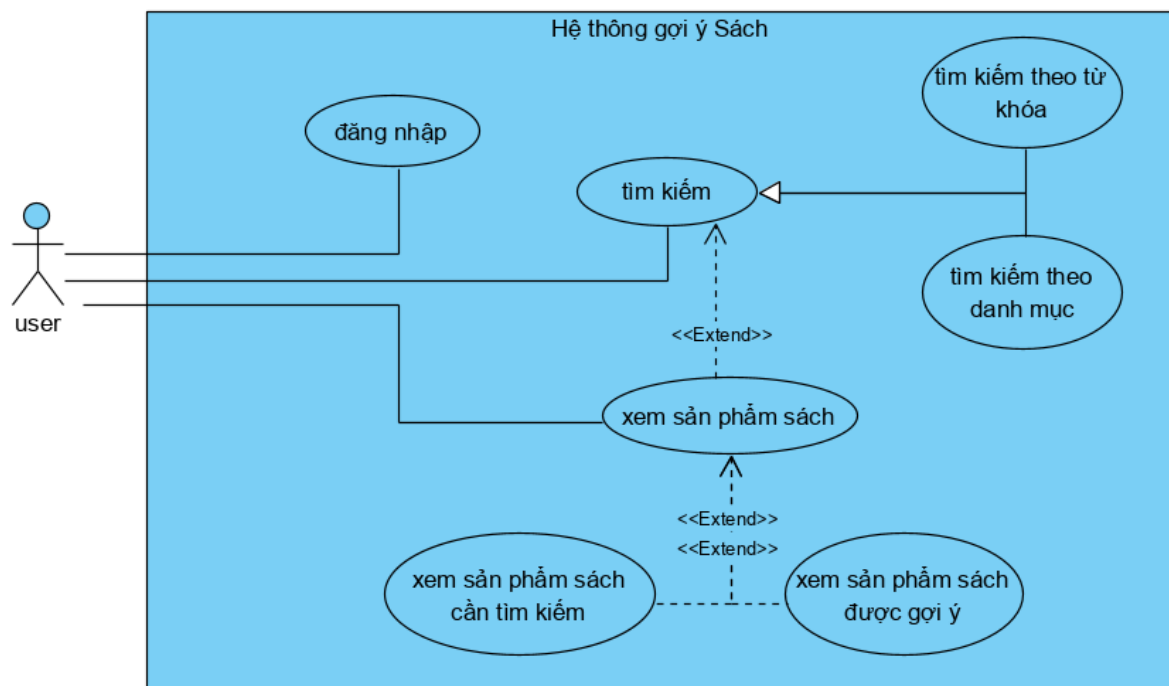
4.1. Mô tả hệ thống

Trang web được xây dựng với mục đích tìm hiểu, nghiên cứu hoạt động của hệ khuyến nghị, biểu diễn kết quả thu được của các mô hình gợi ý đã xây dựng ở chương 3. Hệ thống giới thiệu và gợi ý sách cho phép khách hàng bất kỳ có thể tìm kiếm hoặc xem các sản phẩm theo thể loại, danh sách các sản phẩm được gợi ý.

Trong quá trình xem sản phẩm, hệ thống sẽ gợi ý cho người dùng các loại sách trong quá trình chọn sản phẩm sử dụng kỹ thuật lọc theo nội dung và kỹ thuật lọc cộng tác, hiển thị các cuốn sách tương tự với cuốn sách mà người dùng đang xem sử dụng các thuộc tính tương tự về thể loại sách của sách đó.

4.2. Phân tích thiết kế hệ thống

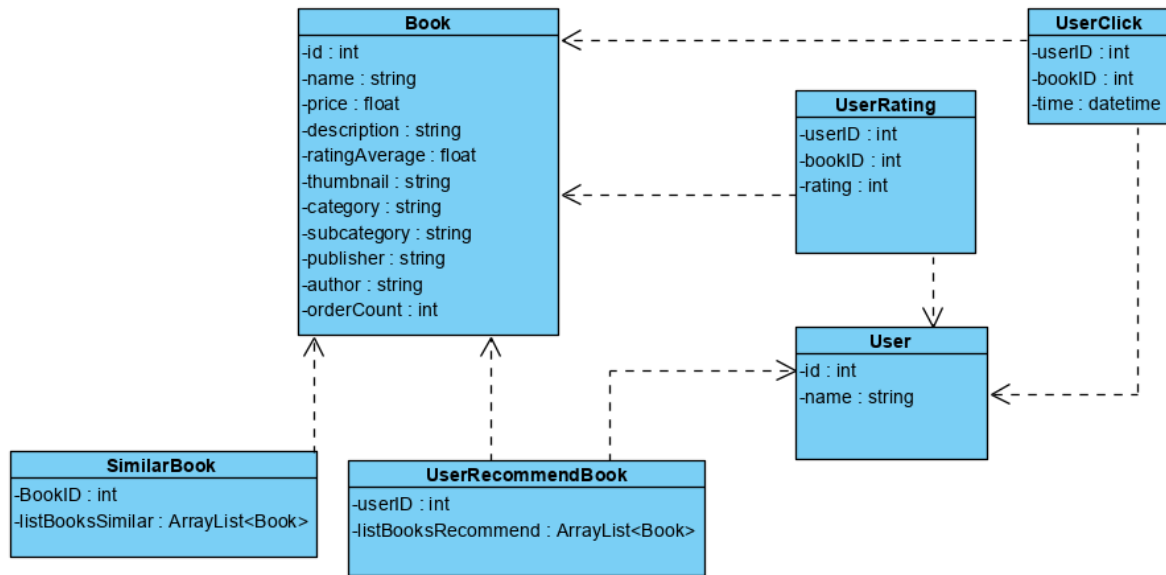
Hình 4.1 biểu diễn sơ đồ usecase hệ thống khuyến nghị sách. Hình 4.2 bao gồm các lớp thực thể trong hệ thống và hình 4.3 là sơ đồ thực thể của hệ thống khuyến nghị sách.



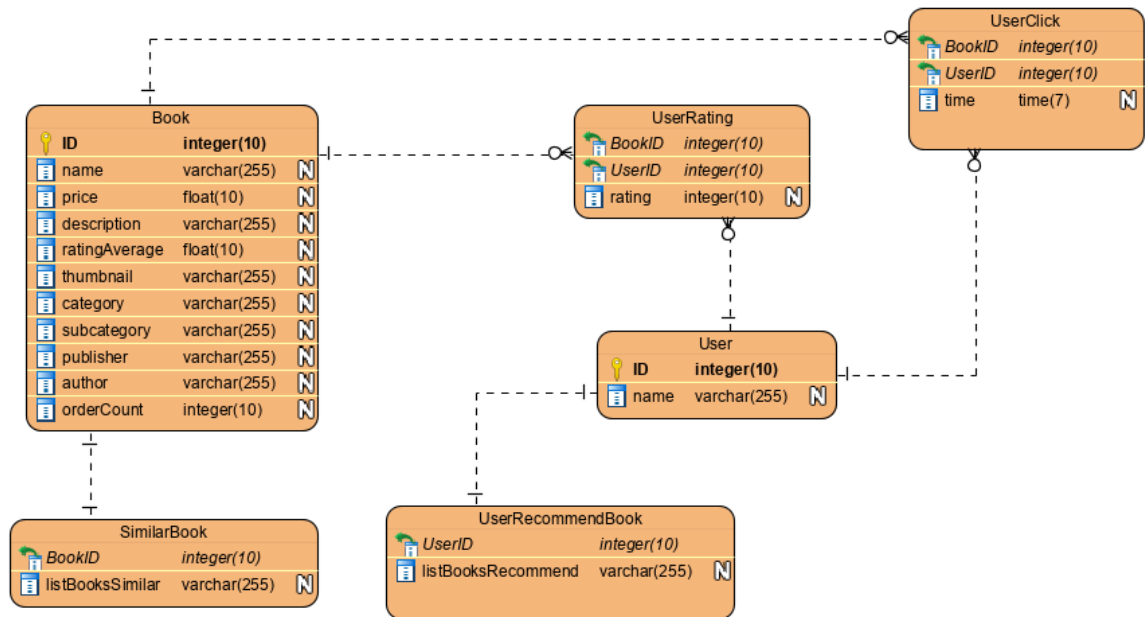
Hình 4.1: Sơ đồ usecase hệ thống khuyến nghị sách

Hệ thống cho phép người dùng thực hiện các chức năng sau:

- Đăng nhập qua Facebook
- Tìm kiếm sách: qua từ khóa hoặc theo danh mục sách
- Xem chi tiết sản phẩm sách
- Xem sản phẩm sách được gợi ý



Hình 4.2: Biểu đồ lớp toàn hệ thống



Hình 4.3: Lược đồ cơ sở dữ liệu

Các lớp thực thể trong hệ thống được mô tả trong bảng 4.1 dưới đây.

Bảng 4.1: Mô tả các lớp trong hệ thống

STT	Tên lớp	Mô tả
1	Book	Lớp Book chứa các đặc trưng về sách: id, tên, giá bán, mô tả, thể loại...
2	User	Lớp User chứa thông tin người dùng: id, tên, email
3	UserClick	Lớp UserClick chứa thông tin sự kiện chọn xem sản phẩm của người dùng

4	UserRating	Lớp UserRating chứa đánh giá của người dùng cho sản phẩm
5	SimilarBook	Lớp SimilarBook chứa danh sách id sách tương tự với một quyển sách tương ứng lớp Book
6	UserRecommendBook	Lớp UserRecommendBook chứa danh sách sách gợi ý cho người dùng tương ứng lớp User

4.3. Thiết kế hệ thống

4.3.1. Các công nghệ sử dụng

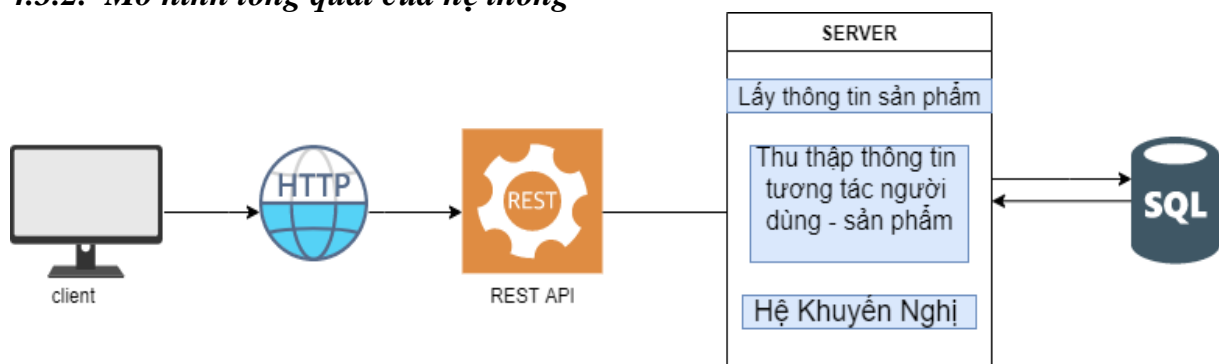
Ngôn ngữ lập trình thuật toán: Python

Ngôn ngữ lập trình server: Python framework Django

Giao diện: HTML, Bootstrap, CSS, Javascripts

Database: SQLite

4.3.2. Mô hình tổng quát của hệ thống



Hình 4.4: Mô hình tổng quát của hệ khuyến nghị sách

Kiến trúc tổng quát của hệ thống như hình 4.4 được thiết kế gồm 3 phần chính như sau:

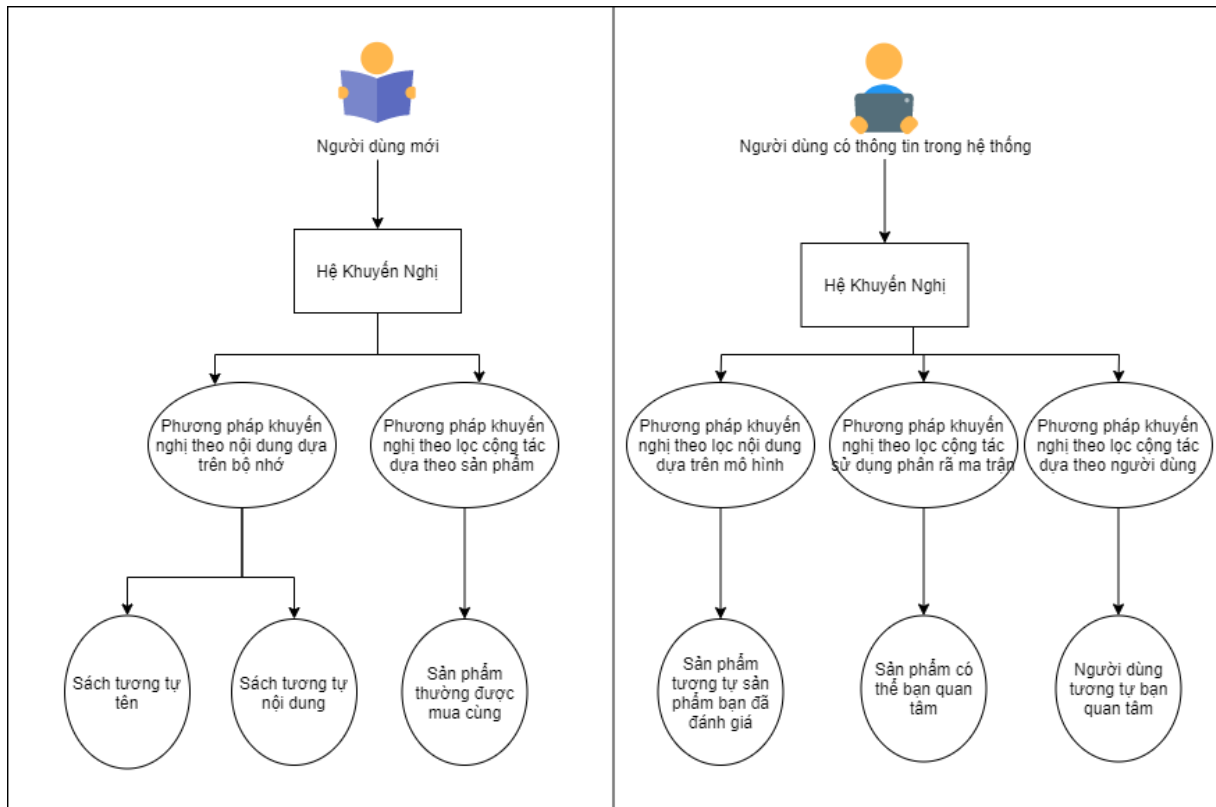
- Phần 1: Giao diện web máy khách (client) hỗ trợ người sử dụng các chức năng như: tìm kiếm sách, xem thông tin sách, xem gợi ý sách
- Phần 2: Ứng dụng máy chủ:
 - + Phụ trách việc truyền nhận thông tin giữa máy khách và máy chủ: truy vấn thông tin sản phẩm từ cơ sở dữ liệu, thu thập thông tin tương tác giữa người dùng và sản phẩm.
 - + Hệ khuyến nghị có chức năng tư vấn các sản phẩm cho người dùng. Các mô hình khuyến nghị thực nghiệm trong chương 3 sẽ tiến hành dự đoán dữ liệu và kết quả được lưu vào cơ sở dữ liệu.
- Phần 3: Cơ sở dữ liệu: lưu trữ thông tin sản phẩm, người dùng, tương tác giữa người dùng và sản phẩm, thông tin khuyến nghị.

Máy chủ và máy khách giao tiếp thông qua API – giao diện lập trình ứng dụng. API cung cấp khả năng truy xuất đến một tập các hàm hay dùng, từ đó có thể trao đổi dữ liệu giữa các ứng dụng (ứng dụng máy khách và máy chủ)

4.3.3. Chương trình demo

Qua quá trình phân tích và xây dựng các mô hình gợi ý, kịch bản gợi ý của hệ thống sách được mô tả như sau:

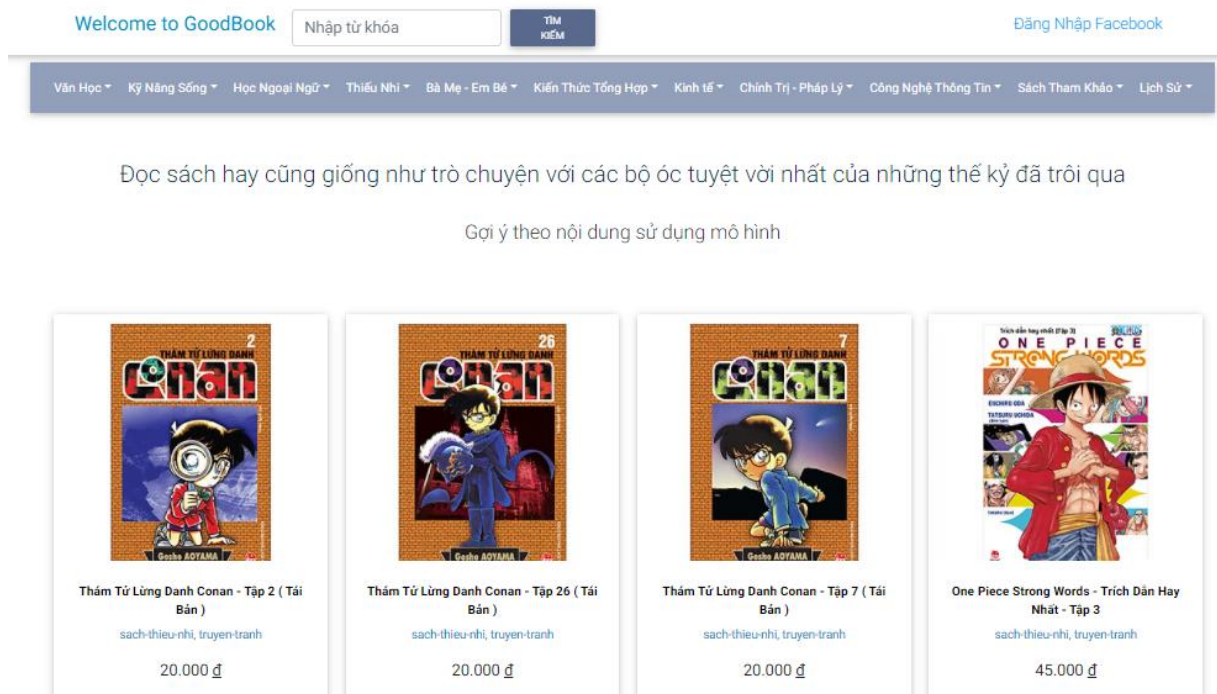
- Đối với người dùng mới: Khi người dùng chọn xem sản phẩm sách, hệ thống sẽ gợi ý các sản phẩm tương tự tên, nội dung, hoặc sản phẩm thường được mua cùng
- Đối với người dùng đã có thông tin trong hệ thống: Trên trang chủ sẽ hiển thị sản phẩm người dùng tương tự sản phẩm người dùng đã đánh giá, sản phẩm có thể người dùng quan tâm hoặc sản phẩm người dùng tương tự quan tâm.



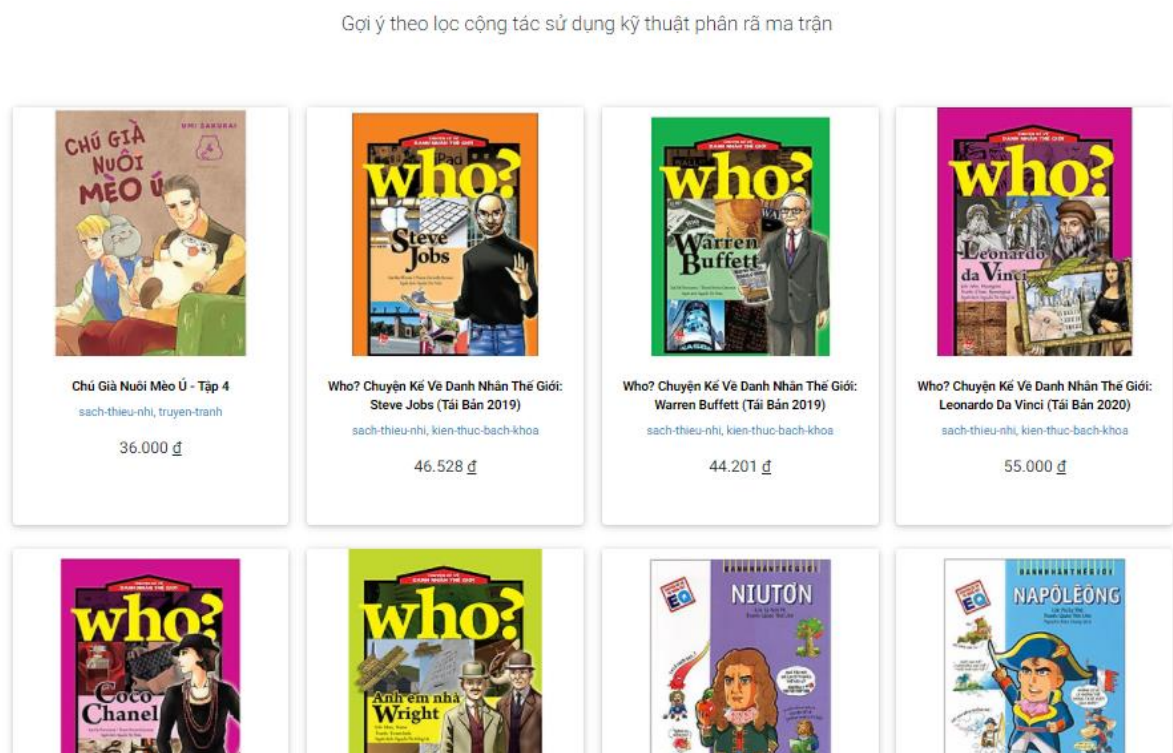
Hình 4.5: Kịch bản demo

Một số giao diện minh họa của hệ thống

- Trang chủ đối với người dùng có thông tin trong hệ thống

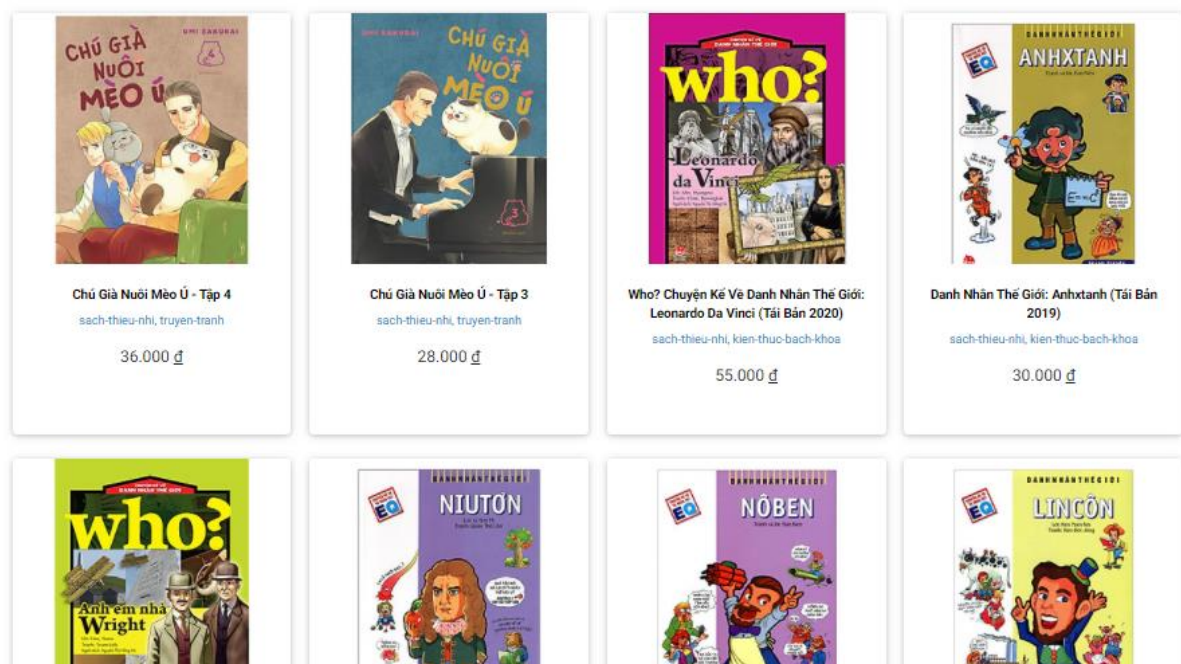


Hình 4.6: Giao diện trang chủ - gợi ý theo lọc nội dung dựa trên mô hình



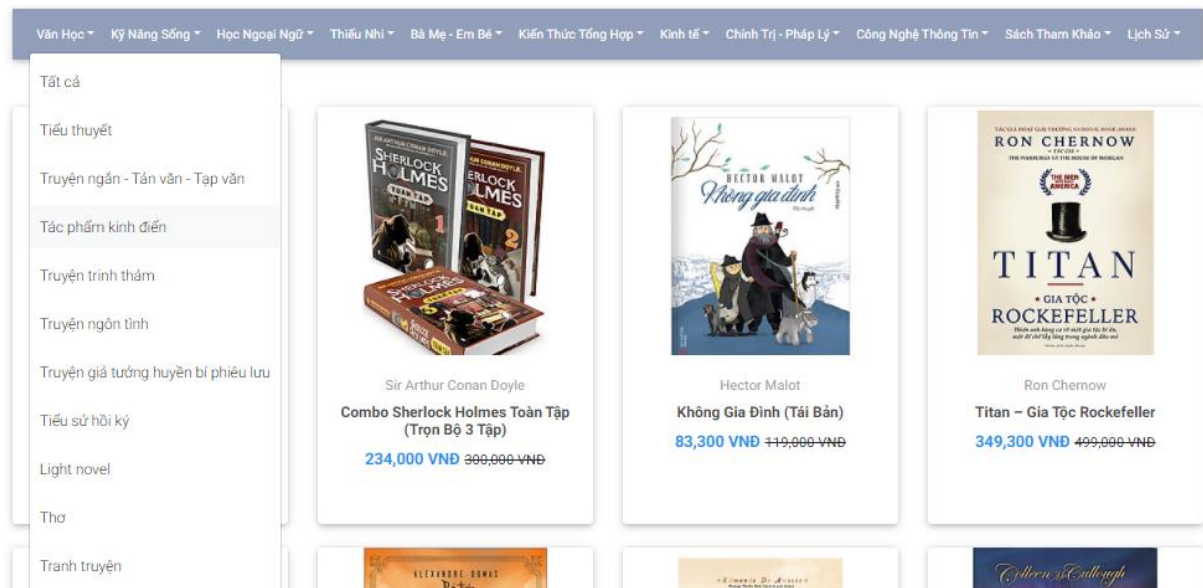
Hình 4.7: Giao diện trang chủ - gợi ý theo SVD

Gợi ý theo lọc cộng tác sử dụng item based



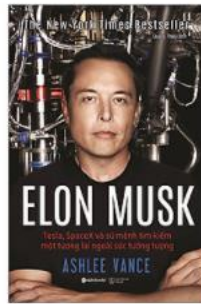
Hình 4.8: Giao diện trang chủ - gợi ý theo lọc cộng tác theo sản phẩm

- Trang hiển thị sách theo phân loại.



Hình 4.9: Giao diện danh mục sách

- Trang chi tiết sản phẩm – Gợi ý cho người dùng mới:
- + Thông tin sản phẩm



Elon Musk: Tesla, SpaceX Và Sứ Mệnh Tìm Kiếm Một Tương Lai Ngoài Sức Tưởng Tượng (Tái Bản 2020)

sach-kinh-te sach-quan-tri

Nhà Xuất Bản: Alphabooks

Tác giả: Ashlee Vance

Rating: 4.8

239000.0 167300.0

TÓM TẮT NỘI DUNG

Trong cuốn Elon Musk: Tesla, SpaceX và sứ mệnh tìm kiếm một tương lai ngoài sức tưởng tượng, nhà báo công nghệ kỳ cựu Ashlee Vance đã mở cánh cửa đầu tiên nhìn vào cuộc sống phi thường của doanh nhân táo bạo nhất thung lũng Silicon. Được viết với độc quyền tiếp cận Musk, gia đình và bạn bè anh, cuốn sách lần theo cuộc hành trình của doanh nhân này, từ thuở nhỏ được nuôi dạy ở Nam Phi đến khi lên đến đỉnh cao của giới kinh doanh toàn cầu. Vance dành hơn 30 giờ trò chuyện với Musk và phỏng vấn gần 300 người để kể những câu chuyện đầy biến động về các công ty đang làm thay đổi thế giới mà Musk thành lập: PayPal, Tesla Motors, SpaceX và SolarCity, và để mô tả một người đàn ông đã tạo sức sống mới cho ngành công nghiệp Mỹ và thổi bùng tinh thần đổi mới ở một cấp độ khác, trong khi cũng tạo nên không ít kẻ thù trên con đường đó.

Hình 4.10: Giao diện thông tin sản phẩm

+ Gợi ý sách tương tự tên

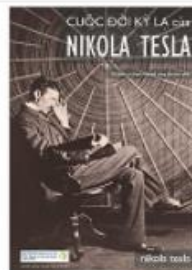
Sách tương tự tên



Một Đồi Quản Trị (Tái Bản)

sach-kinh-te, sach-quan-tri

147.900 đ



Cuộc đời kỳ lạ của Nikola Tesla (tái bản 2018)

sach-kinh-te, sach-doanh-nhan

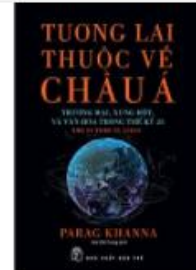
54.900 đ



Vị Giám Đốc Một Phút (Tái Bản 2019)

sach-kinh-te, sach-quan-tri

33.600 đ



Tương Lai Thuộc Về Châu Á

sach-kinh-te, sach-kinh-te-hoc

157.500 đ

Hình 4.11: Giao diện sản phẩm - gợi ý sản phẩm tương tự tên

+ Gợi ý sách tương tự về nội dung



Hình 4.12: Giao diện sản phẩm - gợi ý sản phẩm tương tự nội dung

+ Gợi ý sách thường được mua cùng



Hình 4.13: Giao diện sản phẩm - sách thường được mua cùng

4.4. Kết chương

Chương 4 tập trung cài đặt ứng dụng của hệ khuyến nghị vào hệ thống bán sách trực tuyến. Một số các kết quả chạy mô phỏng đã thể hiện được những ưu điểm của hệ khuyến nghị theo lọc nội dung cũng như lọc cộng tác. Những gợi ý đó được sinh ra nhằm mở rộng sự lựa chọn của người dùng, từ đó tăng khả năng người dùng sẽ mua sản phẩm đó.

KẾT LUẬN

Đồ án đã trình bày được một số nội dung sau:

- **Giới thiệu hệ khuyến nghị và trình bày 2 phương pháp lọc nội dung và lọc cộng tác:** Trình bày về các kỹ thuật gợi ý theo bộ nhớ và theo mô hình của mỗi phương pháp. Chỉ ra những ưu điểm, nhược điểm của mỗi phương pháp.
- **Thử nghiệm và đánh giá:** Giới thiệu bộ dữ liệu thử nghiệm cho các thuật toán đã trình bày trong đồ án. Thử nghiệm cài đặt trên Python cho từng thuật toán ứng với mỗi bộ dữ liệu có các mức độ thưa thớt khác nhau. Kết quả thử nghiệm được đánh giá theo các tiêu chí về độ chính xác, độ nhạy, sai số trung bình và thời gian thực hiện khuyến nghị.
- **Xây dựng ứng dụng hệ khuyến nghị sách:** Áp dụng mô hình đề xuất và các phương pháp lọc nội dung và lọc kết hợp đã trình bày trong đồ án vào xây dựng thành công một hệ khuyến nghị sách trên nền tảng website cho phép người dùng có thể xem thông tin sách, xem các gợi ý sách liên quan.

Hướng phát triển của đồ án:

- Hiện tại, các thuật toán gợi ý đang được nghiên cứu và phát triển, phải kể đến các kỹ thuật sử dụng đồ thị hoặc mạng neuron. Trong tương lai, tôi sẽ thử nghiệm bộ dữ liệu với các mô hình mới, từ đó có cái nhìn tổng quan hơn về các mô hình khuyến nghị.
- Theo lý thuyết, để đánh giá một hệ khuyến nghị có thực sự tốt còn cần quan tâm đến các tính khác của gợi ý như tính đa dạng, tính mới... Tuy nhiên, trong phạm vi trang web chưa hoàn thiện và chưa được triển khai dưới dạng online, nên chưa thể cho đánh giá định tính về hệ gợi ý. Trang web bán sách tích hợp hệ gợi ý được kỳ vọng sẽ triển khai online trong thời gian tới nhằm đánh giá trực quan và cải tiến cho các mô hình khuyến nghị.

DANH MỤC TÀI LIỆU THAM KHẢO

- [1] "Recommender System" website https://en.wikipedia.org/wiki/Recommender_system.
- [2] “Bán sách online và sách điện tử lên ngôi mùa Covid-19”, Báo Nhân Dân, 4 2020, website <https://nhandan.com.vn/dong-chay/ban-sach-online-va-sach-dien-tu-len-ngoi-mua-covid-19-455732/>.
- [3] N. T. Phương, “Nghiên cứu kỹ thuật lọc cộng tác và ứng dụng xây dựng hệ thống gợi ý bán sách trực tuyến” *Luận văn Thạc sĩ khoa học*, 2016.
- [4] "TF-IDF - Xử lý ngôn ngữ tự nhiên" website <https://codetudau.com/bag-of-words-tf-idf-xu-ly-ngon-ngu-tu-nhien/index.html>.
- [5] "Cosine similarity" Wikipedia, website https://en.wikipedia.org/wiki/Cosine_similarity.
- [6] T. Nguyễn, "machinelearningcoban.com" <https://machinelearningcoban.com/2017/05/17/contentbasedrecommendersys/>.
- [7] N. H. Tiệp, “machinelearningcoban.com”, Collaborative Filtering <https://machinelearningcoban.com/2017/05/24/collaborativefiltering/>.
- [8] N. H. Tiệp, “machinelearningcoban.com”, Matrix Factorization <https://machinelearningcoban.com/2017/05/31/matrixfactorization/>.
- [9] Scikit-learn, website <https://scikit-learn.org/stable/>.
- [10] Surprise, website <http://surpriselib.com/>.
- [11] Trần Nguyễn Minh Thư, Phạm Xuân Hiền, “Các phương pháp đánh giá hệ thống gợi ý” *Tạp chí khoa học Trường Đại học Cần Thơ*, 2016.
- [12] N. D. Phương, “Luận án lọc cộng tác và lọc nội dung” *Luận án*, Học Viện Công Nghệ Bưu Chính Viễn Thông.